UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA

JULIO CESAR LAURENTINO ALVES

MÁQUINA DE VETORES DE SUPORTE APLICADA A DADOS DE ESPECTROSCOPIA NIR DE COMBUSTÍVEIS E LUBRIFICANTES PARA O DESENVOLVIMENTO DE MODELOS DE REGRESSÃO E CLASSIFICAÇÃO

TESE DE DOUTORADO APRESENTADA AO
INSTITUTO DE QUÍMICA DA UNICAMP PARA
OBTENÇÃO DO TÍTULO DE DOUTOR EM CIÊNCIAS

ORIENTADOR: PROF. DR. RONEI JESUS POPPI

ESTE EXEMPLAR CORRESPONDE A VERSÃO FINAL DA TESE DEFENDIDA POR JULIO CESAR LAURENTINO ALVES E ORIENTADA PELO PROF. DR. RONEI JESUS POPPI.

_____

PROF. DR. RONEI JESUS POPPI

CAMPINAS, 2012

FICHA CATALOGRÁFICA ELABORADA POR SIMONE LUCAS - CRB8/8144 - BIBLIOTECA DO INSTITUTO DE QUÍMICA DA UNICAMP

AL87m

Alves, Julio Cesar Laurentino (1978-).
Máquina de vetores de suporte aplicada a dados de espectroscopia NIR de combustíveis e lubrificantes para o desenvolvimento de modelos de regressão e classificação / Julio Cesar Laurentino Alves. – Campinas, SP: [s.n.], 2012.

Orientador: Ronei Jesus Poppi.
Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Química.

1. Máquina de vetores de suporte. 2. Espectroscopia de infravermelho próximo. 3. Óleo diesel. 4. Óleo lubrificante. 5. Método de calibração multivariada não linear. I. Poppi, Ronei Jesus. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Informações para Biblioteca Digital

Título em inglês: Support vectors machine applied to NIR spectroscopy data of fuels and lubricants for development of regression and classification models

Palavras-chave em inglês:
Support vectors machine
Near infrared spectroscopy
Diesel oil
Lubricant oil
Nonlinear multivariate calibration method

Área de concentração: Química Analítica

Titulação: Doutor em Ciências

Banca examinadora:
Ronei Jesus Poppi [Orientador]
Marco Aurélio Zezzi Arruda
Fabio Augusto
Jez Willian Batista Braga
Luiz Alexandre Sacorague

Data de defesa: 15/02/2012

Programa de pós-graduação: Química

Agradecimentos

- Ao Prof. Dr. Ronei J. Poppi, pela oportunidade da realização desse trabalho, pela confiança e pela valorosa orientação;
- Aos meus pais, meu irmão e família, pelo apoio;
- Aos colegas do Laboratório de Quimiometria em Química Analítica - LAQQA;
- Ao Instituto de Química da Universidade Estadual de Campinas;
- A Petróleo Brasileiro S.A. - PETROBRAS e especialmente a Claudete B. Henriques, da refinaria de Paulínia – REPLAN, pelo fornecimento de amostras e espectros de óleo diesel e ao Leandro S. F. Martins, da refinaria de Duque de Caxias - REDUC, pelo fornecimento de amostras de óleos básicos;
- A Petrobras Distribuidora S.A. e especialmente ao Silvio Moraes, da base de distribuição de Barueri-SP, pelo fornecimento de amostras de óleo diesel A S50 e biodiesel (B100) de óleo de soja e ao Mauro Noronha, da fábrica de lubrificantes, pelo fornecimento de amostras de óleos básicos;

Curriculum Vitae

Julio Cesar Laurentino Alves julio@iqm.unicamp.br

- Formação Acadêmica

Doutor em ciências – área química analítica, 2012
Universidade Estadual de Campinas – UNICAMP, Campinas-SP

Mestre em química - área química analítica, 2008
Universidade Estadual de Campinas – UNICAMP, Campinas-SP

Bacharel em química com atribuições tecnológicas, 2006
Universidade Estadual de Campinas – UNICAMP, Campinas-SP

- Experiência profissional

Petrobras Distribuidora S.A.
Químico
2009 - atual

Petróleo Brasileiro S.A. - PETROBRAS - Refinaria de Paulinia - REPLAN
Técnico de operação
2004 - 2007

- Atividades acadêmicas
    - Artigos publicados

"Diesel oil quality parameters determination using support vector regression and NIR spectroscopy for HDT feedstock monitoring"
Julio C. L. Alves, Claudete B. Henriques, Ronei J. Poppi,
(submetido)

"Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system"
Julio C. L. Alves, Claudete B. Henriques, Ronei J. Poppi,
(submetido)

"Pharmaceutical analysis in solids using front face fluorescence spectroscopy and multivariate calibration with matrix correction by piecewise direct standardization"
Julio C. L. Alves, Ronei J. Poppi,
(submetido)

"Simultaneous determination of acetylsalicylic acid, paracetamol and caffeine using solid-phase molecular fluorescence and Parallel Factor Analysis"
Julio C. L. Alves, Ronei J. Poppi,
Analytica Chimica Acta, 642 (2009) 212

    - Trabalhos apresentados em congressos nacionais e internacionais

"Classificação das frações que compõem o *pool* de óleo diesel através de dados de espectroscopia NIR e SVM"
Julio C. L. Alves (PG), Claudete B. Henriques (PQ), Ronei J. Poppi (PQ)
16º Encontro nacional de química analítica - ENQA 2011 – Campos do Jordão-SP, 2011

“Support vectors machine and infrared spectroscopy for on-line determination of quality parameters of diesel oil”
Julio C. L. Alves (PG), Danilo A. Maretto (PG), Ronei J. Poppi (PQ)
12º International Conference on Chemometrics for Analytical Chemistry - CAC 2010 – Antwerp – Belgium, 2010

“Espectroscopia no infravermelho próximo e máquina de vetores de suporte para obtenção de parâmetros de qualidade em linha na produção de óleo diesel”
Julio C. L. Alves (PG), Claudete B. Henriques (PQ), Ronei J. Poppi (PQ)
Rio Oil & Gas Conference 2010 – Rio de Janeiro-RJ, 2010

“Determinação de parâmetros de qualidade do óleo diesel através de espectroscopia no infravermelho próximo e máquina de vetores de suporte”
Julio C. L. Alves (PG), Ronei J. Poppi (PQ)
X Seminário de química Petrobras – Rio de Janeiro-RJ, 2010

“Determinação do índice de cetano do óleo diesel através de espectroscopia no infravermelho próximo e máquina de vetores de suporte”
Julio C. L. Alves (PG), Ronei J. Poppi (PQ)
3º Seminário de laboratório da indústria do petróleo - IBP - Rio de Janeiro-RJ, 2010

“Determinação do ponto de anilina do óleo diesel através de espectroscopia no infravermelho próximo e máquina de vetores de suporte”
Julio C. L. Alves (PG), Claudete B. Henriques (PQ), Ronei J. Poppi (PQ)
33º Reunião anual da Sociedade Brasileira de Química – Aguas de Lindóia-SP, 2010

“Resolvendo problemas de efeito de matriz em análise de fármacos por fluorescência de sólidos utilizando transferência de calibração”
Julio C. L. Alves (PG), Ronei J. Poppi (PQ)
15º Encontro nacional de química analítica - ENQA 2009 – Salvador-BA, 2009

“Simultaneous determination of acetylsalicylic acid, paracetamol and caffeine using solid-phase molecular fluorescence and PARAFAC”
Julio C. L. Alves (PG), Ronei J. Poppi (PQ)
11º International Conference on Chemometrics for Analytical Chemistry - CAC 2008 – Montpellier – France, 2008

“Determinação simultânea de AAS, paracetamol e cafeína através de fluorescência molecular em fase sólida e UPLS”
Julio C. L. Alves (PG), Ronei J. Poppi (PQ)
31º Reunião anual da Sociedade Brasileira de Química – Águas de Lindóia-SP, 2008

“Investigação sobre a utilidade do fosfato de vanadila como sensor para gases, reatividade frente a amônia gasosa”.
Julio C. L. Alves (IC), Robson F. Farias (PQ), Cláudio Airoldi (PQ)
26º Reunião anual da Sociedade Brasileira de Química – Poços de Caldas-MG, 2003

- Iniciação ciêntífica

“Investigação sobre a utilidade do fosfato de vanadila como sensor para gases, reatividade frente a amônia gasosa”.
orientador: Prof. Dr. Cláudio Airoldi – IQ-UNICAMP
2002

Resumo

Máquina de vetores de suporte aplicada a dados de espectroscopia NIR de combustíveis e lubrificantes para o desenvolvimento de modelos de regressão e classificação

Modelos lineares de regressão e classificação por vezes proporcionam um desempenho insatisfatório no tratamento de dados de espectroscopia no infravermelho próximo de produtos derivados do petróleo. A máquina de vetores de suporte (SVM), baseada na teoria do aprendizado estatístico, possibilita o desenvolvimento de modelos de regressão e classificação não lineares que podem proporcionar uma melhor modelagem dos referidos dados, porém ainda é pouco explorada para resolução de problemas em química analítica. Nesse trabalho demonstra-se a utilização do SVM para o tratamento de dados de espectroscopia na região do infravermelho próximo de combustíveis e lubrificantes. O SVM foi utilizado para a solução de problemas de regressão e classificação e seus resultados comparados com os algoritmos de referência PLS e SIMCA. Foram abordados os seguintes problemas analíticos relacionados a controle de processos e controle de qualidade: (i) determinação de parâmetros de qualidade do óleo diesel utilizados para otimização do processo de mistura em linha na produção desse combustível; (ii) determinação de parâmetros de qualidade do óleo diesel que é carga do processo de HDT, para controle e otimização das condições de processo dessa unidade; (iii) determinação do teor de biodiesel na mistura com o óleo diesel; (iv) classificação das diferentes correntes que compõem o *pool* de óleo diesel na refinaria, permitindo a identificação de adulterações e controle de qualidade; (v) classificação de lubrificantes quanto ao teor de óleo naftênico e/ou presença de óleo vegetal. Demonstram-se o melhor desempenho do SVM em relação aos modelos desenvolvidos com os métodos quimiométricos de referência (métodos lineares). O desenvolvimento de métodos analíticos rápidos e de baixo custo para solução de problemas em controle de processos e controle de qualidade, com a utilização de modelos de regressão e classificação mais exatos, proporcionam o monitoramento da qualidade de forma mais eficaz e eficiente, contribuindo para o aumento das rentabilidades nas atividades econômicas de produção e comercialização dos derivados do petróleo estudados.

Abstract

Support vectors machine applied to NIR spectroscopy data of fuels and lubricants for development of regression and classification models

Linear regression and classification models can produce a poor performance in processing near-infrared spectroscopy data of petroleum products. Support vectors machine (SVM), based on statistical learning theory, provides the development of models for nonlinear regression and classification that can result in better modeling of these data but it is still little explored for solving problems in analytical chemistry. This work demonstrates the use of the SVM for treatment of near-infrared spectroscopy data of fuels and lubricants. The SVM was used to solve regression and classification problems and its results were compared with the reference algorithms PLS and SIMCA. The following analytical problems related to process control and quality control were studied: (i) quality parameters determination of diesel oil, used for optimization of in line blending process; (ii) quality parameters determination of diesel oil which is feed-stock of HDT unit for optimization of process control; (iii) quantification of biodiesel blended with diesel oil; (iv) classification of different streams that make up the pool of diesel oil in the refinery, enabling identification of adulteration and quality control; (v) classification of lubricants based on the content of naphthenic oil and/or the presence of vegetable oil. It is shown the best performance of the SVM compared to models developed with the reference algorithms. The development of fast and low cost analytical methods used in process control and quality control, with the use of more accurate regression and classification models, allows monitoring quality parameters in more effectiveness and efficient manner, making possible an increase in profitability of economic activities of production and business of petroleum derivatives studied.

Lista de abreviaturas

ALS	*alternating least squares*
ANN	*artificial neural network*
API	*american petroleum institute*
ASTM	*american society for testing and materials*
ATR	*attenuated total reflectance*
ERM	*empirical risk minimization*
FAME	*fatty acid methyl ester*
FCC	*fluid catalytic cracking*
FTIR	*Fourier transformed infrared*
GA	*genetic algorithm*
GC-MS	gas *chromatography mass spectrometry*
KNN	*K-nearest neighbor*
LDA	*linear discriminant analysis*
LSSVM	*least squares support vectors machine*
MIR	*mid infrared*
MLP	*multilayer perceptron*
MLR	*multiple linear regression*
NIR	*nearest infrared*
NMR	*nuclear magnetic ressonance*
OSH	*optimal separation hiperplane*
PAT	*process analytical technology*
PCA	*principal component analysis*
PCR	*principal component regression*
PEV	ponto de ebulição verdadeiro
PIE	ponto inicial de ebulição
PLS	*partial least squares*
PNN	*probabilistic neural network*
QAV	querosene de aviação
QDA	*quadratic discriminant analysis*
QI	querosene de iluminação
RBF	*radial basis function*
RDA	*regularized discriminant analysis*
RMSEC	*root mean square error of calibration*
RMSEP	*root mean square error of prediction*
RON	*research octane number*
SAE	*society of automotive engineers*
SCR	*selective catalytic reduction*
SIMCA	*soft independent modeling of class analogy*
SNV	*standard normal variate*
SRM	*structural risk minimization*
SVC	*support vectors classification*
SVD	*singular value decomposition*
SVM	*support vectors machine*
SVR	*support vectors regression*
ULSD	*ultra low sulfur diesel*
WLS	*weighted least squares*

Índice de tabelas

Índice de figuras

Sumário

Seção I – Introdução

1 - Introdução

O monitoramento de parâmetros durante a produção e o controle de qualidade final de produtos derivados do petróleo é frequentemente realizada com a utilização de métodos espectroscópicos de análise combinados com algoritmos de regressão. O algoritmo de Mínimos Quadrados Parciais - PLS[1] é a técnica mais utilizada para tais fins, devido as suas características de simplicidade, velocidade de tratamento computacional, sua satisfatória performance para um grande número de problemas e a possibilidade da fácil interpretação dos *scores* e *loadings.* No entanto, o PLS apresenta deficiência para a modelagem de relações não lineares presentes nos conjuntos de dados.

Aplicações do algoritmo Máquina de Vetores de Suporte – SVM[2] vêm demonstrando maior eficácia num grande número de modelos de regressão e classificação que podem ser aplicados como ferramenta de Tecnologia Analítica de Processos – PAT[3] e controle de qualidade final. A construção de modelos de regressão e classificação por SVM pode ser bastante apropriada devido a possibilidade de modelar relações não lineares e com elevado poder de generalização.

Nesse trabalho propõe-se explorar o potencial do algoritmo SVM para o desenvolvimento de modelos de regressão e de classificação de múltiplas classes utilizando dados de espectroscopia no infravermelho próximo de combustíveis e lubrificantes para a solução de problemas relacionados a controle de processos e controle de qualidade.

Demonstra-se a aplicação do SVM para resolver problemas de regressão na determinação de alguns parâmetros de qualidade do óleo diesel e na quantificação do biodiesel em mistura com óleo diesel, utilizando dados de espectroscopia NIR, e comparam-se os resultados assim obtidos com os proporcionados por modelos construídos com o algoritmo PLS. Também demonstra-se a aplicação do SVM para resolver problemas de classificação de múltiplas classes com a utilização de dados de espectroscopia NIR de amostras de diferentes correntes que compõem o *pool* de óleo diesel de uma refinaria, para classificação de até sete classes, e na identificação de óleo naftênico e/ou óleo vegetal em misturas de óleos básicos e em lubrificantes de

motor automotivo, para classificação de até quatro classes, comparando-se os resultados com os obtidos com a utilização do algoritmo SIMCA.

Nos cinco diferentes problemas abordados aplicam-se a análise por espectroscopia NIR de combustíveis ou lubrificantes e o tratamento dos dados com a utilização do SVM para a obtenção de onze modelos de regressão (ν-SVR) e oito modelos de classificação (C-SVC e ν-SVC).

Esse trabalho organiza-se em três seções em que são abordados aspectos teóricos e resultados obtidos, como segue:

- Seção I – Introdução:
  - capítulo 1: motivações e objetivos do trabalho e revisão bibliográfica;
  - capítulo 2: conceitos sobre a espectroscopia NIR;
  - capítulo 3: conceitos relativos aos métodos quimiométricos utilizados;
  - capítulo 4: aspectos relevantes sobre a qualidade e processo produtivo do óleo diesel;
  - capítulo 5: aspectos sobre a utilização e qualidade do biodiesel; e
  - capítulo 6: aspectos sobre a produção, utilização e qualidade de óleos lubrificantes de motor automotivo.
- Seção II – Modelos de regressão:
  - capítulo 7: aplicação 1, que refere-se ao desenvolvimento de modelos de calibração mais eficazes para os parâmetros de qualidade ponto de fulgor e número de cetano do óleo diesel, para utilização no analisador *on line* que fornece informação ao sistema otimizador do misturador em linha no processo de produção de óleo diesel nas refinarias;
  - capítulo 8: aplicação 2, que trata do desenvolvimento de modelos de calibração mais eficazes para o monitoramento dos parâmetros de qualidade ponto de anilina, índice de cetano, temperatura de destilação para PIE, 50%, 85% e 90 % recuperados e densidade do óleo diesel que é carga do processo de hidrotratamento para produção de óleo diesel nas refinarias;
  - capítulo 9: é feita uma análise dos resultados obtidos com os nove modelos de regressão desenvolvidos utilizando SVM para otimização e controle de processos de produção de óleo diesel, mostrados nos dois capítulos anteriores; e

o capítulo 10: aplicação 3, que trata da determinação do teor de biodiesel misturado ao óleo diesel utilizando dados de espectroscopia NIR. Nesse trabalho realizaram-se estudos para obtenção de modelos de calibração mais eficazes nas faixas de concentração entre 0% e 35% e entre 0% e 100% de biodiesel adicionado a mistura, uma vez que na prática pode-se utilizar até 30% de biodiesel misturado ao óleo diesel comum sem comprometer a operação ou sem a necessidade de adaptações do motor e pode-se utilizar o biodiesel puro em motores adequadamente adaptados.

- Seção III – Modelos de classificação:

o capítulo 11: aplicação 4, que demonstra a utilização do SVM para classificação de dados de espectroscopia NIR de múltiplas classes de amostras provenientes de diferentes correntes que compõem o *pool* de óleo diesel da refinaria de Paulinia - REPLAN, incluindo-se nos modelos o diesel externo (produto de outra refinaria) e o diesel interno produto final na REPLAN. Nesse trabalho foram utilizadas até sete diferentes classes para desenvolver cinco modelos de classificação; e

o capítulo 12: aplicação 5, que trata do desenvolvimento de modelos de classificação para identificação da presença de óleo naftênico em teor além do comum e/ou de óleo vegetal em misturas de óleos básicos lubrificantes e em óleo lubrificante de motor automotivo.

1.1 – Métodos analíticos para amostras de petróleo e derivados

Embora os métodos de referência para análise de amostras de petróleo e derivados descritos em normas brasileiras (NBR-ABNT) e estrangeiras (ASTM; ISO; etc.) sejam conhecidos e amplamente aceitos, em muitos casos exibem algumas desvantagens tais como: custo operacional relativamente elevado, excessivo tempo de ensaio, utilização de elevado volume de amostra, utilização de reagentes e/ou solventes orgânicos, dificuldade ou impossibilidade para implementação de análises *on line*, etc. Alguns desses métodos envolvem também avaliações subjetivas, com elevada propensão a erros operacionais.

Por outro lado, a indústria do petróleo demanda métodos rápidos e eficazes, seja para análise em laboratório ou análise em linha de produção, como ferramenta de

Tecnologia Analítica de Processo – PAT, o que não é possível com a maioria dos métodos de referência.

Uma alternativa para contornar essas limitações é a utilização de técnicas instrumentais, espectroscópicas e cromatográficas, em conjunto com técnicas quimiométricas, para obtenção de modelos de regressão e classificação para amostras de petróleo e derivados.

Assim, a indústria do petróleo e derivados necessita constantemente buscar soluções no ramo da química analítica capazes de fornecer vantagens competitivas as suas operações. Seja para caracterização do petróleo em reservas para estimar a viabilidade da exploração, para caracterização do petróleo como matéria prima nas refinarias e adaptação das condições do processo de refino, ou no controle de qualidade dos derivados durante e após o processamento. Os avanços tecnológicos na química analítica vêm sendo explorados e agregam valor as operações dessa indústria.

Nesse sentido, as técnicas analíticas instrumentais, principalmente as técnicas espectroscópicas e cromatográficas têm encontrado amplo campo para utilização e vêm sendo regularmente exploradas.

Para extrair o máximo de informação analítica dessas técnicas na análise de amostras de petróleo e derivados, visando a obtenção de métodos eficazes e eficientes, a quimiometria vem ampliando sua aplicação nessa área nos últimos anos.

A complexidade da amostra e as características das técnicas analíticas, principalmente as espectroscópicas, determinam a utilização de algoritmos de calibração com habilidade para tratar complicados problemas de regressão e classificação, linear e não linear, em química do petróleo e derivados.

1.1.1 - Métodos quimiométricos lineares e não lineares: PLS, SIMCA, ANN e SVM

O método mínimos quadrados parciais – PLS[1] é atualmente o mais popular algoritmo quimiométrico para construção de modelos de calibração. Trata-se de um método linear de análise de dados capaz de modelar suaves não linearidades pela escolha adequada do número de variáveis latentes. Existem algumas modificações

desse algoritmo que o tornam apropriado para tratar dados não lineares: Poly-PLS e Spline-PLS[4]. A única diferença entre esses dois algoritmos e o PLS linear está em uma etapa, na qual a função linear é substituída por uma função polinomial (no poly-PLS) ou por uma função *spline*, uma junção de várias polinomiais lado a lado (no spline-PLS). As funções polinomial e *spline* podem ter qualquer ordem. Os cálculos implementados por tais algoritmos não lineares demandam elevado tempo de processamento, similar as redes neurais[4,5].

A ocorrência de uma relação não linear pode ser constata através dos resíduos produzidos pelo modelo desenvolvido com métodos lineares como o PLS. Constatada essa ocorrência o ideal é empregar métodos não lineares como as redes neurais artificiais (ANN)[6] e a máquina de vetores de suporte (SVM)[2,7].

Para os problemas de classificação o algoritmo SIMCA está entre os métodos supervisionados mais utilizados. Trata-se de um método paramétrico e linear em que a análise de componentes principais – PCA é realizada para o conjunto de dados e pode-se ajustar o adequado número de componentes principais utilizadas para cada modelo de classe. A ocorrência de elevados erros de classificação, amostras classificadas em mais de uma classe ou em nenhuma classe evidenciam a falta de ajuste do modelo e nesses casos métodos de classificação supervisionados não lineares, como o ANN e o SVM, podem proporcionar melhores resultados.

Algumas peculiaridades importantes dos métodos não lineares incluem:[6,8]

- Flexibilidade: por se tratar de um método não paramétrico, não requer a existência de um modelo definido, que possa ser entendido do ponto de vista químico. Essa é uma vantagem ainda maior no caso de calibrações com dados de espectroscopia NIR, visto que a excessiva sobreposição de bandas espectrais impede o desenvolvimento desse modelo bem definido. Entretanto, se não houver cuidado, a flexibilidade apresentada pode produzir um sobreajuste nos dados de calibração, resultando em modelos menos robustos.
- Robustez dos modelos desenvolvidos: As calibrações desenvolvidas por ANN ou SVM são bastante robustas no sentido de que seu desempenho sofre pequena alteração na presença de quantidades crescentes de ruído, ao contrário das calibrações desenvolvidas por PLS, por exemplo.

- Dificuldade de interpretação: Os modelos desenvolvidos por ANN ou SVM são de difícil interpretação, especialmente quando comparados a métodos lineares como o PLS. Isso se deve ao fato de que as diferentes operações realizadas sucessivamente nas diversas camadas da rede impedem a dedução de expressões analíticas simples entre os parâmetros de entrada e saída.
- Longo tempo de processamento requerido: as calibrações por ANN demandam maior tempo de processamento por se tratar de um método iterativo, especialmente para a solução de problemas difíceis em redes com configurações mais complexas. No SVM a otimização paramétrica necessária também pode demandar tempo de processamento significativo.

1.1.2 - Comparação entre os métodos quimiométricos

Uma análise crítica sobre os diferentes métodos quimiométricos utilizados para tratamento de dados analíticos de petróleo e derivados deve considerar as peculiaridades da natureza complexa desse tipo de amostra, das técnicas utilizadas para obter a informação analítica desejada e das técnicas quimiométricas utilizadas para explorar toda a informação relevante contida nos dados.

O petróleo é uma mistura complexa de hidrocarbonetos, principalmente n-alcanos, i-alcanos, cicloalcanos e aromáticos. Compostos contendo nitrogênio, enxofre e oxigênio, como resinas e asfaltenos, e traços de espécies metálicas, como vanádio, níquel e cobre também estão presentes. Devido a esta heterogeneidade, o petróleo obtido de diferentes campos de produção mostram características distintas e amplamente variadas. Além disso, o petróleo processado nas refinarias é normalmente uma mistura de petróleos de diferentes origens.

As frações obtidas do processamento do petróleo, carregam consigo algumas características do petróleo de origem e, além disso, frações semelhantes obtidas por diferentes processos também têm características diferentes. Normalmente essas frações idênticas provenientes de diferentes processos na refinaria são por fim misturadas adequadamente para a obtenção das especificações requeridas para o produto derivado.

A dificuldade em construir modelos que englobem toda a variabilidade presente nas amostras de petróleo e derivados além das particularidades dessas misturas que afetam os métodos espectroscópicos devem ser consideradas na construção de uma metodologia para determinação de parâmetros dessas amostras. É importante considerar a possibilidade de ter que tratar adequadamente elevado grau de não linearidade na relação do conjunto de dados com o parâmetro em estudo e a necessidade de prever amostras com características diversas das utilizadas no conjunto de calibração.

Métodos quimiométricos são frequentemente empregados para obtenção de melhores correlações de dados cromatográficos e principalmente espectroscópicos, sendo que os métodos lineares como ALS, MRL, PCR e PLS nem sempre proporcionam um modelo eficiente e eficaz. Métodos que implementam uma modelagem não linear dos dados como polynomial-PLS, spline-PLS, ANN e SVM podem, em muitos casos, proporcionar modelos de regressão mais eficazes.

Muitas estratégias vêm sendo utilizadas para a calibração de sistemas não lineares. Pode-se citar: pré-processamento dos dados (como a transformação dos dados e seleção de variáveis); o uso de métodos lineares (apenas para discretas não linearidades); o uso de modelagem local; a adição de variáveis extras; a utilização de técnicas de calibração não lineares. Dentre essas abordagens, a última é a única capaz de construir modelos de calibração robustos, uma vez que tal modelo de calibração tem o potencial de modelar não linearidades intrínsecas severas que podem ser encontradas em complexos (ex.: multicomponentes) sistemas.[5,9]

Estudos, comparativos ou não, utilizando métodos lineares e/ou não lineares para tratamento de dados analíticos de petróleo e derivados vêm sendo realizados e alguns deles são citados a seguir. Na maioria dos casos relatados na literatura ANN, SVM e/ou PLS não linear proporcionam resultados superiores ao PLS. De outra forma, para alguns conjuntos de dados o PLS proporciona resultados idênticos ou melhores que ANN.

Li *et al.*[10] utilizaram PLS e SVM para tratar um conjunto de dados de espectroscopia NIR obtidas de amostras de óleo diesel, com intuito de obter um modelo de regressão para a temperatura de destilação de 50 % recuperados (T50). O autor utilizou GA para realização de uma otimização paramétrica e obtenção dos parâmetros

ν e C do modelo SVM. O valor do RMSEP obtido utilizando o modelo SVM é 9,5 % menor em relação ao resultado obtido utilizando o algoritmo PLS. Os modelos proporcionaram bons coeficientes de determinação: R^2=0,9835 e R^2=0,9737, para os modelos SVM e PLS, respectivamente. Utilizando o mesmo conjunto de espectros de óleo diesel do trabalho anteriormente citado Zou *et al.*[11] realizaram a calibração de um parâmetro através de PLS e LS-SVM. Utilizando seleção de variáveis, os modelos PLS e LS-SVM obtiveram resultados semelhantes.

Santos Jr. *et al.*[12] realizaram um estudo para determinação de parâmetros de óleo diesel utilizando dados de espectroscopia FTIR, FTNIR, e FT-Raman aliado a métodos de regressão com PLS e ANN. Os parâmetros determinados foram índice de cetano (CI), densidade, viscosidade, temperatura de destilação para 50% e 85% recuperados (T50 e T85) e enxofre total. O algoritmo de regressão não linear proporcionou os melhores resultados para calibração de todos os parâmetros. Obtiveram-se valores de R^2 acima de 0,9 para todos os modelos utilizando ANN e as três técnicas espectroscópicas. Além disso, calculou-se a concordância dos resultados dos métodos propostos com os obtidos pelo método de referência, através do procedimento descrito pela norma ASTM E 1655-05, obtendo-se concordância para todos os modelos com ANN, exceto para o da densidade. Os modelos que utilizam PLS embora tenham obtido valores de R^2 abaixo de 0,7 para calibração de enxofre total e abaixo de 0,8 para a calibração do CI, também proporcionaram resultados passíveis de utilização, segundo os RMSEP obtidos e em comparação com as reprodutibilidades estabelecidas pelos métodos de referência.

Pasadakis *et al.*[13] propuseram a determinação da curva de destilação, ponto de névoa e ponto de fluidez do óleo diesel através de espectroscopia no infravermelho médio e ANN utilizando para determinação desses parâmetros os métodos de referência ASTM. O autor optou por desenvolver um modelo independente com ANN para cada um dos 21 pontos da curva de destilação, a fim de tentar melhorar a eficácia da previsão. Todos os modelos utilizaram o mesmo conjunto de treinamento e validação e mesma arquitetura, consistindo de entrada, duas camadas escondidas e uma camada de saída. A exatidão obtida por esses modelos são adequadas a repetibilidade exigida pelo método de referência. Para construção dos modelos para o ponto de névoa e ponto de fluidez utilizou-se como entrada da rede as 6 primeiras componentes

principais obtidas de uma análise por PCA dos espectros na região de 1700-600 cm^{-1}, de forma a reduzir significativamente a dimensão dos dados. Os valores de RMSEP obtidos para esse dois parâmetros ficaram abaixo do estabelecido pelo método de referência.

Balabin *et al.*[14] realizaram um estudo para classificação de amostras de gasolina utilizando dados de espectroscopia NIR e nove diferentes métodos de classificação multivariada: LDA, QDA, RDA, SIMCA, PLS, KNN, PNN, e ANN-MLP e SVM. Três conjuntos de espectros NIR foram utilizados para classificar amostras de gasolina em 3, 6 e 3 classes, de acordo com sua origem, processo de produção e tipo (segundo a especificação russa). Os conjuntos possuiram 150, 117 e 115 amostras, respectivamente. Com base na eficácia de classificação obtida nesse estudo os métodos podem ser ordenados da seguinte forma:

PNN≥SVM≥KNN>>ANN-MLP≥PLS>SIMCA≥RDA≈LDA≈QDA

Assim, uma vez que o conjunto de dados em estudo contém certo grau de não linearidade na classificação, associada ao tipo de amostra e técnica utilizada, os métodos não lineares e com maior poder de generalização tiveram melhor desempenho. Com base nesses resultados é importante ressaltar que na escolha do método apropriado para as análises desse tipo de dados não só a eficácia do método deve ser considerada, mas também o tempo para desenvolver o modelo e a necessidade de experiência e treinamento do pesquisador para trabalhar com métodos como PNN, ANN-MLP e SVM. Desse modo, o excelente desempenho do método que utiliza KNN pode ser considerado o mais eficiente.

O trabalho realizado por Brudzewski *et al.*[15] demonstra a utilização de SVM para classificar amostras de gasolina em 6 classes baseadas em intervalos do parâmetro RON. Nesse caso utilizou-se espectroscopia NIR e obteve-se eficácia de 100% na classificação das amostras. Considerando a complexidade do conjunto de dados, verifica-se a habilidade do algoritmo SVM para classificação desse complicado problema de classificação não linear.

Os estudos acima mencionados mostram que embora métodos que utilizem o PLS para tratar dados de espectroscopia FT-IR, FT-NIR ou FT-Raman obtenham resultados razoáveis, a utilização de algoritmos não lineares como ANN e SVM proporcionam a obtenção de modelos de regressão ou classificação mais eficazes. No

entanto deve-se considerar que tais métodos demandam maior tempo de processamento computacional e necessitam pré-processamento adequado dos dados.

Falla *et al.*[16] propuseram um método para determinação da curva PEV de petróleo através de espectroscopia NIR e calibração com ANN. Nesse caso, foi estabelecida a correlação com os valores obtidos através de destilação simulada por GC, segundo método ASTM de referência. As variáveis de entrada para a rede neural são os espectros de absorção, em 20 comprimentos de onda no intervalo de 5600 a 6000 cm^{-1} e o grau API. As saídas são os percentuais em massa em 20 temperaturas, entre 50℃ e 647℃. Os valores de RMSEP obtidos var iam de 0,66% a 3,5% em massa, o que segundo o autor indica pequena diferença em comparação com os valores medidos. O valor de R^2 calculado foi de 0,9931, indicando a boa correlação estabelecida pelo modelo.

Métodos cromatográficos são também amplamente utilizados para análise de petróleo e derivados com[15] ou sem[17] a utilização de técnicas quimiométricas.

É importante lembrar que para uma comparação entre a eficácia de determinados métodos quimiométricos para tratamento de dados analíticos de petróleo e derivados e seleção de um método para utilização devemos considerar: número de amostras disponíveis; características dos dados analíticos; sensibilidade desejada; exigência de experiência e treinamento do analista; espaço amostral; origens das amostras; algoritmo quimiométrico utilizado; entre outros.

Em função dessas condições, podem ser obtidos bons resultados para um mesmo tipo de amostra, mesmo parâmetro e mesmo algoritmo de calibração para um conjunto de dados e resultados ruins com outro conjunto de dados (ao utilizar diferentes técnicas analíticas ou amostras de origens diferentes, etc.).

O desenvolvimento de um método quimiométrico para tratamento de dados analíticos de petróleo e derivados deve iniciar por um estudo a partir de métodos lineares, verificando a existência de não linearidade na correlação ou classificação dos dados e a aceitabilidade dos resultados e caso não seja o ideal, deve-se então partir para utilização de métodos não lineares.

De uma forma geral, os métodos não lineares proporcionam maior eficácia no tratamento de dados analíticos de petróleo e derivados, uma vez que a característica de aproximadores universais permite um melhor tratamento para esse tipo de caso. Ainda,

considerando a frequente necessidade de previsão de amostras com características diversas das utilizadas no conjunto de calibração, a utilização de um algoritmo com excelente capacidade de generalização e capaz de evitar sobreajustes, como SVM, pode proporcionar modelos mais eficazes.

2 - Espectroscopia no infravermelho próximo

A espectroscopia no infravermelho próximo (NIR) utiliza a energia do fóton (hν) no intervalo de $2{,}65.10^{-19}$ a $7{,}96.10^{-20}$ J, o que corresponde ao intervalo de comprimento de onda entre 750 e 2500 nm (13300 a 4000 cm^{-1}). A energia situada nesse intervalo é mais do que o necessário para promover transições de modos vibracionais de energia de ligações químicas desde seu estado vibracional menos energético e menos do que a energia tipicamente necessária para excitar elétrons em uma molécula.

Os métodos analíticos que utilizam a espectroscopia NIR mostram como características mais importantes: simplicidade e rapidez de análise (menos de um minuto por amostra); não destrutiva; boa penetração do feixe de radiação; possibilidade de análise on-line, in-line e off-line; aplicável a qualquer molécula contento ligações C-H, N-H, S-H ou O-H; demanda mínima ou nenhuma preparação da amostra; boa reprodutibilidade; elevada razão sinal ruído; custo relativamente baixo.

O espectro NIR consiste de sobretons e bandas de combinação de absorção de modos vibracionais fundamentais presentes na região do infravermelho médio e o espectro consiste de bandas de absorção sobrepostas e mal resolvidas. Assim, técnicas quimiométricas são comumente empregadas para realização de análises qualitativas e quantitativas.

Uma descrição detalhada sobre os princípios da espectroscopia no infravermelho próximo e instrumentação utilizada pode ser obtida na literatura.[18-21] Cabe aqui ressaltar os principais aspectos relativos a utilização da espectroscopia NIR na análise de amostras de petróleo e derivados.

Observa-se que a maioria dos métodos analíticos desenvolvidos para petróleo e derivados utiliza dados obtidos com técnicas espectroscópicas, com ampla predominância de espectroscopia FT-IR e FT-NIR, como exemplificado nos artigos citados no capítulo 1 desse texto.

A indústria do petróleo tem se beneficiado grandemente da espectroscopia NIR, especialmente devido às informações espectrais correspondentes às vibrações das ligações C-H. A figura 2.1 exemplifica a capacidade da espectroscopia NIR em diferenciar hidrocarbonetos de fórmulas parecidas e/ou com diferentes tipos de cadeias carbônicas. São mostrados espectros de quatro diferentes hidrocarbonetos: um de

cadeia carbônica alifática ou parafínico (octano), um de cadeia carbônica ramificada ou iso-parafínico (pentano,2,2,4-trimetil), um de cadeia carbônica aromática (benzeno) e um de cadeia carbônica cíclica (ciclohexano), no intervalo de 10500 a 4000 cm^{-1} (953 a 2500 nm), região espectral na qual há ocorrência das bandas de combinação, primeiro sobreton, segunda região de bandas de combinação (ou sobreton de combinação) e segundo sobreton.[18,19]

As bandas de absorção no infravermelho próximo apresentam menor resolução e maior sobreposição quando comparadas aos sinais obtidos com infravermelho médio. Além disso, para análise de amostras de petróleo cru é difícil obter boa informação por sinais de transmitância no NIR, sendo comumente empregada a técnica ATR-FTIR.

É importante ter em mente que os mesmos fatores limitantes encontrados em outras regiões espectrais restrigem a concordância com a Lei de Beer em sistemas reais em um extenso intervalo analítico. Esses fatores limitantes têm origens instrumentais e/ou são características dos constituintes das amostras, como por exemplo, a não linearidade do sistema de detecção e mudanças nos padrões das ligações de hidrogênio conforme as concentrações das espécies presentes na amostra sofrem mudanças relativas de concentrações.[19]

Variações nas ligações de hidrogênio manifestam-se como mudanças nas constantes de força das ligações X-H. Geralmente as bandas deslocam em freqüência e largura devido a formação das ligações de hidrogênio. Como as bandas de combinação resultam da soma de dois ou mais modos vibracionais fundamentais, e sobretons ocorrem como resultado de múltiplos das vibrações fundamentais, os deslocamentos de freqüência relacionados às ligações de hidrogênio têm um maior efeito relativo nas bandas de combinação e sobretons do que nas bandas dos modos de vibração fundamentais. Essa característica da região espectral do infravermelho próximo alerta para a importância dos efeitos relativos das ligações de hidrogênio devido a variações no solvente e temperatura.

A precisa atribuição de bandas é dificultada na região do NIR porque uma única banda pode ser atribuída a muitas possíveis combinações de vibrações fundamentais e sobretons, todas severamente sobrepostas. A influência do aumento das ligações de hidrogênio resulta em deslocamentos de bandas para freqüências menores e a diminuição das ligações de hidrogênio devido a diluição e maiores temperaturas resulta

em deslocamentos de bandas para freqüências mais elevadas. Deslocamentos de bandas com magnitude de 10-100 cm^{-1}, correspondente a pouco menos de 50 nm, podem ser observados. O importante efeito das ligações de hidrogênio deve ser considerado quando da preparação do conjunto de amostras de calibração para desenvolvimento do método analítico.

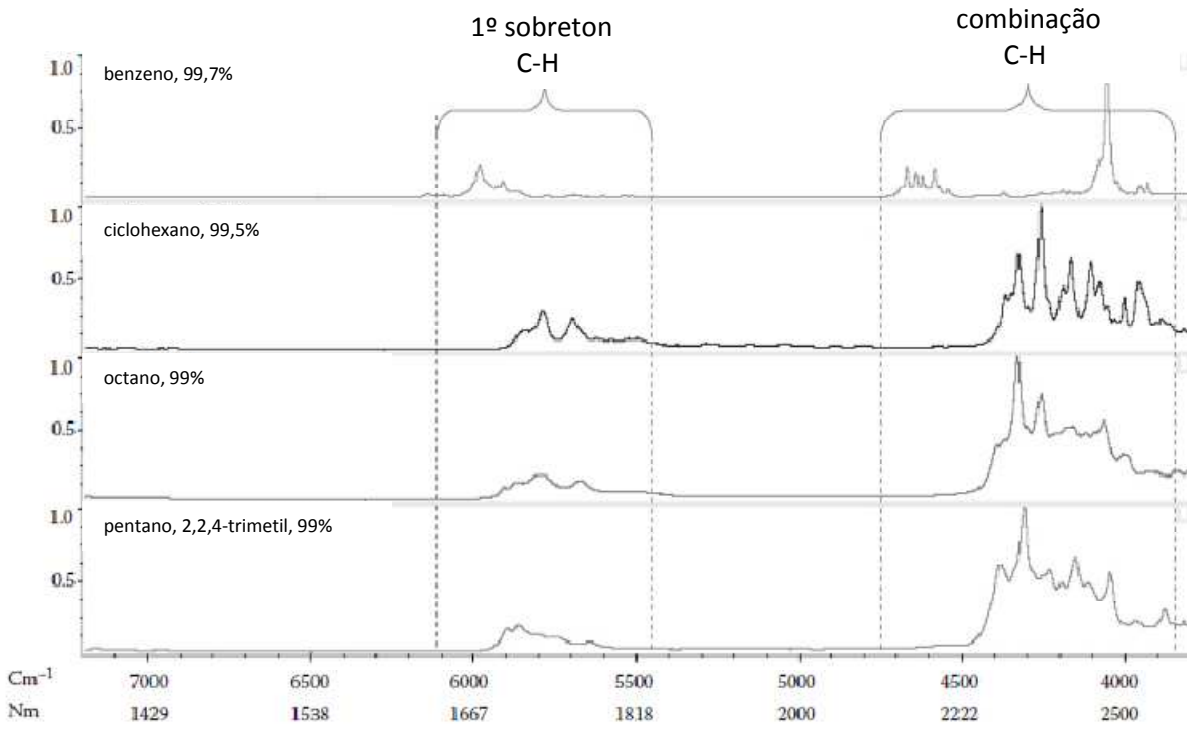
Para melhor aproveitar a informação analítica oferecida pela espectroscopia NIR e MIR, inclusive tratando adequadamente possíveis relações não lineares, associadas a fatores extrínsecos como o efeito causado pela variação na temperatura da amostra ou a fatores intrínsecos como os desvios da Lei de Beer,[4,9,21] diferentes pré-tratamentos dos dados e métodos quimiométricos lineares e não lineares, para regressão e classificação são utilizados.[9,22,23]

2.1 – Atribuições de bandas de absorção no NIR

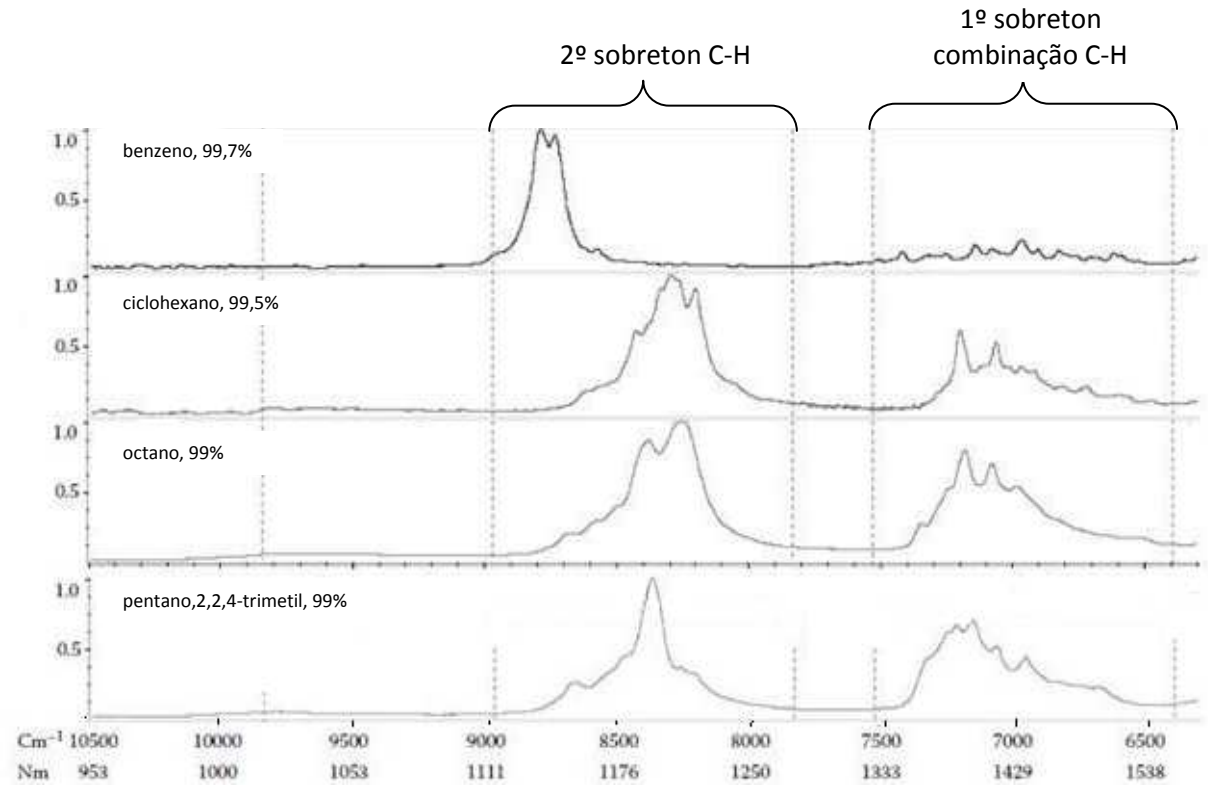
As regiões de primeiro e segundo sobretons dos modos de vibração das ligações C-H de compostos de cadeia linear, cíclica e aromática possuem enorme conteúdo de informação, sendo muito da informação observada também na região dos modos de vibração fundamental. Além disso, as combinações dessas vibrações com outros modos vibracionais contribuem muito com a estrutura do espectro na região do infravermelho próximo.

Considera-se que a região de ocorrência do primeiro sobreton do modo vibracional de estiramento C-H está entre 5555 e 5882 cm^{-1} (1800 a 1700 nm), a região de segundo sobreton entre 8264 e 8696 cm^{-1} (1210 a 1150 nm) e a região do terceiro sobreton entre 10929 e 11364 cm^{-1} (915 a 880 nm). As regiões de ocorrência das mais importantes bandas de combinação de modos vibracionais C-H são entre 4000 e 4545 cm^{-1} (2500 a 2200 nm) e também entre 6666 e 7690 cm^{-1} (1500 a 1300 nm).[18]

Nos hidrocarbonetos parafínicos de cadeia linear, os grupos metila estão presentes nas extremidades da cadeia carbônica e os grupos metileno na parte interna



(a)



(b)

Figura 2.1 – Espectros NIR de hidrocarbonetos parafínico, isoparafínico, cíclico e aromático. (a) região das bandas de combinação e primeiro sobreton e (b) segunda região de bandas de combinação e segundo sobreton (adaptado de J. Workman Jr, L.Weyer[18])

da cadeia. Nos hidrocarbonetos de cadeia ramificada uma proporção maior de grupos metila é observada, em relação ao composto de cadeia carbônica linear correspondente. As variações quanto à intensidade das bandas de absorção no NIR encontradas nessas classes de hidrocarbonetos são atribuídas a alterações nas proporções entre os grupos metila e metileno.

De outra forma, os espectros de absorção no infravermelho próximo de hidrocarbonetos aromáticos têm bandas de absorção que não são encontradas nos espectros dos compostos parafínicos, conforme ilustra a figura 2.1.

A tabela 2.1 cita algumas das atribuições das bandas de absorção dos modos vibracionais das ligações C-H de alcanos e cicloalcanos, ligações C-H e C=C de alcenos e cicloalcenos e ligações C-H de aromáticos, nas diferentes regiões do espectro infravermelho próximo utilizadas nesse trabalho.[18]

Entre as classes de hidrocarbonetos de cadeia linear e cadeia linear ramificada, boas correlações entre as intensidades dos picos e o percentual de grupos funcionais metil e metileno presentes são obtidas. Em geral, a absortividade do grupo metil terminal é maior do que para o grupo metil interno ou ramificado.[24]

As bandas de absorção na região do infravermelho próximo de diversos hidrocarbonetos puros mostram que há diferenças significativas entre as bandas de absorção na região de combinação e nas regiões de sobretons. Como exemplo pode-se citar a menor sobreposição das bandas de absorção atribuídas ao estiramento C-H do anel aromático na região de combinação e segundo sobreton em relação àquelas observadas na região de primeiro sobreton, conforme pode ser observado nos espectros da figura 2.1. Essas características podem ser de grande importância na construção de modelos quimiométricos para determinação de parâmetros físico-químicos, onde os teores de hidrocarbonetos de cadeia linear, ramificada, cíclica, aromática e insaturada têm influência.

Para determinação do teor de biodiesel em óleo diesel e de óleo vegetal em lubrificantes além das bandas de ligação C-H e C=C anteriormente citadas é necessário a utilização das bandas atribuídas a ligação C=O de éster.

A capacidade da espectroscopia NIR em diferenciar entre o óleo diesel A e o biodiesel, bem como entre os óleos básicos parafínico, naftênico e óleo vegetal é ilustrada pelos espectros mostrados nas figuras 2.2 e 2.3, respectivamente.

Tabela 2.1 – Atribuições de bandas de vibração de alcanos e cicloalcanos na região do NIR (adaptado de J. Workman Jr, L.Weyer[18]) *não citado na literatura consultada

ligação C-H - alcanos e cicloalcanos				
Região espectral do NIR	comprimento de onda (nm)	número de onda (cm^{-1})	grupo funcional	vibrações
Bandas de combinação 4000 - 4545 cm^{-1} 2500 - 2200 nm	2220	4545	metileno, ciclopropano	C-H (ν) + C-H (δ)
	2270	4400	metil	C-H ($3\delta_a$)
	2275	4395	metil	C-H (ν) + C-H (δ)
	2306	4336	metileno, butano	C-H (ν_a) + C-H (δ)
	2307	4334	metileno, pentano	C-H (ν_a) + C-H (δ)
	2308	4332	metileno, heptano	C-H (ν_a) + C-H (δ)
	2346	4262	metileno, pentano	C-H (ν_s) + C-H (δ)
	2347	4259	metileno, heptano	C-H (ν_s) + C-H (δ)
	2349	4257	metileno, butano	C-H (ν_s) + C-H (δ)
	2363	4232	metileno, cadeia ramificada	C-H ($2\nu_s$) + C-H (δ)
	2439	4100	metil	C-H ($3\delta_s$)
	2458	4068	metil, cadeia linear alifática	C-H (ν) + C-H (3δ)
	2470	4049	metil, cadeia ramificada alifática	C-H (ν) + C-H (3δ)
Primeiro sobreton 5555 - 5882 cm^{-1} 1800 – 1700 nm	1630	6135	metileno, ciclopropano	C-H ($2\nu_s$)
	1693	5905	metil, alifático terminal	C-H (2ν)
	1701	5876	metil, alifático terminal	C-H (2ν)
	1702	5872	metil, alifático ramificado	*
	1714	5834	metileno, ciclopentano	*
	1723	5800	metileno, alifático linear	C-H ($2\nu_a$) + C-H ($2\nu_s$)
	1727	5791	metileno, ciclohexano	C-H ($2\nu_s$)
	1745	5730	metileno, ciclopentano	*
	1755	5697	metileno, ciclohexano	C-H ($2\nu_s$)
	1762	5680	metileno, alifático linear	C-H ($2\nu_s$)
Bandas de combinação (segunda região) 6666 - 7690 cm^{-1} 1500 – 1300 nm	1391	7186	metileno	C-H (2ν) + C-H (δ)
	1400	7100	metil	C-H (2ν) + C-H (δ)
	1412	7080	metileno	C-H (2ν) + C-H (δ)
	1440	6938	metil	C-H (2ν) + C-H (δ)
				*
Segundo sobreton 8264 - 8696 cm^{-1} 1210 - 1150 nm	1097	9116	metileno, ciclopropano	C-H (3ν)
	1153	8673	metileno	*
	1176	8503	metileno	*
	1185	8434	metileno,ciclopentano	C-H (3ν)
	1191	8396	metil, pentano ou hexano	*
	1192	8388	metil, heptano	*
	1194	8378	metil, decano	*
	1195	8365	metil, alcano cadeia longa	*
	1206	8290	metileno, ciclohexano	C-H (3ν)
	1207	8284	metileno, pentano	C-H (3ν)
	1209	8271	metileno, hexano	C-H (3ν)
	1210	8264	metileno	*
	1211	8256	metileno, heptano	C-H (3ν)
	1212	8247	metileno, decano	C-H (3ν)

Tabela 2.2 – Atribuições de bandas de vibração de alcenos e cicloalcenos na região do NIR (adaptado de J. Workman Jr, L.Weyer[18]) *não citado na literatura consultada

ligação C-H e C=C - alcenos e cicloalcenos				
Região espectral do NIR	comprimento de onda (nm)	número de onda (cm^{-1})	grupo funcional	vibrações
Bandas de combinação 4000 - 4545 cm^{-1} 2500 - 2200 nm	1667	6000	dupla ligação, ciclopenteno	*
	2090	4670	dupla ligação C-H, vinil	C-H (ν) + C-H (δ)
	2140	4673	dupla ligação cis, vinil	*
	2140	4670	dupla ligação, ciclohexeno	*
	2145	4660	dupla ligação, ciclopenteno	*
	2170	4600	dupla ligação C-H, vinil	C-H (ν) + C-H (δ)
	2180	4587	dupla ligação cis, vinil	*
	2230	4482	dupla ligação C-H, vinil	C-H (ν) + C-H (δ)
Primeiro sobreton 5555 - 5882 cm^{-1} 1800 – 1700 nm	1620	6173	dupla ligação C-H, vinil	C-H (2ν)
	1629	6139	aleno, $H_2C=C=CH_2$	C-H (2ν)
	1635	6120	dupla ligação, hexeno	C-H (2ν)
	1677	5963	dupla ligação cis, vinil	C-H (2ν)
	2100	4761	dupla ligação C-H, vinil	C-H (2ν)
Bandas de combinação (segunda região) 6666 - 7690 cm^{-1} 1500 – 1300 nm	1290	7750	dupla ligação C	*
	1360	7346	dupla ligação C	*
Segundo sobreton 8264 - 8696 cm^{-1} 1210 - 1150 nm	1080	9260	metileno terminal	C-H (3ν)
	1140	8777	dupla ligação, cadeia cíclica	*
	1160	8620	aleno, $H_2C=C=CH_2$	C-H (3ν)
	1180	8474	dupla ligação cis, vinil	C-H (3ν)

Tabela 2.3 – Atribuições de bandas de vibração de aromáticos na região do NIR (adaptado de J. Workman Jr, L.Weyer[18]) *não citado na literatura consultada

ligação C-H - aromáticos				
Região espectral do NIR	comprimento de onda (nm)	número de onda (cm^{-1})	grupo funcional	vibrações
Bandas de combinação 4000 - 4545 cm^{-1} 2500 - 2200 nm	2146	4660	C-H aromático, benzeno	C-H (ν) + C-C (ω)
	2154	4642	C-H aromático, benzeno	C-H (ν) + C-C (ν)
	2167	4615	C-H aromático, benzeno	C-H (ν) + C-C (ν)
	2188	4570	C-H aromático, benzeno	C-H (ν) + C-C (ν)
	2206	4532	C-H aromático, benzeno	C-H (ν) + C-C (ν)
	2300	4300	metil, tolueno	C-H (ν) + C-C (ν)
	2387	4190	C-H aromático, benzeno	C-H (ν) + C-C (δ)
	2407	4155	C-H aromático, benzeno	C-H (ν) + C-H (δ)
	2440	4099	C-H aromático, benzeno	C-H (ν) + C-C (δ)
	2469	4050	C-H aromático, benzeno	C-H (ν) + C-C (δ)
	2513	3980	C-H aromático, benzeno	C-H (ν) + C-C (δ)
	2525	3960	C-H aromático, benzeno	C-H (ν) + C-C (ω)
Primeiro sobreton 5555 - 5882 cm^{-1} 1800 – 1700 nm	1671	5985	C-H aromático, benzeno	C-H (ν) + C-H (ν)
	1680	5952	C-H aromático, benzeno	C-H (2ν)
	1689	5920	C-H aromático, benzeno	C-H (ν) + C-H (ν)
	1736	5760	metil, ligado a aromático	C-H (2ν)
	1744	5735	metil, ligado a aromático	C-H (2ν)
	1770	5650	metil, ligado a aromático	C-H (2ν)
Bandas de combinação (segunda região) 6666 - 7690 cm^{-1} 1500 – 1300 nm	1370	7300	metil, ligado a aromático	C-H (2ν) + C-H (δ)
	1417	7057	C-H, aromático	*
	1446	6916	C-H, aromático	*
Segundo sobreton 8264 - 8696 cm^{-1} 1210 - 1150 nm	1132	8834	C-H, aromático	C-H (3ν)
	1142	8754	C-H, aromático	C-H (3ν)
	1143	8749	C-H, aromático	C-H (3ν)
	1145	8734	C-H, aromático alquilado	C-H (3ν)

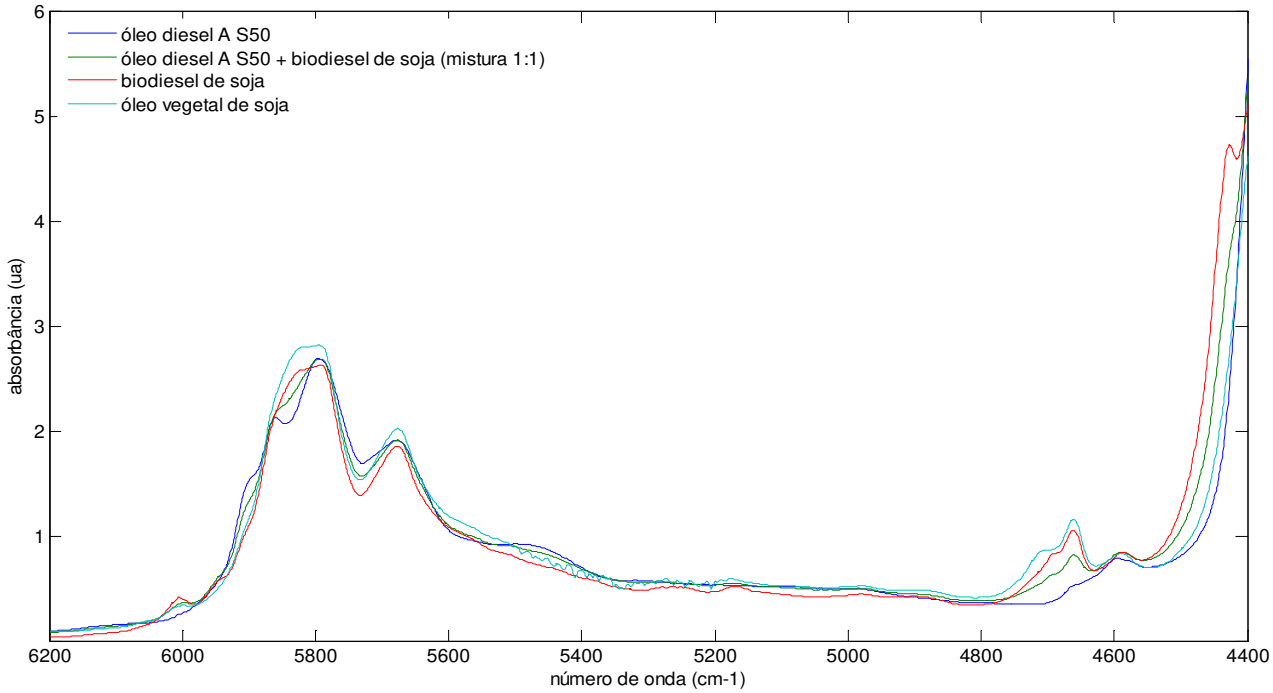


Figura 2.2 – Espectros NIR característicos de óleo diesel, biodiesel e óleo vegetal

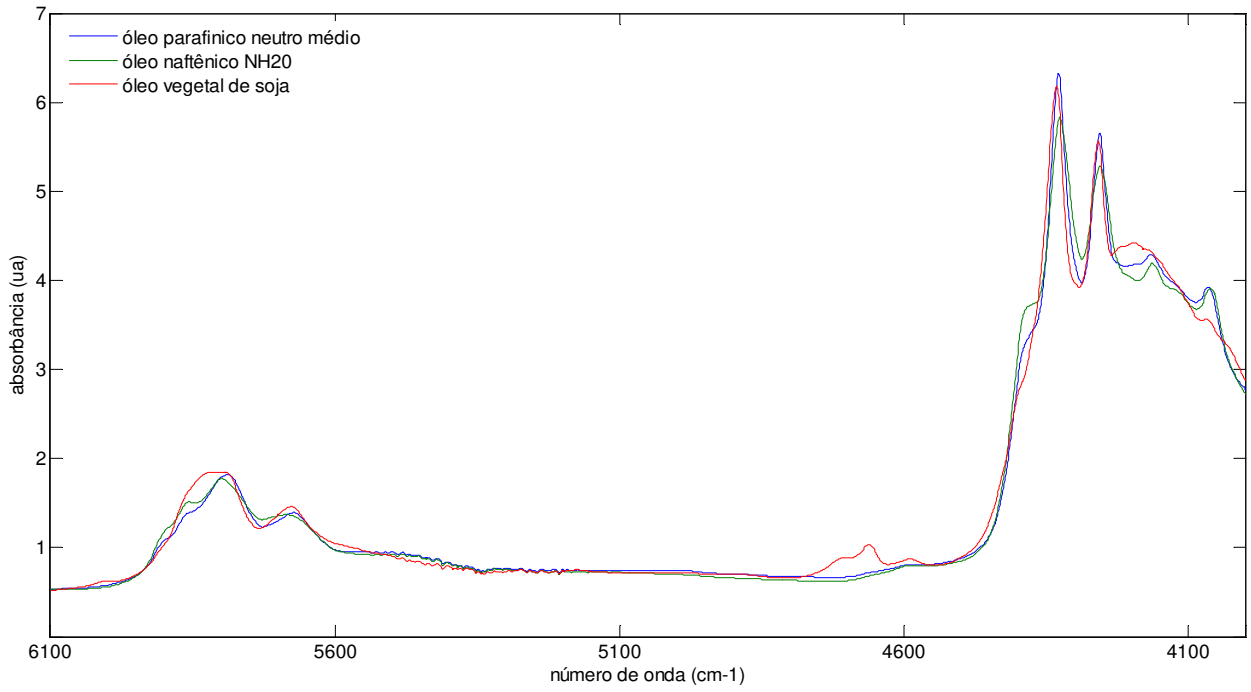


Figura 2.3 – Espectros NIR característicos de óleo parafínico, óleo naftênico e óleo vegetal

A banda atribuída ao estiramento C=O de éster é muito forte na região do infravermelho médio e inclusive o primeiro sobreton ocorre na região do MID. O segundo sobreton pode ser observado na região do NIR em 5160 cm^{-1} (1940 nm) e uma banda de combinação do estiramento C=O com estiramento C-H ocorre em 4650 cm^{-1} (2150 nm).[18]

A espectroscopia NIR possibilita diferenciar entre triacilglicerídeos e seus correspondentes ésteres metílicos, os quais apresentam duas regiões que possibilitam essa distinção. A espectroscopia NIR foi anteriormente investigada[25,26] para monitorar a reação de transesterificação na produção de biodiesel e a base para quantificação foi a peculiaridade dos espectros NIR nas regiões de 6005 cm^{-1} (1660 nm) e 4425 cm^{-1} (2250 nm) onde os ésteres metílicos apresentam picos, provavelmente relacionados ao grupo metílico terminal adicionado na reação de transesterificação, enquanto os triacilglicerídeos exibem apenas ombros. O emprego da absorção em 6005 cm^{-1}, ao invés daquela observada próximo a 4425 cm^{-1}, forneceu os melhores resultados quantitativos em estudos anteriores. As mesmas regiões também proporcionam distinguir entre ésteres metílicos e etílicos em função das diferentes formas das bandas observadas nessas regiões.[25,26]

3 – Métodos quimiométricos para regressão e classificação

A Calibração Multivariada é um meio para construir modelos de previsão em que se considera a correlação entre muitas variáveis analisadas simultaneamente, permitindo a extração de uma grande quantidade de informação, podendo fornecer melhores resultados em relação a utilização de uma única variável x quando esta não proporciona seletividade suficiente para uma boa previsão do parâmetro de interesse y. Assim, em vez de utilizar dados e modelos de calibração de ordem zero, a utilização de dados de primeira ordem, obtidos com técnicas analíticas que podem fornecer um vetor de dados para cada amostra, permite desenvolver modelos com melhor desempenho.

Essa informação pode ser comprimida, com a redução da dimensionalidade dos dados, de modo a reter a informação essencial e facilitar a sua utilização. A análise de componentes principais – PCA[27] trata desse problema e é base para os algoritmos de regressão e classificação mais comumente utilizados atualmente.

A seguir será feita uma descrição sucinta dos métodos quimiométricos de referência utilizados nesse trabalho para regressão e classificação, PLS e SIMCA, respectivamente. Conceitos básicos dos algoritmos PLS e SIMCA são apresentados, iniciando com a PCA.

Após, é apresentado o algoritmo SVM, abordando alguns aspectos teóricos do seu desenvolvimento e funcionamento. No algoritmo SVM o aumento da dimensionalidade dos dados é a base para o seu atrativo desempenho.

3.1 - Análise de Componentes Principais – PCA

A análise de componentes principais – PCA[27] tem por finalidade básica a redução da dimensionalidade de dados a partir de combinações lineares das variáveis originais. O PCA decompõe uma matriz de dados **X** (onde as *m* linhas são as amostras e as *n* colunas, as variáveis) de posto (*rank*) *h*, em uma soma de h matrizes de posto igual a 1, como na equação 3.1:

$$\boldsymbol{X} = \boldsymbol{M_1} + \boldsymbol{M_2} + \boldsymbol{M_3} + \ldots + \boldsymbol{M_h} \qquad (3.1)$$

onde o posto expressa o número de vetores linearmente independentes de uma matriz. Essas novas matrizes de posto 1, são produtos de vetores chamados escores, $\mathbf{t_h}$, e pesos, $\mathbf{p_h}$. Estes escores e pesos podem ser calculados por um ajuste de mínimos quadrados. A operação é equivalente ao cálculo de autovetores e autovalores de uma matriz pela Decomposição em Valores Singulares – SVD. A equação pode ser representada na forma vetorial,

$$\boldsymbol{X} = \boldsymbol{t_1 p_1'} + \boldsymbol{t_2 p_2'} + \ldots + \boldsymbol{t_h p_h'} \qquad (3.2)$$

e na forma matricial,

$$\boldsymbol{X} = \boldsymbol{T P'} \qquad (3.3)$$

Para exemplificar $\mathbf{t_h}$ e $\mathbf{p'_h}$, a figura 3.1 ilustra no plano bidimensional duas variáveis x_1 e x_2. A figura 3.1 (a) mostra uma componente principal (CP), a reta que aponta na direção de maior variabilidade das amostras. Os escores $\mathbf{t_h}$ são as projeções das amostras na direção da CP e os pesos $\mathbf{p'_h}$ são os cossenos dos ângulos formados entre a CP e cada variável original.

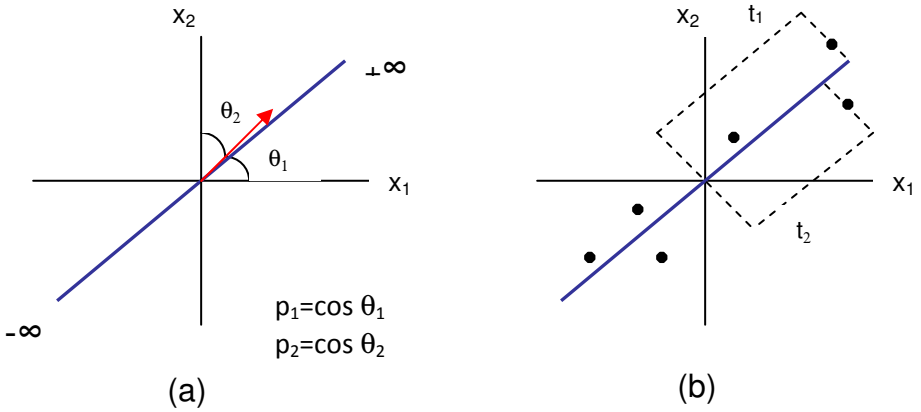


Figura 3.1. Representação de uma Componente Principal (CP) no caso de duas variáveis: (a) os pesos são os cossenos dos ângulos do vetor direção; (b) os escores são as projeções das amostras na direção da CP (os dados estão centrados na média).

As novas variáveis, as componentes principais (CP), são ortogonais entre si e, portanto, não correlacionadas. Normalmente, as primeiras CP explicam a maior parte da variância total contida nos dados e podem ser usadas para representá-los. A Análise de Fatores é adotada em boa parte da literatura como sinônimo de PCA. Porém alguns

autores definem esses termos como métodos diferentes, com base no modo como os fatores (ou CP's) são definidos. Na PCA os fatores devem explicar o máximo da variância contida em todas as variáveis observadas, enquanto que na Análise de Fatores, eles devem explicar o máximo da intercorrelação entre as variáveis. Neste trabalho, de acordo com a maioria da literatura, ambos os termos, fatores e CP, serão tratados como equivalentes.[1,28,29]

3.2 – Mínimos quadrados parciais - PLS

A Calibração Multivariada tem como princípio básico a utilização simultânea de muitas variáveis $x_1,x_2,...x_n$ (como valores de intensidade de absorbância a vários comprimentos de onda), para quantificar alguma outra variável de interesse y (como concentração). O mínimos quadrados parciais – PLS[1], é o método mais usado em calibração multivariada e utiliza a informação de *y* no cálculo das chamadas variáveis latentes. As matrizes **X** (relacionada aos espectros) e **Y** (relacionada às concentrações) são decompostas simultaneamente em uma soma de L componentes principais, como nas equações a seguir:

$$\boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E} = \sum_{i=1}^{L} \boldsymbol{t_i}\, \boldsymbol{p_i'} + \boldsymbol{E} \qquad (3.4)$$

$$\boldsymbol{Y} = \boldsymbol{UQ}' + \boldsymbol{F} = \sum_{i=1}^{L} \boldsymbol{u_i}\, \boldsymbol{q_i'} + \boldsymbol{F} \qquad (3.5)$$

onde, as matrizes **E** e **F** contém a informação de **X** e **Y**, respectivamente, que não é explicada pelo modelo; **T** e **U** são as matrizes de escores; **P** e **Q** as matrizes de pesos e L é o número de variáveis latentes utilizadas pelo modelo.

A correlação entre os dois blocos **X** e **Y** é simplesmente uma relação linear obtida pelo coeficiente de regressão linear **b**, tal como descrito abaixo:

$$\boldsymbol{u_i} = \boldsymbol{b_i t_i} \qquad (3.6)$$

para L variáveis latentes, sendo que os valores de $\mathbf{b_i}$ são agrupados na matriz diagonal **B**, que contém os coeficientes de regressão entre a matriz de escores **T** e **U**. A melhor relação linear possível entre os escores desses dois blocos é obtida através de

pequenas rotações das variáveis latentes dos blocos de **X** e **Y**. A matriz **Y** pode ser calculada de $\mathbf{u_i}$,

$$\boldsymbol{Y} = \boldsymbol{TBQ}' + \boldsymbol{F} \quad (3.7)$$

e a concentração de novas amostras prevista a partir dos novos escores, **T***, substituídos na equação acima

$$\boldsymbol{Y} = \boldsymbol{T}^{*}\boldsymbol{BQ}' \quad (3.8)$$

Nesse processo é necessário identificar o melhor número de variáveis latentes, o que normalmente é feito através do procedimento de validação cruzada, para obtenção de um baixo erro de previsão.

3.3 – *Soft independent modeling of class analogy* - SIMCA

Frequentemente, conjuntos de dados consistem de amostras que pertencem a várias diferentes classes. Classes podem diferir de diversos modos, incluindo diferenças nos tipos de compostos químicos (aromático, alifático, carbonílicos, etc.) entre outros.

O SIMCA[30-32] é um método de classificação ou um método de reconhecimento de padrões supervisionado e paramétrico. Um método supervisionado necessita de um conjunto de dados consistindo de amostras com seus atributos (variáveis) e suas respectivas classes. Uma distinção pode ser feita entre técnicas de reconhecimento de padrões (supervisionado) que consideram a informação sobre a distribuição da população e aquelas que não consideram. As técnicas não paramétricas (ex.: KNN) não fazem suposição sobre a distribuição da população (ou seja, não levam em consideração informação sobre a distribuição da população) enquanto técnicas paramétricas (ex.: SIMCA) fazem. As técnicas paramétricas são baseadas em uma distribuição bem definida.

No modelo SIMCA uma classe consiste em um plano ou hiperplano linearmente definido e restrito no espaço. Os dados de cada objeto podem ser definidos como:

$$\boldsymbol{X}_k = \boldsymbol{M}_q + \boldsymbol{E}_{kq} \quad (3.9)$$

em que $\boldsymbol{M}_q$ é a parte que pode ser explicada pelo modelo para a classe q , ou seja, a parte determinística, e $\boldsymbol{E}_{kq}$ é a parte não explicada pelo modelo para a classe q , ou seja, a parte devida aos fatores que não são controlados pelo modelo. $\boldsymbol{E}_{kq}$ pode ser utilizado para medir a distância entre o objeto e a classe do modelo.

No algoritmo SIMCA uma fronteira de classe é construída ao redor de cada modelo de classe. Isso pode ser considerado como um tipo de intervalo de confiança, de forma que a dispersão da população no espaço é estimado. Esse intervalo de confiança pode ser calculado tanto baseado em uma suposição quanto a distribuição da população como baseado na distribuição das distâncias observadas entre os objetos de treinamento em relação ao modelo de classe (ou seja, baseado no $\boldsymbol{E}_{kq}$). Os intervalos de confiança podem ser construídos com diferentes níveis de significância (α). No presente estudo considerou-se a utilização de 95 % de confiança.

No SIMCA uma classe é modelada através da análise PCA. Isso significa que a partir de L variáveis L componentes principais (PCs) são definidas. Se as variáveis são fortemente correlacionadas então quase toda variabilidade de uma classe pode ser representada no espaço definido pelas poucas componentes principais iniciais. O número de componentes significantes A pode ser definido através de um procedimento de validação cruzada. O modelo de classe M é então definido como:

$$\boldsymbol{M} = \overline{\boldsymbol{X}} + \sum_{a=1}^{A} \boldsymbol{TP} \tag{3.10}$$

enquanto os dados dos objetos pertencentes a classe são descritos pelas equações:

$$\boldsymbol{m}_{ik} = \overline{\boldsymbol{x}}_i + \sum_{a=1}^{A} \boldsymbol{t}_{ak}\boldsymbol{p}_{ia} \tag{3.11}$$

$$\boldsymbol{x}_{ik} = \boldsymbol{m}_{ik} + \boldsymbol{e}_{ik} \tag{3.12}$$

em que: $\overline{\boldsymbol{x}}_i$ é a média da variável i na classe; $\boldsymbol{t}_{ak}$ são os escores que descrevem a situação do objeto k com respeito a componente principal a; $\boldsymbol{p}_{ia}$ são os pesos, indicando a importância da variável i na direção da componente principal a; $\boldsymbol{m}_{ik}$ é a parte da i-ésima medida do objeto k que pode ser explicada pelo modelo de classe; $\boldsymbol{e}_{ik}$ são os resíduos que descrevem a parte não sistemática dos dados.

A combinação das componentes principais significantes definem um subespaço A-dimensional no espaço. A localização do modelo de classe no espaço é definido em

seguida pelo intervalo dos valores de $\boldsymbol{t}_{ak}$ dos objetos de treinamento. Ao longo de cada componente principal os limites de classe superiores ($\boldsymbol{t}_{a,max}$) e inferiores ($\boldsymbol{t}_{a,min}$) são respectivamente definidos como:

$$\boldsymbol{t}_{a,max} = \max(\boldsymbol{t}_{ak}) + 0{,}5 s_{t,a} \tag{3.13}$$

$$\boldsymbol{t}_{a,min} = \min(\boldsymbol{t}_{ak}) - 0{,}5 s_{t,a} \tag{3.14}$$

em que $s_{t,a}$ é o desvio padrão dos escores dos objetos de treinamento na CP a. De acordo com o número de componentes utilizadas em cada modelo de classe, a forma do modelo de classe pode ser um ponto $(A = 0)$, um segmento de linha $(A = 1)$, um retângulo $(A = 2)$, e outras formas. A figura 3.2 mostra uma representação gráfica de um modelo SIMCA. Com base nos resíduos um intervalo de confiança é construído ao redor do modelo de classe da seguinte forma. O desvio padrão dos resíduos para uma classe é dada por:

$$s_0 = \sqrt{\sum_{k=1}^{n} \sum_{i=1}^{L} \boldsymbol{e}_{ik}^2 / [(L - A)(n - A - 1)]} \tag{3.15}$$

s_0 é uma medida da distância média entre os objetos que pertencem a classe e o modelo de classe. O desvio padrão dos resíduos s_k para o objeto k é definido como:

$$s_k = \sqrt{\sum_{i=1}^{L} \boldsymbol{e}_{ik}^2 / (L - A)} \tag{3.16}$$

s_k é uma medida do quão adequadamente o objeto k é explicado pelo modelo de classe. Geometricamente ele corresponde a distância ortogonal entre o objeto e o plano definido pelas CPs significantes. Para objetos situados fora de um ou mais intervalos que restringem o modelo de classe ao longo das componentes principais significantes (o que somente ocorre para objetos não pertencentes ao conjunto de treinamento), a distância em relação ao modelo de classe é definido como:

$$s_k = (\sum_{i=1}^{L} \boldsymbol{e}_{ik}^2 / (L - A) + \sum_{b=1}^{B} (\boldsymbol{t}_{bk} - \boldsymbol{t}_{b,lim})^2 s_0^2 / s_{t,b}^2)^{1/2} \tag{3.17}$$

onde B diz respeito aquelas CPs para as quais k está situado fora do intervalo que restringe o modelo de classe ao longo de uma CP. O termo de correção $\frac{s_0^2}{s_{t,b}^2}$ é

introduzido para alongar o modelo de classe proporcionalmente a dispersão da classe na direção de uma PC. Quanto maior a parte referente a variância dentro de uma classe explicada por essa determinada PC (ou seja, quanto maior o $s_{t,b}^2$) mais a forma (da caixa) da classe será alongada nessa direção.

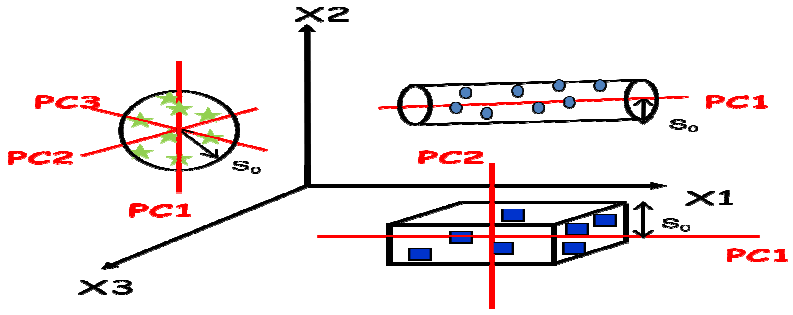


Figura 3.2 – Representação gráfica de um modelo SIMCA

Como os resíduos devem possuir uma distribuição normal, a razão:

$$F = \frac{s_k^2}{s_0^2} \tag{3.18}$$

dos objetos pertencentes a classe, deverão ter uma distribuição F com $(L - A)$ e $(L - A)(n - A - 1)$ graus de liberdade (n é o número de objetos utilizados no modelo de classe). Isso nos permite identificar o valor máximo que F deve possuir (F_{crit}) para que um objeto seja classificado na classe. A partir do F_{crit} o valor s_{crit} pode ser obtido, o qual é a distância que define as fronteiras da classe.

$$s_{crit} = \sqrt{F_{crit}\, s_0^2} \tag{3.19}$$

O termo *soft* refere-se ao fato de que na previsão o classificador pode identificar amostras como pertencentes a uma, a várias ou a nenhuma classe e o bom ajuste do modelo pode ser observado através do número de casos em que amostras são classificadas dessas formas. Amostras anômalas, ou seja, as que não se encaixam em nenhuma classe, podem ocorrer devido a um erro nas medidas, a uma má rotulação, a um fenômeno químico e/ou físico anômalo ou desconhecido ou, ainda, essa amostra pode pertencer a uma classe que não foi inclusa no conjunto de calibração usado na construção dos modelos de cada classe. [33,34]

3.4 - Máquina de vetores de suporte - SVM

As máquinas de vetores de suporte – SVM[2] representam uma adaptação do algoritmo *Generalized Portrait* desenvolvido na Rússia nos anos sessenta[35]. A estratégia de aprendizado baseada em vetores de suporte é solidamente fundamentada na teoria do aprendizado estatístico, ou teoria VC (Vapnik-Chervonenkis)[36-39], que vem sendo desenvolvida ao longo das últimas três décadas visando a proposição de técnicas de aprendizado de máquina que buscam maximizar a capacidade de generalização. Uma máquina de aprendizagem é um processo para seleção de uma função apropriada a partir de um conjunto de funções, para correlação de um conjunto de dados. As funções pertencentes a esse conjunto de funções são denominadas funções hipótese, funções indicadoras ou ainda funções classificadoras.

Em sua formulação mais recente, a abordagem SVM foi concebida nos laboratórios da AT&T por Vapnik *et al.*,[40] sendo orientada a aplicações práticas. Tanto acerca da aplicação de SVM em problemas de classificação de padrões[10,41,42] como também na sua extensão para o tratamento de problemas de regressão[7,10,43] a abordagem SVM mostrou-se altamente competitiva.

O problema de modelagem de dados empíricos é pertinente a muitas aplicações em química e engenharia. Em modelagem de dados empíricos, é usado um processo de indução, em que um modelo matemático capaz de expressar as relações de entrada-saída é construído e a partir do qual são deduzidas respostas ainda não observadas.

Os métodos estatísticos clássicos, largamente utilizados para resolver problemas em química, estão baseados na lei dos grandes números, segundo a qual quando o número de observações tende ao infinito, a função de distribuição empírica $F_e(x)$ converge para a função de distribuição real F(x). Em outras palavras, para obter um modelo matemático coerente utilizando aprendizado de máquina, seria necessário dispor de um conjunto de dados de treinamento incluindo um número infinito de amostras. No entanto, na prática, para os trabalhos de modelagem de dados em química o número de amostras de treinamento é relativamente pequeno.

A quantidade e qualidade dos dados disponíveis governam o desempenho deste modelo empírico. Por corresponder a uma técnica de aprendizado baseada na

utilização de um espaço amostral (número finito de amostras), temos consequentemente um espaço de entrada esparso. Assim, o problema de aprendizado tende a ser mal condicionado, ou seja, não há dependência contínua dos dados e o processo de indução de modelos não possui solução única. Quando uma multiplicidade de soluções candidatas são igualmente admissíveis a capacidade de generalização dos modelos resultantes passa a representar um critério de qualidade capaz de atenuar o efeito do mal condicionamento. No entanto, modelos matemáticos com capacidade de aproximação universal, como as redes neurais artificiais[6] ainda não são dotados de algoritmos de treinamento capazes de maximizar a capacidade de generalização de uma forma sistemática, o que pode levar a um sobreajuste do modelo aos dados. Por operar no espaço original dos dados, em que as não linearidades presentes e a complexidade intrínseca do problema não são conhecidas a priori, os algoritmos de otimização para ajuste de parâmetros e as ferramentas estatísticas adotadas para seleção de modelos podem induzir modelos com baixa capacidade de generalização.

Este é o cenário que torna atrativo a proposição de uma máquina de vetores de suporte - SVM em que a formulação engloba o princípio da minimização do risco estrutural - SRM, tendo-se demonstrado que este princípio é superior ao princípio da minimização do risco empírico – ERM,[38,39] sendo que este último é empregado no projeto de redes neurais artificiais. O princípio SRM envolve a minimização de um limite superior sobre o erro de generalização, enquanto que o princípio ERM envolve a minimização do erro sobre os dados de treinamento. Logo, modelos de aprendizado de máquina baseados no princípio SRM tendem a apresentar uma maior habilidade para generalizar bem frente a dados não observados, sendo este um dos principais propósitos do aprendizado estatístico.

Embora o princípio SRM tenha sido originalmente desenvolvido para solucionar problemas de classificação, ele foi estendido com sucesso para tratar problemas de regressão.[44,45]

Uma das maiores vantagens do SVM é a sua flexibilidade. Utilizando os conceitos básicos de maximização de margem, representação dual ou dualidade e produto interno kernel, pode-se adaptar o problema de classificação binária (apenas com duas classes), que foi a abordagem que originou a formulação da SVM, para resolver outros tipos de problemas como classificação em múltiplas classes e

regressão. Na regressão, há uma modificação simples na formulação original da função objetivo para uma em que o erro, medido pela distância do valor estimado em relação ao valor real, é igual a zero para valores pequenos desta distância, e de valor crescente quando a distância ao valor real é maior do que um determinado limiar [44] .

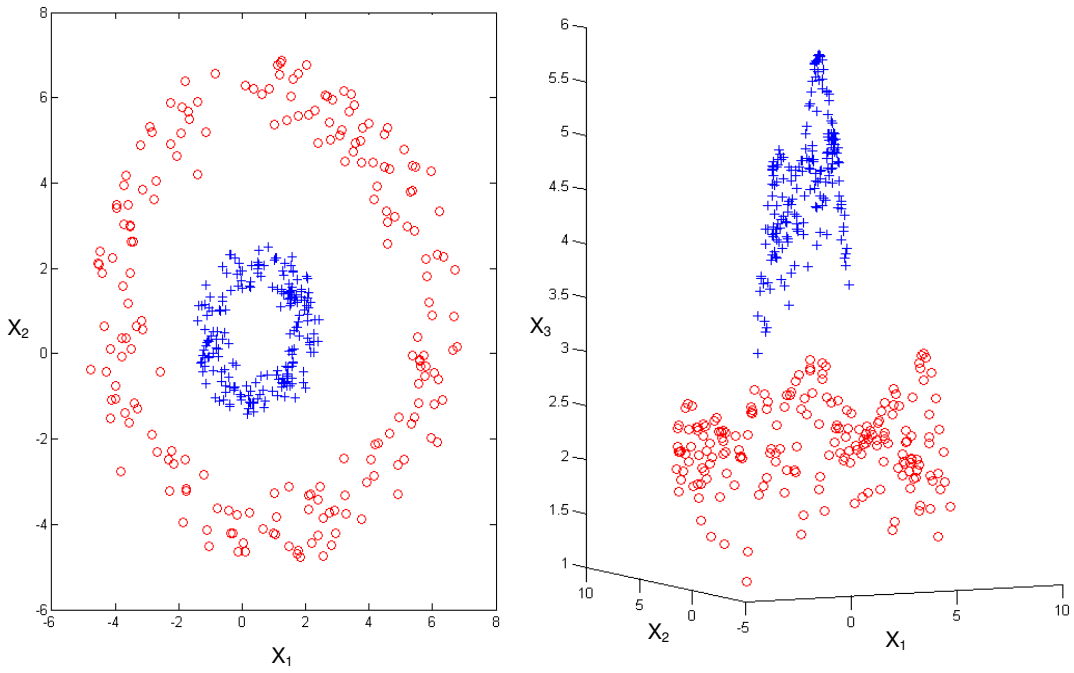
Quando comparado com outros algoritmos utilizados para modelagem de dados em química, o SVM proporciona algumas vantagens: ele pode ser utilizado tanto para classificação – SVC, como para regressão – SVR; é capaz de tratar adequadamente conjuntos de dados lineares e não lineares; tem excelente habilidade de generalização, especialmente para problemas de pequeno conjunto de amostras; em comparação com as ANN, o SVM não possui problema de mínimo local e a solução é única.

Do ponto de vista da teoria do aprendizado estatístico, o princípio do SVM é muito diferente daqueles dos métodos comumente utilizados para modelagem de dados em química, como PCA e PLS. Sabe-se bem que nas etapas iniciais de processamento dos dados com esses métodos há uma redução da dimensionalidade dos dados, de modo a superar a maldição da dimensionalidade e fazer o modelo matemático mais confiável, mas, no algoritmo SVM a tarefa mais importante é elevar a dimensionalidade, utilizando-se uma função kernel para mapear os pontos amostrais do espaço de entrada em um espaço de características de elevada dimensão através de uma transformação não linear. Nesse espaço de características de elevada dimensão, pontos amostrais não linearmente separáveis no espaço de entrada podem ser linearmente separáveis com a máxima margem de separação, e um algoritmo de separação linear pode ser utilizado para obter um modelo matemático com excelente capacidade de previsão.[46]

Nos próximos tópicos abordam-se as questões da dimensionalidade superior e funções kernel e em seguida são introduzidos alguns conceitos relativos à teoria do aprendizado estatístico, com destaque para risco funcional, minimização do risco empírico, minimização do risco estrutural, dimensão VC e margem de separação, para em seguida descrever a formulação básica do SVM para classificação e regressão.

3.4.1 – Dimensionalidade superior e separabilidade linear

Consideremos um problema de classificação binária em um espaço de duas dimensões, como ilustrado na figura 3.3 (a), e seja a i-ésima amostra denotada por $\boldsymbol{x}_i = (x_{i1}, x_{i2})$. Os dados em cada classe são distribuídos em duas regiões circulares e cada amostra pertence a uma das duas diferentes classes, que não podem ser linearmente separadas em um espaço de duas dimensões. Uma forma de resolver tal problema é construindo um complicado modelo não linear, tal como ANN, no entanto, deve-se ter em mente que o ajuste de um modelo não paramétrico geralmente requer bastante tempo e trabalho, e, além disso, normalmente tais funções não lineares não são suficientemente robustas.



(a) (b)

Figura 3.3 – Ilustração da superioridade dimensional em um problema de classificação binária: (a) espaço de entrada e (b) espaço de características

A outra solução factível e efetiva é simplesmente aumentar a dimensionalidade dos dados, acrescentando uma dimensão a cada amostra. O valor da terceira dimensão pode ser calculada como $x_{i3} = x_{i1}^2 + x_{i2}^2$. Então, num espaço de três dimensões a i-ésima amostra pode ser denotada por $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})$. Essa operação realiza um

mapeamento não linear dos dados originais do espaço de entrada para o espaço característico (espaço de elevada dimensão) e é o elemento chave na construção de modelos com SVM, como será demonstrado adiante. Na figura 3.3 (b) vemos que as amostras que não podem ser linearmente separadas em um espaço 2D, podem ser separadas por um hiperplano linear de maneira efetiva. Essa é a dimensionalidade superior, que deve ser aqui enfatizada, pois possui uma significante importância no algoritmo SVM. Pode-se concluir que os dados certamente contêm muito mais informação que pode ser aproveitada com o aumento da dimensionalidade e o incremento de informação é necessário para o SVM, de modo que seja possível caracterizar cada amostra em um determinado conjunto de dados.

A função kernel é a responsável por transformar o espaço de entrada em um espaço característico de alta dimensão onde a solução do problema pode ser representada como sendo um problema linear.[10]

3.4.2 – A função kernel

Muitos dos conjuntos de dados utilizados para resolver problemas em química têm característica não linear, enquanto o algoritmo SVM está baseado em uma máquina de aprendizado linear com utilização de margem de separação. Assim, torna-se necessário a utilização de alguma técnica para mapeamento não linear de modo a tornar possível o tratamento dos conjuntos de dados com SVM. As funções kernel[47,48] são uma ferramenta efetiva para esse propósito, pois possibilitam mapear o espaço de entrada original em um espaço de características de elevada dimensão onde uma relação não linear pode ser modelada de forma linear.

Uma máquina não linear pode ser construída em duas etapas:

(i) os dados são submetidos a uma transformação não linear pré-determinada, em um espaço de características de elevada dimensão;

(ii) uma máquina linear é utilizada para classificação ou regressão no espaço de características.

Vamos considerar o conjunto de dados da figura 3.3. Nesse caso para obter uma separação linear das duas classes de dados primeiro projetou-se todas as amostras em

um espaço característico 3D pelo incremento de uma dimensão. A função kernel pode solucionar o problema de projetar os dados em um espaço de dimensionalidade muito maior e pode dessa forma transformar um conjunto de dados não linearmente separáveis em um conjunto de dados linearmente separáveis.

Consideremos uma medida de similaridade na forma:

$$K: X.X \rightarrow \mathbb{R}$$
$$(x_i, x_j) \rightarrow K(x_i.x_j) \qquad 3.20$$

que é uma função em que dados dois objetos x_i e x_j que pertencem ao espaço de dados original X, retorna um número real característico de sua similaridade. Um tipo simples de medida de similaridade que é de interesse matemático particular é o produto interno ou produto escalar. Dados dois vetores $\boldsymbol{x}_i$ e $\boldsymbol{x}_j \in \mathbb{R}^N$ o produto interno canônico é definido como

$$(x_i, x_j) = \sum_{i,j=1}^{N} \boldsymbol{x}_i.\boldsymbol{x}_j \qquad 3.21$$

A interpretação geométrica do produto interno canônico é que ele calcula o cosseno do ângulo entre os vetores $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$, considerando os mesmos normalizados para comprimento 1. Além disso, ele permite o cálculo do comprimento (ou norma) do vetor $\boldsymbol{x}$ como na equação 3.22 e a distância entre dois vetores é computada como o cumprimento do vetor diferença.

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}.\boldsymbol{x})} \qquad 3.22$$

Para podermos utilizar o produto interno como medida de similaridade é necessário primeiro representar os objetos como vetores em um espaço de características H com a utilização de uma função $\Phi(x)$ apropriada : $\Phi(x): X \rightarrow H$. A função kernel K é uma função que recebe dois pontos $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$ do espaço de entradas e computa o produto interno desses dados no espaço de características. Dada uma matriz de dados originais **X**, de ordem m x 2 (m amostras e duas variáveis) em que

cada amostra é denotada por $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ e seja $\Phi(\boldsymbol{x})$ uma função com a seguinte forma:

$$\Phi(\boldsymbol{x}) = \{x_1^2, x_2^2, \sqrt{2x_1x_2}\} \tag{3.23}$$

que pode mapear de forma não linear as amostras do espaço 2D em um espaço característico 3D. Para a i-ésima e j-ésima amostra temos:

$$\boldsymbol{x}_i = [x_{i1}, x_{i2}], \qquad \Phi(\boldsymbol{x}_i) = [x_{i1}^2, x_{i2}^2, \sqrt{2x_{i1}x_{i2}}]$$
$$\boldsymbol{x}_j = [x_{j1}, x_{j2}], \qquad \Phi(\boldsymbol{x}_j) = [x_{j1}^2, x_{j2}^2, \sqrt{2x_{j1}x_{j2}}] \tag{3.24}$$

O produto interno dessas equações é definido como:

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = x_{i1}x_{j1} + x_{i2}x_{j2} \tag{3.25}$$

E o produto interno no espaço característico é:

$$K\left(\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j)\right) = x_{i1}^2x_{j1}^2 + x_{i2}^2x_{j2}^2 + \sqrt{2x_{i1}x_{i2}}\sqrt{2x_{j1}x_{j2}} = (x_{i1}x_{j1} + x_{i2}x_{j2})^2 \tag{3.26}$$

Substituindo a equação 3.25 em 3.26 temos:

$$K\left(\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j)\right) = (D(\boldsymbol{x}_i, \boldsymbol{x}_j))^2 = (\boldsymbol{x}_i.\boldsymbol{x}_j)^2 \tag{3.27}$$

É interessante notar que o produto interno, que é uma medida da similaridade entre duas amostras num espaço característico de elevada dimensionalidade, pode ser calculado diretamente no espaço de entrada original pela introdução da função kernel.[10,39] As funções kernel mais comumente usadas são:

Linear: $$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = a(\boldsymbol{x}_i.\boldsymbol{x}_j) + b \tag{3.28}$$

Polinomial: $$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (a(\boldsymbol{x}_i.\boldsymbol{x}_j) + b)^n \tag{3.29}$$

Função de base radial (RBF): $$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \parallel \boldsymbol{x}_i - \boldsymbol{x}_j \parallel^2) \tag{3.30}$$

Sigmoidal: $$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(a(\boldsymbol{x}_i.\boldsymbol{x}_j) + b) \tag{3.31}$$

Para as funções kernel linear, polinomial e sigmoidal “a” e “b” são constantes. Para o kernel RBF, ”γ” é um parâmetro de ajuste que controla a forma do hiperplano de separação (a largura da função gaussiana), e deve ser adequadamente selecionado para obtenção de uma boa performance de generalização.[43,49] Para o kernel polinomial o parâmetro *n* representa o grau do polinômio.

É possível demonstrar que qualquer função que satisfaça as condições do Teorema de Mercer pode ser usada como um kernel.[42,44-48] O teorema de Mercer nos diz simplesmente quando uma função candidata a kernel é de fato um produto interno kernel em algum espaço determinado e portanto admissível para ser utilizada no treinamento de SVM. Porém este teorema não indica como obter as funções $\Phi(\boldsymbol{x})$.

Uma matriz kernel $\boldsymbol{K}$ de ordem m x m deve ser primeiramente calculada com seu elemento $\boldsymbol{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ utilizando a função kernel para a construção do modelo SVM. Como mencionado anteriormente, $\boldsymbol{K}_{ij}$ é uma medida da similaridade no espaço de características de alta dimensão da i-ésima e j-ésima amostra. Logicamente, os dados transformados pela função kernel não são representados individualmente, mas apenas através de um conjunto de pares de comparação. As funções kernel estão baseadas nos objetos de treinamento, ou seja, um dos dois objetos do kernel é sempre um objeto de treinamento. Um objeto de teste é classificado por comparação do mesmo com todos os objetos de treinamento com peso diferente de zero.[39]

Além disso, essa representação não depende da natureza dos objetos para ser analisada. Muitos objetos, como imagens, moléculas, etc., podem ser representadas dessa maneira.

De forma sucinta, a essência da utilização das funções kernel diz respeito a três pontos:

(i) ela projeta os dados originais em um espaço de mais alta dimensão através da adição de dimensões;

(ii) ela simultaneamente proporciona um meio eficiente para calcular o produto interno no espaço característico.

(iii) nos permite tratar com os objetos geometricamente, possibilitando estudar os algoritmos de aprendizado em termos de álgebra linear e geometria analítica.

3.4.3 - Princípios da teoria do aprendizado estatístico

Nos métodos estatísticos clássicos, enfatiza-se que o erro de treinamento deve ser minimizado. Isso é importante devido a necessidade de se minimizar a falta de ajuste para os modelos construídos, mas, a prática da aplicação de ANN e a argumentação teórica da teoria do aprendizado estatístico[36-39] mostram que esse conceito não abrange toda a problemática de modelagem de dados em química e engenharia. Para maximizar a habilidade de predição dos modelos matemáticos obtidos, devemos evitar simultâneamente a falta de ajuste e o sobreajuste no processamento dos dados. Na teoria do aprendizado estatístico, o erro de treinamento é chamado de "risco empírico", denotado por R_{emp}.

Seja f um classificador e F o conjunto de todos os classificadores que um determinado algoritmo de aprendizado de máquina pode gerar. Esse algoritmo, durante o processo de aprendizado, utiliza um conjunto de treinamento T composto de N pares $(\boldsymbol{x_i}, y_i)$, para gerar um classificador particular f' que pertence a F. A teoria do aprendizado estatístico estabelece condições matemáticas que auxiliam na escolha de um classificador particular f' a partir de um conjunto de dados de treinamento. Essas condições levam em conta o desempenho do classificador no conjunto de treinamento e a sua complexidade, com o objetivo de obter um bom desempenho também para novos dados do mesmo domínio.

Normalmente utiliza-se o princípio da indução para inferir uma função f' que minimize o erro sobre os dados de treinamento e espera-se que esse procedimento leve também a um menor erro sobre os dados de teste. O risco empírico de f, mede o desempenho do classificador nos dados de treinamento, por meio da taxa de classificações incorretas obtidas em T.

Esse processo de indução com base nos dados de treinamento conhecidos constitui o princípio de minimização do risco empírico. Assintoticamente, com $N \rightarrow \infty$, é possível estabelecer condições para o algoritmo de aprendizado que garantam a obtenção de classificadores cujos valores de risco empírico convergem para o risco esperado. Para conjuntos de dados menores, porém, geralmente não é possível

determinar esse tipo de garantia. Embora a minimização do risco empírico possa levar a um menor risco esperado, nem sempre isso ocorre.

A noção expressa nesses argumentos é a de que, permitindo que f' seja escolhida a partir de um conjunto de funções amplo F é sempre possível encontrar f com pequeno risco empírico. Porém, nesse caso os exemplos de treinamento podem se tornar pouco informativos para a tarefa de aprendizado, pois o classificador induzido pode se sobre ajustar a eles. Deve-se então restringir a classe de funções da qual f' é extraída. Existem diversas abordagens para tal e a teoria do aprendizado estatístico trata essa questão considerando a complexidade (ou capacidade) da classe de funções que o algoritmo de aprendizado é capaz de obter. Pode-se então obter diversos limites no risco esperado de uma função de classificação, os quais podem ser empregados na escolha do classificador. [38,39,45]

3.4.3.1 - Controle do ajuste de modelos, dimensão VC e margem de separação

De acordo com o princípio da minimização do risco estrutural – SRM é necessário diminuir o erro de treinamento, mas, apenas isso não é o suficiente, uma vez que o risco de predição contém ainda outro termo para o risco devido ao sobreajuste. [33,34,43]

Um limite importante fornecido pela teoria do aprendizado estatístico relaciona o risco esperado de uma função ao seu risco empírico e a um termo de capacidade. Esse limite, apresentado na inequação 3.32, é garantido com probabilidade $1-\eta$, em que $\eta \in [0,1]$.

$$R_{pred} \leq R_{emp} + \sqrt{\frac{h\left(ln\frac{2N}{h}+1\right)-\ln\left(\frac{\eta}{4}\right)}{N}} \qquad (3.32)$$

Nessa equação, h denota a dimensão Vapnik-Chervonenkis (VC) da classe de funções F à qual f pertence, N representa a quantidade de objetos no conjunto de treinamento T e a parcela de raiz na soma é referenciada como termo de capacidade.

A dimensão VC h mede a capacidade do conjunto de funções F. Quanto maior o seu valor, mais complexas são as funções de classificação que podem ser induzidas a partir de F. Dado um problema de classificação binário, essa dimensão é definida como o número máximo de objetos que podem ser corretamente separados em duas classes pelas funções contidas em F, para todas as possíveis combinações binárias desses dados.

Para ilustrar esse conceito, considere os três dados apresentados na figura 3.4. Pode-se verificar que, para qualquer conformação arbitrária dos objetos das duas classes que esses dados possam assumir, é possível determinar retas capazes de separá-los. Porém, para os quatro pontos em R^2 ilustrados na figura 3.4, existem rótulos para os dados que podem ser separados por uma reta, mas também é possível definir rótulos tal que uma só reta seja incapaz de realizar a separação em classes. Para uma divisão binária arbitrária desses quatro pontos em R^2, deve-se então recorrer a funções de complexidade superior à das retas.

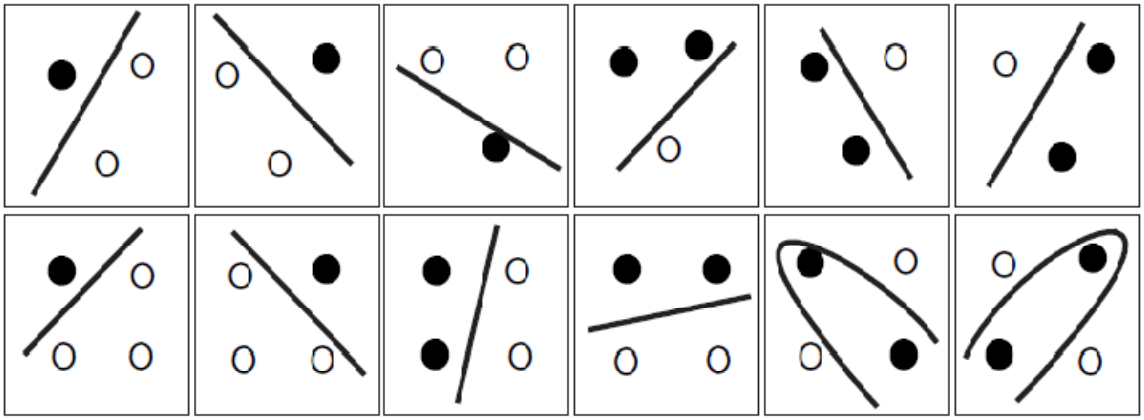


Figura 3.4 – Em um plano todas as combinações de três pontos podem ser separados por uma reta. Quatro pontos não podem ser separados por um classificador linear

Essa observação se aplica a quaisquer quatro pontos no espaço bidimensional. Portanto, a dimensão VC do conjunto de funções lineares no espaço bidimensional é 3, uma vez que existe (pelo menos) uma configuração de três pontos nesse espaço que pode ser separada por retas em todas as $2^3 = 8$ combinações binárias de rótulos.

A contribuição principal da inequação 3.32 está em afirmar a importância de se controlar a capacidade do conjunto de funções F do qual o classificador é extraído.

Interpretando-a em termos práticos, tem-se que o risco esperado pode ser minimizado pela escolha adequada, por parte do algoritmo de aprendizado, de um classificador f' que minimize o risco empírico e que pertença a uma classe de funções F com baixa dimensão VC. Com esses objetivos, definiu-se um princípio de indução denominado minimização do risco estrutural – SRM.

Como no limite apresentado o termo de capacidade diz respeito à classe de funções F e o risco empírico refere-se a um classificador particular f, para minimizar ambas as parcelas divide-se inicialmente F em subconjuntos de funções com dimensão VC crescente. É comum referir-se a esse processo como introduzir uma estrutura em F, sendo os subconjuntos definidos também denominados estruturas. Minimiza-se então o limite sobre as estruturas introduzidas.

Consideram-se subconjuntos F_i da seguinte forma: $F_0 \subset F_1 \subset \ldots \subset F_i \subset F$. Como cada F_i é maior com o crescimento do índice *i*, a capacidade do conjunto de funções que ele representa também é maior à medida que *i* cresce, ou seja, $h_0 < h_1 < \ldots < h_i < h$. Para um subconjunto particular F_k seja f_k' o classificador com o menor risco empírico. A medida que *k* cresce, o risco empírico de f_k' diminui, uma vez que a complexidade do conjunto de classificadores é maior. Porém, o termo de capacidade aumenta com *k*. Como resultado, deve haver um valor ótimo *k** em que se obtém uma soma mínima do risco empírico e do termo de capacidade, minimizando assim o limite sobre o risco esperado. A escolha da função f_{k^*}' constitui o princípio da minimização do risco estrutural.

Embora o limite representado na inequação 3.32 tenha sido útil na definição do procedimento de minimização do risco estrutural, na prática surgem alguns problemas. Em primeiro lugar, computar a dimensão VC de uma classe de funções geralmente não é uma tarefa trivial. Soma-se a isso o fato de que o valor de h poder ser desconhecido ou infinito.

Para funções de decisão lineares do tipo $f(x) = w.x$, entretanto, existem resultados alternativos que relacionam o risco esperado ao conceito de margem.

Uma das mais importantes realizações de Vapnik e seus colaboradores foi o conceito de máxima margem de separação. Verificou-se que a dimensão VC pode ser sensivelmente reduzida se os pontos amostrais das diferentes classes podem ser mapeadas em outro espaço de características para obtenção da máxima margem de

separação entre as classes. Assim, foi proposta a seguinte estratégia de processamento: primeiramente os pontos amostrais do espaço de entrada são mapeados em um espaço de características com elevada dimensão através de uma transformação linear ou não linear, de modo que a distribuição dos pontos amostrais seja feita com a obtenção da máxima margem de separação e então um hiperplano ótimo de separação é usado para descrever o critério de classificação entre as diferentes classes. Desse modo, os modelos matemáticos para os conjuntos de dados linearmente ou não linearmente separáveis podem ser obtidos com boa habilidade de previsão. [38,39,45,50]

A margem de um objeto tem relação com sua distância à fronteira de decisão induzida, sendo uma medida da confiança da previsão do classificador. Para um problema binário, em que $y_i \in \{-1,1\}$ dada uma função f e um objeto x_i, a margem $\varrho(f(x_i), y_i)$ com que esse dado é classificado por f pode ser calculada pela equação 3.33. Logo, um valor negativo de $\varrho(x_i, y_i)$ denota uma classificação incorreta.

$$\varrho(f(x_i), y_i) = y_i f(x_i) \tag{3.33}$$

Para obter a margem geométrica de um dado x_i em relação a uma fronteira linear $f(x) = w.x + b$, a qual mede efetivamente a distância de x_i à fronteira de decisão, divide-se o termo à direita da Equação 3.33 por $\| w \|$. Para objetos incorretamente classificados, o valor obtido equivale à distância com sinal negativo. Para realizar uma diferenciação, a margem da Equação 3.33 será referenciada como margem de confiança.

A partir do conceito introduzido, é possível definir o erro marginal de uma função $f(R_\mu(f))$ sobre um conjunto de treinamento. Esse erro fornece a proporção de objetos de treinamento cuja margem de confiança é inferior a uma determinada constante $\mu > 0$.

$$R_\mu(f) = \frac{1}{N}\sum_{i=1}^{N} I(y_i f(x_i) < \mu \tag{3.34}$$

onde: $I(q) = 1$ *se* q é $verdadeiro$ *e* $I(q) = 0$ *se* q é $falso$

Existe uma constante *c* tal que, com probabilidade $1 - \eta \in [0,1]$, para todo $\mu > 0$ e F correspondendo à classe de funções lineares $f(x) = w.x$ com $\| x \| \leq R$ e $\| w \| \leq 1$, o seguinte limite se aplica:

$$R(f) \leq R_{\mu}(f) + \sqrt{\frac{c}{N}}\left(\frac{R^2}{\mu^2}\, log^2\left(\frac{N}{\mu}\right) + \log\left(\frac{1}{\eta}\right)\right) \tag{3.35}$$

Como na expressão 3.32, tem-se na expressão 3.35 novamente o erro esperado limitado pela soma de uma medida de erro no conjunto de treinamento, neste caso o erro marginal, a um termo de capacidade. A interpretação do presente limite é de que uma maior margem μ implica em um menor termo de capacidade. Entretanto, a maximização da margem pode levar a um aumento na taxa de erro marginal, pois torna-se mais difícil obedecer à restrição de todos os dados de treinamento estarem distantes de uma margem maior em relação ao hiperplano separador. Um baixo valor de μ, em contrapartida, leva a um erro marginal menor, porém aumenta o termo de capacidade. Deve-se então buscar um compromisso entre a maximização da margem e a obtenção de um erro marginal baixo.

Como conclusão tem-se que, na geração de um classificador linear, deve-se buscar um hiperplano que tenha margem elevada e cometa poucos erros marginais, minimizando assim o erro sobre os dados de teste e de treinamento, respectivamente. Esse hiperplano é denominado ótimo.

Existem diversos outros limites reportados na literatura, assim como outros tipos de medida de complexidade de uma classe de funções. Os limites apresentados anteriormente, embora possam ser considerados simplificados, provém uma base teórica suficiente à compreensão do SVM. [38,39]

3.4.4 – Máquina de vetores de suporte para classificação – SVC

O SVM foi originalmente desenvolvido por Vapnik para o reconhecimento de padrões. Embora o SVC tenha sido inicialmente desenvolvido para classificação binária, ele vem sendo também utilizado para problemas de classificação em múltiplas classes.

Como a natureza da classificação em múltiplas classes é a mesma da classificação binária, discutiremos apenas o algoritmo da classificação binária. A idéia básica do SVC pode ser sintetizada em dois itens:

(i) realiza o mapeamento não linear dos dados originais em um espaço característico de alta dimensão com a utilização de uma função kernel;

(ii) constrói um hiperplano ótimo de separação – OSH, no espaço característico, de forma a maximizar a margem de separação entre as duas classes.

Os problemas de classificação binária podem ser divididos em dois casos: linearmente separável e linearmente inseparável. A solução para o primeiro é facilmente obtida, mas no segundo caso uma função kernel deve ser utilizada para resolver o problema.

Suponhamos que haja um conjunto de dados T com duas classes de amostras, no qual cada amostra é denotada por $\boldsymbol{x}_i \in X$ e cada classe por $y_i \in Y$, tal como,

$$\boldsymbol{x}_i \in R^n, \quad y_i \in \{-1,1\}, \quad i = 1, 2, 3, \dots, N$$

Aqui, $\mathbf{x_i}$ é um vetor n-dimensional com o correspondente valor y_i igual a 1 se ele pertence a classe positiva e -1 se pertence a classe negativa. Veremos a seguir como é derivada a função de decisão para ambos os casos mencionados acima.

3.4.4.1 - Algoritmo para o caso de separação linear

A equação de um hiperplano é apresentada na equação 3.36, em que $\boldsymbol{w}.\boldsymbol{x_i}$ é o produto interno entre os vetores. Sendo que $\boldsymbol{w} \in X$ é o vetor normal ao hiperplano descrito (vetor de pesos normalizado, com as mesmas dimensões de $\mathbf{x_i}$) e $\frac{b}{\|\boldsymbol{w}\|}$ corresponde a distância do hiperplano em relação a origem, com $b \in R$ (b é o bias normalizado do hiperplano).

$$f(\boldsymbol{x_i}) = \boldsymbol{w}.\boldsymbol{x_i} + b \tag{3.36}$$

Essa equação divide o espaço dos dados X em duas regiões: $\boldsymbol{w}.\boldsymbol{x_i} + b > 0$ e $\boldsymbol{w}.\boldsymbol{x_i} + b < 0$. Uma função sinal $g(x) = \mathrm{sgn}(f(x))$ pode então ser utilizada na obtenção das classificações, conforme abaixo

$$g(x) = \mathrm{sgn}\big(f(x)\big) = \begin{matrix} +1 \quad se \quad \boldsymbol{w}.\boldsymbol{x_i} + b > 0 \\ -1 \quad se \quad \boldsymbol{w}.\boldsymbol{x_i} + b < 0 \end{matrix} \tag{3.37}$$

A partir de $f(x)$ é possível obter um número infinito de hiperplanos equivalentes, pela multiplicação de $\boldsymbol{w}$ e b por uma mesma constante. Define-se o hiperplano canônico em relação ao conjunto T como aquele em que $\boldsymbol{w}$ e b são escalonados de forma que os exemplos mais próximos ao hiperplano, descrito pela equação 3.36, satisfaçam a equação 3.38

$$|\boldsymbol{w}.\boldsymbol{x_i} + b| = 1 \tag{3.38}$$

O SVC seleciona o hiperplano que maximiza a margem, ou seja, maximiza a distância da margem para os dados de treinamento, de modo que para o hiperplano de separação ótimo a distância da margem para a fronteira da classe positiva é igual a distância da margem para a fronteira da classe negativa. No caso linearmente separável, qualquer hiperplano f(x) deve seguir as seguintes condições:

$$f(\boldsymbol{x}_i) = \boldsymbol{w}.\boldsymbol{x}_i + b \geq 1, y_i = 1 \tag{3.39}$$

$$f(\boldsymbol{x}_i) = \boldsymbol{w}.\boldsymbol{x}_i + b \leq -1, y_i = -1 \tag{3.40}$$

As duas restrições acima podem ser combinadas em uma forma:

$$(\boldsymbol{w}.\boldsymbol{x}_i + b)y_i \geq 1 \tag{3.41}$$

Seja $\boldsymbol{x}_1$ um ponto no hiperplano H_1 e $\boldsymbol{x}_2$ um ponto no hiperplano H_2, conforme ilustra a figura 3.5. Projetando $\boldsymbol{x}_1 - \boldsymbol{x}_2$ na direção de $\boldsymbol{w}$, perpendicular ao hiperplano separador OSH, é possível obter a distância entre os hiperplanos H_1 e H_2.

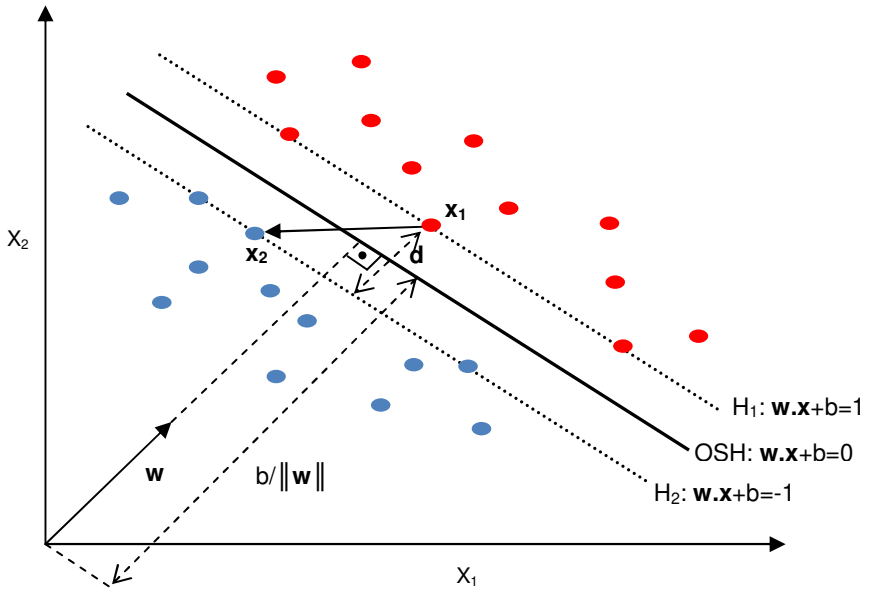


Figura 3.5 – Expressão do OSH e a margem para o caso linearmente separável

Devemos atentar para o fato de o escalonamento de $\boldsymbol{w}$ e b fazer f(x) igual a 1 ou -1 se $\boldsymbol{x}_i$ está na fronteira de classe. Então, a distância d, correspondente a largura da margem entre dois hiperplanos paralelos (linhas pontilhadas na figura 3.5) pode ser escrito como:

$$d = 2\frac{|\boldsymbol{w}.\boldsymbol{x} + b|}{\| \boldsymbol{w} \|} = 2\frac{1}{\| \boldsymbol{w} \|} = \frac{2}{\| \boldsymbol{w} \|} \tag{3.42}$$

Como $\boldsymbol{w}$ e b foram escalonados de forma a não haver exemplos entre H_1 e H_2, tem-se que $\frac{1}{\|\boldsymbol{w}\|}$ é a distância mínima entre o hiperplano separador e os dados de treinamento. Essa distância é definida como a margem geométrica do classificador linear.

O objetivo do SVC é localizar o OSH que maximiza a margem, submetido às restrições da equação 3.41. Isso implica em minimizar $\| \boldsymbol{w} \|$. Portanto, a construção do OSH pode ser convertido no seguinte problema de otimização:

$$\begin{aligned} &minimizar\text{: } \frac{1}{2} \| \boldsymbol{w} \|^2 \\ &restrição\text{: } \quad (\boldsymbol{w}.\boldsymbol{x}_i + b)y_i \geq 1 \end{aligned} \tag{3.43}$$

Com a ajuda do método dos multiplicadores de Lagrange,[51] esse problema pode ser novamente convertido para minimizar a seguinte função objetivo:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \| \boldsymbol{w} \|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) - 1] \quad (3.44)$$

onde $\alpha_i (\alpha_i \geq 0)$ é o chamado multiplicador de Lagrange. A função Lagrangiana deve ser minimizada, o que implica em maximizar as variáveis α_i e minimizar $\boldsymbol{w}$ e b. Tem-se então um ponto de cela no qual:

$$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{x}_i = 0 \quad (3.45)$$

$$\frac{\partial L(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial b} = \sum_{i=1}^{N} y_i \alpha_i = 0 \quad (3.46)$$

E as soluções para as duas equações acima são:

$$\boldsymbol{w} = \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{x}_i \quad (3.47)$$

$$0 = \sum_{i=1}^{N} y_i \alpha_i \quad (3.48)$$

Colocando a equação 3.47 e a equação 3.48 na função de Lagrange (equação 3.44) obtemos o seguinte problema de otimização:

$$maximizar, \alpha: \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j (\boldsymbol{x}_i.\boldsymbol{x}_j) \quad (3.49)$$

$$restrição \quad \alpha_i \geq 0, i = 1, \ldots, N \quad e \quad \sum_{i=1}^{N} y_i \alpha_i = 0$$

A solução do problema e a obtenção do valor ótimo de α_i pode ser calculado através de um algoritmo de programação quadrática.[51]

Essa formulação é denominada forma dual, enquanto o problema original é denominado forma primal.[51] A forma dual possui os atrativos de apresentar restrições mais simples e permitir a representação do problema de otimização em termos de produtos internos entre dados, o que será útil na resolução de problemas não lineares com SVM. É interessante observar também que o problema dual é formulado utilizando apenas os dados de treinamento e os seus rótulos.

Seja α a solução do problema dual e $\boldsymbol{w}$ e b as soluções da forma primal, obtido o valor de α, $\boldsymbol{w}$ pode ser determinado pela Equação 3.47. Para o problema dual formulado tem-se:

$$\alpha_i(y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) - 1) = 0 \quad (3.50)$$

Observa-se nessa equação que o α_i pode ser diferente de 0 somente para os dados que se encontram sobre os hiperplanos de fronteira de classe. Estes são os exemplos que se situam mais próximos ao hiperplano separador (OSH). Para os outros casos, a condição apresentada na equação 3.50 é obedecida apenas com $\alpha_i = 0$. Esses pontos não participam então do cálculo de $\boldsymbol{w}$ (equação 3.47). Os dados que possuem $\alpha_i > 0$ são denominados vetores de suporte e podem ser considerados os dados mais informativos do conjunto de treinamento, pois somente eles participam da equação do hiperplano separador (equação 3.52).

Geralmente os vetores de suporte são uma pequena porcentagem das amostras de treinamento (chegando até a aproximadamente 2/3 das amostras de calibração em alguns casos), o que é denominado esparsidade da solução para maximização da função da equação 3.49. Então, a esparsidade faz o modelo SVC depender apenas desses poucos vetores de suporte, e o OSH é determinado apenas por esses vetores.

O valor de b é calculado a partir dos vetores de suporte e das condições representadas na equação 3.50. Calcula-se a média apresentada na equação 3.51 sobre todos $\boldsymbol{x}_j$ tal que $\alpha_j > 0$, ou seja, todos os vetores de suporte.

$$b = \frac{1}{n_{SV}} \sum_{\boldsymbol{x}_j \in SV} \frac{1}{y_j} - \boldsymbol{w}.\boldsymbol{x}_j \quad (3.51)$$

onde, n_{sv} é o número de vetores de suporte e SV representa o conjunto dos vetores de suporte. E $\boldsymbol{w}$ é obtido pela equação 3.47.

Por fim, a função de decisão otimizada (OSH) pode ser escrita na forma:

$$f(x) = sgn(\boldsymbol{w}.\boldsymbol{x} + b) = sgn[(\sum_{i=1}^{N} y_i \alpha_i (\boldsymbol{x}_i.\boldsymbol{x}) + b] \quad (3.52)$$

$0 \leq \alpha_i$

Esta função linear representa o hiperplano que separa os dados com maior margem, considerando aquele com melhor capacidade de generalização de acordo com a teoria do aprendizado estatístico. Essa característica difere as SVMs lineares de margens rígidas das redes neurais perceptron, em que o hiperplano obtido na separação dos dados pode não corresponder ao de maior margem de separação.[10,38,39]

3.4.4.2 – Algoritmo para o caso não separável linearmente

Em muitos casos os dados não são linearmente separáveis, devido a ruído experimental, não linearidade intrínseca, etc. Como mostrado na figura 3.6, não é possível construir um hiperplano que separe linearmente as duas classes sem erros.

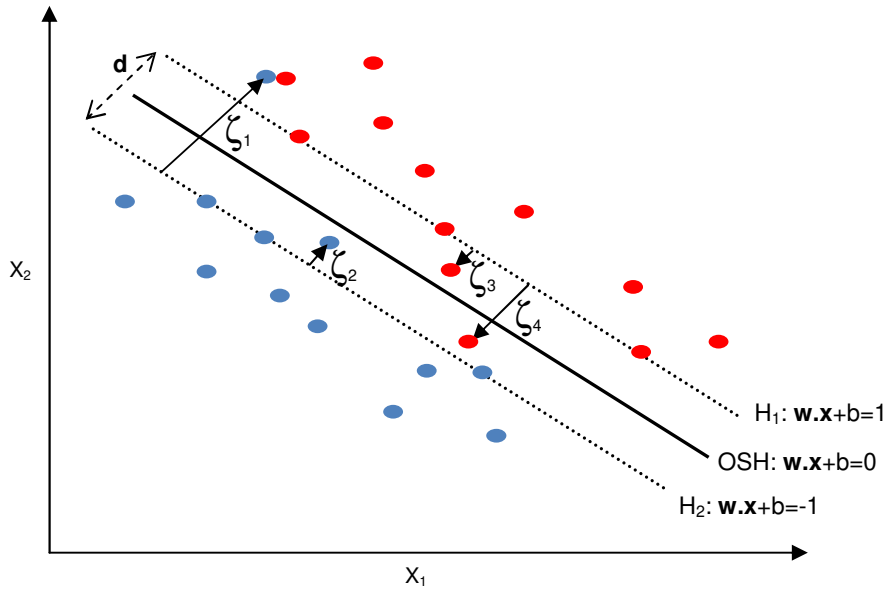


Figura 3.6 – As variáveis de folga e o OSH para o caso não separável linearmente

A seguir são mostradas as técnicas para tratar desse problema.

3.4.4.2.1 – Variáveis de folga e técnica da margem suave

As variáveis de folga foram introduzidas para construir o OSH levando em consideração os inevitáveis erros introduzidos por alguns objetos. A inequação de restrição para o OSH pode ser expresso na seguinte forma:

$$(\boldsymbol{w}.\boldsymbol{x_i} + b)y_i \geq 1 - \xi_i, \xi_i \geq 0 \qquad i = 1,2,\dots,N \tag{3.53}$$

onde ξ_i é a variável de folga, que é uma medida do quanto uma amostra ultrapassa o hiperplano de fronteira de classe mostrado na figura 3.6 e sua quantidade total contabilizada deve ser minimizada. A aplicação desse procedimento suaviza as margens do classificador linear, permitindo que alguns dados permaneçam entre os hiperplanos H_1 e H_2 e também a ocorrência de alguns erros de classificação. Um classificador com boa capacidade de generalização é então obtido tanto pelo controle da capacidade do classificador (via $\| \boldsymbol{w} \|$) como pela soma das variáveis de folga $\sum_i^N \xi_i$. Pode-se demonstrar que isso proporciona um limite no número de erros de treinamento.

A construção do OSH pode ser expresso como o seguinte problema de otimização:

$$\begin{aligned} &minimizar: && \tfrac{1}{2} \| \boldsymbol{w} \|^2 + \tfrac{C}{N} \textstyle\sum_i^N \xi_i \\ &restrição: && (\boldsymbol{w}.\boldsymbol{x_i} + b)y_i \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \tag{3.54}$$

onde a constante C > 0 é um termo de regularização que impõe um peso à minimização dos erros no conjunto de treinamento em relação à minimização da complexidade do modelo. Mais precisamente, ele controla o custo entre dois objetivos conflitantes: quando C é pequeno a maximização da margem é enfatizada, enquanto quando C é grande a minimização do erro é enfatizada proporcionando um estreitamento da margem.

Com a ajuda do método do multiplicador de Lagrange e de um algoritmo de programação quadrática a solução otimizada para $\boldsymbol{w}$ e b pode ser calculada. Novamente tornando suas derivadas parciais nulas, a função de decisão correspondente ao OSH pode ser escrito como:

$$f(x) = sgn(\boldsymbol{w}.\boldsymbol{x} + b) = sgn[(\textstyle\sum_{i=1}^N y_i \alpha_i (\boldsymbol{x}_i.\boldsymbol{x}) + b] \tag{3.55}$$

$$0 \leq \alpha_i \leq C$$

onde α_i é o multiplicador de Lagrange para cada amostra. Observa-se que essa formulação é igual a apresentada para o SVC de margens rígidas, a não ser pela restrição nos α_i, que agora são limitados pelo valor de C.

3.4.4.2.2 – Mapeamento não linear e utilização das funções kernel

Como citado anteriormente a superioridade dimensional pode ajudar a resolver problemas de classificação eficientemente. Uma técnica para lidar com casos de separação não linear é primeiramente proceder ao mapeamento não linear dos dados originais em um espaço característico de maior dimensão, e então construir o OSH que maximiza a margem, como no caso da separação linear. Com esse intuito os dados são inicialmente projetados em um espaço de características com a utilização de uma função de mapeamento $\Phi(\boldsymbol{x})$. O SVM linear é então aplicado nesse espaço de características, onde idealmente os dados podem ser linearmente separados.

O uso desse procedimento é motivado pelo teorema de Cover. Dado um conjunto de dados não linear no espaço de entradas X, esse teorema afirma que X pode ser transformado em um espaço de características de elevada dimensão no qual com alta probabilidade os dados são linearmente separáveis. Para isso duas condições devem ser satisfeitas. A primeira é que a transformação seja não linear, enquanto a segunda é que a dimensão do espaço de características seja suficientemente alta.[38]

A função kernel pode, implicitamente, não apenas realizar esse mapeamento não linear dos dados originais em um espaço característico de mais alta dimensão, mas também proporciona uma ferramenta matemática eficiente para o cálculo do produto interno no espaço característico.

Dado que a função kernel selecionada $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ está associada com a correspondente função $\Phi(\boldsymbol{x})$, a qual pode mapear os dados originais em um espaço característico de H dimensões, temos as seguintes condições:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \Phi(\boldsymbol{x}_i).\Phi\left(\boldsymbol{x}_j\right) \qquad i, j = 1,2,\dots,N \tag{3.56}$$

Agora, no espaço característico, o OSH linear pode ser construído usando o mesmo procedimento utilizado no caso da separação linear, apenas pela substituição da amostra $\boldsymbol{x}_i$ por $\Phi(\boldsymbol{x}_i)$. Assim o problema de otimização no espaço característico pode ser escrito de forma similar a equação 3.49, como segue,

$$\begin{aligned} &maximizar, \alpha: && \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j (\Phi(\boldsymbol{x}_i).\Phi(\boldsymbol{x}_j)) \\ &restrição: && 0 \le \alpha_i \le C \qquad i = 1,2,\dots,N\,; e \quad \sum_{i=1}^{N} y_i \alpha_i = 0 \end{aligned} \tag{3.57}$$

Pela substituição da equação 3.56 na equação 3.57, obtemos:

$$\begin{aligned} &maximizar, \alpha: && \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &restrição: && 0 \le \alpha_i \le C \qquad i = 1,2,\dots,N\,; e \quad \sum_{i=1}^{N} y_i \alpha_i = 0 \end{aligned} \tag{3.58}$$

A função de decisão pode ser calculada como:

$$f(\boldsymbol{x}) = sgn[\sum_{i=1}^{N} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b] \tag{3.59}$$

e b pode ser calculado usando a seguinte equação:

$$b = \frac{1}{n_{SV}: \alpha < C} \sum_{\boldsymbol{x}_j \in SV:\, \alpha_j < C} (\frac{1}{y_j} - \sum_{x_i \in SV} \alpha_i\, y_i\, \Phi(\boldsymbol{x}_i).\Phi(\boldsymbol{x}_j)) \tag{3.60}$$

Percebe-se pelas equações 3.57, 3.58 e 3.59 que a única informação necessária sobre o mapeamento é de como realizar o cálculo de produtos internos entre os dados no espaço de características, pois tem-se sempre $\Phi(\boldsymbol{x}_i).\Phi(\boldsymbol{x}_j)$, para dois dados $\boldsymbol{x}_i, \boldsymbol{x}_j$, em conjunto.

Por fim, para uma melhor compreensão do modelo SVC, sua arquitetura é mostrada na figura 3.7 que ilustra a amostra de previsão $\mathbf{x_t}$ e os vetores de suporte $\mathbf{x_1}$ ... $\mathbf{x_n}$ que são mapeados em um espaço de características com a função não linear Φ, e os produtos internos são determinados pela função kernel.

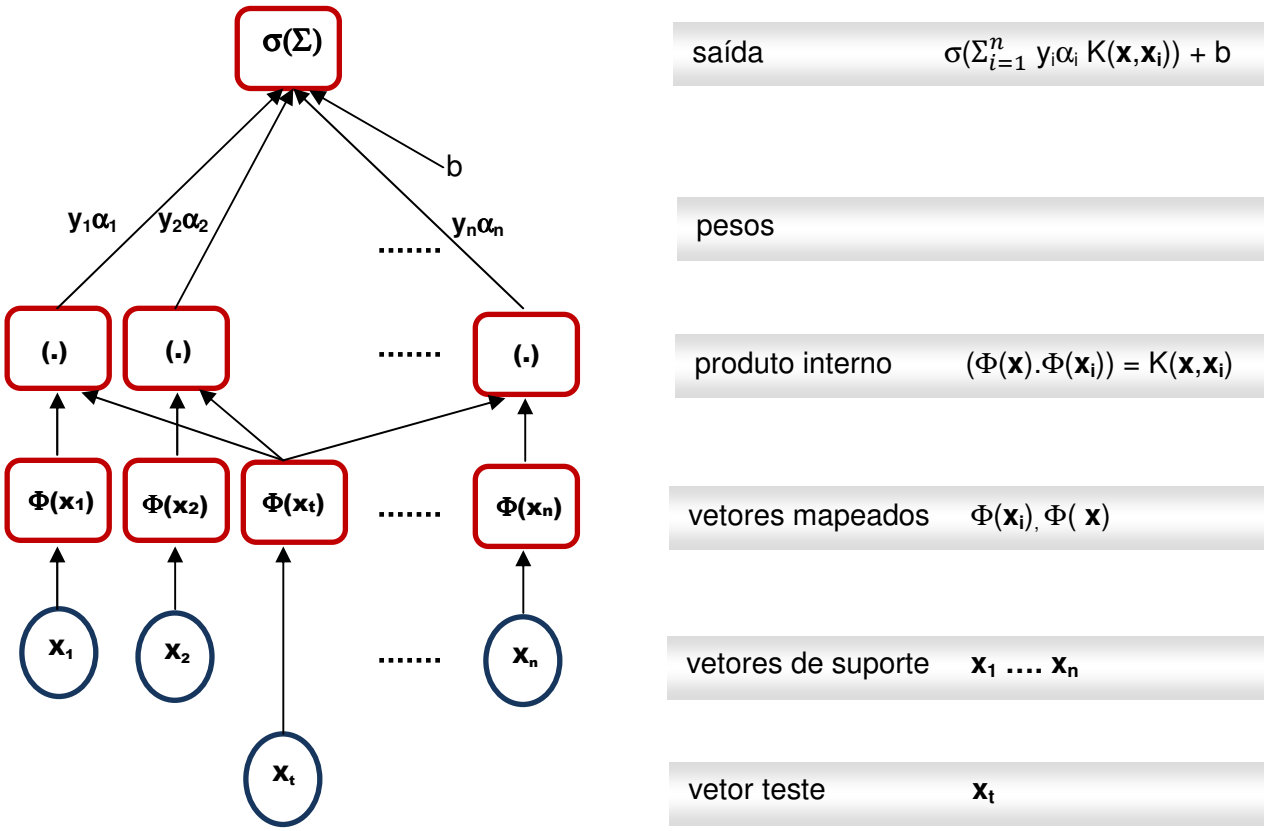


Figura 3.7 – A arquitetura da máquina de vetores de suporte

A função de decisão final é uma combinação linear das saídas transformadas com base na função kernel mais o vetor de pesos e o *bias* b. A combinação linear é obtida pela função $\sigma(x) = sgn(x + b)$. Do ponto de vista da sua arquitetura, ele é muito semelhante a rede neural artificial, por isso é também chamado de rede de vetores de suporte.[2,38,39]

3.4.4.3 - Seleção de parâmetros para o SVC

A otimização dos parâmetros C (fator de penalização) e γ (parâmetro do kernel RBF) é de fundamental importância no SVC uma vez que seus valores combinados determinam a largura e a complexidade da margem e consequentemente o desempenho do modelo de classificação. Quanto menor o valor de C maior a ênfase na maximização da margem, determinada por uma fronteira menos complexa. Com o

aumento do valor de C a minimização do erro é enfatizada, proporcionando uma margem mais estreita e determinada por uma fronteira mais complexa. Quanto maior o valor de γ tanto mais simples será a margem, tendendo para uma margem linear e levando a falta de ajuste no caso de conjuntos de dados com relações não lineares, e quanto menor o valor de γ tanto mais complexa a fronteira se torna.

Para realizar essa otimização, diferentes métodos podem ser empregados, como *grid search* e algoritmo genético. Ambos os métodos são baseados em uma validação cruzada para avaliação da performance do modelo e minimização do risco de sobreajuste. Nesse sentido, além da adequada seleção dos parâmetros do modelo SVM é importante também observar o número de vetores de suporte utilizados no modelo, o que não pode ser avaliado apenas pela validação cruzada. Eventualmente diferentes combinações de parâmetros podem fornecer o mesmo desempenho do modelo na validação cruzada. Nesses casos, uma adequada escolha entre os modelos ajustados pode ser feita através do número de vetores de suporte utilizados em cada modelo. Aquele com o menor número de vetores de suporte proporciona minimizar o efeito de um potencial sobreajuste.[43,52,53]

Nesse estudo, para o desenvolvimento dos modelos de classificação com SVM e com utilização do kernel RBF, utilizou-se o *grid search* por se tratar de um método mais simples e rápido, fornecendo bons resultados quando apenas dois parâmetros necessitam ser selecionados.

3.4.5 – Máquinas de vetores de suporte para regressão (SVR)

O SVM foi inicialmente desenvolvido para resolver problemas de classificação e depois foi estendido para resolver problemas de regressão, mantendo as principais propriedades que caracterizam o algoritmo de máxima margem, como a dualidade, esparsidade, utilização de função kernel e convexidade. Porém, diferentemente, o algoritmo de regressão por vetores de suporte introduziu uma função de perda que ignora erros que estão além de uma certa distância dos valores considerados válidos. Esse tipo de função é denominada função de perda ε insensível (análogo a margem

suave no SVC), proposta por Vapnik e controla um parâmetro que é equivalente ao parâmetro de margem para separação de hiperplanos. A outra motivação para utilização da função de perda ε insensível é que ela garante a esparsidade da variável dual (conjunto esparso de vetores de suporte), como ocorre no algoritmo de classificação com SVM.[44,46]

Assim, no SVR dois conceitos importantes devem ser primeiramente comentados, o ε-tubo e a função de perda ε- insensível. Como no SVC, os dados originais são primeiramente mapeados não linearmente em um espaço de características de alta dimensão e então uma função linear é ajustada para aproximar a função latente entre $\boldsymbol{X}$ e y.

3.4.5.1 - ε-tubo e função de perda ε- insensível

A teoria do aprendizado estatístico pode também ser aplicada para problemas de regressão pela introdução de uma nova função de perda, de modo a evitar a modelagem de erro experimental e o sobreajuste dos modelos: a função de perda ε-insensível.

Na regressão com vetores de suporte, o propósito é encontrar uma função f(x) que tem no máximo o desvio ε para as respostas y_i encontradas para todo o conjunto de treinamento e ao mesmo tempo permite que o ε-tubo seja o mais delgado possível. Em outras palavras, não queremos tratar com erros menores que ε, mas, não se aceita nenhum desvio além desse valor sem que haja penalização.

O ε-tubo, mostrado na figura 3.8 pode ser obtido pelo movimento da linha sólida com um deslocamento de ε para cima e para baixo. Aqui ε é um número positivo pré-definido.

A função de perda ε- insensível tem a seguinte forma:

$$L(y - f(x), \varepsilon) = \begin{cases} |y - f(x)| - \varepsilon\,, & se \quad |y - f(x)| \geq \varepsilon \\ 0, & se \quad de\ outra\ maneira \end{cases} \tag{3.61}$$

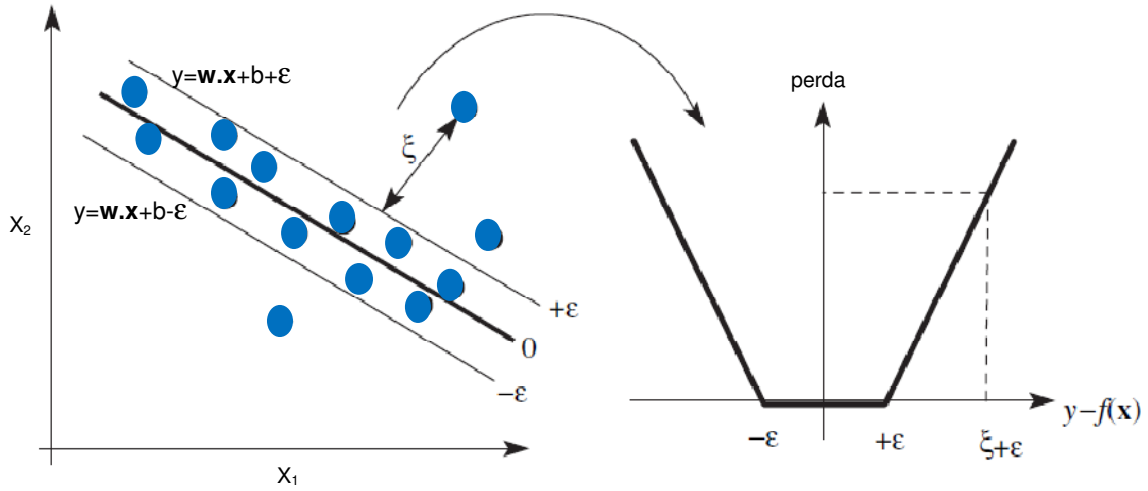


Figura 3.8 – Ilustração do processo de penalização da função de perda ε-insensível para uma SVM com kernel linear.

De fato, essas restrições definem um tubo de raio ε ao redor da função de regressão hipotética, de forma que se um ponto está posicionado dentro do tubo a função de perda é igual a 0, enquanto se um ponto está posicionado fora do tubo, a perda é proporcional a distância euclidiana entre o ponto e o raio ε do tubo.

3.4.5.2 - ε-SVR linear

Considere o problema de aproximar um conjunto de dados de treinamento, composto por: $T = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, onde $\boldsymbol{x}_i$ representa o vetor de amostras, y_i as respostas correspondentes e N é o número total de amostras, com uma função linear como a da equação 3.36, então, a partir dos dados de treinamento, o algoritmo do ε-SVR linear resolve o problema de otimização, o qual pode ser escrito da seguinte forma, com um termo ε- insensível:

$$minimizar: \quad \frac{1}{2} \| \boldsymbol{w} \|^2 + \frac{C}{N} \sum_{i=1}^{N} L(y_i - f(\boldsymbol{x}_i), \varepsilon) \tag{3.62}$$

onde minimizar $\frac{1}{2} \| \boldsymbol{w} \|^2$ corresponde a obtenção do ε-tubo o mais delgado possível e a constante C > 0 (um parâmetro pré-definido de regularização) determina a penalização

para minimização do erro de treinamento ou risco empírico e o consequente custo para obtenção da região delgada do ε-tubo, considerando o termo de complexidade do modelo $\| \boldsymbol{w} \|^2$ e o limite para o qual desvios maiores que ε são tolerados.

De forma análoga a utilização da margem suave no SVC, agora iremos introduzir as variáveis de folga ξ_i , ξ_i^* para tratar com outro tipo de restrição de resolução necessária ao problema de otimização. Então, o problema de minimização acima pode também ser expresso da seguinte forma, com as variáveis de folga ξ_i , ξ_i^* representando as restrições dos dois lados do hiperplano, conforme as expressões abaixo:

$$
\begin{aligned}
&minimizar: && \tfrac{1}{2} \| \boldsymbol{w} \|^2 + \tfrac{C}{N} \textstyle\sum_{i=1}^{N} (\xi_i + \xi_i^*) && && (3.63)\\
&restrição: && (\boldsymbol{w}.\boldsymbol{x_i} + b) - y_i \leq \varepsilon + \xi_i, && i = 1,2,\dots,N\\
& && y_i - (\boldsymbol{w}.\boldsymbol{x_i} + b) \leq \varepsilon + \xi_i^*, && i = 1,2,\dots,N\\
& && \xi_i, \xi_i^* \geq 0, && i = 1,2,\dots,N
\end{aligned}
$$

Somente as amostras fora do ε-tubo na figura 3.8 contribuem para o funcional da inequação 3.63, isto é, as variações que excedam ε são penalizadas, neste caso de forma linear.

As variáveis de folga ξ_i , ξ_i^* são introduzidas quando os valores nominais excedem, com respeito a origem no espaço de dados original, o limite superior de ε (ξ_i) e o limite inferior de ε (ξ_i^*) . Os pontos fora do ε tubo são chamados de vetores de suporte, porque estabelecem os fundamentos para a função de regressão estimada. Isso significa que todos os outros pontos não são incluídos no modelo e podem ser removidos após a construção do modelo SVR. Assim, normalmente, muito menos objetos de treinamento constituem o modelo de regressão, e por esse motivo a solução é definida como sendo esparsa. Tal solução é inerente ao algoritmo SVM inicialmente desenvolvido para resolução de problemas de reconhecimento de padrões.

Com o auxilio do método do multiplicador de Lagrange e um algoritmo de programação quadrática a função de regressão da equação 3.63 pode ser reformulada segundo o formalismo de um problema dual, da seguinte forma:

$$f(x) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)(\boldsymbol{x_i}.\boldsymbol{x}) + b \quad (3.64)$$

$$b = y_j - \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)(\boldsymbol{x_i}.\boldsymbol{x}) + \varepsilon \quad (3.65)$$

onde α_i^* e α_i são os multiplicadores de Lagrange otimizados.

Os vetores de suporte vão corresponder às amostras para as quais exatamente um dos multiplicadores de Lagrange é maior que zero, ou seja, as amostras fora do tubo ε insensível. Quando ε = 0, obtém-se então uma função de perda de Laplace, e o problema de otimização é simplificado.[39,44]

3.4.5.3 - ε-SVR com utilização da função kernel

Frequentemente, um modelo não linear é necessário para modelar adequadamente os dados. É necessário estender o ε-SVR linear para a regressão não linear. Com a introdução de uma função kernel, os dados de entrada originais são primeiramente não linearmente mapeados no espaço de características de elevada dimensão, onde uma regressão linear pode ser utilizada para tratar com complicados problemas de regressão não linear em química.

Como o procedimento de derivação da função de decisão final é muito parecido com o caso linear, mostra-se a seguir apenas a forma final.

$$f(x) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)K(\boldsymbol{x}_i\boldsymbol{x}) + b \quad (3.66)$$

$$b = y_j - \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)K\left(\boldsymbol{x}_i\boldsymbol{x}_j\right) + \varepsilon \quad (3.67)$$

onde α_i^* e α_i são os multiplicadores de Lagrange otimizados.

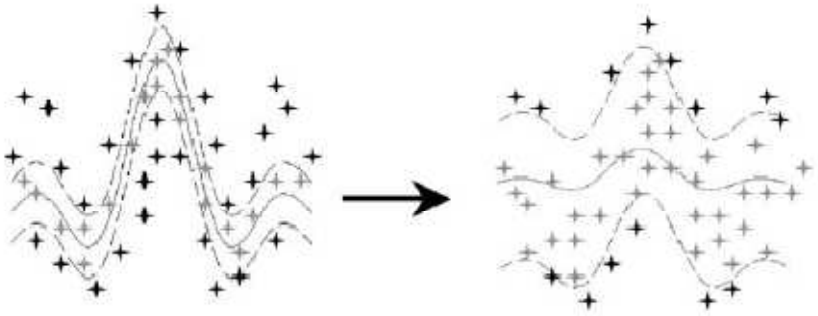
A figura 3.7 pode agora ser utilizada para fornecer uma visão gráfica dos diferentes passos envolvidos na regressão com SVR com as modificações pertinentes ao problema de regressão em que os pesos são os valores de $\alpha_i^* - \alpha_i$ e a combinação linear é obtida pela função $\sigma(x) = (x + b)$.

O parâmetro do kernel e também os citados parâmetros C e ε necessitam ser adequadamente selecionados pelo usuário, porque a capacidade de generalização do modelo SVR depende sensivelmente da correta utilização desses parâmetros.[39,44]

3.4.5.4 – Seleção de parâmetros para o ε-SVR

Conforme mencionado, o parâmetro do kernel e os parâmetros ε e C necessitam ser otimizados.

O parâmetro ε regula o raio do tubo ε ao redor da função de regressão e, então, o número de vetores de suporte que serão selecionados para construção da função de regressão (proporcionando uma solução esparsa). Um valor muito grande de ε resulta em poucos vetores de suporte (mais pontos serão ajustados no interior do tubo ε) e, consequentemente, uma função de regressão menos robusta, como mostra a figura 3.9(b). Nesse caso a função de regressão resultante não será sempre aplicável devido a obtenção de elevados erros de previsão. Sabe-se que o valor de ε está relacionado com a amplitude do ruído presente no conjunto de treinamento. Desde que a exata contribuição do ruído presente na real informação em um conjunto de dados é normalmente desconhecida, ε pode variar em um intervalo proposto pelo usuário.



(a) (b)

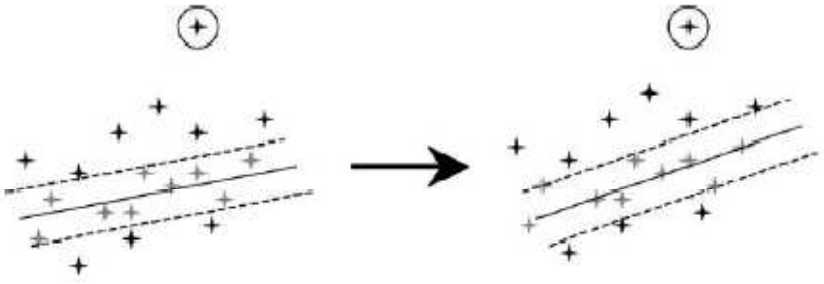
Figura 3.9 – Influência do valor de ε no ajuste do modelo SVR . Em (a) com um valor menor de ε e em (b) com um valor maior de ε.

Na figura 3.9 (a) ilustra-se que um valor pequeno de ε permite que mais pontos estejam externos ao tubo ε (linha tracejada) e resulta em mais vetores de suporte

(pontos negros). Na figura 3.9 (b) um valor grande de ε resulta em menos vetores de suporte.

O parâmetro C determina a penalização para os casos em que a função de regressão aceita objetos com desvios maiores que o valor de ε. Dessa forma, a robustez do modelo de regressão depende da escolha do valor de C, porque os valores mais altos de α_i e α_i^* são, por definição (de acordo com o procedimento de otimização de Lagrange) iguais a C. Isso significa que a escolha do valor de C influencia o significado de objetos individuais do conjunto de treinamento. Por exemplo, um alto valor de C resulta em vetores de suporte com uma grande diferença entre os valores de α_i e α_i^*. Nesse caso os vetores de suporte com os mais elevados valores de α_i e α_i^* estão dominando a construção da função de regressão. Por outro lado, um valor pequeno de C pode resultar em vetores de suporte com pequena diferença ou mesmo valores similares de α_i e α_i^*. Nesse caso os objetos selecionados como vetores de suporte com valores similares de α_i e α_i^* contribuem igualmente para a função de regressão.

Assim, uma escolha apropriada de C em combinação com ε deve resultar em uma boa e robusta performance do modelo de regressão, que é também insensível a presença de possíveis *outliers*, como ilustra a figura 3.10.



(a) (b)

Figura 3.10 – Influência do valor de C no ajuste do modelo SVR. Em (a) com um valor menor de C e em (b) com um valor maior de C.

Uma boa seleção de ambos os parâmetros também previne um sobreajuste do modelo e isso pode ser verificado através da realização de validações cruzadas

internas durante o ajuste do mesmo. De acordo com a teoria do SVR, C pode assumir valores entre 0 e ∞ . Nesse trabalho esse valor varia entre 1 e 10000. [43,44]

Na figura 3.10 (a) ilustra-se que no caso de um valor de C pequeno os vetores de suporte mais distantes têm a mesma contribuição. Por essa razão, a função de regressão é mais robusta a possíveis *outliers* (vetor de suporte dentro do circulo). Na figura 3.10 (b) o aumento do valor de C resulta em uma linha de regressão, que é mais influenciada por possíveis *outliers*. A distância entre o *outlier* e o tubo ε irá diminuir em comparação com o caso de um menor valor de C selecionado.

3.4.6 – ν-SVM

No ν–SVM o parâmetro ν permite um controle mais eficaz do número de vetores de suporte utilizados nos modelos. No caso da classificação utiliza-se o parâmetro ν em lugar do parâmetro de regularização C e no caso da regressão utilizam-se os parâmetros ν e C em lugar de ε e C.[54]

O parâmetro $\nu \in (0,1]$ determina um limite superior para a fração do conjunto de treinamento que contribui para erros na determinação da fronteira de margem ou do ε-tubo e determina também um limite inferior para o número de vetores de suporte.[54]

Para um problema de classificação, o ν-SVC, remove a constante C e introduz um novo parâmetro ν que controla o número de vetores de suporte e os erros de treinamento,

$$\begin{aligned} &minimizar: \quad \frac{1}{2} \| \boldsymbol{w} \|^2 - \nu\rho + \frac{1}{N} \sum_i^N \xi_i \\ &restrição: \quad (\boldsymbol{w}.\boldsymbol{x_i} + b) y_i \geq \rho - \xi_i, \xi_i \geq 0; e\ \rho \geq 0 \end{aligned} \tag{3.68}$$

em que a restrição envolve o parâmetro de margem ρ que também é uma variável no problema de otimização.

Quanto maior o valor de ν, maior o número de objetos permitidos internos a margem. Utilizando a técnica dos multiplicadores de Lagrange é possível obter a resolução desse problema de otimização, obtendo as funções finais mostradas nas equações 3.55 e 3.59.

Para a regressão, de forma similar, o ν-SVR utiliza o parâmetro ν para controlar o número de vetores de suporte. No entanto, em lugar de substituir a constante C, aqui ν substitui o parâmetro ε, do ε-SVR. Após selecionar o ν, o valor de ε é automaticamente selecionado pelo algoritmo.[54]

No ν-SVR embora o parâmetro ε controle a esparsidade da solução, ele o faz apenas de uma forma indireta. Devido a escassez de informação inicial quanto a exatidão dos valores de y, isso torna difícil a obtenção de um valor razoável de ε inicialmente. Em vez disso é preferível especificar o grau da esparsidade e deixar o algoritmo computar ε a partir dos dados. Essa é a idéia do ν-SVR, uma modificação do ε-SVR original. Ele preconiza que para obtenção da máxima capacidade de generalização, o parâmetro de esparsidade $\nu \in (0,1]$ deve ser selecionado de acordo com o ruído presente nos valores de y.

No ν-SVR a grandeza de ε não é definida inicialmente, mas é também uma variável. Seu valor é ajustado com base na complexidade do modelo e nas variáveis de folga, através do parâmetro $\nu \in (0,1]$:

$$minimizar: \qquad \frac{1}{2} \| \boldsymbol{w} \|^2 + C(\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*)) \tag{3.69}$$

sujeita as restrições da inequação 3.63.

Quanto maior o valor de ν mais pontos são permitidos externos ao tubo. Por exemplo, se na figura 3.9 (a) e (b) temos valores de ν = 0,8 e ν = 0,2, respectivamente, o algoritmo ajusta automaticamente o valor de ε = 0,04 e ε = 0,22 para os casos da figura 3.9 (a) e 3.9 (b), respectivamente.

Utilizando a técnica dos multiplicadores de Lagrange é possível obter a resolução desse problema de otimização convexo com um mínimo global, obtendo as funções de regressão mostradas nas equações 3.64 e 3.66.

Uma completa descrição quanto a formulação do ν-SVM para classificação e regressão pode ser obtida na literatura.[54]

3.4.7 - Máquina de vetores de suporte por mínimos quadrados – LSSVM

Essa variação do algoritmo SVM foi proposta com a modificação da metodologia SVM através da introdução da função de perda por quadrados mínimos, em lugar da função de perda ε-insensível, e restrições de igualdades, em lugar de restrições de desigualdades. Em lugar de resolver um problema de programação quadrática, a solução é obtida a partir de um conjunto de equações lineares, reduzindo significantemente a complexidade e o esforço computacional.

Mas, apesar das características atrativas computacionalmente, a solução baseada em LSSVM também tem algumas limitações. Uma delas está associada a perda da natureza esparsa do problema. Os valores suporte (α_i) são proporcionais aos erros dos correspondentes dados de treinamento. Isso implica que usualmente todos os pontos dos dados de treinamento são vetores de suporte e a propriedade de esparsidade não mais existe na formulação do LSSVM. Valores de suporte elevados, indicam uma maior contribuição do ponto do conjunto de treinamento na fronteira de decisão.

Também, o uso da função custo baseada no erro quadrático sem regularização pode conduzir a estimativas que são menos robustas principalmente quando há *outliers* nos dados de treinamento. Para atenuar tais limitações, foi proposta uma versão ponderada do LSSVM. Nesta versão, introduz-se uma variável que pondera a importância dos erros cometidos pelo modelo na função custo.[53]

3.4.8- Seleção de parâmetros para o SVM com algoritmo genético e *grid search*

Uma vez que uma boa performance de generalização dos modelos SVM para classificação e regressão depende de uma apropriada seleção dos seus parâmetros, a principal questão é encontrar a melhor seleção para um determinado conjunto de dados, e, não há uma definição geral de como selecioná-los. O problema da otimização da seleção de parâmetros torna-se mais complicado devido a necessidade de utilização conjunta dos mesmos (interação dos parâmetros) pelo algoritmo SVM. Isso significa

que a otimização de forma separada de cada parâmetro não é suficiente para encontrar o melhor modelo de regressão.

Nesse trabalho tanto o algoritmo genético como o método *grid search* foram utilizados para otimização paramétrica no desenvolvimento de modelos com SVM.

O algoritmo genético – GA[55] é um algoritmo de seleção baseado no princípio da evolução natural. O GA é uma técnica estocástica de seleção que pode ser utilizada para encontrar uma solução ótima global em um espaço de seleção multidimensional. O algoritmo inicia com um conjunto randômico de soluções denominado população. Uma solução individual é representada de forma codificada (código binário) e denomina-se cromossomo. Cada cromossomo é formado por uma sequência de estruturas individuais denominadas genes, que representam os parâmetros a serem otimizados. As soluções de uma população são utilizadas para gerar a população seguinte. Isso é motivado pela idéia de que a nova população, na média, será melhor do que a população anterior. Para criar a nova população o GA utiliza os operadores genéticos de mutação e permutação cruzada ou *crossover* e um processo de seleção. Os operadores genéticos são utilizados para criar as novas soluções a partir do conjunto de soluções atual. Essa seleção reflete o princípio da melhor aptidão das novas soluções e é o mecanismo para encontrar melhores soluções a partir da população atual. Durante o processo de seleção, as soluções são selecionadas de acordo com os seus valores para a função objetivo. O GA irá repetir esse processo até que a condição de término seja satisfeita. A melhor solução é definida como a solução otimizada.

O processo do GA pode ser resumida nas seguintes etapas:

i) obter uma população gerada randomicamente;
ii) calcular a aptidão de cada cromossomo na população;
iii) criar a nova população com os operadores genéticos: seleção, mutação e permutação cruzada;
iv) verificar a condição de terminação. Se a nova população não satisfaz a condição de término as etapas (ii) a (iv) são repetidas a partir da nova geração.

Nesse trabalho a função objetivo a ser otimizada pelo GA, e definido inicialmente, consistiu na utilização dos valores obtidos pela validação cruzada no conjunto de treinamento, buscando-se o menor valor do erro de validação cruzada.

A otimização paramétrica através do método *grid search*[56] também é comumente utilizada e proporciona bons resultados com um tempo de processamento razoável para seleção de dois parâmetros, buscando-se o menor valor do erro de validação cruzada.

Os modelos de regressão e classificação com SVM desenvolvidos nesse trabalho utilizaram o algoritmo genético e o método *grid search*, respectivamente, para otimização paramétrica.

4 - Petróleo e produção de óleo diesel

O petróleo é uma mistura que consiste predominantemente de hidrocarbonetos e derivados orgânicos sulfurados, nitrogenados e oxigenados. O petróleo bruto está comumente acompanhado por quantidades variáveis de substâncias estranhas tais como água, matéria inorgânica e gases. Nessas condições ele encontra poucas aplicações práticas, servindo quase que tão somente como óleo combustível. Para que ele tenha seu potencial energético e não energético plenamente aproveitado é necessário que através de um adequado processo de refino, baseado em processos de separação, conversão e tratamento, obtenham-se a as diversas frações do petróleo. Essas frações podem ser adequadamente misturadas para obtenção de produtos derivados do petróleo de grande valor agregado, segundo especificações estabelecidas.

Entre as moléculas de hidrocarbonetos constituintes do petróleo encontram-se desde o metano até cadeias com inúmeros átomos de carbono, como os asfaltenos. A predominância de determinadas classes destes compostos conferem características distintas aos petróleos. Por exemplo, enquanto o óleo encontrado na península arábica é considerado leve e parafínico (predominância de hidrocarbonetos saturados de menor massa molecular) a maior parte do petróleo produzido no Brasil atualmente é pesado e com grande quantidade de hidrocarbonetos aromáticos, naftênicos, resinas e asfaltenos.

Estas diferentes composições levam a modos distintos de refino de petróleo a fim de obter os produtos derivados. De um modo geral, no refino de petróleos mais pesados utiliza-se mais energia para a obtenção dos derivados além da necessidade da implementação de processos adicionais de remoção de contaminantes, como o enxofre (cujo teor encontrado aumenta com o aumento da massa molecular média do petróleo) e como o nitrogênio encontrado em maior teor nos óleos brasileiros quando comparados a óleos provenientes de outros países. [57,58]

4.1 – Panorama da indústria do petróleo e derivados no Brasil

A importância do desenvolvimento da química analítica para obtenção de soluções na indústria do petróleo pode ser ilustrado por alguns números relativos ao desenvolvimento desse setor divulgados pela Agência Nacional do Petróleo – ANP em 2011,[59] quanto a exploração e produção, e, refino e processamento.

Em 2010, a produção nacional diária de petróleo - incluindo óleo cru e condensado, mas excluindo líquido de gás natural (LGN) e óleo de xisto - aumentou 5,6% e chegou a 750 milhões de barris o que elevou o Brasil à 12ª colocação no ranking mundial de produtores de petróleo. Nos últimos 10 anos, o crescimento médio anual da produção brasileira foi de 5,3%.

Em 2010, as 16 refinarias nacionais (12 pertencem a Petrobras e responderam por 98,1% da capacidade total) – não incluindo a Superintendência de Industrialização do Xisto (SIX) – somaram uma capacidade de refino de 332,7 mil m³/dia. A capacidade de refino, considerando utilização de 95%, foi de aproximadamente 316,1 mil m^3/dia.

Em 2010, a produção brasileira de derivados de petróleo foi de 110,1 milhões m³, 0,6% a mais que em 2009. Do volume total de derivados produzido no Brasil, o óleo diesel teve participação de 37,6% ou 41,4 milhões m³, e a gasolina A de 20,9% ou 23,1 milhões m³. Entre os derivados não-energéticos, destacou-se a nafta, responsável por 6,6% da produção total de derivados e 42,3% da produção de não energéticos.

A figura 4.1 ilustra a distribuição percentual da produção de derivados do petróleo energéticos em 2010.

Para complementar o suprimento nacional, o Brasil importou 123,6 milhões de barris de petróleo em 2010, 13,8% a menos que no ano anterior.

Em 2010, as importações de derivados de petróleo totalizaram 27,4 milhões m³, volume 71,8% superior ao registrado em 2009. Os derivados energéticos representaram 53,8% das importações, após um aumento de 100,2% em relação ao ano anterior. Já os não-energéticos tiveram crescimento de 47,4% e atingiram 12,7 milhões m³ ou 46,2% do total. Os derivados energéticos importados em maior quantidade foram óleo diesel, GLP e QAV com, respectivamente, 32,9%, 11,4% e 7% do volume total. Dentre os não-energéticos, a nafta se sobressaiu com 24,5%.

Em 2010, a exportação de derivados de petróleo totalizou 13,8 milhões m³, um decréscimo de 9,1% em relação a 2009. Os derivados energéticos representaram 94,3% do total exportado, com destaque para o óleo combustível, com 4,9 milhões m³ ou 35,8% do total.

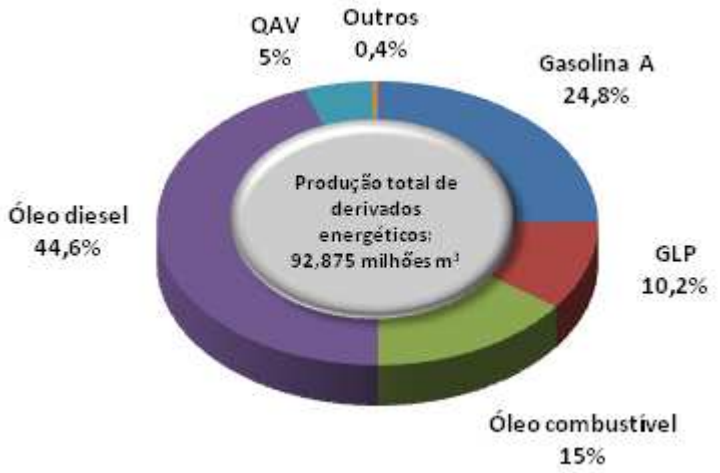


Figura 4.1 - Distribuição percentual da produção de derivados do petróleo energéticos em 2010 (adaptado de Anuário estatístico ANP 2011 [59])

Em 2010, o Brasil teve um resultado superavitário no comércio internacional de petróleo e derivados, reafirmando a auto suficiência de abastecimento de petróleo e derivados alcançada em 2006. A exportação líquida de petróleo bruto foi de 46,5 mil m³/dia. Por outro lado, a importação líquida de derivados foi de 37,2 mil m³/dia.[59]

4.2 – Processos de produção dos derivados do petróleo

Os processos comumente encontrados nas refinarias podem ser divididos em três grandes classes:

(i) Processos de Separação, quando os constituintes existentes na carga do processo são separados de acordo com alguma propriedade física que os caracterize, tal como ponto de ebulição (destilação), solubilidade (desaromatização, desasfaltação), ponto de fusão (desparafinação) e outros. Nestes processos não ocorre nenhuma transformação química dos constituintes da carga;

(ii) Processos de conversão, quando os hidrocarbonetos constituintes da carga são transformados em outros hidrocarbonetos através de processos químicos, catalíticos ou não. Estes processos são complementados por processos físicos, para separar as frações obtidas pela transformação dos constituintes da carga;
(iii) Processos de tratamento, quando não ocorre alteração física nem química nos hidrocarbonetos, havendo, no entanto, a remoção ou transformação dos contaminantes da carga através de processos químicos ou físicos, como tratamento cáustico, o tratamento Merox, o hidrotratamento e outros. Os processos de tratamentos são via de regra usados em sequência aos processos de separação e de conversão, sendo por isso algumas vezes chamados de processos de acabamento.

Tabela 4.1 – Aplicações comerciais das frações de destilação do petróleo (adaptado de Valle [60])

Fração	Faixa de destilação (ºC)	Principais aplicações comerciais
Gás combustível	Abaixo de – 44	Gás combustível; petroquímica
Gás liquefeito do petróleo	- 44 a 0	Combustível doméstico e industrial; petroquímica
Nafta leve	30 a 90	Gasolina; petroquímica; solventes
Nafta pesada	90 a 170	Gasolina; petroquímica; obtenção de aromáticos
querosene	170 a 270	QI; QAV; óleo diesel, detergentes (parafinas)
Gasóleo leve atmosférico	270 a 320	Óleo diesel; óleo de aquecimento
Gasóleo pesado atmosférico	320 a 390	Óleo diesel; gasóleo petroquímico
Gasóleo leve de vácuo	390 a 420	Lubrificantes; óleo diesel
Gasóleo pesado de vácuo	420 a 550	Carga de FCC; lubrificantes
Resíduo de vácuo	Acima de 550	Óleo combustível; lubrificantes; asfaltos, coque

O refino do petróleo se inicia através da sua separação física nas frações básicas do refino por destilação atmosférica e a vácuo, separadas de acordo com suas faixas de temperaturas de ebulição, conforme mostrado na tabela 4.1. A destilação é realizada a pressão atmosférica até aproximadamente 400 ºC (não há degradação) após o que é feito vácuo no sistema, mantendo-se a temperatura de fundo a 400 ºC. Essas frações são encaminhadas para tanques de estocagem finais, onde irão compor misturadas ou não a outras frações os derivados finais. As frações básicas podem ainda ser enviadas a tanques intermediários, de onde seguem para outros processos de separação, transformação ou acabamento, de onde seguirão ainda para outros processos ou para tanques de produtos acabados.[60]

Assim, os derivados do petróleo podem ser compostos por frações de diversos processos de refino, constituindo o chamado *pool*, conjunto de frações que fazem parte de um derivado de petróleo.[57,58]

4.3 – Óleo diesel

Conforme mostrado na figura 4.1, o óleo diesel é o derivado de petróleo de maior produção no Brasil, devido a elevada demanda por este combustível uma vez que o principal modal de transporte de cargas no país é o transporte rodoviário. Apesar da auto suficiência brasileira em petróleo, o óleo brasileiro produzido atualmente é predominantemente pesado e, por isso, não é totalmente processado no país, pois o parque de refino, concebido na década de 70 para o processamento de óleos leves, não é capaz de produzir a quantidade de derivados necessária ao consumo interno, sobretudo de óleo diesel, a partir deste tipo de petróleo. Desta forma o Brasil exporta o excedente de óleo pesado e importa óleo leve além de importar óleo diesel a fim de garantir o abastecimento deste combustível no país.

O óleo diesel é o derivado do petróleo constituído por hidrocarbonetos de 10 a 18 átomos de carbono com faixa de destilação comumente situada entre 220 e 380℃. Este derivado apresenta um conjunto de propriedades físico-químicas adequadas para utilização como combustível em motores ciclo diesel. No óleo diesel os hidrocarbonetos com moléculas de cadeia carbônica linear e sem ramificações, como as parafinas,

proporcionam as propriedades desejáveis para a ignição e um bom funcionamento do motor ciclo diesel. Por outro lado, a presença de moléculas com cadeias carbônicas ramificadas ou que possuem estruturas cíclicas estáveis, como naftênicos e aromáticos, prejudicam as características desejáveis do óleo diesel e consequentemente requerem maior temperatura e pressão para entrar em ignição.[61]

O *pool* de óleo diesel, denominação dada ao conjunto de correntes de uma refinaria que compõem a mistura do óleo diesel, pode ser composto das mais diversas correntes. Um exemplo das correntes que podem compor um *pool* pode ser visualizado na tabela 4.2.[58]

Tabela 4.2 – Exemplos de correntes utilizadas no *pool* de óleo diesel (adaptado de Bezerra *et al.* [58])

Característica	Componentes do *pool* de óleo diesel					
	Nafta pesada	Querosene	Diesel leve	Diesel leve HDT	Diesel pesado HDT	Diesel exterior
Enxofre total (mg/kg)	106	918	2197	30	111	1779
T10 (℃)	112,9	168,2	219,4	228,8	244,5	185,1
T50 (℃)	122,7	198,0	251,5	269,9	316,9	279,6
T85 (℃)	144,4	238,8	287,6	326,6	380,0	362,1
Densidade (g/mL)	750,4	807,6	842,5	864,1	882,1	845,4
Ponto de fulgor (℃)	17	44,5	81,5	-	-	47,5

Apenas com a mistura adequada de diferentes correntes torna-se viável atender as especificações do óleo diesel, mostradas na tabela 4.4.

A ocorrência de diferentes hidrocarbonetos nas frações de diversos tipos de petróleos é mostrada na tabela 4.3, que fornece uma estimativa média quanto ao teor dos diferentes tipos de hidrocarbonetos em função das faixas de temperaturas de destilação, nas quais se inserem as diferentes frações do petróleo que compõem o óleo diesel.

Tabela 4.3 – Teores típicos médios de ocorrência de hidrocarbonetos em petróleos de 10 a 40 ºAPI em função das faixas de temperaturas

Faixa de temperatura de ebulição (ºC)	Parafínicos (%)	Cicloparafínicos (%)	Aromáticos totais (%)
5 – 150	59,1	29,6	10,1
150 – 250	21,6	42,9	33,3
250 – 400	25,9	44,8	25,9
400 – 550	15,1	48,1	31,9

Para o óleo diesel produto final os teores médios de hidrocarbonetos parafínicos, naftênicos e aromáticos totais são em torno de 27 %, 45 % e 28 %, respectivamente.

4.3.1 - Parâmetros de qualidade do óleo diesel

A Resolução ANP nº 65 de dezembro de 2011 especifica atualmente as características do óleo diesel S10, S50, S500 e S1800, para uso nas regiões estabelecidas pela mesma. A seguir a tabela 4.4 mostra algumas especificações bem como os métodos de referência ASTM estabelecidas por esta Resolução e os parâmetros estudados nesse trabalho são comentados. Como os modelos de regressão para determinação de parâmetros do óleo diesel desenvolvidos nesse trabalho têm aplicação em análises realizadas durante o processo de produção, alguns parâmetros discutidos não são especificados na legislação. Também, à época da coleta de dados para realização desse trabalho estava vigente a Resolução ANP nº 42 de dezembro de 2009 em que não consta especificações para o óleo diesel S10, bem como não havia produção do mesmo.

- Destilação : as características de destilação (volatilidade) de hidrocarbonetos têm um importante efeito na sua segurança e desempenho, especialmente no caso de combustíveis e solventes. A volatilidade de um combustível é o principal determinante

da tendência de um hidrocarboneto de produzir vapores potencialmente explosivos e afeta criticamente as condições de operação do motor nas fases de partida a frio e operação a elevada temperatura.

- 10% recuperados (T10) – indica a quantidade adequada de hidrocarbonetos leves responsáveis pela boa condição de partida e aquecimento do motor, a fim de que o motor a frio entre em pleno funcionamento com o menor número de rotações possíveis, favorecendo a partida fácil e rápida.

Tabela 4.4 – especificação do óleo diesel de uso rodoviário

Característica	**Limite**				**Método ASTM**
	Tipo A e B				
	S10	**S50**	**S500**	**S1800**	
Enxofre total (mg/kg)	10	50	500	1800	D2622, D5453 D7039, D7212 D7220
Destilação - T10 (°C)	180 (min)	anotar			D86
Destilação - T50 (°C)	245,0 a 295,0	245,0 a 310,0			D86
Destilação - T85, máx. (°C)	-	-	360,0	370,0	D86
Destilação – T90 (°C)	-	360,0 (máx)	anotar	anotar	D86
Destilação – T95 (°C)	370,0 (máx)	-	-	-	D86
Densidade a 20 °C (g/mL)	0,820 a 0,850		0,820 a 0,865	0,820 a 0,880	D1298 D4052
Ponto de fulgor, mín. (°C)	38,0				D56 D93 D3828
Viscosidade a 40 °C (mm^2/s)	2,0 a 4,5	2,0 a 5,0			D445
Número de cetano, mín.	48	46	42	42	D613
Outros: aspecto, cor, teor de biodiesel, ponto de entupimento de filtro a frio, resíduo de carbono, cinzas, corrosividade ao cobre, contaminação total, água e sedimentos, hidrocarbonetos policíclicos aromáticos, estabilidade à oxidação, índice de neutralização, lubricidade, condutividade elétrica					

- 50% recuperados (T50) – indica a quantidade adequada de hidrocarbonetos responsáveis pela rápida evaporação com o motor quente. A temperatura registrada do

destilado aos 50 % deve ser tal que assegure uma volatilidade média necessária para permitir uma utilização máxima de potência pelo motor.

- 85% e 90 % recuperados (T85 e T90) – limitam o teor de hidrocarbonetos de alto ponto de ebulição que podem resultar em quantidades importantes não vaporizadas, que se condensando nas paredes do cilindro descem para o cárter, diluindo o óleo lubrificante, além de propiciar a formação de depósitos de carvão na câmara de combustão.

- Densidade: esse parâmetro proporciona uma indicação da composição do combustível e da performance relacionada. A densidade tem importância na performance de motores ciclo diesel porque a injeção de combustível ocorre através de um sistema de medida de volume. Uma alteração na densidade influencia na operação do motor devido a diferença em massa de combustível injetado e combustíveis com densidade mais elevada tendem a proporcionar maior emissão de fumaça e maior consumo de energia.

- Ponto de fulgor: o ponto de fulgor representa a menor temperatura na qual o produto se vaporiza em quantidade suficiente para formar com o ar uma mistura capaz de se inflamar momentaneamente, quando se incide uma centelha sobre a mesma. No ponto de fulgor a quantidade de vapor formada não é suficiente para sustentar a combustão da amostra, isso só ocorre quando se atinge o Ponto de combustão, o qual representa a menor temperatura em que a amostra se vaporiza em quantidade tal que proporciona sua combustão contínua, por um período de no mínimo 5 segundos. O Ponto de fulgor está relacionado com o limite inferior de explosividade, indicando os riscos envolvidos no manuseio, armazenamento e transporte do combustível, isto é, considera o aspecto da segurança. O Ponto de Fulgor é inversamente proporcional a presença de frações leves sendo tanto menor quanto maior for o teor destas frações.

- Número de cetano: esse parâmetro do combustível para máquinas que operam segundo o ciclo diesel representa a qualidade de combustão medida pela sua facilidade de ignição quando submetido a elevadas pressões e alta turbulência da mistura combustível e ar no motor diesel. A qualidade de ignição do óleo diesel é medida por comparação com padrões de boa qualidade (parafínicos) e de má qualidade (aromáticos). O número de cetano do óleo diesel é definido como o número inteiro mais próximo do valor determinado por cálculo do valor da porcentagem em volume de n-

hexadecano (NC=100) em uma mistura com heptametilnonano (NC=15) que iguala a qualidade de combustão do combustível ensaiado em um motor de teste padronizado, quando comparado por este método. Para se obter o número de cetano utiliza-se a equação:

$$NC = \% \text{ hexadecano} + 0{,}15\ (\% \text{ heptametilnonano}) \quad (4.1)$$

Quanto menor o retardo de ignição, mais alto será o número de cetano do combustível. Quanto maior o teor de compostos parafínicos saturados de cadeia normal no óleo diesel melhores serão suas características de ignição para essas máquinas.

- Índice de cetano: a determinação do número de cetano pelo ensaio motor ASTM requer equipamento especial, de elevado custo, e demanda tempo e mão de obra especializada. Foram desenvolvidos métodos alternativos para estimar o número de cetano, sendo que os cálculos são baseados em equações que envolvem outras características do combustível. Um dos métodos mais utilizados é o ASTM D976 que calcula o índice de cetano através da equação 4.2 que envolve o ponto médio de ebulição da destilação (T50) e a densidade (d).

$$IC = 454{,}74 - 1641{,}416\ d + 774{,}74\ d^2 - 0.554\ T50 + 97{,}803\ (\log T50)^2 \quad (4.2)$$

O índice de cetano calculado apresenta algumas limitações: i) não é aplicável a combustíveis contendo aditivo para elevar o número de cetano; ii) a equação é válida para óleo diesel convencional e apresenta desvios se utilizada para resíduos, frações leves ou produtos de craqueamento catalítico ou térmico.

- Ponto de anilina: a legislação não especifica os limites desse parâmetro no óleo diesel, porém, ele é importante na caracterização da carga do processo de HDT. Este ensaio, realizado pelo método de referência ASTM D611-07, serve para avaliar a predominância de hidrocarbonetos aromáticos ou parafínicos de um produto. Ele se baseia na maior ou menor facilidade com que a anilina (amina aromática) se mistura com o produto. O ponto de anilina é a menor temperatura na qual uma mistura de anilina e amostra, em volumes iguais, apresenta condições de completa miscibilidade. Sendo a anilina (fenilamina) um composto aromático, quanto maior a aromaticidade da amostra, mais baixo será seu ponto de anilina.

4.3.2 – Mistura em linha e obtenção da qualidade final na produção do óleo diesel

O óleo diesel e os demais derivados do petróleo obtidos através do processo de refino são quase sempre uma mistura de correntes oriundas de diferentes unidades de processo na refinaria e de correntes externas. O processo de mistura pode ser considerado um processo em batelada em que o volume misturado é fixado pela programação de produção da refinaria.

Existem três formas de realizar operações de mistura de frações de petróleo a fim de compor um derivado: i) bateladas em tanques; ii) automatizadas em linhas a partir de frações armazenadas em tanques; iii) automatizadas em linhas a partir de frações das unidades de refino (*on-line*).

A produção em bateladas demanda maior tempo de operação e mais instalações, pode resultar em correções e re-certificações, além de não otimizar o processo e exigir mais mão-de-obra. Por outro lado, as misturas em linhas a partir de tanques, apesar de serem mais eficazes que aquelas em bateladas, requerem maior tancagem e necessidade de maior número de operações do que aquelas a partir das unidades de refino. Quanto mais eficaz for a operação, maior será seu nível de automação e maior o investimento em sistemas automatizados de controle. No caso da refinaria de Paulinia – Replan, que forneceu as amostras de óleo diesel para esse trabalho, a mistura é feita em bateladas e automatizada em linhas a partir de frações armazenadas em tanques. Serão apresentados mais detalhadamente a seguir estes dois modos de produção:

- Bateladas em tanques: esta é a forma mais utilizada em todas as refinarias para todos os produtos de um modo geral, pois não necessita de investimento em sistemas de controle. É a forma menos eficiente, pois leva mais tempo e pode requerer ajustes para obter a especificação. Normalmente utiliza medições mais imprecisas que outros métodos e apresenta maior custo associado ao desperdício de qualidade - *quality giveaway* - para evitar re-certificações. Todos os cálculos são baseados em correlações e regras de mistura aplicadas aos componentes para obter a especificação.
- Mistura em linha a partir de tanques: na Replan o óleo diesel é também produzido desta forma, que é um método mais eficiente, pois utiliza medições mais precisas,

análise em linha após o misturador e um sistema de controle avançado. Desta forma, o produto já sai especificado na linha, diminuindo o tempo de preparo da mistura.

O Sistema Digital de Controle Distribuído – SDCD que utiliza as informações do analisador em linha diminui o problema da necessidade de novos ajustes da mistura final e minimiza o custo associado ao desperdício de qualidade do produto final. Assim, são necessárias correlações e regras de mistura confiáveis para relacionar as variáveis manipuladas (as vazões das frações) às variáveis controladas (parâmetros de qualidade do produto) e, desta forma, melhorar o controle e obter maior rentabilidade. Na figura 4.2 ilustra-se um esquema de mistura em linha de óleo diesel.

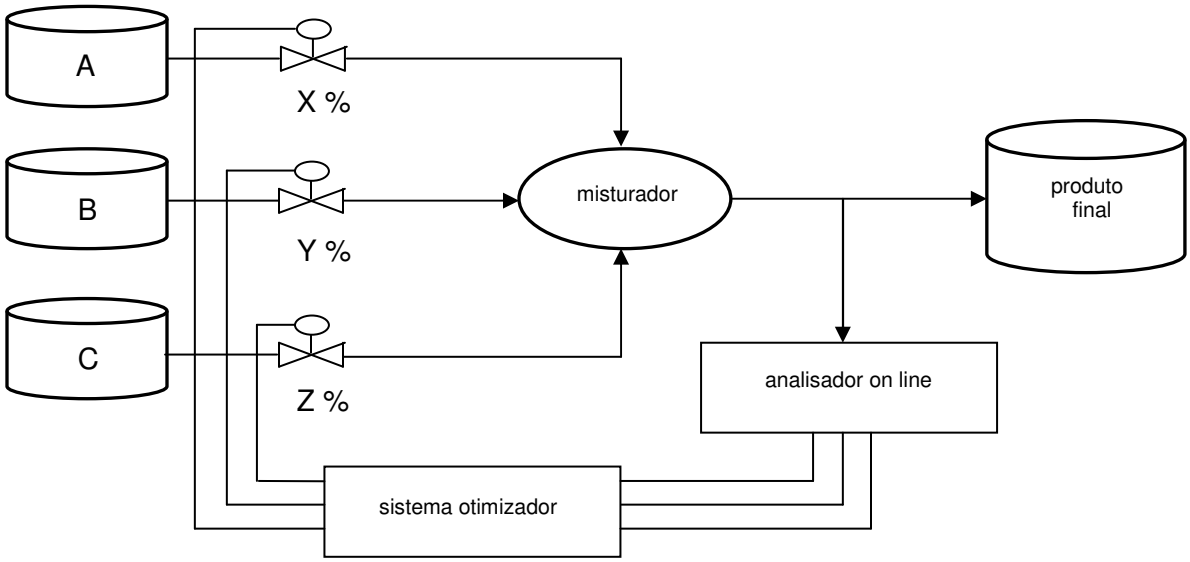


Figura 4.2 – Esquema de mistura em linha para produção de óleo diesel

Os diferentes tipos de óleo diesel produzidos na Replan são compostos, basicamente, por 6 tipos de componentes. No misturador em linha são recebidas as seguintes correntes provenientes de tanques: óleo diesel externo, recebido de oleoduto; diesel hidrotratado; nafta pesada, querosene, diesel leve e diesel pesado destilados. Entretanto, uma análise dos fluxos à montante dos sistemas de transferência revela que os cinco componentes internos se desdobram em 16 correntes intermediárias: as cargas dos hidrotratamentos, compostas por duas correntes de óleo leve de reciclo de craqueamento catalítico, óleo vegetal e seis correntes de coqueamento retardado

(gasóleos médios, gasóleos leves e naftas pesadas); nos destilados, uma mistura de naftas pesadas de destilação e outras seis correntes de destilados médios (querosene, diesel leve e diesel pesado).

Existem muitas combinações possíveis para a produção deste combustível na Replan, desde o controle de qualidade de cada componente nas unidades de destilação, craqueamento catalítico e coqueamento retardado, passando pela composição da carga de cada unidade de hidrotratamento, até a mistura final de cada tipo de óleo diesel.

Dezenas de tipos de petróleo podem participar de forma significativa da composição da carga da Replan anualmente. A adaptação da programação à constante mudança da composição do petróleo é um exercício diário, já que a produção do combustível é restrita por 11 limites principais de propriedades em cada especificação: ponto de fulgor mínimo, temperatura de destilação mínima e máxima relativa à recuperação de 50% (T50) da amostra, temperatura de destilação máxima relativa à recuperação de 85% (T85) da amostra, massa específica mínima e máxima, viscosidade mínima e máxima, ponto de entupimento máximo, teor de enxofre máximo e número de cetano mínimo.

Na realização do *blend*, o valor dos produtos intermediários podem sofrer bastante impacto, positivo ou negativo, conforme são combinadas as diferentes correntes intermediárias para se compor o produto final. O objetivo, nesta fase, é fazer os produtos com a qualidade requerida, utilizando-se os componentes mais baratos e dentro dos limites de flexibilidade dos mercados e produzir quantidade maior de produtos mais nobres, atingindo-se a otimização global.

As métricas normalmente utilizadas para se verificar a otimização da solução de um problema de *blending* são: o máximo lucro, o mínimo custo, o mínimo custo do desperdício de qualidade – *quality giveaway*, o mínimo desvio de uma receita requerida e o máximo de proporção de determinados componentes.[57,58,62]

Assim, os objetivos de qualquer controle de mistura de óleo diesel e sistema de otimização é maximizar a rentabilidade da mistura enquanto satisfaz as especificações de qualidade e as restrições operacionais de mistura (disponibilidade das correntes, tancagem, etc.).

Normalmente um sistema de automação de mistura de óleo diesel é construído em três níveis: otimização *off-line* ou programação de produção, otimização on-line, e controle das especificações.

No topo da hierarquia está a programação da produção que planeja as operações da refinaria em: longo prazo (meses) e considera cenários de mercado, objetivos estratégicos da companhia, etc.; médio prazo (semanas) que considera as demandas mais imediatas do mercado, a disponibilidade de cargas para processamento na refinaria, etc., podendo também ser útil para revisar a programação de longo prazo; e curto prazo (dias) que considera a unidade de processamento e planeja operações de mistura no período de um ou dois dias.

A otimização *on-line* utiliza a informação da qualidade da mistura para modificar a receita original durante a operação de mistura e proporciona a adequada instrução a ser executada pelo controlador.

Devido ao fato de ser o estágio final no processo de refino, a otimização desse processo é muito importante. Independentemente da excelente eficiência dos processos das unidades a montante do misturador, a produtividade e a qualidade podem ser prejudicadas se uma mistura mal otimizada produz um combustível fora das especificações ideais. Em muitos aspectos esse é o mais importante processo a otimizar e pode proporcionar importante contribuição em termos de lucro da refinaria. Como o óleo diesel é um dos mais importantes produtos da refinaria, um rígido controle do processo de mistura pode proporcionar o limite final para definição do lucro da refinaria.[57,62,63]

Atualmente as determinações das propriedades da mistura são comumente feitas através de correlações empíricas consagradas na indústria do petróleo. Existem também modelos de abordagem termodinâmica para a previsão de propriedades. No entanto, eles funcionam apenas quando se trabalha com componentes puros. No Brasil onde cada vez mais o processamento dos óleos nacionais (bem mais pesados que a média mundial) é prioridade na estratégia de refino, nem sempre estas correlações apresentam resultados satisfatórios.

Entre as propriedades limitantes, as quais impedem que certas correntes sejam adicionadas indiscriminadamente ao *pool* de óleo diesel, uma vez que tal adição pode tirar a mistura de especificação há o ponto de fulgor, que limita a adição de frações

muito leves, como é o caso da nafta pesada, que é um dos componentes do *pool* de óleo diesel da Replan e também há o número de cetano.[58]

Nos analisadores que estão no *shelter*, no setor de transferência e estocagem são executadas as análises durante o processo de mistura. Uma análise por espectroscopia NIR informa o número de cetano e o analisador de ponto de fulgor determina esta propriedade através do ensaio de referência ASTM, com duração de aproximadamente 30 minutos. Estas informações são transferidas ao sistema otimizador do misturador em linha que então modifica as contribuições de cada corrente de acordo com a necessidade da especificação até completar o tanque.

4.3.2.1 - Determinação de parâmetros de qualidade em linha

Nas refinarias de petróleo o monitoramento *on-line* de parâmetros de qualidade é uma ferramenta constantemente utilizada para controle de processos e otimização da produção de derivados. No caso da produção de óleo diesel isso pode ser obtido pelo desenvolvimento de eficazes modelos de calibração de parâmetros de qualidade para utilização pelo otimizador do misturador em linha. A rapidez do método e a qualidade dos resultados fornecidos pelo modelo de calibração permitem sua utilização para determinação de parâmetros *on-line*, proporcionando um ganho de produtividade, quando utilizados no controle de processos, em que são necessárias medições suficientemente rápidas e exatas para permitirem uma resposta do sistema central, na malha de controle, para o caso de perturbações ou tendências do *setpoint* ou valor desejado.

A espectroscopia FT-NIR pode ser utilizada pelos analisadores em linha para obtenção dos parâmetros utilizados pelo otimizador do misturador em linha. Tais parâmetros, ponto de fulgor e número de cetano, podem ser obtidos através de um modelo de calibração multivariada e devem utilizar conjuntos de amostras que permitam a maior abrangência possível da variabilidade encontrada no processo de mistura final que inclui diferentes correntes provenientes do refino de petróleo de diferentes origens. Por exemplo, a mistura final deve ter ponto de fulgor em torno de 40 ºC, porém, se

ocorrer uma situação em que se altere drasticamente a qualidade do diesel, como quando a linha recebe uma amostra rica em diesel pesado, cujo ponto de fulgor é em torno de 75 ºC, é importante que o modelo seja capaz de realizar boas previsões em todo esse intervalo.

Considerando a dificuldade em fazer com que a variabilidade amostral incluída no conjunto de calibração contemple toda a variabilidade presente no processo de produção e a possibilidade de ocorrência de relações não lineares entre o sinal analítico e o parâmetro estudado no espaço amostral considerado, é importante a utilização de um modelo de calibração que possua habilidade em modelar relações não lineares e com elevado poder de generalização.

Também, quanto maior a agilidade do ensaio para determinação desses parâmetros maior o ganho em produtividade de óleo diesel adequado às especificações requeridas e menor o custo associado ao desperdício de qualidade – *quality giveaway* - desse produto.

Com o desenvolvimento de modelos de calibração mais eficazes, eles poderão ser utilizados em programas de otimização de produção on-line e off-line. O capítulo 7 desse trabalho aborda esse problema.

4.3.3– O processo de hidrorrefino

O processo de hidrorrefino (HDR) consiste na mistura de frações de petróleo com hidrogênio em presença de um catalisador, sob certas condições operacionais determinadas em função do objetivo que se tem com esta etapa do refino.

As unidades de hidrogenação são classificadas na literatura conforme sua finalidade:

- unidades de hidrotratamento (HDT): têm a finalidade de melhorar as propriedades da carga a ser hidrogenada e proteger catalisadores de processos subseqüentes. O produto da unidade tem essencialmente a mesma faixa de destilação da carga, embora possa existir a produção secundária de produtos mais leves por hidrocraqueamento. As

cargas típicas dessas unidades variam desde a faixa da nafta até a de gasóleo pesado de vácuo.

- unidades de hidroconversão (HC): têm o objetivo de produzir frações mais leves do que a carga e eventualmente melhorar a qualidade da fração não convertida. As primeiras unidades de hidroconversão construídas foram as de hidrocraqueamento de alta severidade (HCC). Posteriormente, surgiram as unidades de hidrocraqueamento brando (MHC) com condições operacionais menos severas. As cargas típicas de HC ficam na faixa de gasóleo pesado de vácuo e resíduo.

4.3.3.1 - Unidades de hidrotratamento

O hidrotratamento da carga durante o processo de refino ocorre principalmente: a) para obter a especificação de produtos, através de reações de hidrodessulfurização, hidrodesnitrogenação, hidrogenação de aromáticos e olefinas, dependendo das características da corrente a ser hidrogenada, e b) para obter a proteção de catalisadores de diversos processos através do pré-tratamento da carga, com a remoção do nitrogênio, enxofre e metais, como Ni e V, que podem reagir com catalisadores utilizados em processos como reforma catalítica ou hidrocraqueamento e diminuir o rendimento do processo.

A redução do teor de enxofre é o principal objetivo do tratamento das correntes de gasóleos leve e pesado da destilação, também chamados de diesel leve e diesel pesado. O gasóleo leve de destilação a vácuo que é incorporado ao *pool* de diesel, também deve ser tratado para remoção de enxofre, compostos nitrogenados e aromáticos polinucleares. As correntes de gasóleo leve de coqueamento retardado e gasóleo leve de craquemanto catalítico fluido (óleo leve de reciclo – LCO), também conhecidas como frações instáveis por se oxidarem facilmente, podem ser adicionadas ao *pool* de diesel e devem ser tratadas para se adequarem quanto à estabilidade a oxidação e número de cetano. Assim, as reações de remoção de enxofre e nitrogênio e a hidrogenação de olefinas e aromáticos polinucleares são igualmente importantes no tratamento.

As principais variáveis operacionais nesse processo são: temperatura de reação, velocidade espacial, pressão parcial de hidrogênio, tipo de catalisador e tipo de carga.

Pertinente a motivação desse trabalho, cabe mencionar que a natureza da carga influencia diretamente nas condições da unidade. Existe uma variedade muito grande do tipo de carga a ser tratada em função da sua faixa de destilação e da sua constituição em termos de olefinas, compostos de enxofre, compostos de nitrogênio, compostos aromáticos e compostos metálicos.

Entre os parâmetros da carga que mais afetam as condições operacionais estão o peso molecular e a faixa de destilação, sendo que uma pequena variação pode causar um impacto considerável na performance da unidade. Quanto mais elevado for o peso molecular e o ponto final de ebulição da carga, maior será o teor de compostos de enxofre e nitrogênio, de cadeias cíclicas, e maior o teor de compostos aromáticos policíclicos.[64]

4.3.3.2 - Caracterização da carga

Toda carga a ser processada deve ser caracterizada de modo que se possa adequar as variáveis do processo. Da mesma forma, a análise do produto é fundamental para verificar se este atende as especificações. O acompanhamento pode ser feito através de análises químicas e físicas das correntes, tais como as especificadas pelos métodos ASTM de referência para determinação do ponto de anilina, índice de cetano, destilação, densidade, etc. Porém, tais métodos de referência muitas vezes podem demandar excessivo tempo de ensaio, custo relativamente elevado, utilização de solventes orgânicos, treinamento de analistas, etc., o que dificulta o monitoramento da carga e do produto da unidade de processamento.

A utilização de espectroscopia NIR aliada a técnicas quimiométricas possibilita a obtenção de modelos de calibração, que tornam possíveis a realização de rápidas análises *on-line*.

O desenvolvimento de modelos de calibração mais eficazes e com possibilidade de utilização em analisadores *on-line* para monitoramento da carga do

processo de hidrotratamento são importantes para o melhor controle do processo e a otimização da operação dessa unidade. Esse problema é abordado no capítulo 8 desse trabalho.

5 - Biodiesel

A utilização do etanol, de óleo diesel produzido a partir de processos fermentativos biológicos (produtos de biologia sintética)[65,66] e do biodiesel, os chamados biocombustíveis ou combustíveis renováveis, obtidos a partir da biomassa, tem aumentado no Brasil e no mundo nos últimos anos devido a atrativos aspectos ambientais, econômicos e sociais.

O biodiesel é atualmente o principal substituto do óleo diesel derivado do petróleo devido a suas semelhanças quanto às propriedades físico-químicas, permitindo sua utilização em motores ciclo diesel comuns em misturas de até 20% (v/v) com óleo diesel de origem fóssil (B20) sem modificações nos motores. Misturas com teores de biodiesel de até 30% (v/v) também são utilizadas em motores diesel comuns porém podem demandar ajustes no sistema de injeção e adaptações em filtros de combustível e materiais elastômeros do motor, dependendo do fabricante. Veículos com motores adequadamente adaptados podem também utilizar como combustível o biodiesel puro a 100 % (B100).

Em muitos países a mistura de uma fração de biodiesel no combustível óleo diesel é realizada. Nos Estados Unidos o mais comum é a utilização da mistura a 20 % (v/v) ou menos e a norma ASTM D7467-08 fornece a especificação de referência para misturas contendo de 6 a 20 % de biodiesel (B6 a B20).[67] Na europa o mais comum atualmente é a utilização da mistura entre 5 e 7 % (v/v) de biodiesel adicionado ao óleo diesel de petróleo embora existam experiências com a utilização de até 30 % (v/v) nas frotas cativas de ônibus e em veículos leves.[68] No Brasil as experiências das cidades de São Paulo e Curitiba se destacam. Na cidade de São Paulo já foi utilizada uma mistura com até 30 % (v/v) de biodiesel (B30) em ônibus da frota cativa do transporte público e atualmente tem aumentado a utilização da mistura com 20 % (v/v) de biodiesel (B20). Em Curitiba ônibus movidos a B100 integram a frota de transporte público da cidade.

No Brasil, além de fatores econômicos como a diminuição da importação de óleo diesel, e sociais ligados ao incentivo a produção agrícola de pequenos produtores, a crescente atenção para a diminuição da emissão dos gases causadores do efeito estufa, enxofre e material particulado na atmosfera das grandes cidades e regiões metropolitanas levou o governo a introduzir o biodiesel na matriz energética brasileira

com a vigência da Lei 11097/2005 e a exigir a utilização do biodiesel em teor de 5% (v/v) misturado ao combustível óleo diesel até 2013. A Resolução ANP 7/2008, alterada pela Resolução ANP 4/2010 antecipou e regulamentou a utilização do referido teor de biodiesel para janeiro de 2010 em todo território nacional. Na cidade de São Paulo, com a vigência da Lei 14933/2009, chamada de Lei municipal da mudança do clima, determinou-se que toda a frota cativa de ônibus deverá substituir gradativamente e integralmente até 2018 a utilização de combustíveis fósseis por combustíveis renováveis. Assim, na cidade de São Paulo atualmente é comum a utilização de 20 % (v/v) de biodiesel no combustível óleo diesel utilizado na frota de transporte público e a tendência é de que a quantidade de biocombustível utilizado aumente nos próximos anos. A Resolução ANP 2/2011 regulamenta e especifica o combustível com biodiesel em teores de 6 a 20% (v/v).

A utilização do biodiesel não implica em modificações na estrutura de distribuição e armazenagem, embora cuidados operacionais comuns ao uso do óleo diesel de petróleo tanto na armazenagem como na operação e manutenção dos motores devam ser cuidadosamente controlados devido às peculiaridades do biodiesel relativas a maior higroscopicidade; menor estabilidade oxidativa, maior poder de solvência em relação ao óleo diesel de petróleo e as suas características físico-químicas a baixas temperaturas.

A presença de água no combustível pode ocasionar algum crescimento microbiano e embora esse fator seja comum ao óleo diesel de petróleo, com o uso do biodiesel em teores elevados esse fator torna-se mais delicado, devido a maior higroscopicidade e maior biodegradabilidade do biodiesel. Este problema está relacionado ao aumento da acidez no combustível, o que contribuirá para a corrosão de acessórios de cujo ajuste depende o bom funcionamento do motor, e a formação de compostos insolúveis (borras) no combustível, os quais formam depósitos de verniz e sedimentos que podem obstruir os filtros de combustível além de acelerar a degradação do produto enquanto armazenado.

Usuários de biodiesel têm percebido que a elevação no teor de biodiesel na mistura com diesel de petróleo de B5 para B20 provoca um sensível aumento na capacidade desse combustível em ressuspender sedimentos insolúveis de produtos de oxidação ou sujeira depositados nas paredes de tanques e tubulações e isso pode causar o entupimento de filtros de combustível durante o período de transição.

As inadequadas propriedades de fluidez a baixas temperaturas são causadas pela presença de ésteres de ácidos graxos saturados no biodiesel. Esses compostos fazem com que ocorra a cristalização da mistura contendo biodiesel em temperaturas acima do comum para o óleo diesel convencional, podendo ocasionar o entupimento de filtros de combustível e causar problemas em partidas do motor a frio, limitando o uso do biodiesel em climas frios.

Tais peculiaridades relativas ao uso do biodiesel podem tornar necessárias adaptações, relativas ao uso de filtros de combustível e elastômeros em motores diesel comuns, dependendo do fabricante. Além disso, em virtude de sua maior densidade (biodiesel de soja, d_{20} = 0,8818 g/ml) pode ser necessário o ajuste do volume de injeção de combustível com elevados teores de biodiesel, visando a economia de consumo e a adequação dos padrões de emissões.

O biodiesel é biodegradável, não contém compostos aromáticos e o biodiesel de óleo de soja contém muito pouco enxofre (menos de 3 mg/kg) . Motores que utilizam o combustível com biodiesel emitem menos monóxido de carbono, hidrocarbonetos, óxidos de enxofre e material particulado. Verificou-se que as emissões médias do biodiesel utilizado em motores ciclo diesel comuns comportam-se relativamente as emissões do óleo diesel convencional conforme mostrado na tabela 5.1.[69-71]

Tabela 5.1 – Emissões médias com utilização de biodiesel em motores ciclo diesel* em relação a utilização do óleo diesel convencional. (adaptado de Relatório EPA [70])

Tipo de emissão	emissão relativa (%)	
	B100	B20
CO	-48	-11
hidrocarbonetos	-67	-21
material particulado	-47	-10
SO_X	-100	-20
NO_X	+10	+2
hidrocarbonetos policíclicos aromáticos (PAH)	-80	-13

*depende da família de motores

Como o biodiesel é produzido a partir de matéria prima vegetal não há incremento da quantidade de dióxido de carbono na atmosfera com a queima desse combustível.

O biodiesel tem número de cetano mais elevado em relação ao diesel convencional, sendo que o biodiesel produzido com matérias-primas saturadas tem número de cetano mais elevado do que aquele produzido com matérias-primas menos saturadas. O biodiesel de soja tem número de cetano frequentemente entre 48 e 52.

O biodiesel tem maior lubricidade em relação ao óleo diesel, o que é especialmente importante para utilização do óleo diesel com baixo teor de enxofre (ULSD) atualmente em uso, tendo-se demonstrado que a utilização de misturas com 2% de biodiesel fornece ao combustível a lubricidade anteriormente proporcionada por teores de enxofre acima de 500 ppm, resultando em maior vida útil aos componentes do motor.

O ponto de fulgor do biodiesel também é mais elevado em relação ao diesel fóssil, o que proporciona maior segurança na utilização do combustível.

A utilização do biodiesel proporciona um aumento no consumo de combustível proporcional ao seu menor poder calorífico mas a eficiência do motor não sofre redução significativa. A utilização do B20 aumenta o consumo de combustível em aproximadamente 2 %.

Motores que consomem biodiesel apresentam um aumento na emissão de óxidos de nitrogênio (NO_x) devido a propriedades físico-químicas desse combustível.[68] As causas para o aumento das emissões de NO_x associado ao uso do biodiesel, ao menos para sistemas de injeção unitários, estão relacionadas a um pequeno deslocamento no intervalo de injeção do combustível que é causado por diferenças nas propriedades mecânicas do biodiesel em relação ao diesel convencional. Devido ao maior módulo de compressibilidade (ou velocidade do som) do biodiesel, há uma transferência mais rápida da onda de pressão da bomba de injeção para a agulha do injetor, resultando na antecipação do levantamento da agulha e na produção de um pequeno avanço no intervalo de injeção. Estudos mostram que o retardamento no intervalo de injeção pode reduzir a emissão de NO_x, paralelamente à perda de alguma eficiência na redução de material particulado e na economia de combustível.[25] No entanto, a utilização de combustível com baixo teor de enxofre ULSD (menor que 50

ppm) permite que essa desvantagem possa ser contornada pela utilização de dispositivos como o SCR (*Selective Catalytic Reduction*) obrigatório no Brasil para atender os níveis de emissões de gases exigidos pelo Programa de Controle da Poluição do Ar por Veículos Automotores – PROCONVE para veículos pesados novos a partir de 2012 (Fase P-7) e também utilizado nos Estados Unidos e Europa (padrão Euro V).

O biodiesel contém um variável teor de ésteres de cadeia insaturada, dependendo da matéria-prima utilizada, que são suscetíveis a reações de oxidação acelerada pela exposição ao oxigênio e a altas temperaturas. A taxa de oxidação aumenta com o número de insaturações nas moléculas de éster do biodiesel: linolênico (três insaturações) > linoléico (duas insaturações) > oléico (uma insaturação) > esteárico (saturado). A adição de um antioxidante aumenta a estabilidade do biodiesel, embora a degradação continue lentamente.

O biodiesel é obtido através da transesterificação (uma reação orgânica na qual um éster é transformado em outro através da troca dos grupos alcóxidos) dos triglicerídeos de óleos e gorduras de origem vegetal ou animal com um mono-álcool de cadeia curta, tipicamente metanol ou etanol, na presença de um catalisador, produzindo uma mistura de ésteres alquílicos de ácidos graxos e glicerol.[72]

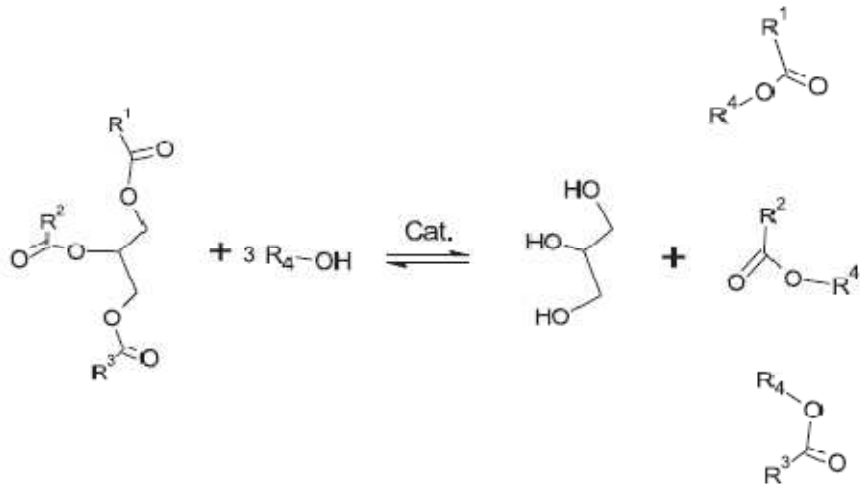


Figura 5.1 – Reação de transesterificação de triglicerídeos

Os óleos vegetais são constituídos predominantemente de substâncias conhecidas como triglicerídeos (também chamadas de triacilgliceróis ou triacilglicerídeos), que são ésteres formados a partir de ácidos carboxílicos de cadeia longa (ácidos graxos) e glicerol.

Os ácidos graxos constituintes dos triglicerídeos mais comuns apresentam 12, 14, 16 ou 18 átomos de carbono. Entretanto, outros ácidos graxos com menor ou maior número de átomos de carbono ou ainda contendo a função álcool também podem ser encontrados em vários óleos e gorduras. A tabela 5.2 mostra os intervalos de valores dos teores de ésteres de diferentes ácidos graxos no biodiesel em função de sua matéria-prima.[25] A composição do biodiesel determina suas características físico-químicas e de especial importância é o teor de compostos saturados devido a sua influência nas propriedades a baixas temperaturas como ponto de fluidez e ponto de congelamento. Verifica-se que a gordura bovina apresenta elevado teor de compostos saturados, razão pela qual o biodiesel derivado dessa matéria-prima apresenta desempenho inadequado a baixas temperaturas.

Tabela 5.2 – Composição aproximada de alguns ésteres de ácidos graxos no biodiesel em função da matéria-prima (adaptado de Knothe *et al*. [25])

Ácido graxo	Número de carbonos	Número de insaturações	teor de ésteres de ácidos graxos no biodiesel (%, m/m))			
			Óleo de soja	Óleo de palma	Óleo de girassol	Gordura bovina
Ácido mirístico	14	0	-	0,5-2,4	-	2,1-6,9
Ácido palmítico	16	0	2,3-13,3	32-47,5	3,5-7,6	25-37
Ácido esteárico	18	0	2,4-6,0	3,5-6,3	1,3-6,5	9,5-34,2
Ácido oléico	18	1	17,7-30,8	36-53	14-43	14-50
Ácido linoléico	18	2	49-57,1	6-12	44-74	26-50
Ácido linolênico	18	3	2-10,5	-	-	-
Outros	-	-	0-0,3	0-0,4	-	-

A mistura de ésteres de ácidos graxos resultante da transesterificação é denominada biodiesel e os principais componentes quantificados no biodiesel, além do

próprio éster metílico em teor não inferior a 96,5% (m/m), segundo a legislação vigente, são a glicerina, o metanol e os mono, di e triglicerídeos.

5.1 – Determinação do teor de biodiesel em óleo diesel

A ANP é o principal órgão governamental responsável pelo monitoramento quanto ao teor de biodiesel exigido nos combustíveis óleo diesel B5 – B20 através de seu Programa de Monitoramento da Qualidade de Combustíveis líquidos - PMQC.

Trabalhos recentes demonstram a aplicação da espectroscopia nas regiões do infravermelho próximo e médio para determinação do teor de biodiesel em misturas com óleo diesel utilizando modelos de calibração nas faixas de 0 - 5 % e 0 – 100 % de biodiesel. [73-76]

Os métodos tradicionais para determinação do teor de biodiesel em óleo diesel utilizam a espectroscopia no infravermelho médio, através de medidas de transmitância ou ATR e calibração com PLS, conforme descrito pelas normas ASTM D 7371-07e ABNT NBR 15568.

O método ASTM sugere a construção de modelos nas faixas de 0 a 10 % (v/v), 10 a 30 % (v/v) e 30 a 100 % (v/v). A reprodutibilidade é estabelecida de acordo com o teor de biodiesel na amostra e varia de 0,76 a 1,66 % (v/v), para amostras com 1 % e 20 % (v/v) de biodiesel, respectivamente.

O método ABNT NBR sugere a construção de modelos nas faixas de 0 a 8 % (v/v) e de 8 a 30 % (v/v). A norma estabelece que esses modelos tenham valores de RMSEP que não podem ser superiores a 0,1 % (v/v) e 1 % (v/v), respectivamente.

A obtenção de modelos de calibração mais eficientes e eficazes, em relação as normas de referência e considerando os problemas analíticos atuais, utilizando-se o SVM aplicado a dados de espectroscopia NIR para determinação do teor de biodiesel em mistura com óleo diesel mineral nas faixas de 0 - 35 % e 0 – 100 % é abordado no capítulo 10.

6 - Óleos lubrificantes

Os óleos lubrificantes são produtos formulados com a utilização de óleos básicos e compostos aditivos adequados para prover ao produto lubrificante as características desejadas para a aplicação e o nível de desempenho específico a que o óleo lubrificante se destina.

Para cada aplicação específica de um óleo lubrificante os óleos básicos devem atender a determinados critérios de desempenho, como o comportamento da viscosidade em função da temperatura, volatilidade, estabilidade a oxidação, estabilidade hidrolítica, tendência a formação de depósitos, solvência dos compostos aditivos, compatibilidade com elastômeros, demulsibilidade, entre outras. Os óleos básicos podem ser divididos em categorias, segundo sua origem: mineral, sintético, biológico e rerrefinados. Os óleos lubrificantes fabricados com a utilização de diferentes óleos básicos exibem diferentes propriedades e são adequados para diferentes aplicações: (i) óleos minerais são derivados do petróleo e os mais comumente utilizados como lubrificantes automotivos e industriais devido a boa relação entre custo e benefício. Podem ser óleos parafínicos ou naftênicos; (ii) óleos sintéticos são produzidos para superarem as especificações de qualidade e as características de desempenho dos óleos minerais, podendo ser aplicados em condições mais severas de serviço, como temperaturas extremas e/ou com intervalos de troca estendidos. Diferentes tipos de óleos básicos sintéticos são produzidos para proporcionar um melhor desempenho do lubrificante em diferentes tipos de aplicações; (iii) óleos biológicos são mais facilmente biodegradáveis do que os óleos minerais e sintéticos e vêm ampliando sua importância no mercado devido às exigências crescentes por produtos ambientalmente amigáveis. Os óleos vegetais, como óleo de soja e óleo de girassol, são os mais utilizados entre os óleos de origem biológica. Aplicações onde o risco de contaminação deve ser baixo são as principais demandas para esse tipo de produto; (iv) os óleos rerrefinados são oriundos de óleos lubrificantes usados e passam por um processo de refino que inclui processos de filtragens, destilação e tratamento ácido. Os óleos rerrefinados são comumente misturados aos óleos minerais, em proporções variáveis e que usualmente não ultrapassa 30 % (v/v), podendo ser utilizados em algumas aplicações semelhantes a dos óleos minerais.

Além do óleo básico, pacotes de aditivos desenvolvidos especialmente para cada tipo de aplicação também são responsáveis pelas características de desempenho do óleo lubrificante. Os principais tipos de aditivos utilizados em formulações para óleos lubrificantes de motor incluem: detergentes, dispersantes, anticorrosivo, antidesgaste, antiespumante, antioxidante, agente de reserva alcalina, melhorador do índice de viscosidade e abaixador do ponto de fluidez.

Um óleo lubrificante de motor típico é composto de aproximadamente 85% (v/v) de óleo básico e 15% (v/v) de aditivos. Para se obter as características de viscosidade e desempenho adequadas para cada tipo de motor podem ser utilizados uma mistura de óleos básicos e um pacote de aditivos apropriado.

6.1 - Classificação dos óleos básicos

A classificação dos óleos básicos segundo a *American Petroleum Institute* - API é tradicionamente feita em cinco diferentes grupos. Essa classificação é feita de acordo com o teor de compostos saturados, teor de enxofre e índice de viscosidade conforme mostrado na tabela 6.1. Os óleos básicos do grupo I são minerais e produzidos pelo tradicional processo de refino da rota solvente, enquanto os óleos básicos dos grupos II e III embora também sejam de origem mineral passam por um processo de refino com hidrotratamento mais severo. Os óleos básicos do grupo IV são as polialfaolefinas (PAO). No grupo V incluem-se todos os demais óleos básicos, como os óleos naftênicos e os óleos sintéticos com exceção das PAO.[77]

Tabela 6.1 – Classificação API para os óleos básicos

Categoria	Enxofre (% m/m)		Saturados (% m/m)		Índice de viscosidade
Grupo I	> 0,03	e/ou	< 90	e	80 a 119
Grupo II	≤ 0,03	e	≥ 90	e	80 a 119
Grupo III	≤ 0,03	e	≥ 90	e	≥ 120
Grupo IV	Polialfaolefinas (PAOs)				
Grupo V	Todos os básicos não incluídos nos grupos I-IV (básicos naftênicos e sintéticos exceto PAOs)				

6.2 – Óleos básicos minerais

O petróleo cru é classificado como parafínico, naftênico e aromático, dependendo da proporção desses compostos na mistura de hidrocarbonetos. Os óleos básicos obtidos pela destilação desses diferentes petróleos são assim igualmente denominados e são constituídos predominantemente por compostos de cadeia carbônica C_{20} a C_{50}. A Portaria ANP 129/1999 especifica os óleos lubrificantes básicos minerais comercializados no Brasil segundo diversos parâmetros de qualidade.

A grande maioria dos óleos básicos fabricados nas refinarias brasileiras atualmente são parafínicos. Os óleos básicos parafínicos apresentam a melhor relação entre custo e benefício para muitas aplicações devido ao bom desempenho proporcionado pelas suas características físico-químicas. Entre essas características ressalta-se a boa estabilidade a oxidação e o alto índice de viscosidade, que é uma medida da variação da viscosidade com a variação da temperatura e quanto maior o índice de viscosidade menor a diminuição da viscosidade com o aumento da temperatura.

Os óleos básicos naftênicos são produzidos em menor quantidade e encontram aplicação, por exemplo, onde se requer um lubrificante com boa característica de fluidez a baixas temperaturas, como em compressores de refrigeração. A quantidade de compostos parafínicos nesse tipo de óleo é reduzida, assim, os óleos naftênicos têm um ponto de fluidez mais baixo, tornando-o mais apropriado para esse tipo de aplicação. Porém, esses óleos apresentam um baixo índice de viscosidade e baixa estabilidade a oxidação, o que compromete sua utilização como lubrificante em muitas aplicações, especialmente em formulações para lubrificação de motores automotivos. Outras aplicações possíveis para óleos naftênicos são como óleo isolante e como fluido para usinagem de metais.

As composições das frações de petróleo são frequentemente expressas em termos das proporções de anéis aromáticos (R_A), anéis naftênicos (R_N), e cadeias parafínicas (C_P) que devem ser constituintes das moléculas hipotéticas presentes no óleo. Alternativamente, a composição pode ser expressa em termos da distribuição de carbonos, ou seja, a porcentagem do número total de átomos de carbono que estão

presentes em estruturas de anéis aromáticos (%C_A), estruturas de anéis naftênicos (%C_N), e cadeias parafínicas (%C_P).[78]

Muitas das moléculas nos óleos minerais misturam estrutura alifática e aromática, como uma longa cadeia parafínica ligada a um anel benzênico. A estrutura dessa molécula é predominantemente parafínica uma vez que a maioria dos átomos de carbono são de natureza parafínica.

É possível caracterizar e quantificar os hidrocarbonetos envolvidos na mistura utilizando técnicas como GC-MS ou NMR H^+, no entanto a técnica mais difundida para caracterização do tipo de óleo básico é baseada na determinação da quantidade de carbonos que fazem parte de estruturas alifáticas ou parafínicas, cíclicas ou naftênicas e aromáticas. Esse tipo de determinação em óleos básicos é feita de forma indireta através de correlações descritas pelo norma ASTM D3238-95 , conhecida como método ndm (n=índice de refração, d=densidade, m=massa molecular) que utiliza as medidas do índice de refração, densidade, viscosidade e teor de enxofre para calcular as quantidades de carbono parafínico (%C_P), carbono naftênico (%C_N) e carbono aromático (%C_A) presentes no óleo, utilizando o conjunto de equações adequadas. Essa técnica se baseia no fato de que as propriedades físico-químicas variam com a quantidade e estrutura das cadeias carbônicas dos hidrocarbonetos presentes na mistura e podem ser correlacionados com os teores dos tipos de carbono presentes na mistura.

Os valores modais de alguns parâmetros de qualidade e valores de referência para distribuição de carbonos determinada pelo método ndm (ASTM D3238) para diferentes óleos básicos parafínicos e um óleo básico naftênico são mostrados na tabela 6.2.

6.2.1 - Tecnologia de refino dos óleos básicos minerais

O processo de refino dos óleos básicos parafínicos provenientes do processo de destilação do petróleo pode ocorrer pela tradicional rota solvente para produção de óleos básicos Grupo I ou através do processo que inclui uma etapa de hidrorrefino mais severa para produção de óleos básicos dos Grupos II, III e V (naftênicos). O tradicional

processo de refino dos óleos básicos pela rota solvente, ilustrado na figura 6.1, ocorre em cinco etapas principais: (i) destilação atmosférica e destilação a vácuo – o produto de fundo do processo de destilação atmosférica é submetido ao processo de destilação a vácuo para obtenção das frações adequadas do petróleo, (ii) desasfaltação - através do processo de extração a propano para recuperação das frações lubrificantes do resíduo da destilação a vácuo, (iii) desaromatização - através do processo de extração líquido-líquido para redução do teor de aromáticos e elevação do índice de viscosidade e da estabilidade a oxidação, (iv) desparafinação - através da mistura com solvente e posterior redução da temperatura para cristalização e remoção de parafinas que criam problemas de fluidez a baixas temperaturas, (v) hidroacabamento - pelo processo de hidrogenação para remoção de insaturações e oxigênio, enxofre e nitrogênio.[77]

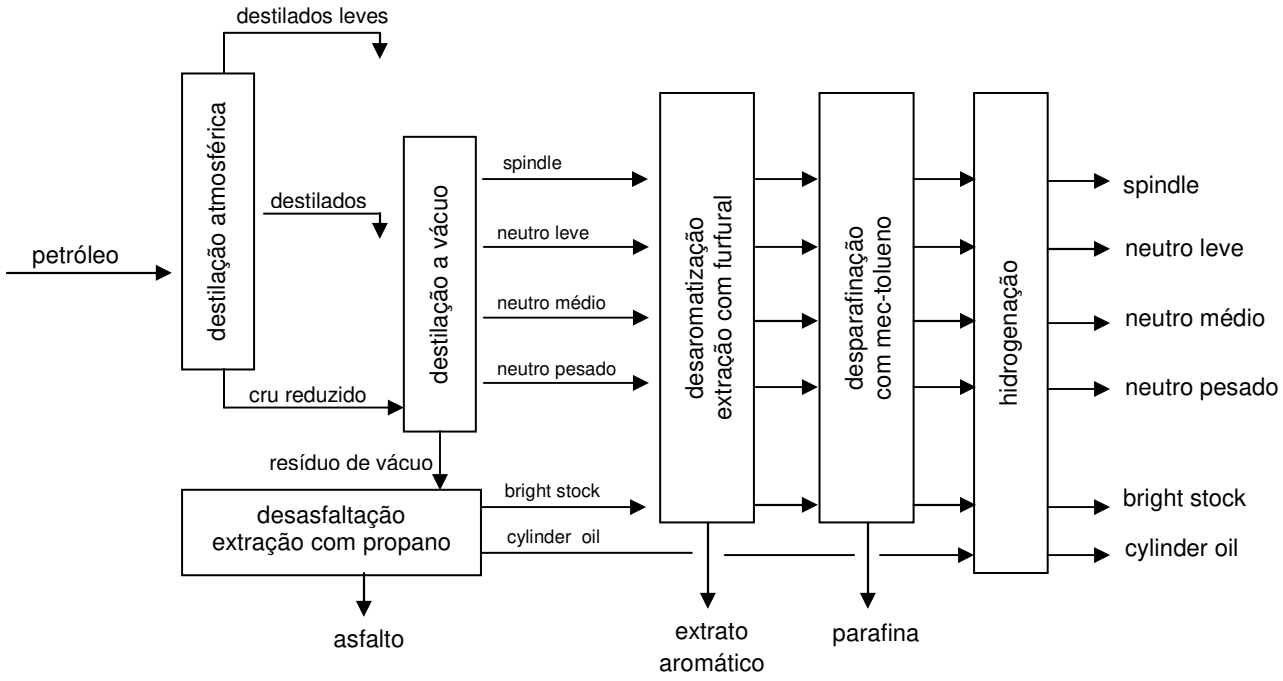


Figura 6.1 – Processo de produção de óleos básicos minerais do grupo I

Os óleos básicos parafínicos produzidos pela rota solvente são os básicos spindle (PSP), neutro leve (PNL), neutro médio, (PNM) neutro pesado (PNP), turbina leve (PTL), turbina pesado (PTP), bright stock (PBS) e cilindro (PCL). Os básicos turbina são oriundos dos mesmos destilados que dão origem aos respectivos básicos neutros (os óleos do grupo I mais utilizados em formulações automotivas), porém passam por um processo mais severo de desaromatização. Essa maior severidade no

processo de refino confere aos básicos turbina qualidade superior, com melhora nas características de estabilidade a oxidação, índice de viscosidade e demulsibilidade. Embora os óleos básicos turbina tenham qualidade inferior a dos óleos hidrorrefinados e sintéticos (grupos II, III e IV) têm qualidade superior à dos óleos básicos neutros.

Tabela 6.2 – Valores modais de parâmetros de qualidade dos óleos básicos

		parafínicos							naftênico
		PSP 09	PNL 30	PNM 55	PNP 95	PTL 25	PTP 85	PBS 33	NH 20
densidade (g/ml)		0,8537	0,8713	0,8705	0,8846	0,8586	0,8765	0,9023	0,9017
ponto de fulgor (ºC)		182	228	242	272	232	268	326	160
ponto de fluidez (ºC)		-12	-6	-6	-6	-6	-6	-9	-33
viscosidade (cSt)	40 ºC	9,90	29,01	52,55	96,35	26,67	83,95	502,1	20,38
	100 ºC	2,61	5,06	7,22	10,8	4,96	10,23	32,46	3,59
índice de viscosidade		96	104	99	97	111	103	95	13
distribuição de carbonos (%)	C_P	64,6	66,3	66,2	67,4	68,4	68,8	68,8	44,8
	C_N	28,1	27,4	26,0	25,4	28,5	26,7	22,9	42,8
	C_A	7,3	6,3	7,8	7,2	3,1	4,5	8,2	12,4

O óleo básico pesado bright stock é oriundo do processo de desasfaltação e pode compor formulações para diversas aplicações contribuindo para o ajuste da viscosidade adequada do óleo lubrificante.

6.3 - Óleos básicos vegetais

A crescente preocupação com a adoção de modos de produção ambientalmente amigáveis vem aumentado a demanda por óleos lubrificantes biodegradáveis. Os óleos vegetais são muito mais biodegradáveis do que os óleos minerais e poderiam ser substitutos de baixo custo para os óleos minerais, porém, algumas propriedades

indesejáveis reduzem sua possibilidade de larga utilização como lubrificante em muitas aplicações.

Embora os óleos vegetais tenham melhor desempenho do que os óleos minerais em propriedades como volatilidade e índice de viscosidade e tenham desempenho similar em algumas outras propriedades, suas grandes deficiências se relacionam a resistência a degradação térmica e oxidativa e propriedades a baixas temperaturas indesejáveis, como alto ponto de fluidez. Esses problemas podem ser apenas parcialmente resolvidos com a utilização de aditivos e aumentam o custo do produto.

Assim, óleos vegetais necessitam ser modificados quimicamente para eliminação dos grupos suscetíveis a oxidação e para impedir a formação de cristais a baixas temperaturas.[79]

A utilização de óleos vegetais em formulações de óleos lubrificantes de motor deve ser adequadamente estudada e testada, uma vez que devido a suas características indesejáveis sua utilização pode ocasionar sérios danos ao equipamento.

6.4 – Lubrificantes automotivos e a qualidade dos produtos no mercado

Os óleos lubrificantes de motor são a principal categoria (em termos de volume demandado no mercado) de lubrificantes automotivos. Os óleos de motor podem ser classificados quanto a sua viscosidade e quanto a suas características de desempenho. A classificação de viscosidade definida pela *Society of Automotive Engineers* – SAE nos EUA na norma SAE J300 e a classificação de desempenho estabelecida pela *American Petroleum Institute* - API são as mais utilizadas.

A classificação de viscosidade SAE considera dois intervalos de graduação para as viscosidades. A graduação acompanhada da letra W é definida para viscosidades a baixas temperaturas e a graduação sem a letra W é definida para viscosidades a temperaturas elevadas. Para cada grau W é definida uma viscosidade máxima a baixas temperaturas e uma temperatura limite de bombeio, assim como uma viscosidade mínima a 100 ºC.

Tabela 6.3 – classificação de viscosidade para óleos de motor SAE J300

grau de viscosidade SAE	partida a baixa temperatura viscosidade máx. (cP) [1]	temperatura limite de bombeio viscosidade Max. (cP) [2]	viscosidade (cSt) a 100 (ºC) [3]		viscosidade HTHS * (cP) a 150 ºC e 10^6 s^{-1} [4]
			mín.	máx.	mín.
0W	6200 a -35 ºC	60000 a -40 ºC	3,8	-	-
5W	6600 a -30 ºC	60000 a -35 ºC	3,8	-	-
10W	7000 a -25 ºC	60000 a -30 ºC	4,1	-	-
15W	7000 a -20 ºC	60000 a -25 ºC	5,6	-	-
20W	9500 a -15 ºC	60000 a -20 ºC	5,6	-	-
25W	13000 a -10 ºC	60000 a -15 ºC	9,3	-	-
20	-	-	5,6	<9,3	2,6
30	-	-	9,3	<12,5	2,9
40	-	-	12,5	<16,3	2,9 (0W40, 5W40, 10W40)
40	-	-	12,5	<16,3	3,7 (15W40, 20W40, 25W40, 40)
50	-	-	16,3	<21,9	3,7
60	-	-	21,9	<26,1	3,7

1 – medida no simulador de partida a frio (ASTM D5293); 2 – medida no miniviscosímetro rotativo (ASTM D4684); 3 – ASTM D445; 4 – ASTM D 4683; *HTHS: alta temperatura/alta taxa de cisalhamento

Um óleo multigrau atende a uma graduação de viscosidade SAE para temperaturas baixas assim como a 100 ºC atende o limite estipulado para esse mesmo grau W. Isso é possível com a adição de um aditivo melhorador do índice de viscosidade ao óleo com viscosidade adequada para baixas temperaturas. Os óleos multigraduados proporcionam a boa lubrificação dos motores a frio e em temperaturas elevadas sua viscosidade ainda é capaz de prover a adequada lubrificação. Os intervalos de viscosidade para cada grau SAE são mostrados na tabela 6.3.

Para obtenção das diferentes viscosidades especificadas pelos fabricantes para utilização nos motores, é geralmente necessário realizar a mistura de óleos básicos com viscosidades distintas. A tabela 6.2 mostra alguns parâmetros de qualidade dos diferentes óleos básicos, incluindo suas viscosidades.

A classificação de desempenho API estabelece parâmetros de desempenho para o óleo aplicado a diferentes tipos de motor, ciclo Otto ou ciclo Diesel. Com o avanço da tecnologia dos motores e o aumento das exigências de caráter ambiental os óleos lubrificantes ficam sujeitos a condições mais severas de serviço e com exigências de intervalos de troca estendidos. Para atender a esses requisitos de desempenho os produtores de lubrificantes utilizam os óleos básicos apropriados e pacotes de aditivos com tecnologias diferenciadas.

A Agência Nacional do Petróleo – ANP mantém o Programa de Monitoramento da Qualidade de Lubrificantes – PMQL que tem por objetivo acompanhar sistematicamente a qualidade dos óleos lubrificantes comercializados no país bem como proporcionar ferramenta importante para o direcionamento das ações de fiscalização da ANP. O PMQL tem como principal alvo os óleos lubrificantes para motores automotivos comercializados no mercado revendedor. Entre as análises realizadas verificam-se a viscosidade, o teor de aditivos e a presença de produtos prejudiciais ao motor como óleo vegetal e óleo básico naftênico.

A ANP divulga periódicamente o Boletim Mensal do Monitoramento dos Lubrificantes onde os resultados do monitoramento da qualidade dos produtos são mostrados. Frequentemente reporta-se a presença de óleo vegetal e óleo básico naftênico nas amostras estudadas, conforme ilustra a figura 6.2, com resultados de agosto de 2010 e dezembro de 2010.[80,81] Atualmente a ANP estuda a proibição da presença de óleos naftênicos, extrato aromático e óleo vegetal em óleos para algumas aplicações específicas.

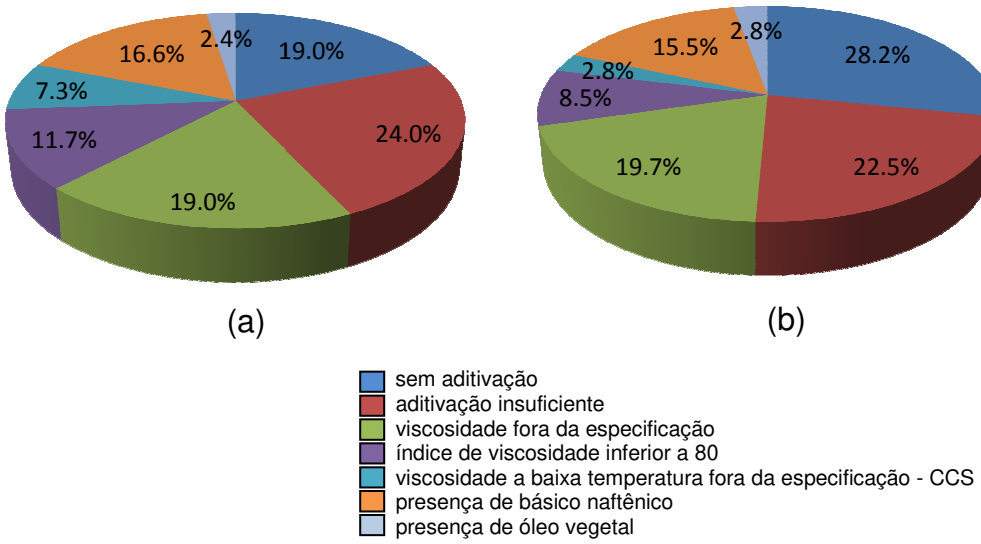


Figura 6.2 – Não conformidades de qualidade – por parâmetros, reportados pelo PMQL em (a) ago. 2010 e (b) dez. 2010

Os óleos naftênicos apresentam um baixo índice de viscosidade e resistência a oxidação inferior a dos óleos parafínicos, assim, quanto maior a proporção de

compostos naftênicos nos óleos de motor, pior o desempenho do óleo para essa aplicação.

Os óleos vegetais embora possuam alto índice de viscosidade, apresentam também alto ponto de fluidez e baixa estabilidade térmica e química o que o torna inapropriado para utilização em óleos de motor.

Existem poucos trabalhos na literatura que reportam a utilização da espectroscopia NIR para análise de óleos básicos e óleos lubrificantes. Balabin *et al.*[82] utilizaram dados de espectroscopia NIR e diferentes métodos de classificação, como SIMCA e MLP para realizar a classificação de lubrificantes por viscosidades. Em outro trabalho realizou também a classificação quanto aos óleos básicos em minerais, sintéticos e semi-sintéticos. Lima *et al.*[83] determinaram parâmetros de qualidade de óleos básicos minerais utilizando dados de espectroscopia NIR e PLS.

Conforme mencionado, o método ASTM D3238-95 (método ndm), utilizado para determinação do $\%C_P$, $\%C_N$ e $\%C_A$ em óleos básicos demanda a obtenção de diversos parâmetros da amostra e a realização de cálculos.

Para o controle de qualidade e identificação de possíveis adulterações em óleos básicos parafínicos e óleo lubrificante de motor automotivo é adequado o desenvolvimento de modelos de classificação simples, eficientes e eficazes que possibilitem identificar teores de compostos naftênicos acima do comum para óleos parafínicos e simultaneamente possibilitam identificar também a presença de óleo vegetal. A possibilidade de utilizar o mesmo modelo de classificação para análise de amostras comerciais de óleo lubrificante de motor automotivo sugere uma excelente ferramenta de controle de qualidade. A obtenção de um modelo de classificação com essas características utilizando dados de espectroscopia NIR e SVM é abordada no capítulo 12.

Seção II – Modelos de regressão

7 – Determinação de parâmetros de qualidade em óleo diesel utilizando espectroscopia NIR e SVM para utilização no sistema otimizador do misturador em linha

Esse trabalho demonstra a utilização do algoritmo SVM aplicado a dados de espectroscopia NIR de óleo diesel para obtenção de modelos de calibração mais eficazes para determinação dos parâmetros ponto de fulgor e número de cetano, podendo ser utilizados pelo analisador *on-line* que fornece informação ao sistema otimizador do misturador em linha de produção de óleo diesel e que se tornam uma importante ferramenta de Tecnologia Analítica de Processos – PAT, substituindo os métodos tradicionais, que demandam excessivo tempo de ensaio e têm custo relativamente elevado.

7.1 – Parte experimental

Foram utilizados dados de espectroscopia NIR de amostras de óleo diesel para as quais realizou-se também os ensaios através dos métodos de referência. Todos os dados foram obtidos pelo setor de otimização e qualidade de produto da refinaria de Paulinia/SP – Replan. Para obtenção dos espectros na região do infravermelho foi utilizado um espectrômetro ABB/Bomen MID/NIR com fonte glowbar (carbeto de silício), detector de sulfato de triglicina deuterada (DTGS), usando uma cubeta de transmitância de CaF_2 de caminho óptico igual a 0,5 mm. Cada espectro foi obtido como uma média de 32 varreduras nas regiões espectrais utilizadas, com resolução de 4 cm^{-1}. O espectrômetro utilizado em laboratório é mostrado na figura 7.1(a) e um instrumento idêntico, do mesmo fabricante, com adaptações para utilização na linha de produção da refinaria é mostrado na figura 7.1(b). A utilização desses equipamentos do mesmo fabricante e de características semelhantes deverá facilitar a obtenção de uma transferência de calibração eficaz.

Foram realizados diferentes pré-processamentos dos dados para verificar qual proporciona a construção do melhor modelo utilizando os algoritmos PLS e SVM. Os pré-processamentos testados foram: correção de linha base WLS (WLS *baseline method*); correção de linha base WLS e centragem na média; SNV; e correção de linha base WLS e primeira derivada (janela com 15 pontos).

(a) (b)

Figura 7.1 – (a) espectrômetro MID/NIR de bancada (b) espectrômetro MID/NIR de processo instalado no abrigo dos analisadores

O algoritmo Kennard-Stone[84] foi aplicado para obtenção dos conjuntos de calibração e validação, após o pré-processamento dos dados. Esse é um algoritmo clássico para selecionar os objetos de cada conjunto de maneira uniforme no espaço amostral. Ele começa selecionando as duas amostras com a maior distância euclidiana entre si. Para cada uma das amostras restantes, calcula-se a distância com respeito às amostras já selecionadas. Feito isso, a amostra com a maior distância é retida, e o

procedimento é repetido até que um determinado número de amostras seja selecionado. Como o algoritmo Kennard-Stone utiliza a comparação entre os espectros para obtenção dos conjuntos, diferentes conjuntos de calibração e validação podem ser obtidos em função de diferentes pré-processamentos utilizados.

O pacote LIBSVM[85] versão 2.88 foi utilizado para o desenvolvimento dos modelos com SVM e o algoritmo genético foi aplicado para realização da otimização paramétrica. Todos os programas são adequados para utilização com Matlab 7.7 da Mathworks.

Para obtenção dos modelos com SVM foram testadas diferentes funções kernel, tais como: linear, RBF, polinomial e sigmoidal.

Para construção dos modelos com SVM os blocos de dados **X** e **y** dos conjuntos de calibração e validação foram previamente escalonados entre [0,1]. Foi utilizado como parâmetro γ do kernel RBF o valor *default* do pacote LIBSVM ($\gamma = 1/k$, onde k é o número de atributos ou variáveis nos dados de entrada) e o grau do polinômio no kernel polinomial igual a 3. Os parâmetros C e ν foram selecionados entre os intervalos de 0 a 10^4 e 10^{-4} a 1, respectivamente, utilizando-se o algoritmo genético (GA). Ainda, para otimização paramétrica com GA estipulou-se a utilização de uma população com 30 indivíduos e um máximo de 15 gerações, pois observou-se que com essa configuração o valor do erro de validação cruzada se estabiliza, não havendo melhora com o aumento do número de gerações. A função objetivo a ser otimizada pelo GA, e definido inicialmente, consistiu simplesmente na utilização dos valores obtidos pela validação cruzada com 5 e 3 subconjuntos do conjunto de treinamento, para calibração do ponto de fulgor e do número de cetano, respectivamente, buscando-se o menor valor do erro de validação cruzada. Os parâmetros a serem otimizados foram tratados como genes no GA.

Como a minimização do erro de validação cruzada no conjunto de treinamento não garante a obtenção dos parâmetros ótimos, eventualmente um *grid search* manual pode ser necessário, a partir dos valores previamente selecionados pelo GA, para refinamento do resultado, sempre considerando a utilização do adequado número de vetores de suporte e valores próximos de RMSEC e RMSEP de modo a evitar um sobreajuste do modelo.

7.1.1 – Procedimento experimental para o modelo de regressão do ponto de fulgor

Todas as 451 amostras utilizadas foram analisadas segundo o método ASTM D56 para a determinação do ponto de fulgor pelo método do vaso fechado tag, com o analisador automático FP56 5G2 da ISL, mostrado na figura 7.2.

Figura 7.2 – Analisador automático para determinação do ponto de fulgor

O intervalo de valores desse parâmetro no espaço amostral varia de 24,5 ºC a 76,5 ºC, o qual permite cobrir toda a variabilidade encontrada nas condições do processo produtivo da refinaria, conforme preconiza a norma ASTM E 1655-05 (*Standard practices for infrared multivariate quantitative analysis*).

Os espectros foram obtidos no intervalo de 3944 cm^{-1} a 4769 cm^{-1} (região das bandas de combinação).

7.1.2 - Procedimento experimental para o modelo de regressão do número de cetano

As 114 amostras utilizadas foram analisadas conforme o método ASTM D 613, para determinação do número de cetano do óleo diesel. A figura 7.3, mostra o motor padrão, do fabricante Waukesha, utilizado para realização dos ensaios.

Os espectros foram obtidos no intervalo de 3500 cm^{-1} a 6129 cm^{-1}.(região das bandas de combinação e do primeiro sobreton).

Figura 7.3 – Motor para o ensaio de determinação de número de cetano

O intervalo de valores desse parâmetro no espaço amostral varia de 37,6 a 48,9, de modo a abranger toda a variabilidade encontrada no processo produtivo, quando da coleta das amostras.

7.2 – Resultados e discussão

Os métodos analíticos de referência citados, tanto o ensaio para determinação do ponto de fulgor como para determinação do número de cetano têm custo relativamente elevado, com utilização de compostos orgânicos padrões, envolvimento de técnicos na condução do ensaio e também demandando um longo tempo para

realização da análise. Estas características praticamente inviabilizam o uso destes métodos de laboratório para utilização em processos de otimização de produção de óleo diesel. Assim, buscou-se a obtenção de métodos que possam ser implantados para medição em linha, que sejam rápidos, de custo relativamente baixo, necessitem de pouco envolvimento de mão de obra após sua instalação, forneçam resultados eficazes e permitam a obtenção de modelos robustos o bastante para cobrir toda a variabilidade do espaço amostral que pode ser encontrada no processo de refino do óleo diesel.

Para comparar se há diferença significante entre os modelos SVM e PLS desenvolvidos foi realizado um teste-F. Também, a validação quanto a concordância entre os resultados obtidos pelo método de referência e pelo modelo de calibração utilizando SVM foi realizado para os parâmetros estudados conforme o procedimento descrito pelo método ASTM E 1655-05 (*Standard practices for infrared multivariate quantitative analysis*).

7.2.1 – Modelos de regressão para o ponto de fulgor

Na construção dos modelos de regressão para o ponto de fulgor foram realizados ensaios para determinação do ponto de fulgor pelo método de referência e obtidos espectros na região do infravermelho próximo para 451 amostras. Obtiveram-se modelos utilizando 350 amostras de calibração e 101 amostras para validação.

O intervalo espectral utilizado foi de 3944 cm^{-1} a 4769 cm^{-1}, que corresponde a região das bandas de combinação. Essa região possui diversas bandas de absorção atribuídas a combinação de modos vibracionais da ligação C-H de grupos metil e metileno e ligação C-H de anéis aromáticos. As atribuições de algumas bandas nessa região espectral podem ser verificadas na tabela 2.1.

Essa região permite a calibração do ponto de fulgor porque este parâmetro está associado a presença de frações leves e pesadas do petróleo na amostra. Quanto maior for o teor de frações mais leves, menor será o ponto de fulgor. Moléculas de alcanos de cadeia longa são os principais constituintes das frações mais leves, enquanto compostos naftênicos e aromáticos estão presentes em maiores proporções nas frações mais pesadas do petróleo.

Os espectros das 451 amostras utilizadas são mostrados na figura 7.4.

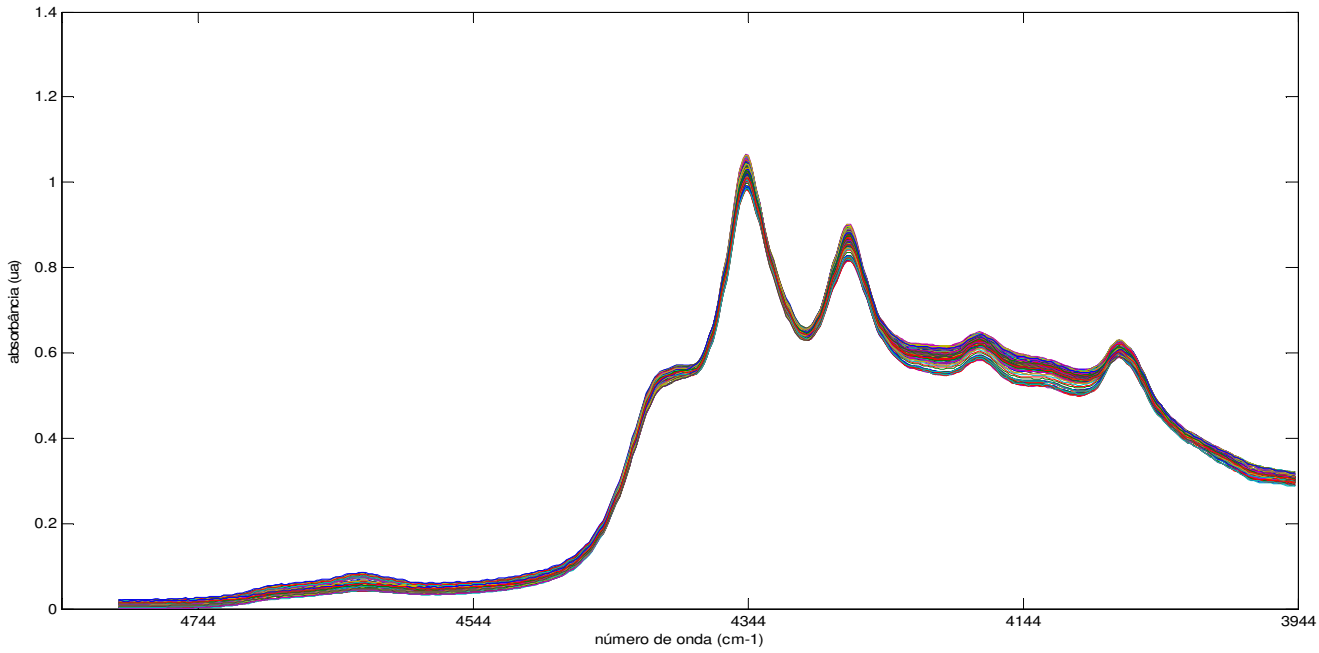


Figura 7.4 – Espectros das 451 amostras utilizadas na calibração e validação do modelo do ponto de fulgor

7.2.1.1 – Modelo PLS

O melhor resultado foi obtido utilizando-se o pré-processamento SNV. Utilizou-se 3 variáveis latentes, que explicam 99,12% da variância dos dados. Os resultados são mostrados na tabela 7.1.

Tabela 7.1 – Resultados do melhor modelo de calibração para o ponto de fulgor obtido com PLS.

Pré-processamento	RMSEC (ºC)	RMSEP (ºC)	R^2
SNV	4,21	3,77	0,698

A figura 7.5 ilustra o resultado do modelo ajustado com o algoritmo PLS e com as 350 amostras de calibração e as 101 amostras de validação.

Apesar de a reprodutibilidade especificada pela norma ASTM D56 ser de 4,3 ºC, o valor obtido para o RMSEP com o modelo PLS construído não é considerado satisfatório, uma vez que o ponto de fulgor do óleo diesel tem um valor mínimo de 38 ºC especificado na legislação brasileira, ou seja, o erro de previsão do modelo fica próximo de 10% desse valor, e, tratando-se de uma propriedade limitante para a especificação do óleo diesel produzido na refinaria é importante a obtenção de um modelo que proporcione resultados mais exatos para utilização pelo otimizador em linha de produção. Dessa forma, novos modelos de previsão foram propostos, como mostrado a seguir.

7.2.1.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a função kernel RBF e os dados com correção de linha base WLS e centrados na média. A figura 7.6 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 7.2. Também são mostrados os resultados obtidos com o pré-processamento SNV, que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

O melhor modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 47% melhor em relação ao valor obtido para o modelo PLS. Esse valor, de 1,98 ºC, torna o modelo construído muito útil para utilização pelo otimizador do misturador em linha nas refinarias, uma vez que está bem abaixo do valor especificado para a reprodutibilidade da análise do ponto de fulgor pelo método de referência.

Verificou-se que as funções kernel polynomial e linear também proporcionam a construção de bons modelos para o fim proposto, uma vez que fornecem valores de RMSEP muito próximos do obtido com a função kernel RBF.

O bom ajuste dos modelos pode ser verificado ao comparar-se os valores de RMSEC e RMSEP obtidos para um modelo específico, os quais não diferem significativamente, indicando não haver sobreajuste do mesmo.

Também, em todos os modelos mostrados (com exceção do que utilizou o pré-processamento SNV e o kernel linear) o número de vetores de suporte utilizados não ultrapassa ou fica bem próximo do máximo de dois terços das 350 amostras de calibração, o que pode ser considerado uma indicação de bom ajuste do modelo, pois considera-se que quanto menor o número de vetores de suporte utilizados, menor é a possibilidade de estar havendo sobreajuste do modelo.

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 7.5 e 7.6, também pode-se verificar o melhor ajuste do modelo SVM em relação ao modelo PLS através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos mostrados nas figuras 7.7 e 7.8, respectivamente. A figura 7.7 evidencia uma piora nos valores previstos pelo modelo PLS com o aumento do ponto de fulgor, caracterizando uma relação não linear no espaço amostral utilizado. A figura 7.8 demonstra que o modelo SVM possibilita uma modelagem sensivelmente melhor nesse espaço amostral.

A tabela 7.3 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 7.7 mostra uma síntese dos resultados obtidos com SVM e PLS e os valores de referência estabelecidos para reprodutibilidade dos métodos (pelos métodos ASTM) e valores limites estabelecidos pela legislação da ANP para o produto final.

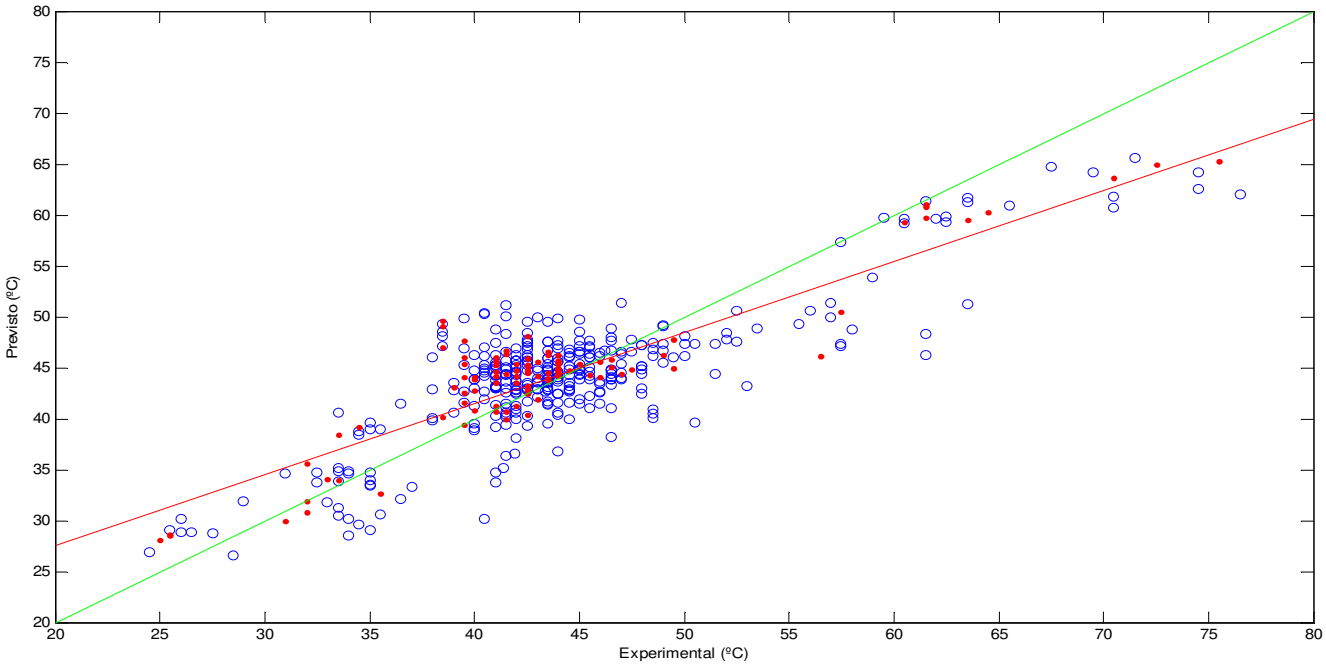


Figura 7.5 – Valores experimentais contra previstos para o modelo PLS para o ponto de fulgor com as 350 amostras de calibração (○) e as 101 amostras de validação (●).

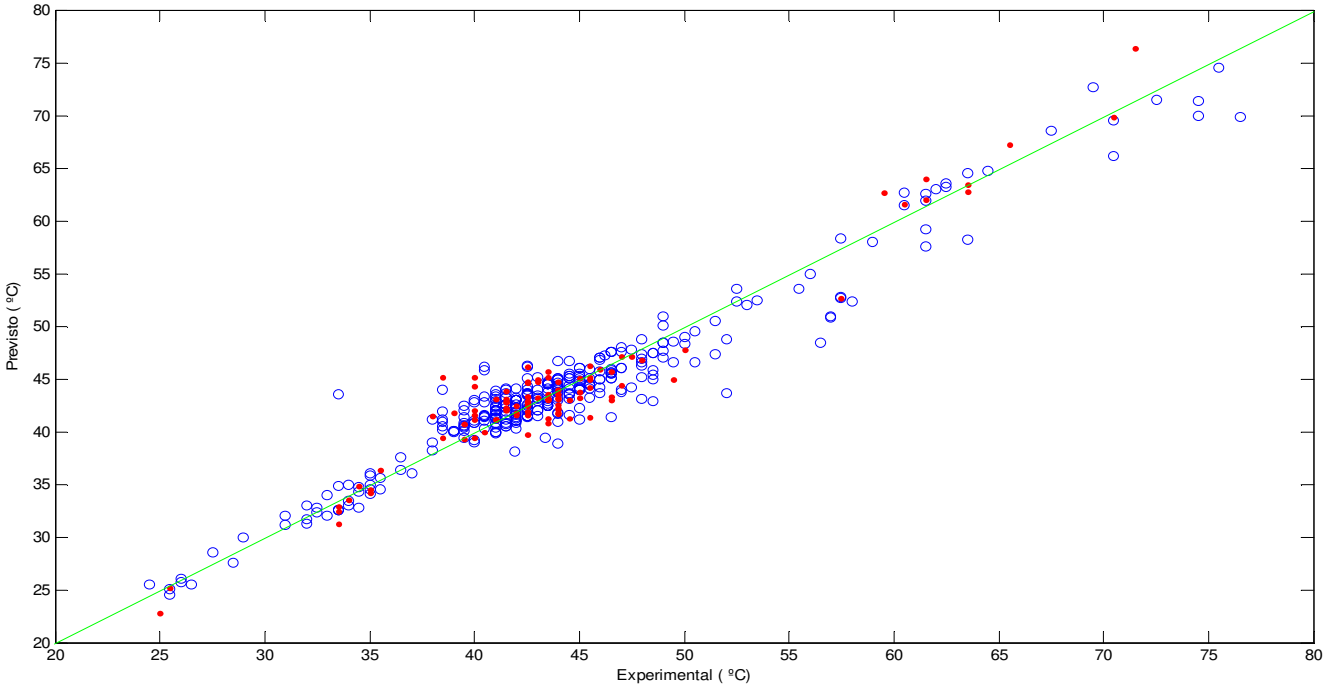
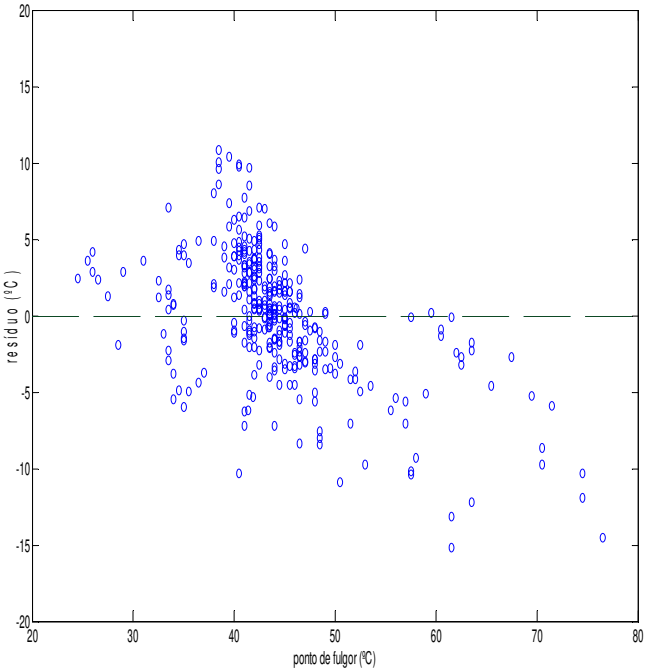


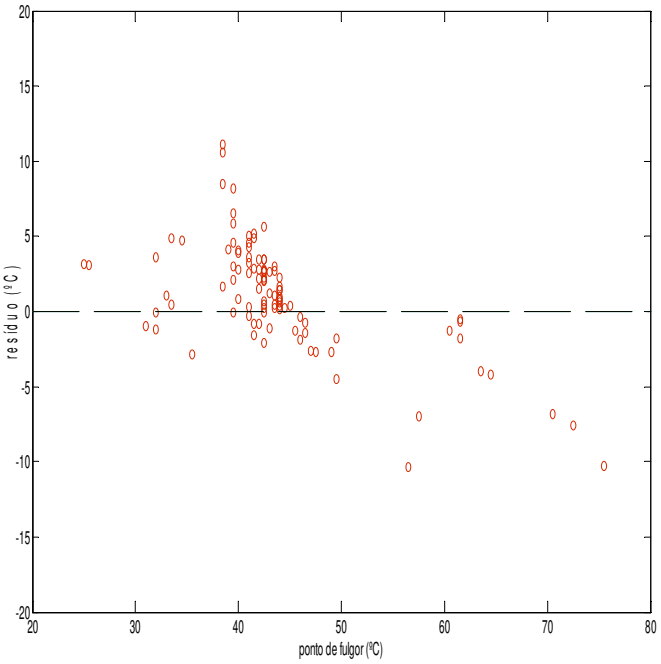
Figura 7.6 – Valores experimentais contra previstos para o modelo SVM para o ponto de fulgor com as 350 amostras de calibração (○) e as 101 amostras de validação (●).

Tabela 7.2 – Modelos de calibração para o ponto de fulgor obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento SNV							
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				
		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	SV's		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	SV's	
RBF	Correção de linha base e centragem na média	255,4	0,4601	1,99	1,98	0,9357	239	SNV	157,3	0,4038	2,02	2,26	0,9297	210	
Polinomial	Correção de linha base e centragem na média	188,6	0,4600	2,67	2,03	0,8948	196		3000,5	0,0981	1,99	2,30	0,9308	115	
Sigmoidal	Correção de linha base e centragem na média	12,5	0,0246	5,49	4,80	0,6166	13		6,3	0,0946	4,67	4,70	0,7033	39	
Linear	Correção de linha base e centragem na média	52,8	0,0296	2,70	2,20	0,8910	214		55,2	0,3501	1,56	3,68	0,9210	265	

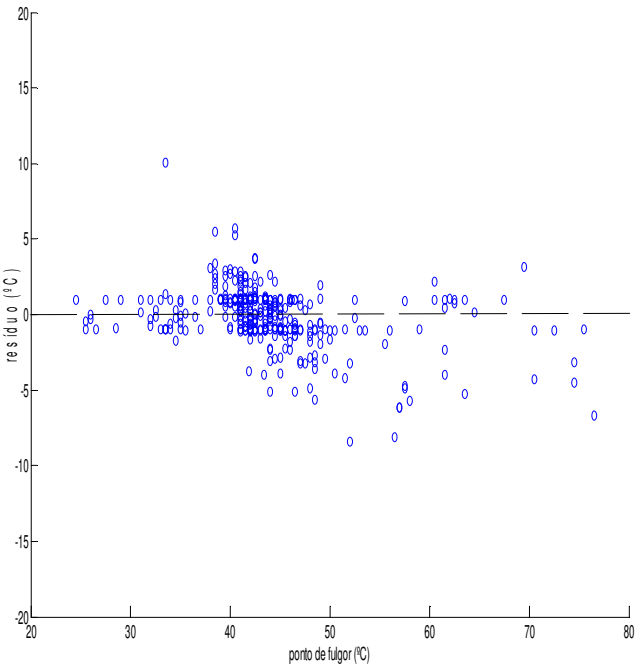


(a)

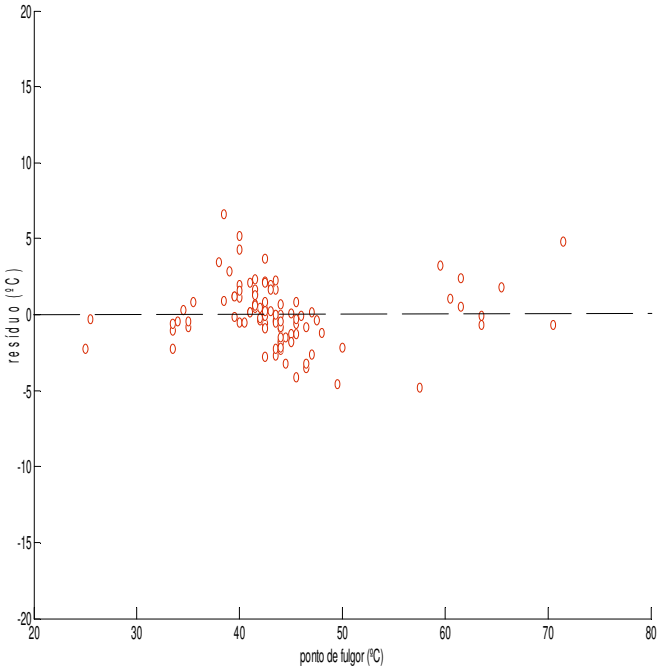


(b)

Figura 7.7 – distribuição dos erros de calibração (a) e validação (b) do modelo PLS para o ponto de fulgor



(a)



(b)

Figura 7.8 – distribuição dos erros de calibração (a) e validação (b) do modelo SVM para o ponto de fulgor

Tabela 7.3 – resultados de previsão dos modelos PLS e SVM para o ponto de fulgor

	PLS			SVM		
	Medido (ºC)	Previsto (ºC)	erro relativo (%)	Medido (ºC)	Previsto (ºC)	erro relativo (%)
1.	25,00	28,12	12,48	25,00	22,74	-9,04
2.	25,50	28,57	12,03	25,50	25,21	-1,14
3.	25,50	28,61	12,19	33,50	32,87	-1,89
4.	31,00	30,00	-3,23	33,50	31,26	-6,70
5.	32,00	30,77	-3,83	33,50	32,43	-3,19
6.	32,00	31,92	-0,25	34,00	33,55	-1,33
7.	32,00	35,58	11,19	34,50	34,82	0,91
8.	33,00	34,06	3,22	35,00	34,53	-1,33
9.	33,50	38,42	14,67	35,00	34,16	-2,40
10.	33,50	33,93	1,29	35,50	36,33	2,35
11.	34,50	39,24	13,73	38,00	41,45	9,08
12.	35,50	32,64	-8,05	38,50	45,13	17,23
13.	38,50	49,61	28,85	38,50	39,42	2,38
14.	38,50	40,13	4,24	39,00	41,85	7,30
15.	38,50	49,09	27,51	39,50	40,68	2,99
16.	38,50	47,00	22,07	39,50	40,69	3,01
17.	39,00	43,14	10,61	39,50	39,34	-0,40
18.	39,50	44,10	11,64	40,00	44,29	10,73
19.	39,50	41,62	5,36	40,00	41,54	3,85
20.	39,50	39,46	-0,11	40,00	45,20	12,99
21.	39,50	42,54	7,69	40,00	41,96	4,89
22.	39,50	45,34	14,79	40,00	39,44	-1,40
23.	39,50	46,02	16,50	40,00	41,15	2,89
24.	39,50	47,67	20,69	40,50	39,96	-1,33
25.	40,00	43,87	9,68	41,00	41,13	0,32
26.	40,00	44,07	10,17	41,00	41,17	0,42
27.	40,00	40,86	2,14	41,00	43,08	5,08
28.	40,00	42,78	6,95	41,50	42,11	1,48
29.	41,00	46,02	12,24	41,50	42,20	1,68
30.	41,00	40,69	-0,77	41,50	42,81	3,15
31.	41,00	43,56	6,23	41,50	41,97	1,14
32.	41,00	41,28	0,68	41,50	43,83	5,60
33.	41,00	44,20	7,80	41,50	43,13	3,93
34.	41,00	45,55	11,10	41,50	42,85	3,24

35.	41,00	45,30	10,48	42,00	41,79	-0,49
36.	41,00	44,59	8,76	42,00	42,46	1,09
37.	41,50	46,35	11,68	42,00	41,59	-0,98
38.	41,50	46,67	12,46	42,50	43,36	2,03
39.	41,50	44,36	6,89	42,50	42,75	0,58
40.	41,50	40,66	-2,03	42,50	42,42	-0,18
41.	41,50	39,96	-3,72	42,50	41,57	-2,19
42.	42,00	44,19	5,22	42,50	42,52	0,05
43.	42,00	45,43	8,16	42,50	44,57	4,87
44.	42,00	44,74	6,52	42,50	46,18	8,66
45.	42,00	43,53	3,64	42,50	43,27	1,82
46.	42,00	41,21	-1,89	42,50	42,93	1,00
47.	42,50	45,95	8,11	42,50	41,91	-1,40
48.	42,50	45,28	6,55	42,50	39,73	-6,51
49.	42,50	45,22	6,40	42,50	42,59	0,21
50.	42,50	42,40	-0,25	42,50	44,69	5,16
51.	42,50	45,15	6,24	43,00	44,68	3,90
52.	42,50	45,94	8,09	43,00	44,93	4,49
53.	42,50	42,72	0,53	43,00	43,19	0,45
54.	42,50	44,51	4,72	43,50	43,51	0,03
55.	42,50	40,39	-4,96	43,50	45,73	5,12
56.	42,50	42,92	0,99	43,50	42,99	-1,18
57.	42,50	43,16	1,56	43,50	43,11	-0,90
58.	42,50	44,58	4,90	43,50	43,06	-1,02
59.	42,50	44,72	5,23	43,50	41,24	-5,21
60.	42,50	48,13	13,24	43,50	45,18	3,85
61.	43,00	45,63	6,12	43,50	40,82	-6,17
62.	43,00	41,89	-2,59	44,00	44,68	1,54
63.	43,00	44,23	2,86	44,00	43,58	-0,95
64.	43,50	46,21	6,22	44,00	42,51	-3,38
65.	43,50	44,53	2,36	44,00	43,21	-1,79
66.	43,50	43,94	1,02	44,00	41,82	-4,96
67.	43,50	43,70	0,47	44,00	42,16	-4,18
68.	43,50	46,53	6,97	44,00	42,26	-3,95
69.	44,00	46,27	5,17	44,00	41,67	-5,29
70.	44,00	45,68	3,83	44,00	44,00	0,01
71.	44,00	44,58	1,31	44,00	43,11	-2,03
72.	44,00	44,83	1,88	44,50	42,97	-3,45
73.	44,00	44,74	1,69	44,50	41,30	-7,19

74.	44,00	44,98	2,23	45,00	45,09	0,20
75.	44,00	45,46	3,31	45,00	43,17	-4,06
76.	44,00	44,37	0,84	45,00	43,73	-2,82
77.	44,00	44,18	0,41	45,50	45,19	-0,69
78.	44,00	45,42	3,23	45,50	44,93	-1,24
79.	44,50	44,74	0,55	45,50	46,29	1,73
80.	45,00	45,37	0,82	45,50	41,37	-9,07
81.	45,50	44,26	-2,73	45,50	44,20	-2,87
82.	46,00	45,64	-0,77	46,00	45,92	-0,17
83.	46,00	44,11	-4,10	46,50	43,28	-6,91
84.	46,50	45,78	-1,54	46,50	45,66	-1,82
85.	46,50	45,05	-3,11	46,50	43,01	-7,50
86.	47,00	44,40	-5,54	47,00	44,41	-5,51
87.	47,50	44,81	-5,67	47,00	47,15	0,33
88.	49,00	46,28	-5,56	47,50	47,11	-0,81
89.	49,50	47,72	-3,59	48,00	46,78	-2,53
90.	49,50	44,99	-9,10	49,50	44,91	-9,28
91.	56,50	46,14	-18,34	50,00	47,79	-4,41
92.	57,50	50,54	-12,10	57,50	52,71	-8,33
93.	60,50	59,25	-2,06	59,50	62,71	5,39
94.	61,50	61,00	-0,81	60,50	61,58	1,78
95.	61,50	60,85	-1,05	61,50	63,93	3,95
96.	61,50	59,70	-2,92	61,50	62,04	0,88
97.	63,50	59,52	-6,27	63,50	62,82	-1,08
98.	64,50	60,28	-6,55	63,50	63,44	-0,09
99.	70,50	63,68	-9,68	65,50	67,27	2,71
100.	72,50	64,92	-10,46	70,50	69,84	-0,94
101.	75,50	65,25	-13,58	71,50	76,31	6,73

7.2.2 – Modelos de regressão para o número de cetano

Os modelos de regressão para o número de cetano foram feitos a partir da realização de ensaios pelo método de referência para determinação do número de cetano e com a obtenção de espectros na região do infravermelho próximo para 114 amostras. Obtiveram-se modelos utilizando 77 amostras de calibração e 37 amostras para validação.

Os espectros foram obtidos no intervalo de 3470 cm^{-1} a 6129 cm^{-1}. As bandas de absorção encontradas na região de combinação de modos vibracionais, mencionada na seção anterior, também são úteis para determinação de número de cetano. Além disso, as bandas de absorção entre 5290 cm^{-1} e 6129 cm^{-1} atribuídas ao primeiro sobreton de modos vibracionais de estiramento da ligação CH dos grupos metil e metileno, ligação CH de anéis aromáticos e ligação CH do grupo metil ligado a anel aromático, também proporcionam informação para determinação do número de cetano. As atribuições de algumas bandas nessa região espectral podem ser verificadas na tabela 2.1.

Ambas as regiões proporcionam importante informação para determinação desse parâmetro porque os compostos parafínicos aumentam os valores do número de cetano, enquanto os cíclicos e aromáticos diminuem os valores desse parâmetro.

De modo a encontrar a melhor região espectral para obtenção dos modelos de calibração foi testada a obtenção de modelos utilizando: (i) toda a região espectral citada; (ii) somente a parte que corresponde a região das bandas de combinação, entre 3500 cm^{-1} a 4678 cm^{-1}; (iii) somente a região do primeiro sobreton, entre 5290 cm^{-1} a 6129 cm^{-1}.

Verificou-se que tanto os modelos com PLS como com SVM forneceram melhores resultados com a utilização somente da região das bandas de combinação, embora também seja possível a calibração utilizando a região do primeiro sobreton.

Os espectros das 114 amostras utilizadas são mostrados na figura 7.9.

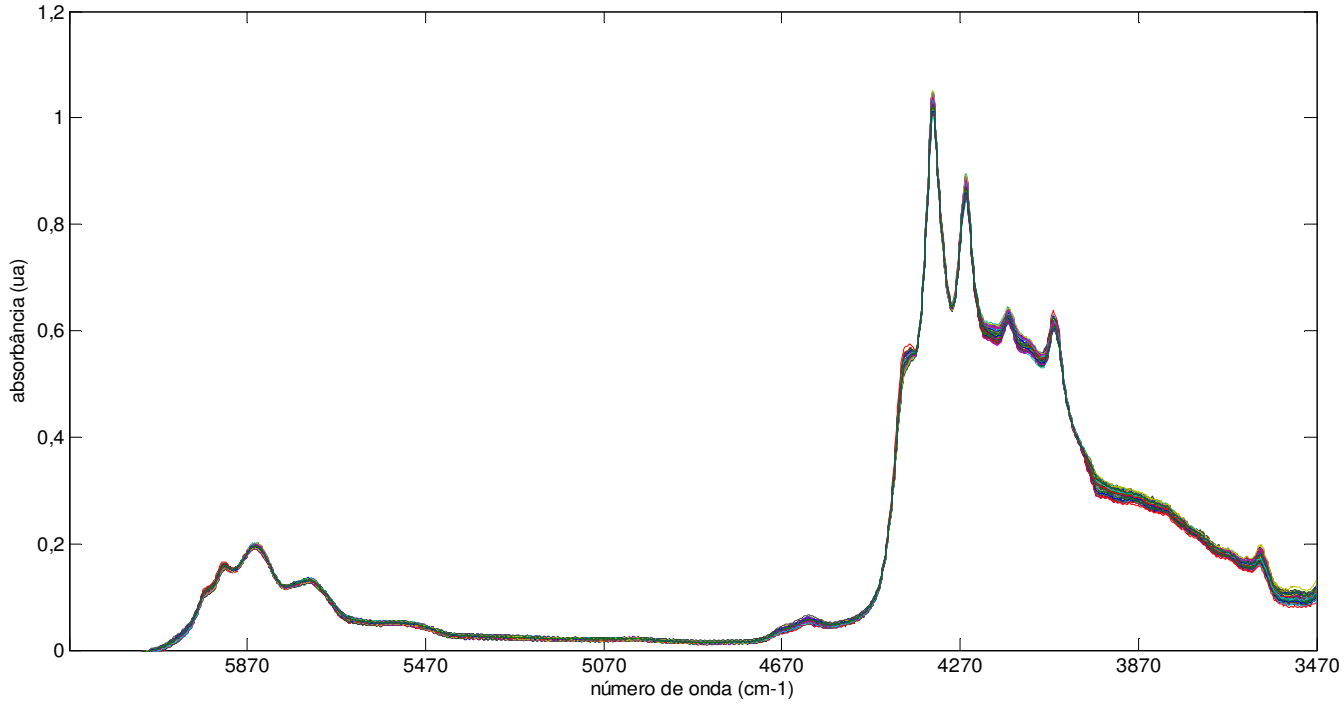


Figura 7.9 - Espectros das 114 amostras utilizadas na calibração e validação do modelo para o número de cetano

7.2.2.1 – Modelo PLS

Obteve-se o modelo que fornece melhores resultados utilizando os dados com correção da linha base WLS e centrados na média. Utilizou-se 5 variáveis latentes, que explicam 97,06 % da variância dos dados. Os resultados são mostrados na tabela 7.4.

Tabela 7.4 – Resultados do melhor modelo de calibração para o número de cetano obtido com PLS.

Pré-processamento	RMSEC (ºC)	RMSEP (ºC)	R^2
Correção da linha base e centragem na média	0,7451	0,5564	0,894

A figura 7.10 ilustra o resultado do modelo ajustado com o algoritmos PLS.

A reprodutibilidade especificada pela norma ASTM D613 é de 2,8, o que torna possível o uso desse modelo de calibração tomando a comparação com o RMSEP obtido como parâmetro de decisão.

No entanto, mais uma vez deve-se considerar a importância desse parâmetro de qualidade do óleo diesel na especificação do produto e na utilização pelo otimizador do misturador em linha de produção. Assim, buscou-se a obtenção de um modelo de calibração com melhor poder de ajuste aos dados, visando diminuir o erro de previsão.

7.2.2.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a função kernel RBF e os dados com correção de linha base WLS e centrados na média. A figura 7.11 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 7.5. Também são mostrados os resultados obtidos para os dados com correção de linha base e centrados na média, pré-processamento que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

Todas as funções kernel utilizadas, com exceção da linear, proporcionaram resultados com RMSEP melhor do que o obtido com o modelo PLS. Aqui mais uma vez verifica-se a consistência dos modelos obtidos, que utilizam um número não excessivo de vetores de suporte e valores próximos de RMSEC e RMSEP, indicando que não há sobreajuste dos modelos.

A utilização da função kernel linear também proporcionou um bom resultado, embora um pouco pior em relação aos demais.

O melhor modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 20% melhor em relação ao valor obtido para o modelo PLS. Esse valor, de 0,4535, torna o modelo construído muito útil para utilização pelo otimizador do misturador em linha nas refinarias, uma vez que está bem abaixo do valor especificado para a reprodutibilidade da análise do número de cetano pelo método de referência.

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 7.10 e 7.11, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 7.12 e 7.13, respectivamente. Verifica-se que o modelo SVM possibilita a diminuição dos erros de previsão nesse espaço amostral.

A tabela 7.6 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 7.7 mostra uma síntese dos resultados obtidos com SVM e PLS e dos valores de referência estabelecidos pelos métodos ASTM e pela legislação da ANP.

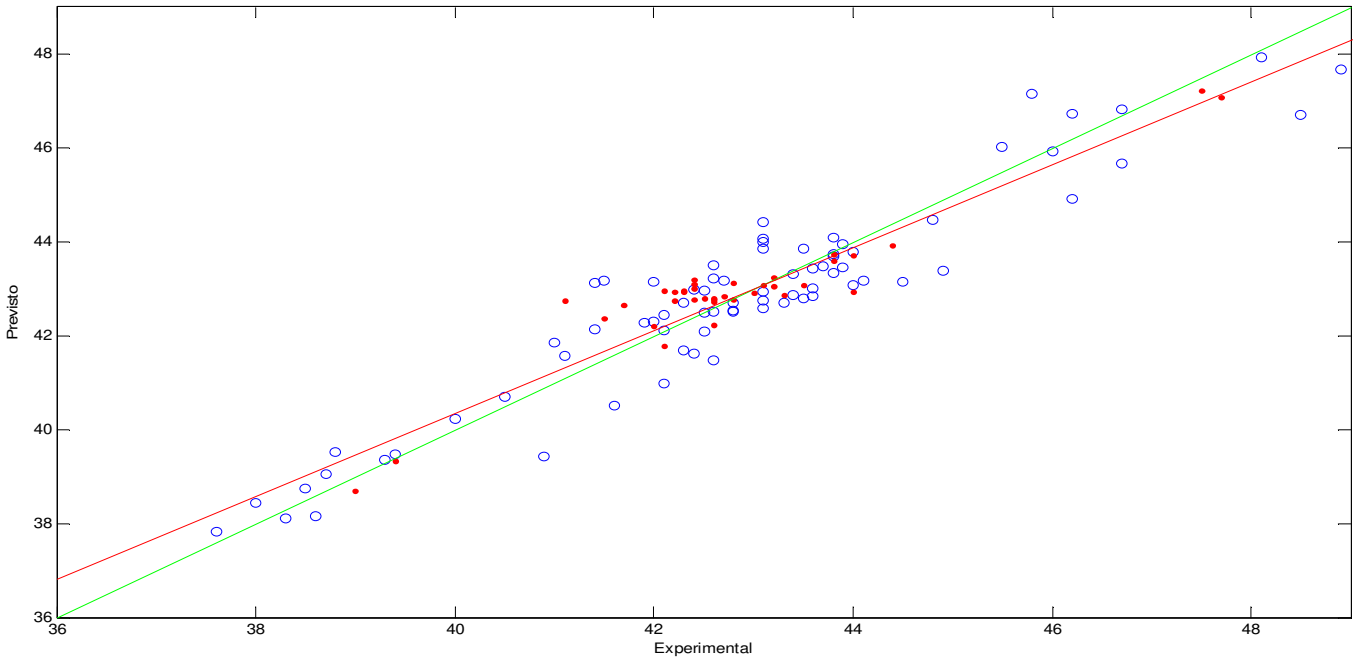


Figura 7.10 – Valores experimentais contra previstos para o modelo PLS para o número de cetano com as 77 amostras de calibração (○) e as 37 amostras de validação (●).

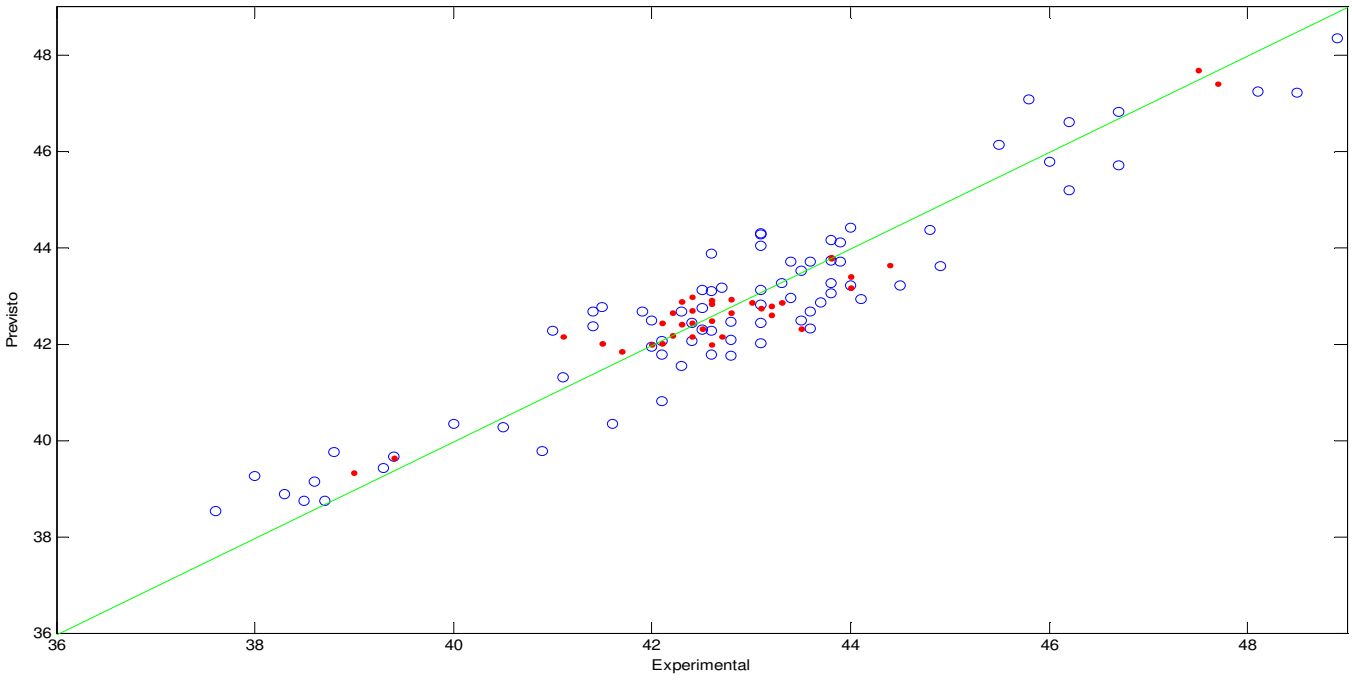
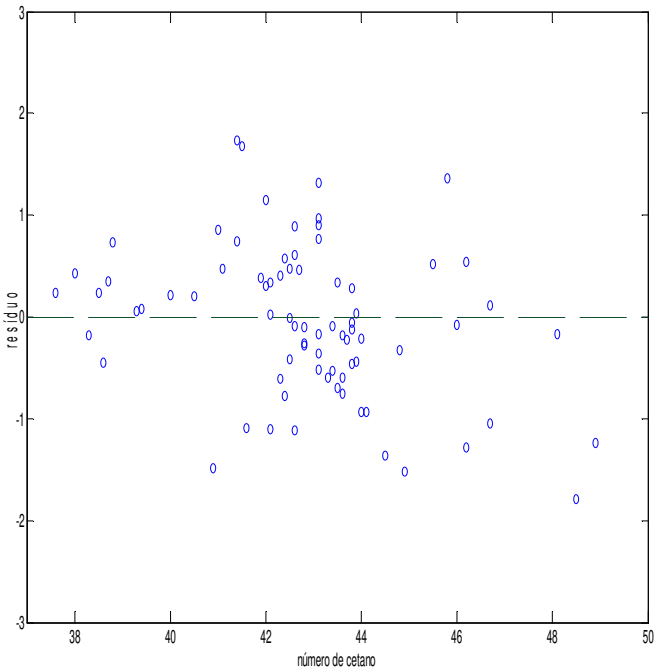


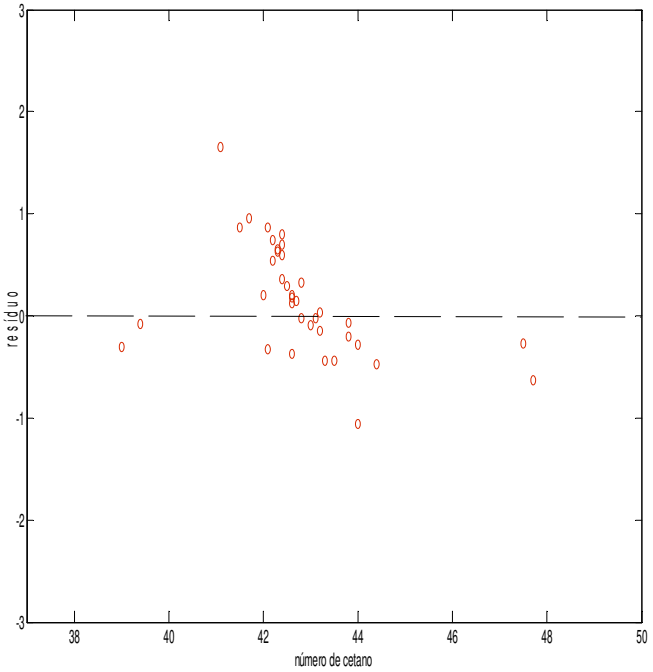
Figura 7.11 – Valores experimentais contra previstos para o modelo SVM para o número de cetano com as 77 amostras de calibração (○) e as 37 amostras de validação (●).

Tabela 7.5 – Modelos de calibração para o número de cetano obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento correção da linha base e centragem na média							
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				
		C	ν	RMSEC	RMSEP	R^2	SV's		C	ν	RMSEC	RMSEP	R^2	SV's	
RBF	Correção de linha base e centragem na média	440,0	0,0026	0,7652	0,4535	0,8946	11	Correção de linha base e centragem na média	440,0	0,0026	0,7652	0,4535	0,8946	11	
Polinomial	Correção de linha base	25,9	0,1761	0,7693	0,4902	0,8923	21		790,7	0,0099	0,7361	0,4993	0,8975	12	
Sigmoidal	Correção de linha base	12,8	0,6462	0,8216	0,5025	0,8767	11		22,8	0,1069	0,7763	0,5280	0,8856	12	
Linear	Correção de linha base e centragem na média	3,8	0,0189	0,6727	0,5689	0,9053	24		3,8	0,0189	0,6727	0,5689	0,9053	24	

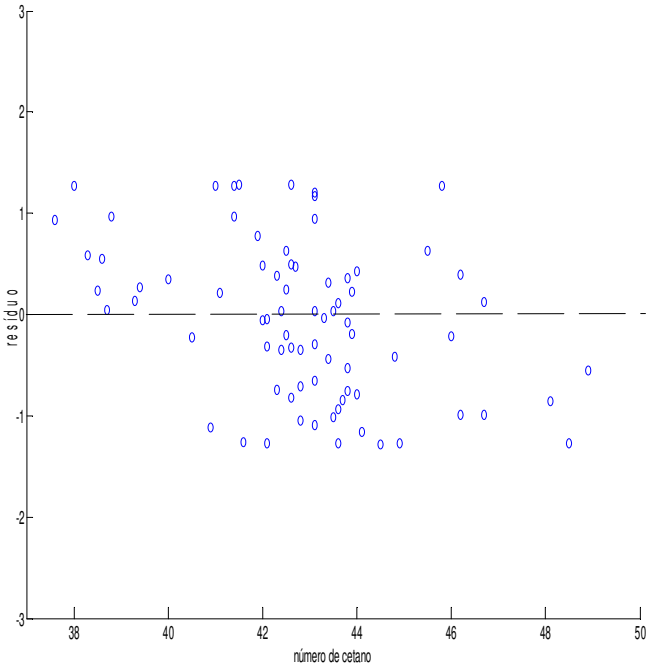


(a)

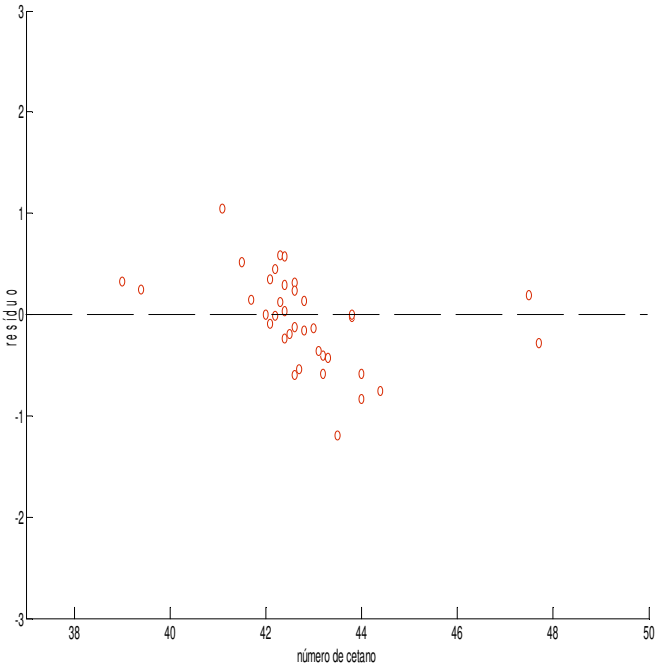


(b)

Figura 7.12 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS para o número de cetano



(a)



(b)

Figura 7.13 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM para o número de cetano

Tabela 7.6 – resultados de previsão dos modelos PLS e SVM para o NC

	PLS			SVM		
	Medido (ºC)	Previsto (ºC)	erro relativo (%)	Medido (ºC)	Previsto (ºC)	erro relativo (%)
1.	39,00	38,70	-0,77	39,00	39,33	0,85
2.	39,40	39,32	-0,20	39,40	39,65	0,63
3.	41,10	42,75	4,03	41,10	42,15	2,57
4.	41,50	42,36	2,08	41,50	42,02	1,25
5.	41,70	42,65	2,29	41,70	41,85	0,35
6.	42,00	42,20	0,48	42,00	42,00	-0,01
7.	42,10	41,78	-0,76	42,10	42,01	-0,21
8.	42,10	42,97	2,06	42,10	42,45	0,82
9.	42,20	42,74	1,27	42,20	42,65	1,07
10.	42,20	42,95	1,77	42,20	42,19	-0,02
11.	42,30	42,96	1,55	42,30	42,89	1,40
12.	42,30	42,93	1,48	42,30	42,42	0,27
13.	42,40	43,19	1,87	42,40	42,97	1,35
14.	42,40	42,76	0,85	42,40	42,16	-0,56
15.	42,40	43,00	1,41	42,40	42,69	0,69
16.	42,40	43,10	1,65	42,40	42,43	0,07
17.	42,50	42,79	0,68	42,50	42,31	-0,44
18.	42,60	42,80	0,47	42,60	42,91	0,74
19.	42,60	42,72	0,28	42,60	42,48	-0,28
20.	42,60	42,23	-0,87	42,60	42,00	-1,40
21.	42,60	42,78	0,42	42,60	42,84	0,56
22.	42,70	42,85	0,35	42,70	42,16	-1,27
23.	42,80	43,13	0,77	42,80	42,64	-0,37
24.	42,80	42,78	-0,04	42,80	42,93	0,30
25.	43,00	42,91	-0,21	43,00	42,87	-0,30
26.	43,10	43,08	-0,05	43,10	42,74	-0,83
27.	43,20	43,05	-0,35	43,20	42,79	-0,95
28.	43,20	43,24	0,09	43,20	42,61	-1,37
29.	43,30	42,86	-1,01	43,30	42,87	-1,00
30.	43,50	43,07	-1,00	43,50	42,31	-2,73
31.	43,80	43,73	-0,16	43,80	43,78	-0,05
32.	43,80	43,60	-0,46	43,80	43,80	0,01
33.	44,00	43,72	-0,64	44,00	43,41	-1,33
34.	44,00	42,94	-2,41	44,00	43,17	-1,90
35.	44,40	43,93	-1,07	44,40	43,65	-1,68
36.	47,50	47,23	-0,56	47,50	47,69	0,40
37.	47,70	47,07	-1,31	47,70	47,42	-0,59

Tabela 7.7 – Resultados dos modelos PLS e SVM e valores de referência dos métodos ASTM e legislação vigente

Conjunto de dados	**Região espectral (cm^{-1})**	**Parâmetro**	**Modelo PLS RMSEP R^2**	**Modelo SVM RMSEP R^2**	**Melhora do RMSEP (%)**	**ASTM Repetibilidade Reprodutibilidade**	**Espaço amostral**	**Especificação Resolução ANP 65/2011**
A	3944-4769	Ponto de fulgor	3,77 ℃ 0,698	1,98 ℃ 0,936	47	D56-05 1,2 ℃ 4,3 ℃	24,5 ℃ a 76,5 ℃	mín. 38 ℃
B	3500-4678	Número de cetano	0,5564 0,894	0,4535 0,895	20	D613-08 0,8 2,8	37,6 a 48,9	48 (S10) 46 (S50) 42 (S500 e S1800)

7.2.3 – Comparação dos resultados dos modelos SVM e PLS – teste F

Em trabalhos experimentais, especialmente no desenvolvimento de um novo procedimento de análise, é comum realizar-se uma avaliação estatística dos resultados obtidos, visando identificar a existência de uma diferença significativa na variância entre este conjunto de dados e outro conjunto obtido por um procedimento de referência. Esta avaliação é feita usando-se o teste F. Este teste usa a razão das variâncias dos dois conjuntos de dados para estabelecer se efetivamente existe uma diferença estatisticamente significativa entre os valores que estão sendo comparados. O valor de F é calculado pela seguinte expressão:

$$F = \frac{s_1^2}{s_2^2} \tag{7.1}$$

O valor de F obtido é comparado a valores críticos calculados para um determinado nível de confiança. Quando o valor experimental de F excede o valor crítico tabelado, então a diferença na variância é tomada como estatisticamente significante[86,87].

Comumente em aplicações com calibração multivariada utiliza-se o parâmetro RMSEP, que expressa o grau de concordância entre os valores estimados por um modelo previamente construído e o valor considerado real ou de referência.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (y_p - y_r)^2}{n}} \tag{7.2}$$

onde y_p são os valores previstos pelo modelo, y_r são os valores de referência e n é o número de amostras utilizadas no conjunto de validação.

Como pode ser observado na equação 7.2, o RMSEP é uma medida de dispersão semelhante ao desvio padrão, mas que mede a dispersão entre os valores estimados pelo modelo e os valores de referência. Outra propriedade que se assemelha à do desvio padrão é que o RMSEP é uma medida que considera apenas erros

aleatórios, que é uma decorrência da elevação dos erros ao quadrado na equação 7.2. Por exemplo, considerando os resultados de dois métodos distintos, supondo que um apresente erros sistemáticos negativos e o outro tenha erros com o mesmo valor em módulo mas que sejam distribuídos de forma aleatória, ambos fornecem os mesmos valores de RMSEP. Assim, a constatação de que dois RMSEP são estatisticamente equivalentes por meio de um teste-F torna possível afirmar que os erros médios na estimativa da propriedade de interesse dos dois métodos são equivalentes não podendo ser utilizada para inferir sobre a exatidão do método.[88]

A avaliação da similaridade dos resultados obtidos através dos dois diferentes algoritmos utilizados foi feita através do teste F, utilizando os valores calculados do RMSEP para cada modelo. Nesse caso o teste F é usado para comparação dos RMSEP:

$$F = \frac{RMSEP_1^2}{RMSEP_2^2} \tag{7.3}$$

O valores críticos de $F_{100,100}$ (ponto de fulgor) e $F_{36,36}$ (número de cetano) ao nível de significância de 95 % são mostrados na tabela 7.8 juntamente com os valores calculados de F.

Tabela 7.8 – Resultados do teste F na comparação dos modelos PLS e SVM

	F calculado	F crítico 95 %
Ponto de fulgor	3,62	1,39
Número de cetano	1,50	1,74

Para o parâmetro ponto de fulgor há evidências, ao nível de significância de 95%, de que o modelo SVM fornece melhores resultados do que o modelo PLS. Para o parâmetro número de cetano o valor calculado está próximo, mas inferior ao valor

crítico de $F_{36,36}$ ao nível de significância de 95%, indicando que nesse nível de confiança não existe diferença entre os modelos com PLS e SVM.

7.2.4 - Comparação dos resultados de referência com os dos modelos PLS e SVM.

A validação quanto a concordância entre os resultados obtidos pelo método de referência e pelo modelo de calibração utilizando SVM foi realizado para os parâmetros estudados conforme o procedimento descrito pelo método ASTM E 1655-05. Nesse método, considera-se a reprodutibilidade (r) do método ASTM de referência e a seguinte equação:

$$y_i' - r < y_i < y_i' + r \qquad (7.4)$$

onde y_i é o valor de referência obtido com o método ASTM de referência e y_i' é o valor previsto pelo modelo de calibração desenvolvido. Considerando a definição de reprodutibilidade, se 95% ou mais dos valores de referência para o conjunto de validação estão no intervalo proposto pela equação 7.4 para uma determinada propriedade, então as previsões do modelo concordam com o método de referência como um segundo laboratório, repetindo o método de referência.

Verifica-se que os resultados de previsão dos modelos SVM para o ponto de fulgor e para o número de cetano podem ser considerados como concordantes com os resultados dos métodos de referência, conforme mostrado na tabela 7.9, enquanto o modelo para o ponto de fulgor com utilização do PLS tem um nível de concordância bem inferior.

Tabela 7.9 – Percentual dos valores de referência que estão no intervalo estabelecido pelo método ASTM E 1655-05 para os modelos PLS e SVM

Parâmetro	Modelo PLS (%)	Modelo SVM (%)
Ponto de fulgor	81,2	95,0
Número de cetano	100	100

7.3 – Conclusões

Com os modelos desenvolvidos com SVM, os valores de RMSEP obtiveram uma melhora de 47% e 21% para o ponto de fulgor e número de cetano, respectivamente, em relação aos modelos desenvolvidos com PLS, e todos os valores de RMSEP são menores do que os valores de reprodutibilidade estabelecidos pelos métodos de referência.

O desenvolvimento de modelos de calibração com SVM proporcionou resultados que possibilitam a aplicação dos mesmos como ferramenta em Tecnologia Analítica de Processos – PAT e controle de qualidade, devido a sua habilidade para modelar relações não lineares e seu elevador poder de generalização.

Com o desenvolvimento de modelos de calibração mais eficazes utilizando o algoritmo SVM e dados de espectroscopia NIR, para determinação *on line* de parâmetros para utilização pelo sistema otimizador do misturador em linha torna-se possível obter um melhor aproveitamento das correntes disponíveis na refinaria de modo a contribuir para a otimização da produção. Esse desenvolvimento é importante porque proporciona uma melhora na produtividade de óleo diesel, reduzindo a importação desse derivado.

8 – Determinação de parâmetros de qualidade em óleo diesel utilizando espectroscopia NIR e SVM para o monitoramento da carga no processo de HDT

Esse trabalho demonstra a utilização do algoritmo SVM aplicado a dados de espectroscopia NIR do óleo diesel que é carga do processo de hidrotratamento na refinaria de petróleo, para obtenção de modelos de calibração mais eficazes para determinação dos parâmetros ponto de anilina, índice de cetano, PIE, T50, T85, T90 e densidade, que se tornam uma importante ferramenta de Tecnologia Analítica de Processos – PAT, substituindo os métodos tradicionais, que demandam excessivo tempo de ensaio e com custo relativamente elevado.

8.1 – Parte experimental

Foram utilizados dados de espectroscopia NIR de amostras de óleo diesel para as quais realizou-se também os ensaios através dos métodos de referência. Todos os dados foram obtidos pelo setor de otimização e qualidade de produto da refinaria de Paulinia/SP – Replan. Para obtenção dos espectros na região do infravermelho foi utilizado um espectrômetro ABB/Bomen MID/NIR com fonte glowbar (carbeto de silício), detector de sulfato de triglicina deuterada (DTGS), usando uma cubeta de transmitância de CaF_2 de caminho óptico igual a 0,5 mm. Cada espectro foi obtido como uma média de 32 varreduras na região espectral utilizada, com resolução de 4 cm^{-1}.

Obtiveram-se espectros na região do infravermelho próximo para 88 amostras de diesel, carga da unidade de hidrotratamento (HDT), a fim de monitorar propriedades que são variáveis importantes do referido processo. A figura 8.1 ilustra os espectros das 88 amostras utilizadas para obtenção dos conjuntos de calibração e validação dos modelos para os parâmetros: ponto de anilina, índice de cetano, temperatura de destilação para PIE, 50%, 85% e 90% recuperados e densidade.

Foram realizados diferentes pré-processamentos dos dados para verificar qual proporciona a construção do melhor modelo utilizando os algoritmos PLS e SVM. Os pré-processamentos testados foram: correção de linha base WLS e centragem na média; SNV; primeira derivada (janela com 15 pontos); e correção de linha base WLS.

Obtiveram-se modelos utilizando 60 amostras de calibração e 28 amostras para validação, sendo que o algoritmo Kennard-Stone[84] foi aplicado para obtenção desses conjuntos, após o pré-processamento dos dados. Como o algoritmo Kennard-Stone utiliza a comparação entre os espectros para obtenção dos dois conjuntos, diferentes conjuntos de calibração e validação podem ser obtidos em função de diferentes pré-processamentos utilizados.

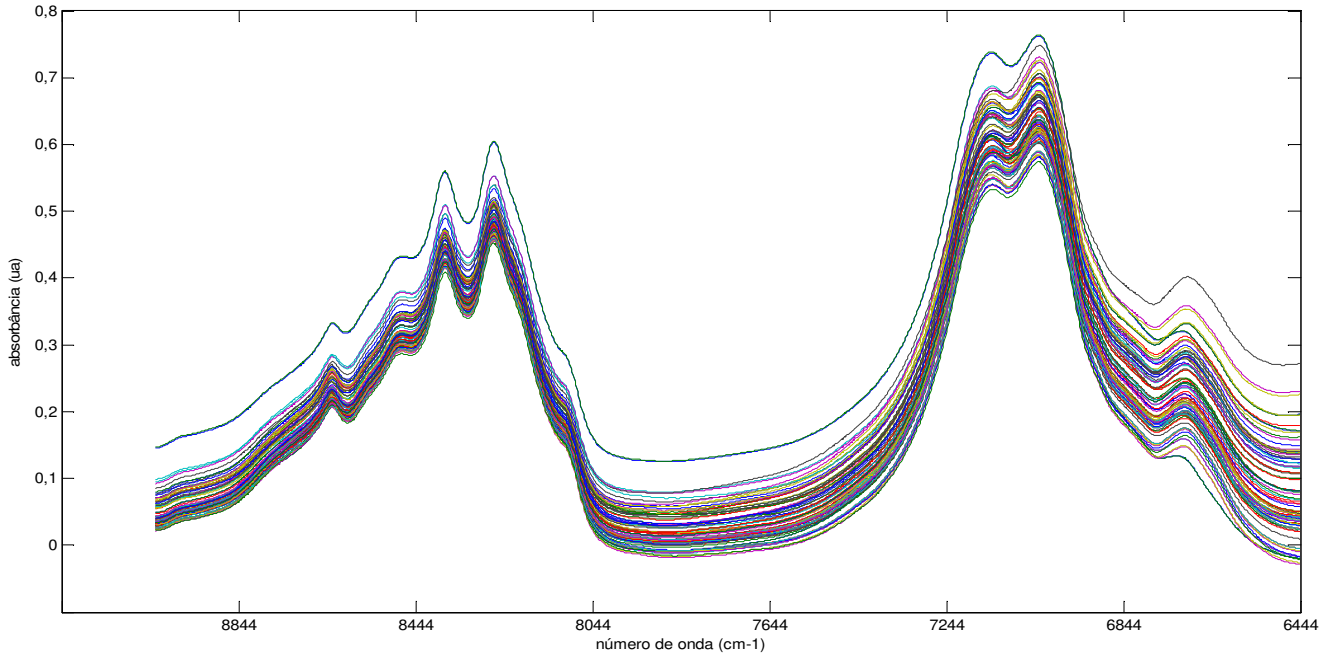


Figura 8.1 – Espectros das 88 amostras utilizadas nos conjunto de calibração e validação dos modelos.

O pacote LIBSVM[85] versão 2.88 foi utilizado para o desenvolvimento dos modelos com SVM e o algoritmo genético foi aplicado para realização da otimização paramétrica. Todos os programas são adequados para utilização com Matlab 7.7 da Mathworks.

Para obtenção dos modelos de regressão com SVM foram testadas as performances de diferentes funções kernel, tais como: RBF, polinomial, sigmoidal e linear.

Para construção dos modelos com SVM os blocos de dados **X** e **y** dos conjuntos de calibração e validação foram previamente escalonados entre [0,1]. Foi utilizado como parâmetro γ do kernel RBF o valor *default* do pacote LIBSVM ($\gamma = 1/k$, onde k é o número de atributos ou variáveis nos dados de entrada) e o grau do polinômio no kernel polinomial igual a 3. Os parâmetros C e ν foram selecionados entre os intervalos de 0 a 10^4 e 10^{-4} a 1, respectivamente, utilizando-se algoritmo genético (GA). Ainda, para otimização paramétrica com GA estipulou-se a utilização de 30 indivíduos e um máximo de 20 gerações, pois observou-se que com essa configuração o valor do erro de validação cruzada se estabilizava, não havendo melhora com o aumento do número de gerações. A função objetivo a ser otimizada pelo GA, e definido inicialmente, consistiu simplesmente na utilização dos valores obtidos pela validação cruzada com 3 subconjuntos do conjunto de treinamento, buscando-se o menor valor do erro de validação cruzada. Os parâmetros a serem otimizados foram tratados como genes no GA.

Como a minimização do erro de validação cruzada no conjunto de treinamento não garante a obtenção dos parâmetros ótimos, eventualmente um *grid search* manual pode ser necessário, a partir dos valores previamente selecionados pelo GA, para refinamento do resultado, sempre considerando a utilização do adequado número de vetores de suporte e valores próximos de RMSEC e RMSEP de modo a evitar um sobreajuste do modelo.

Para obtenção dos modelos nesse trabalho foi testada a utilização de dois intervalos espectrais: (i) de 6444 cm^{-1} a 8936 cm^{-1}, que corresponde a região do primeiro sobreton dos grupos N-H e O-H, sobreton de combinação do grupo C-H e segundo sobreton do grupo C-H; e (ii) de 8044 cm^{-1} a 8444 cm^{-1} , que corresponde somente a região do segundo sobreton do grupo C-H.

8.1.1 – Procedimento experimental para obtenção dos valores de referência

As 88 amostras utilizadas foram analisadas para cada parâmetro em estudo segundo os métodos das normas ASTM de referência citados na tabela 8.22.

O espaço amostral utilizado para cada parâmetro em estudo, bem como os valores de reprodutibilidade especificados pelas respectivas normas ASTM são também mostrados na tabela 8.22. O espaço amostral utilizado para cada parâmetro permite cobrir toda variabilidade encontrada nas condições do processo produtivo da refinaria, conforme preconiza a norma ASTM E 1655.

Figura 8.2 – Analisador automático para determinação do ponto de anilina

Para a determinação do ponto de anilina do óleo diesel foi utilizado o analisador automático AP611 da ISL, mostrado na figura 8.2.

8.2 – Resultados e discussão

Os melhores modelos para os sete parâmetros citados foram obtidos com a utilização do intervalo espectral de 6444 cm^{-1} a 8936 cm^{-1}.

Bandas do primeiro sobreton dos grupos N-H e O-H e de sobreton de combinação e segundo sobreton C-H dos grupos metil e metileno de compostos

parafínicos, naftênicos e aromáticos são observadas nesse intervalo. As atribuições de algumas bandas nessa região espectral foram apresentadas na tabela 2.1.

As frações de tais compostos presentes nas amostras são responsáveis pelos valores dos parâmetros determinados, como segue:

- ponto de anilina: quanto maior a fração de compostos parafínicos na amostra, maior será o valor do ponto de anilina, ou, quanto maior a fração de compostos aromáticos na amostra, menor será o valor desse parâmetro.
- índice de cetano: os compostos parafínicos aumentam os valores do número de cetano (e do índice de cetano) enquanto que as cadeias carbônicas ramificadas e compostos cíclicos e aromáticos diminuem esses valores.
- densidade e temperatura de destilação: verifica-se que para o mesmo número de átomos de carbono, os compostos aromáticos apresentam maiores valores de densidade e de temperatura de ebulição e os parafínicos são os que apresentam os menores valores para essas propriedades. Assim, quanto maior a presença de compostos cíclicos e aromáticos maiores serão os valores desses parâmetros.

Para comparar se há diferença significante entre os modelos SVM e PLS foi realizado um teste-F para todos os parâmetros estudados. A validação quanto a concordância entre os resultados obtidos pelos métodos de referência e pelos modelos de calibração desenvolvidos foi realizado para todos os parâmetros estudados conforme o procedimento descrito pelo método ASTM E 1655-05.

8.2.1 – Modelos de regressão para o ponto de anilina

Foram realizados ensaios utilizando o método de referência ASTM D611-07 para determinação do ponto de anilina e obtidos espectros na região do infravermelho próximo para as 88 amostras.

8.2.1.1 – Modelo PLS

O melhor resultado foi obtido utilizando-se como pré-processamento dos dados a primeira derivada. Utilizou-se 5 variáveis latentes, que explicam 99,9% da variância dos dados. Os resultados são mostrados na tabela 8.1.

Tabela 8.1 – Resultados do melhor modelo de calibração para o ponto de anilina obtido com PLS

Pré-processamento	RMSEC (ºC)	RMSEP (ºC)	R^2
primeira derivada	1,00	1,02	0,758

A figura 8.3 ilustra o resultado do modelo ajustado com o algoritmo PLS e com as 60 amostras de calibração e as 28 amostras de validação

Considerando que a reprodutibilidade especificada pela norma ASTM D 611é de 0,5 ºC o valor obtido para o RMSEP com o modelo PLS desenvolvido não é considerado satisfatório. Dessa forma, novos modelos de previsão foram propostos, como mostrado a seguir.

8.2.1.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a função kernel polinomial e os dados com correção da linha base WLS e centrados na média. A figura 8.4 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 8.2. Também são mostrados os resultados obtidos com o pré-processamento da primeira derivada aplicada aos dados, que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

As funções kernel RBF e sigmoidal também proporcionam a obtenção de bons modelos de calibração, com valor de RMSEP melhor e igual, respectivamente, ao obtido com o modelo PLS. Por outro lado, a função kernel linear, não proporciona a obtenção de um modelo de calibração para os dados utilizados, pois fica evidente a ocorrência de sobreajuste do modelo.

O melhor modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 53 % melhor em relação ao valor obtido com o modelo PLS. Esse valor, de 0,54 ºC, torna o modelo construído muito útil para utilização no monitoramento e controle da carga do processo de hidrotratamento, uma vez que atinge o valor especificado para a reprodutibilidade da análise do ponto de anilina pelo método de referência.

O bom ajuste dos modelos pode ser verificado ao comparar-se os valores de RMSEC e RMSEP obtidos para um modelo específico, os quais não diferem significativamente, indicando não haver sobreajuste do mesmo. Além disso, há também uma significativa melhora nos valores dos coeficientes de determinação dos modelos com SVM.

Também, para o melhor modelo SVM obtido, o número de vetores de suporte utilizado não é considerado excessivo, pois verificou-se que a diferença entre os valores de RMSEC e RMSEP, manteve-se praticamente constante, após a utilização de dois terços das amostras de calibração, o que pode ser considerado uma indicação de bom ajuste do modelo, sem ocorrência de sobreajuste.

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 8.3 e 8.4, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 8.5 e 8.6, respectivamente. Verifica-se que o modelo SVM possibilita uma modelagem sensivelmente melhor nesse espaço amostral, com a redução dos resíduos de previsão.

A tabela 8.3 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 8.22 mostra uma síntese dos resultados obtidos com SVM e PLS e os valores de referência estabelecidos pelos métodos ASTM.

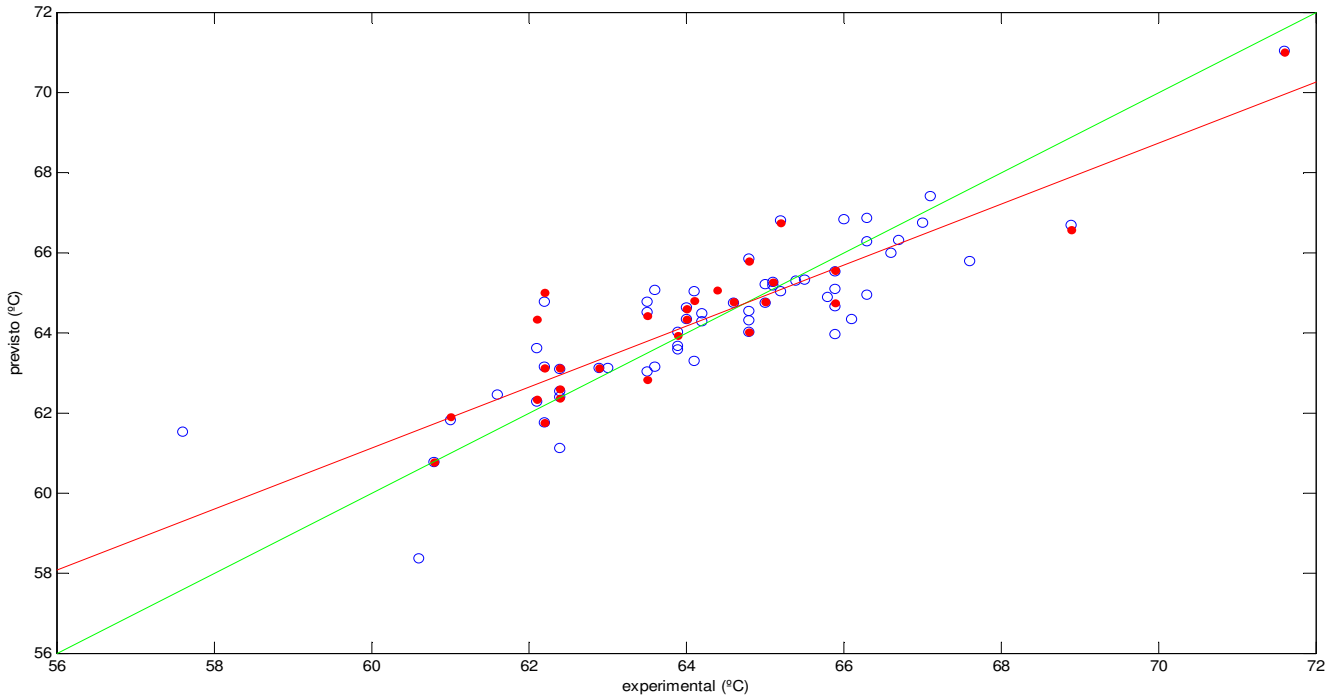


Figura 8.3 – Valores experimentais contra previstos para o modelo PLS para o ponto de anilina com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

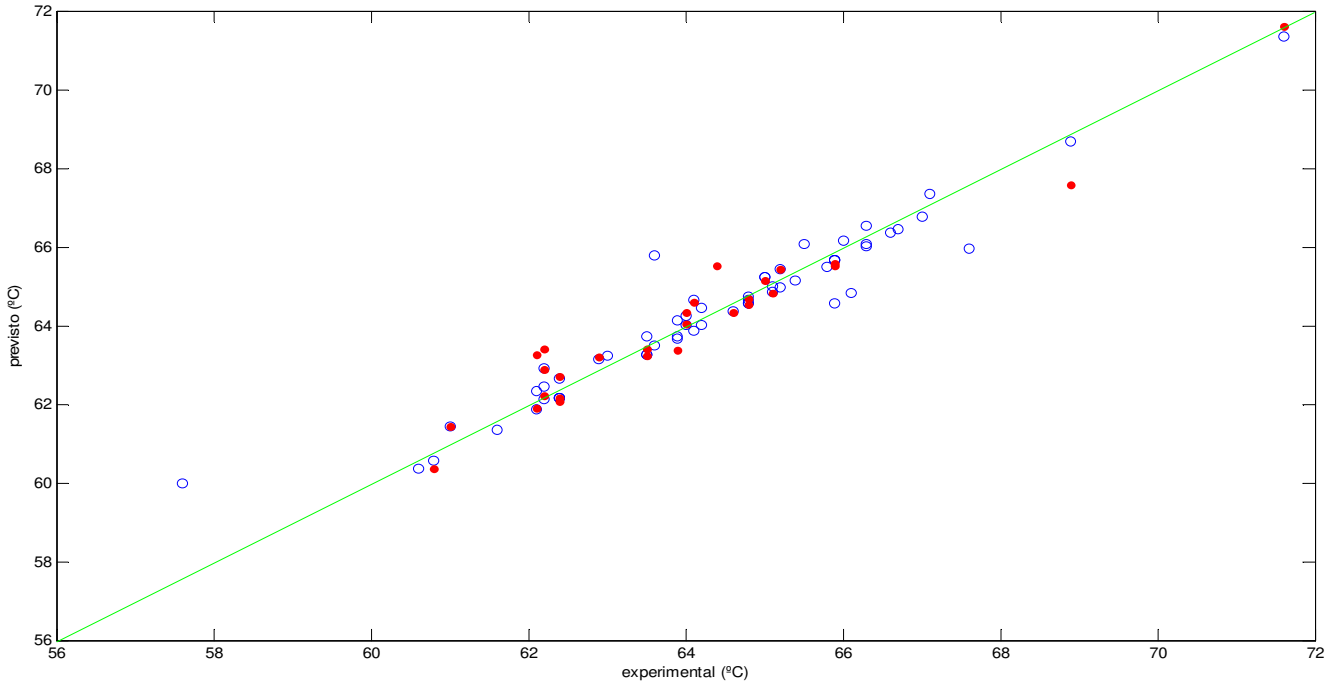
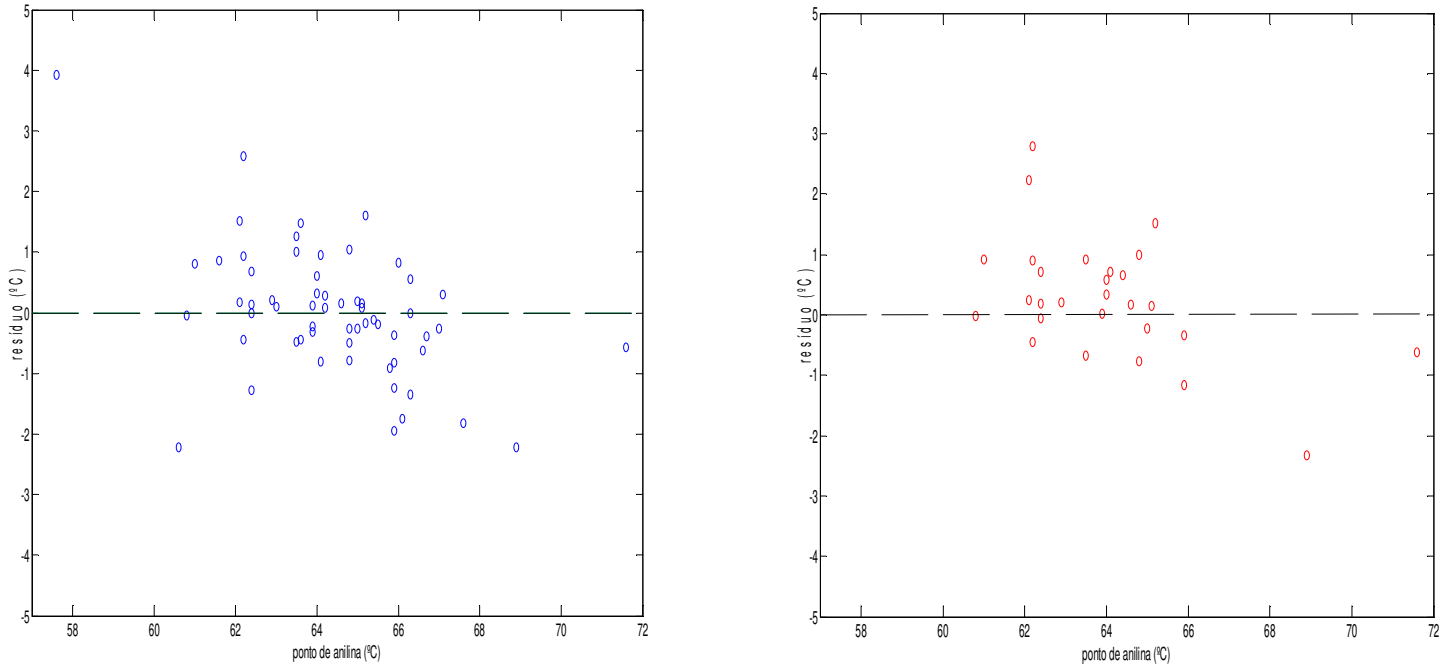


Figura 8.4 – Valores experimentais contra previstos para o modelo SVM para o ponto de anilina com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

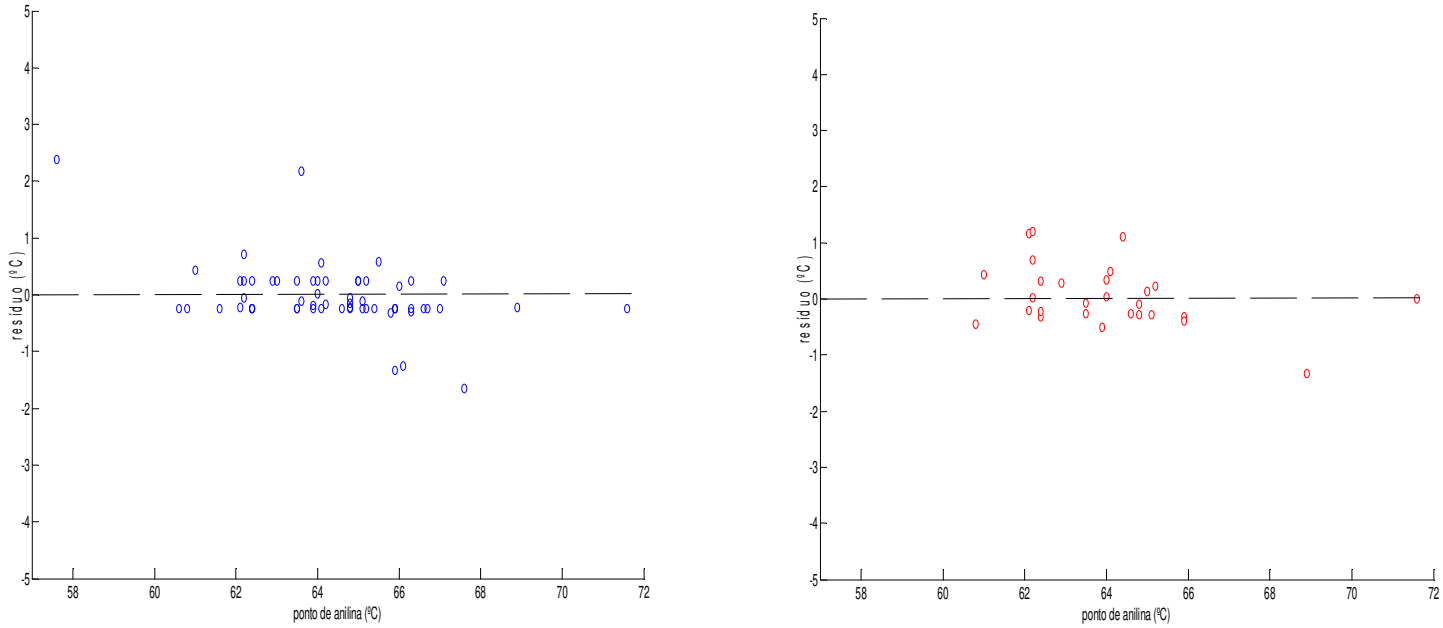
Tabela 8.2 – Modelos de calibração para o ponto de anilina obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento primeira derivada SavGol							
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				
		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	SV's		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	SV's	
RBF	primeira derivada	500,0	0,0120	0,73	0,88	0,8814	24	primeira derivada	500,0	0,0120	0,73	0,88	0,8814	24	
Polinomial	Correção de linha base e centragem na média	2000,0	0,4150	0,58	0,54	0,9346	49		1000,0	0,0229	0,95	1,10	0,8087	20	
Sigmoidal	primeira derivada	22,0	0,1357	1,00	1,02	0,7913	14		22,0	0,1357	1,00	1,02	0,7913	14	
Linear	-	-	-	-	-	-	-		-	-	-	-	-	-	



(a) (b)

Figura 8.5 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS para o ponto de anilina



(a) (b)

Figura 8.6 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM para o ponto de anilina

Tabela 8.3 – resultados de previsão dos modelos PLS e SVM para o ponto de anilina

	PLS			SVM		
	Medido (ºC)	Previsto (ºC)	erro relativo (%)	Medido (ºC)	Previsto (ºC)	erro relativo (%)
1.	60,80	60,77	-0,04	60,80	60,35	-0,74
2.	61,00	61,91	1,49	61,00	61,43	0,71
3.	62,10	64,33	3,59	62,10	63,27	1,88
4.	62,10	62,34	0,38	62,10	61,89	-0,33
5.	62,20	61,76	-0,71	62,20	62,21	0,01
6.	62,20	64,99	4,49	62,20	63,40	1,93
7.	62,20	63,11	1,46	62,20	62,90	1,12
8.	62,40	62,35	-0,08	62,40	62,71	0,49
9.	62,40	63,11	1,14	62,40	62,08	-0,51
10.	62,40	62,58	0,30	62,40	62,17	-0,36
11.	62,90	63,11	0,34	62,90	63,19	0,46
12.	63,50	62,82	-1,08	63,50	63,24	-0,42
13.	63,50	64,42	1,45	63,50	63,42	-0,13
14.	63,90	63,92	0,03	63,90	63,39	-0,80
15.	64,00	64,59	0,92	64,00	64,33	0,52
16.	64,00	64,34	0,53	64,00	64,03	0,05
17.	64,10	64,81	1,11	64,10	64,58	0,75
18.	64,40	65,06	1,02	64,40	65,51	1,72
19.	64,60	64,77	0,26	64,60	64,33	-0,41
20.	64,80	64,03	-1,19	64,80	64,52	-0,42
21.	64,80	65,80	1,54	64,80	64,70	-0,16
22.	65,00	64,77	-0,36	65,00	65,14	0,22
23.	65,10	65,25	0,23	65,10	64,82	-0,44
24.	65,20	66,73	2,34	65,20	65,43	0,35
25.	65,90	64,73	-1,77	65,90	65,59	-0,47
26.	65,90	65,56	-0,51	65,90	65,51	-0,59
27.	68,90	66,57	-3,38	68,90	67,57	-1,94
28.	71,60	70,99	-0,86	71,60	71,60	0,00

8.2.2 – Modelos de regressão para o índice de cetano

Os modelos de regressão para o índice de cetano foram feitos a partir dos valores calculados conforme o método de referência ASTM D976-06 e com a obtenção de espectros na região do infravermelho próximo para 88 amostras.

8.2.2.1 – Modelo PLS

Obteve-se o modelo que fornece melhores resultados utilizando os dados pré-processados com correção de linha base WLS e centrados na média. Utilizou-se 6 variáveis latentes, que explicam 99,5 % da variância dos dados. Os resultados são mostrados na tabela 8.4.

Tabela 8.4 – Resultados do melhor modelo de calibração para o índice de cetano obtido com PLS

Pré-processamento	RMSEC	RMSEP	R^2
correção de linha base e centragem na média	1,014	1,148	0,774

A figura 8.7 ilustra o resultado do modelo ajustado com o algoritmo PLS e com as 60 amostras de calibração e as 28 amostras de validação.

A precisão especificada pela norma ASTM D 976 é de que exista uma correlação entre o índice de cetano calculado e o número de cetano de no mínimo ± 2 número de cetano em ao menos 75% das amostras. Dessa forma, um modelo de calibração para aquele parâmetro deve ter um valor de RMSEP abaixo de 2 número de cetano.

Considerando esse argumento, o modelo PLS obtido é aceitável para determinação do índice de cetano, no entanto, em vista da importância desse parâmetro de qualidade do óleo diesel na especificação da carga do processo de

hidrotratamento buscou-se a obtenção de um modelo de calibração com melhor poder de ajuste aos dados, visando diminuir o erro de previsão.

8.2.2.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido com a função kernel polinomial e os dados com correção de linha base WLS e centrados na média. A figura 8.8 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 8.5. Também são mostrados os resultados obtidos para os dados com correção de linha base WLS e centrados na média, pré-processamento que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

Todas as funções kernel utilizadas, com exceção da sigmoidal, proporcionaram resultados com RMSEP melhor do que o obtido com o modelo PLS. Aqui mais uma vez verifica-se a consistência dos modelos obtidos, que utilizam um número não excessivo de vetores de suporte e valores próximos de RMSEC e RMSEP, indicando que não há sobreajuste dos modelos.

O melhor modelo SVM obtido fornece um RMSEP de 0,9066, que é aproximadamente 21% melhor em relação ao valor obtido com o modelo PLS.

Além dos valores de RMSEP e dos gráficos de valores calculados contra previstos das figuras 8.7 e 8.8, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 8.9 e 8.10, respectivamente. Verifica-se que o modelo SVM possibilita uma melhor modelagem nesse espaço amostral, com a redução dos resíduos de previsão.

A tabela 8.6 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 8.22 mostra uma síntese dos resultados obtidos com SVM e PLS e dos valores de referência estabelecidos pelos métodos ASTM.

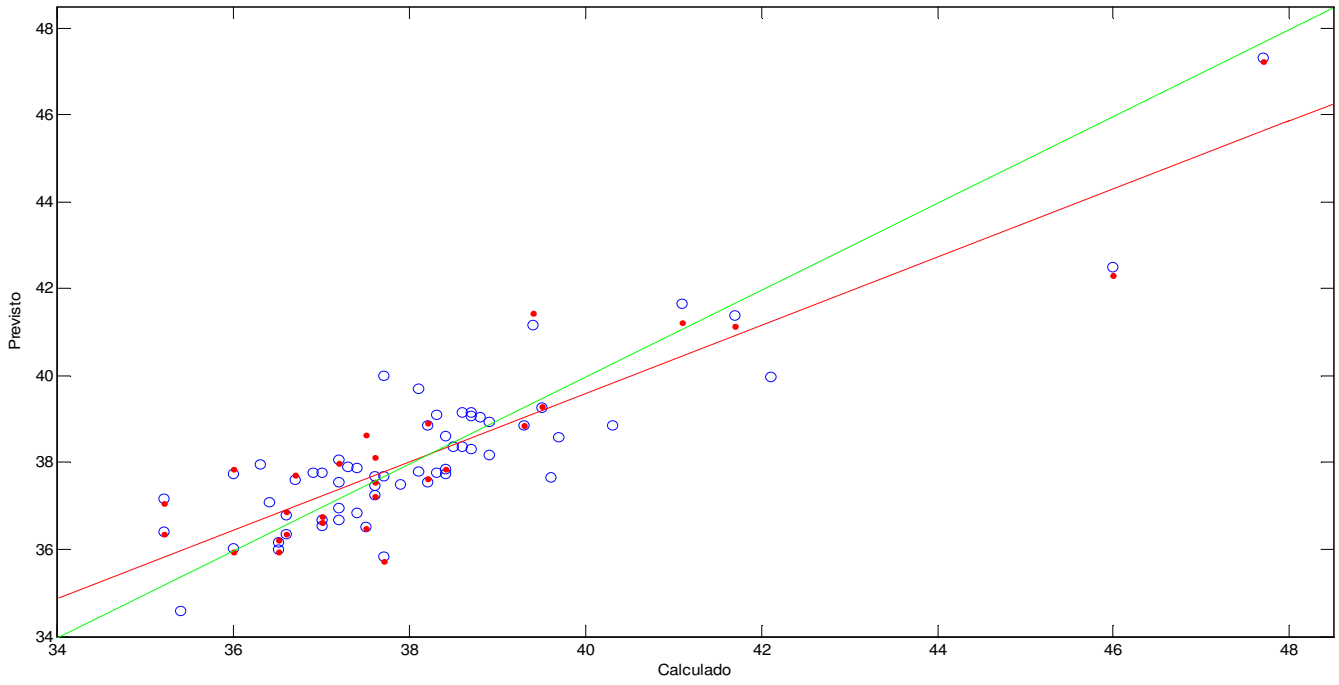


Figura 8.7 – Valores experimentais contra previstos para o modelo PLS para o índice de cetano com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

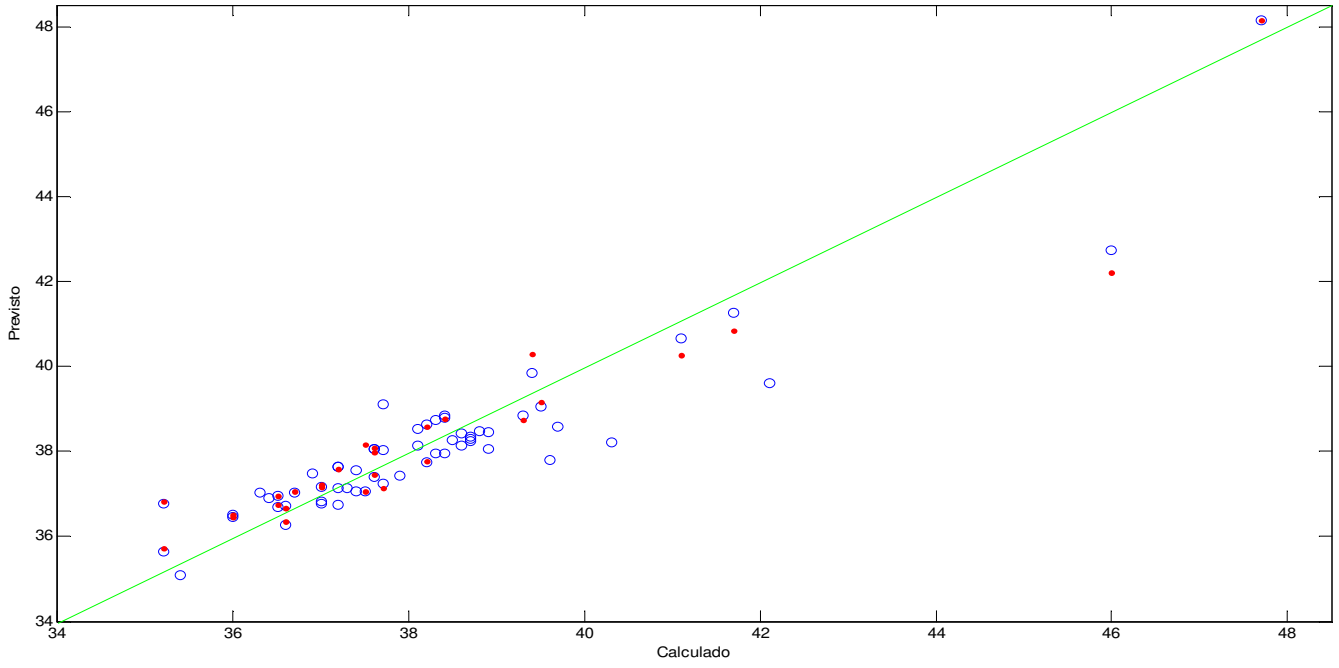
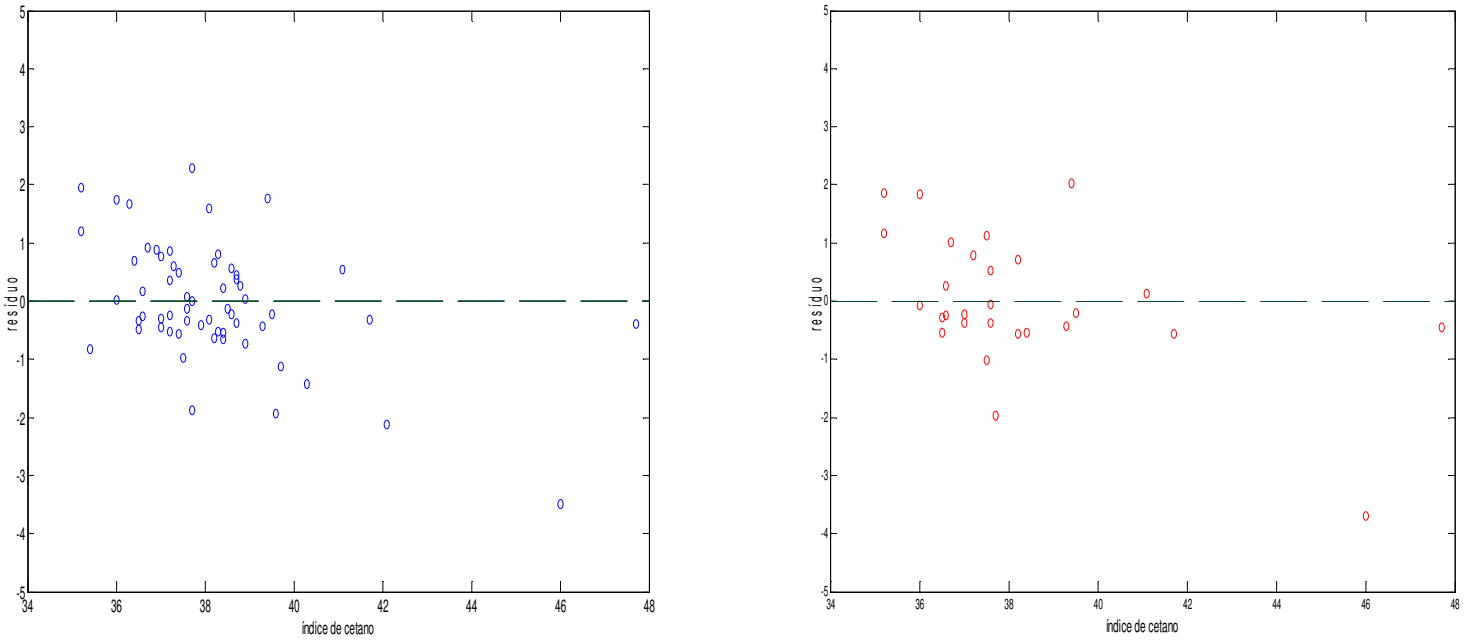


Figura 8.8 – Valores experimentais contra previstos para o modelo SVM para o índice de cetano com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

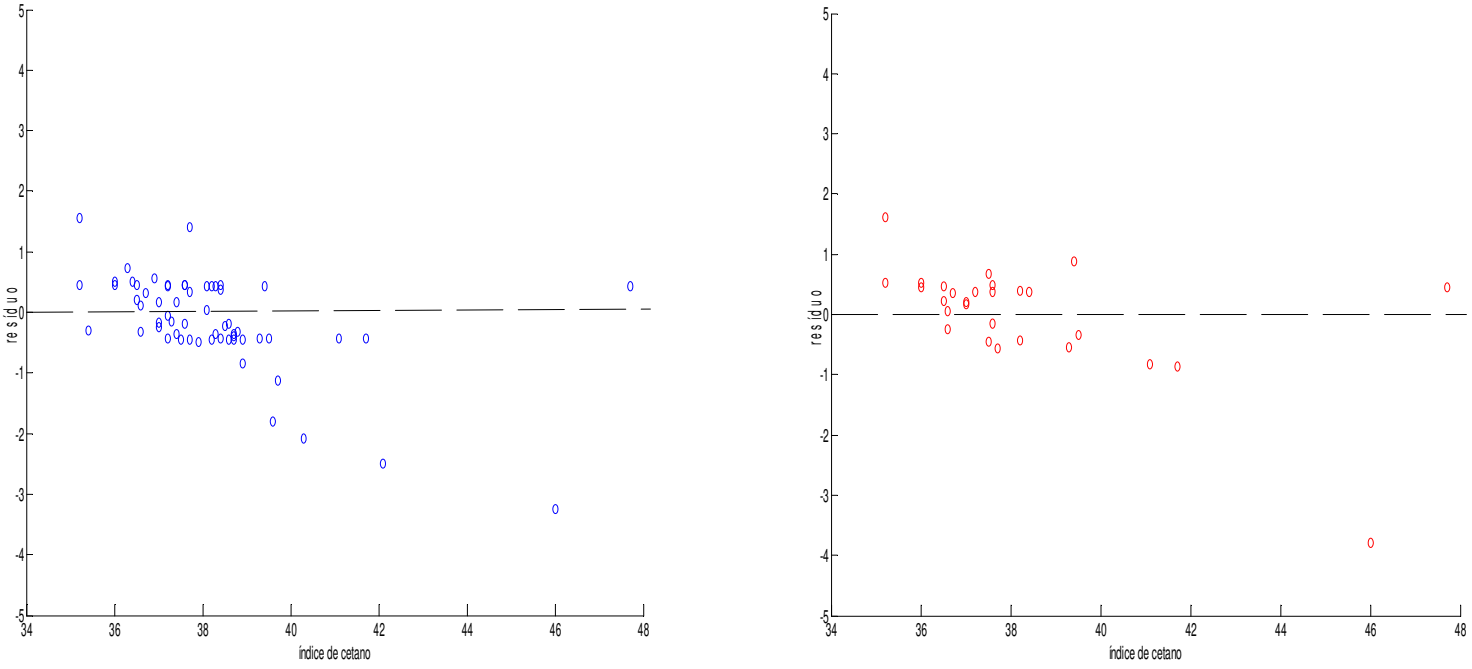
Tabela 8.5 – Modelos de calibração para o índice de cetano obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento correção de linha base e centragem na média						
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM			
		C	ν	RMSEC	RMSEP	R^2	nSV		C	ν	RMSEC	RMSEP	R^2	nSV
RBF	Correção de linha base e centragem na média	60	0,325	0,8845	0,9864	0,8596	33	Correção de linha base e centragem na média	60	0,325	0,8845	0,9864	0,8596	33
Polinomial	Correção de linha base e centragem na média	700	0,35	0,8061	0,9066	0,8840	38		700	0,35	0,8061	0,9066	0,8840	38
Sigmoidal	Correção de linha base e centragem na média	200	0,003	1,4013	1,5105	0,6592	4		200	0,003	1,4013	1,5105	0,6592	4
Linear	Correção de linha base e centragem na média	90	0,003	0,7274	0,9944	0,8846	28		90	0,003	0,7274	0,9944	0,8846	28



(a) (b)

Figura 8.9 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS para o índice de cetano



(a) (b)

Figura 8.10 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM para o índice de cetano

Tabela 8.6 – resultados de previsão dos modelos PLS e SVM para o índice de cetano

	PLS			SVM		
	Calculado (ºC)	Previsto (ºC)	erro relativo (%)	Calculado (ºC)	Previsto (ºC)	erro relativo (%)
1.	35,20	36,36	3,30	35,20	35,72	1,48
2.	35,20	37,06	5,28	35,20	36,81	4,58
3.	36,00	37,84	5,10	36,00	36,45	1,26
4.	36,00	35,93	-0,19	36,00	36,52	1,44
5.	36,50	36,23	-0,75	36,50	36,96	1,25
6.	36,50	35,95	-1,51	36,50	36,73	0,64
7.	36,60	36,86	0,70	36,60	36,35	-0,68
8.	36,60	36,36	-0,67	36,60	36,66	0,17
9.	36,70	37,72	2,78	36,70	37,06	0,98
10.	37,00	36,77	-0,61	37,00	37,16	0,42
11.	37,00	36,62	-1,03	37,00	37,21	0,57
12.	37,20	37,99	2,12	37,20	37,58	1,02
13.	37,50	38,63	3,02	37,50	38,17	1,78
14.	37,50	36,49	-2,71	37,50	37,05	-1,19
15.	37,60	37,22	-1,01	37,60	38,08	1,28
16.	37,60	37,55	-0,13	37,60	37,45	-0,40
17.	37,60	38,12	1,40	37,60	37,98	1,01
18.	37,70	35,73	-5,23	37,70	37,14	-1,48
19.	38,20	37,64	-1,47	38,20	37,76	-1,15
20.	38,20	38,91	1,86	38,20	38,59	1,01
21.	38,40	37,85	-1,42	38,40	38,77	0,96
22.	39,30	38,86	-1,12	39,30	38,75	-1,39
23.	39,40	41,43	5,16	39,40	40,28	2,23
24.	39,50	39,29	-0,54	39,50	39,16	-0,85
25.	41,10	41,23	0,32	41,10	40,27	-2,03
26.	41,70	41,14	-1,35	41,70	40,84	-2,05
27.	46,00	42,30	-8,05	46,00	42,21	-8,23
28.	47,70	47,24	-0,96	47,70	48,15	0,95

8.2.3 – Modelos de regressão para a temperatura de destilação do PIE

Os modelos de regressão para o PIE foram feitos a partir da realização de ensaios pelo método de referência ASTM D86-09 para determinação desse parâmetro e com a obtenção de espectros na região do infravermelho próximo para 88 amostras.

8.2.3.1 – Modelo PLS

Obteve-se o modelo que fornece melhores resultados utilizando os dados com correção da linha base WLS e centrados na média. Utilizou-se 9 variáveis latentes, que explicam 99,8 % da variância dos dados. Os resultados são mostrados na tabela 8.7.

Tabela 8.7 – Resultados do melhor modelo de calibração para o PIE obtido com PLS.

Pré-processamento	RMSEC (ºC)	RMSEP (ºC)	R^2
Correção da linha base e centragem na média	6,583	8,558	0,771

A figura 8.11 ilustra o resultado do modelo ajustado com o algoritmo PLS e com as 60 amostras de calibração e as 28 amostras de validação.

A reprodutibilidade especificada pela norma ASTM D 86 é de 8,5 ºC, o que torna possível o uso desse modelo de calibração tomando a comparação com o RMSEP obtido como parâmetro de decisão.

Mais uma vez devemos considerar a importância desse parâmetro de qualidade do óleo diesel na especificação da carga do processo de hidrotratamento. Assim, buscou-se a obtenção de um modelo de calibração com melhor poder de ajuste aos dados, visando diminuir o erro de previsão.

8.2.3.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a função kernel polinomial e os dados com pré-processamento primeira derivada. A figura 8.12 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 8.8. Também são mostrados os resultados obtidos para os dados com correção de linha base WLS e centrados na média, pré-processamento que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

Todas as funções kernel utilizadas, com exceção da função sigmoidal, proporcionaram resultados com RMSEP melhor do que o obtido com o modelo PLS. A função kernel sigmoidal proporciona modelos com valor de R^2 abaixo de 0,6 e seus resultados não são considerados na tabela 8.8.

Aqui mais uma vez verifica-se a consistência dos modelos obtidos, que utilizam um número não excessivo de vetores de suporte e valores próximos de RMSEC e RMSEP, indicando que não há sobreajuste dos modelos.

O melhor modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 32 % melhor em relação ao valor obtido com o modelo PLS. Esse valor, de 5,78 ºC, torna o modelo construído muito útil para utilização, uma vez que está bem abaixo do valor especificado para a reprodutibilidade da análise do PIE pelo método de referência.

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 8.11 e 8.12, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 8.13 e 8.14, respectivamente. Verifica-se que o modelo SVM possibilita uma modelagem sensivelmente melhor nesse espaço amostral, com a redução dos resíduos de previsão.

A tabela 8.9 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 8.22 mostra uma síntese dos resultados obtidos com SVM e PLS e dos valores de referência estabelecidos pelos métodos ASTM.

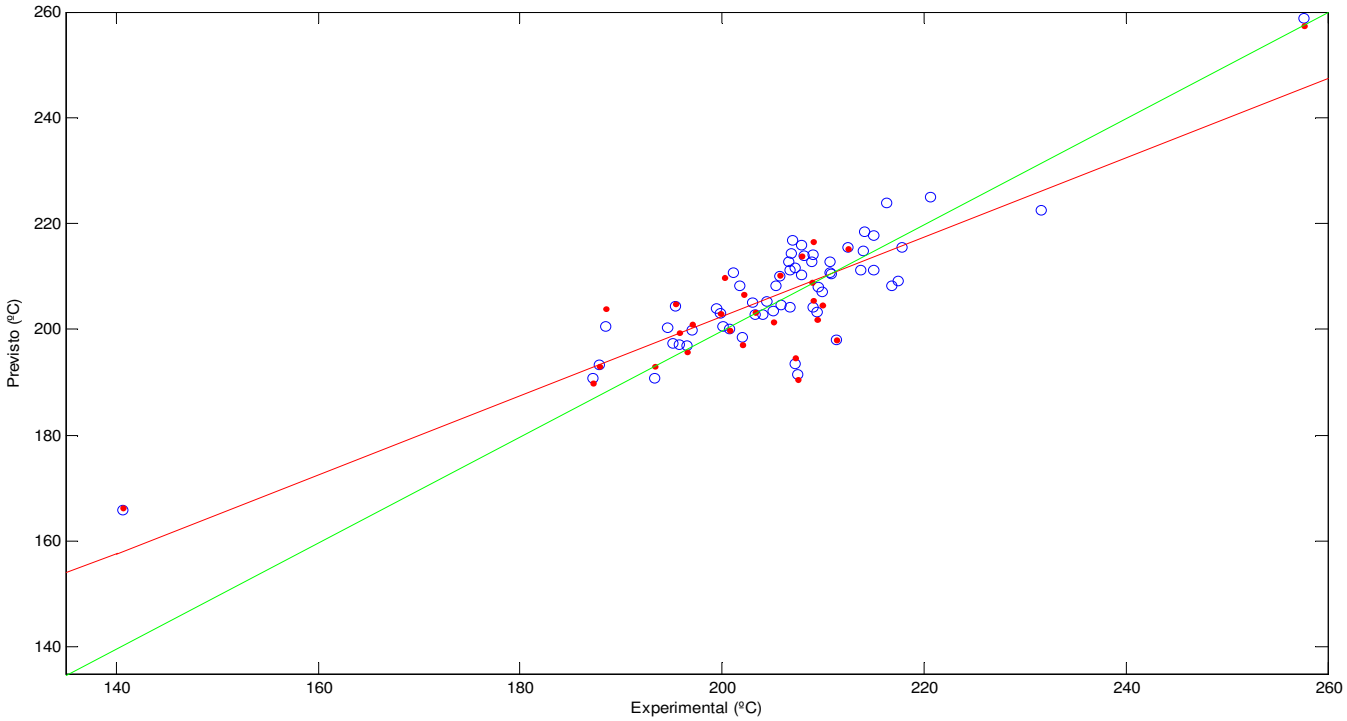


Figura 8.11 – Valores experimentais contra previstos para o modelo PLS para o PIE com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

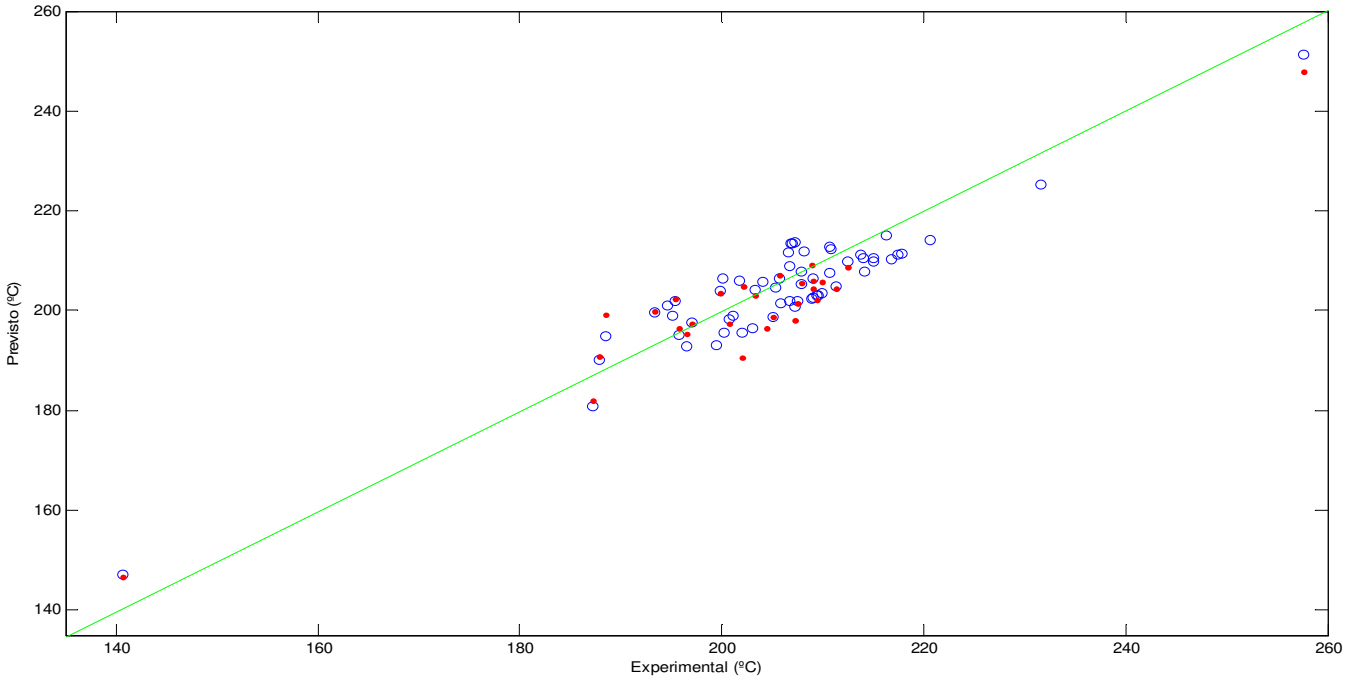
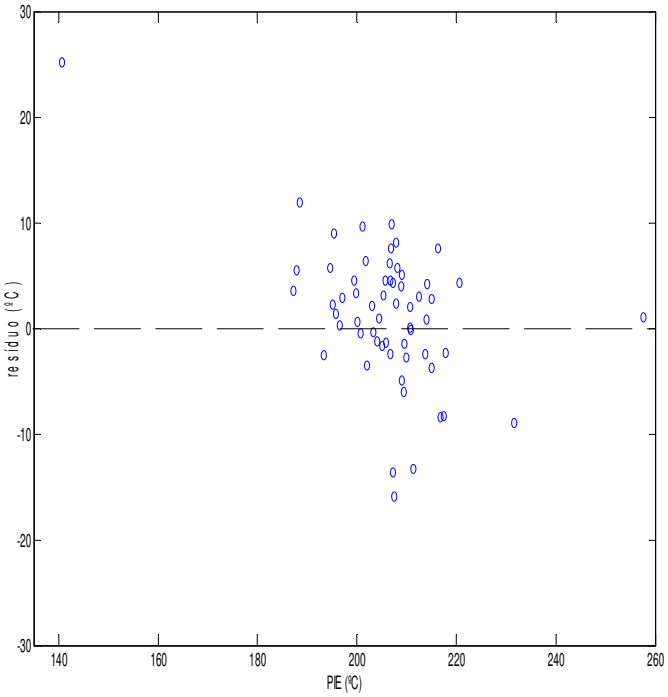


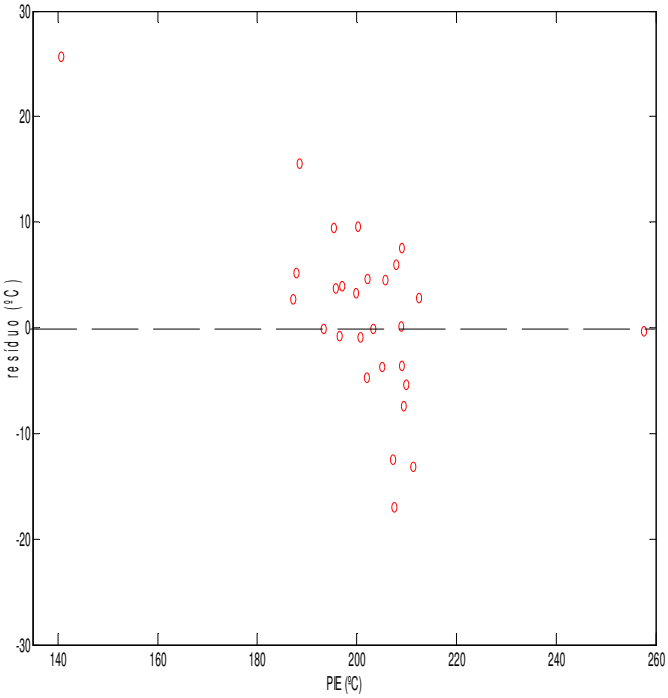
Figura 8.12 – Valores experimentais contra previstos para o modelo SVM para o PIE com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

Tabela 8.8 – Modelos de calibração para o PIE obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento correção da linha base e centragem na média							
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				
		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	nSV		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	nSV	
RBF	SNV	7400	0,0910	4,8315	5,8950	0,8825	36	Correção de linha base e centragem na média	800,0	0,0700	4,8561	6,5794	0,8691	32	
Polinomial	primeira derivada	5990	0,0050	4,9638	5,7851	0,8909	28		5300	0,1240	4,9896	6,1946	0,8717	30	
Sigmoidal	-	-	-	-	-	-	-		-	-	-	-	-	-	
Linear	Correção de linha base e centragem na média	20,00	0,0060	6,4598	8,6892	0,7770	22		20,00	0,0060	6,4598	8,6892	0,7770	22	

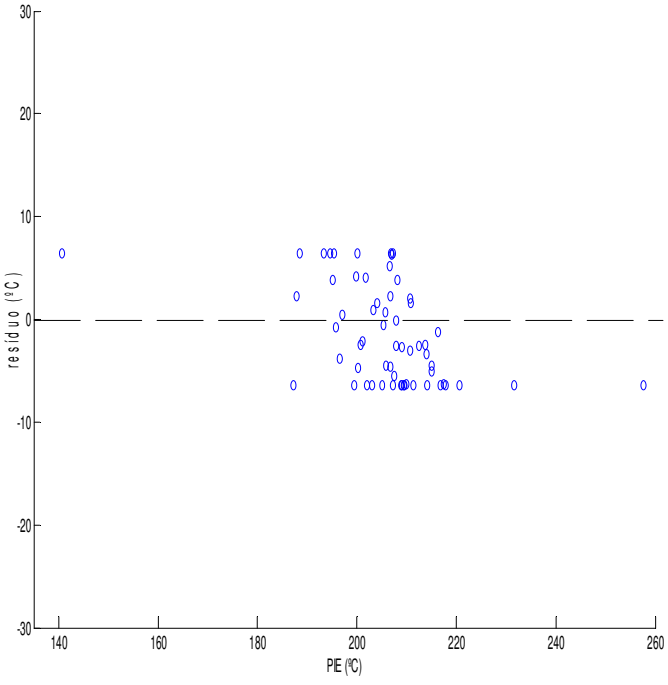


(a)

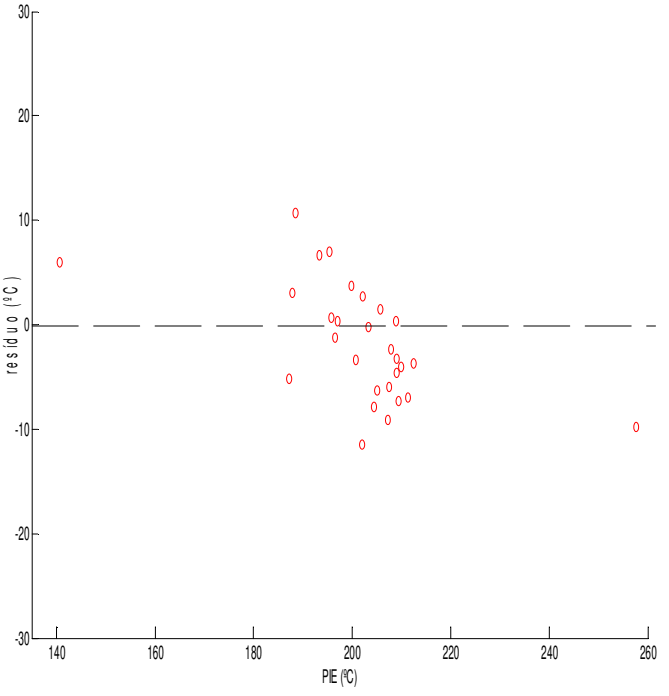


(b)

Figura 8.13 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS para o PIE



(a)



(b)

Figura 8.14 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM para o PIE

Tabela 8.9 – resultados de previsão dos modelos PLS e SVM para o PIE

	PLS			SVM		
	Medido (ºC)	Previsto (ºC)	erro relativo (%)	Medido (ºC)	Previsto (ºC)	erro relativo (%)
1.	140,70	166,37	18,25	140,70	146,67	4,25
2.	187,20	189,94	1,47	187,20	182,00	-2,78
3.	187,90	193,07	2,75	187,90	190,90	1,60
4.	188,50	204,02	8,23	188,50	199,17	5,66
5.	193,30	193,18	-0,06	193,30	199,99	3,46
6.	195,40	204,88	4,85	195,40	202,40	3,58
7.	195,80	199,50	1,89	195,80	196,42	0,31
8.	196,60	195,76	-0,43	196,60	195,35	-0,63
9.	197,10	201,00	1,98	197,10	197,44	0,17
10.	199,80	203,02	1,61	199,80	203,53	1,86
11.	200,30	209,92	4,80	200,70	197,30	-1,69
12.	200,70	199,79	-0,45	202,00	190,52	-5,68
13.	202,00	197,24	-2,36	202,10	204,84	1,36
14.	202,10	206,71	2,28	203,30	203,11	-0,09
15.	203,30	203,20	-0,05	204,40	196,51	-3,86
16.	205,10	201,43	-1,79	205,10	198,78	-3,08
17.	205,70	210,23	2,20	205,70	207,21	0,73
18.	207,20	194,66	-6,05	207,20	198,04	-4,42
19.	207,50	190,55	-8,17	207,50	201,48	-2,90
20.	207,90	213,81	2,84	207,90	205,53	-1,14
21.	208,90	208,96	0,03	208,90	209,24	0,16
22.	209,00	205,45	-1,70	209,00	204,38	-2,21
23.	209,10	216,61	3,59	209,10	205,89	-1,53
24.	209,40	201,92	-3,57	209,40	202,04	-3,52
25.	209,90	204,55	-2,55	209,90	205,87	-1,92
26.	211,30	198,10	-6,25	211,30	204,30	-3,31
27.	212,50	215,27	1,30	212,50	208,74	-1,77
28.	257,60	257,29	-0,12	257,60	247,82	-3,80

8.2.4 – Modelos de regressão para a temperatura de destilação de 50% recuperados

Os modelos de regressão para a T50 foram feitos a partir da realização de ensaios pelo método de referência ASTM D86-09 para determinação desse parâmetro e com a obtenção de espectros na região do infravermelho próximo para 88 amostras.

8.2.4.1 – Modelo PLS

Obteve-se o modelo que fornece os melhores resultados utilizando os dados com correção da linha base WLS e centrados na média. Utilizou-se 8 variáveis latentes, que explicam 99,8 % da variância dos dados. Os resultados são mostrados na tabela 8.10.

Tabela 8.10 – Resultados do melhor modelo de calibração para a T50 obtido com PLS.

Pré-processamento	RMSEC (ºC)	RMSEP (ºC)	R^2
Correção da linha base e centragem na média	5,12	5,41	0,810

A figura 8.15 ilustra o resultado do modelo ajustado com o algoritmo PLS e com as 60 amostras de calibração e as 28 amostras de validação.

A reprodutibilidade especificada pela norma ASTM D 86 é de 3,0 ºC, o que não possibilita o uso desse modelo de calibração tomando a comparação com o RMSEP obtido como parâmetro de decisão.

Assim, buscou-se a obtenção de um modelo de calibração com melhor poder de ajuste aos dados, visando diminuir o erro de previsão.

8.2.4.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a função kernel RBF e os dados com correção de linha base WLS e centrados na média. A figura 8.16 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 8.11. Também são mostrados os resultados obtidos para os dados com correção de linha base WLS e centrados na média, pré-processamento que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

Todas as funções kernel utilizadas, com exceção da função sigmoidal, proporcionaram resultados com RMSEP melhor do que o obtido com o modelo PLS. Aqui mais uma vez verifica-se a consistência dos modelos obtidos, que utilizam um número não excessivo de vetores de suporte e valores próximos de RMSEC e RMSEP, indicando que não há sobreajuste dos modelos.

O melhor modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 54 % melhor em relação ao valor obtido com o modelo PLS. Esse valor, de 2,47 ºC, torna o modelo construído muito útil para utilização, uma vez que está bem abaixo do valor especificado para a reprodutibilidade da análise da T50 pelo método de referência.

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 8.15 e 8.16, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 8.17 e 8.18, respectivamente. Verifica-se que o modelo SVM possibilita uma modelagem sensivelmente melhor nesse espaço amostral, com a redução dos resíduos de previsão.

A tabela 8.12 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 8.22 mostra uma síntese dos resultados obtidos com SVM e PLS e dos valores de referência estabelecidos pelos métodos ASTM.

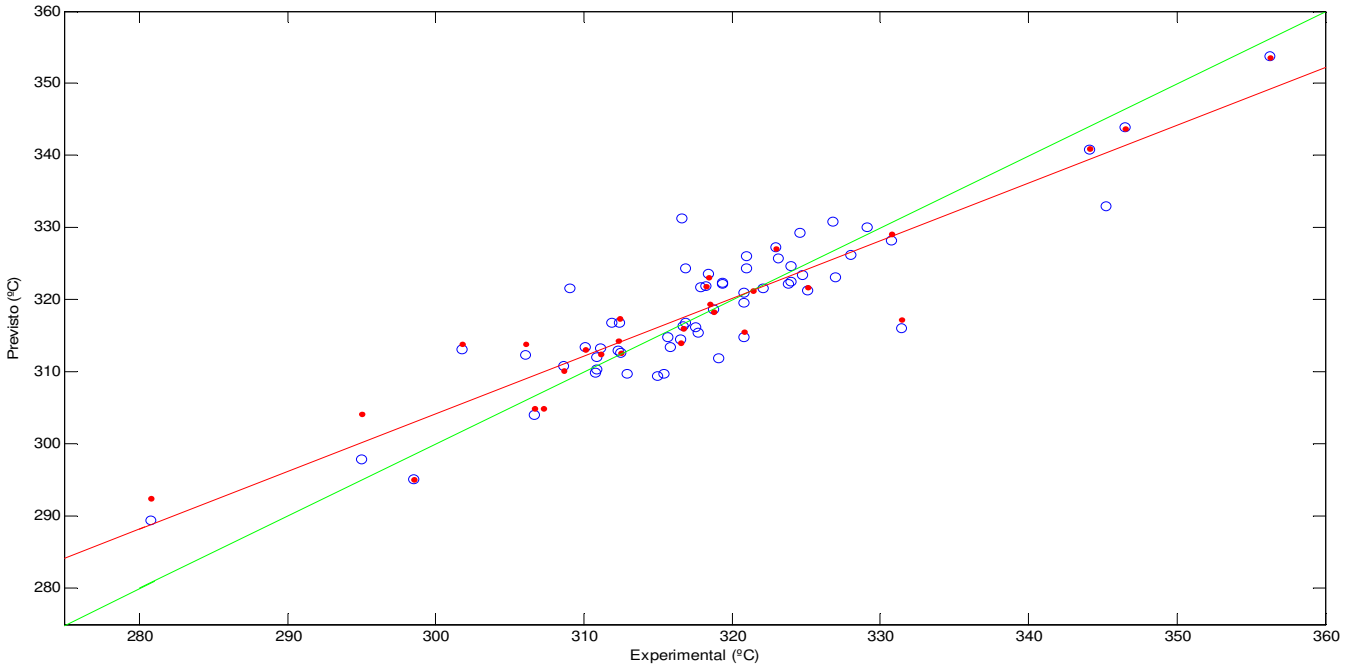


Figura 8.15 – Valores experimentais contra previstos para o modelo PLS para a T50 com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

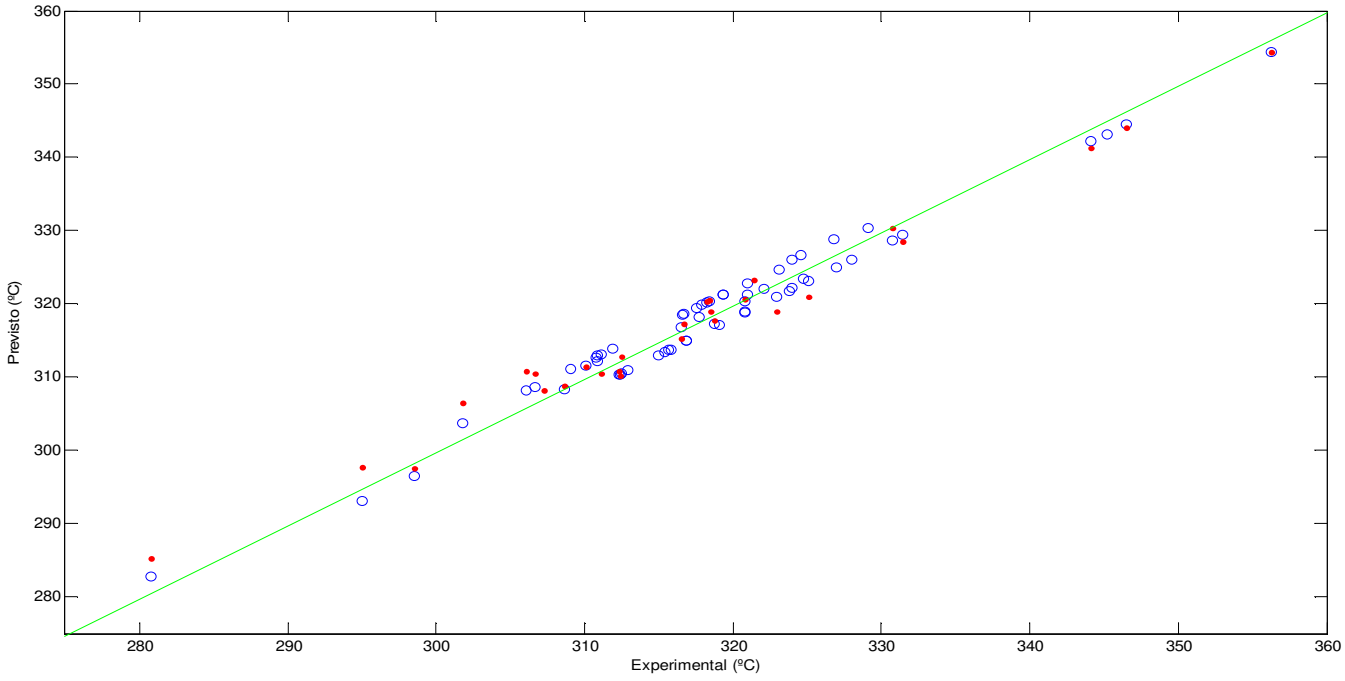
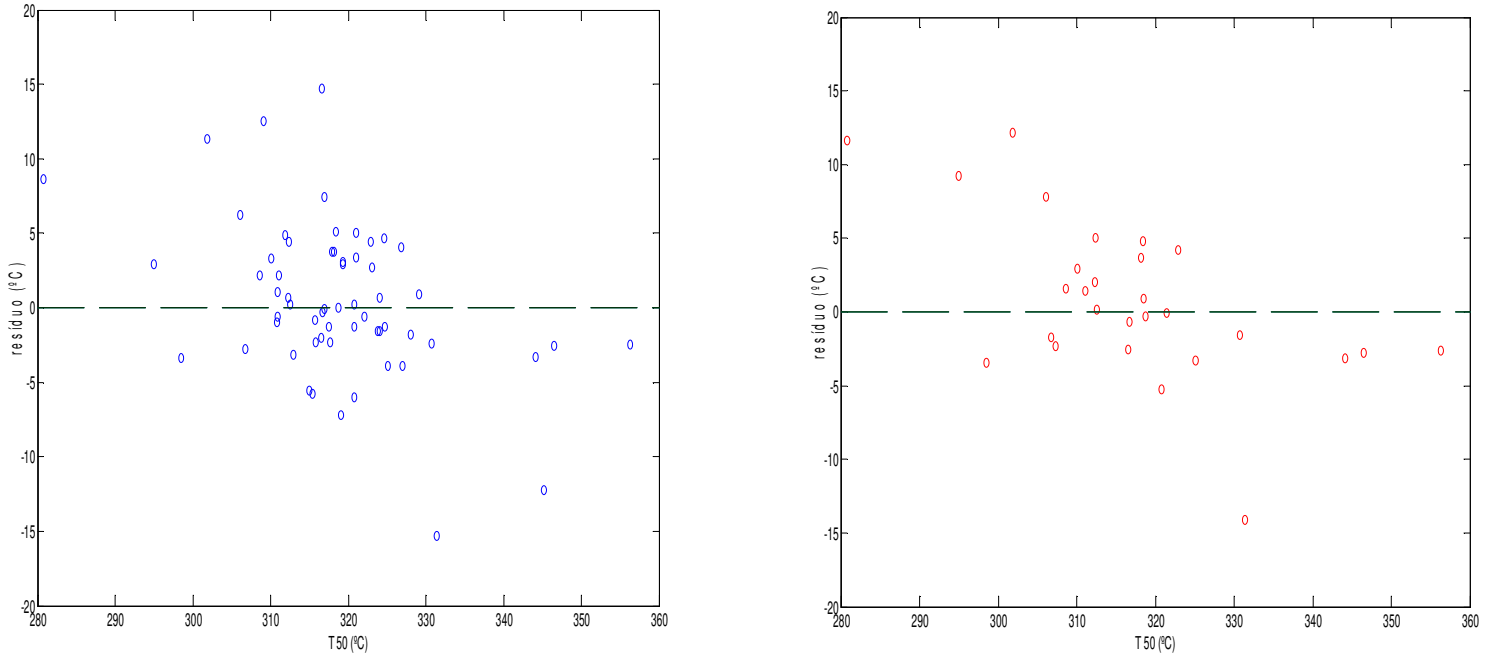


Figura 8.16 – Valores experimentais contra previstos para o modelo SVM para a T50 com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

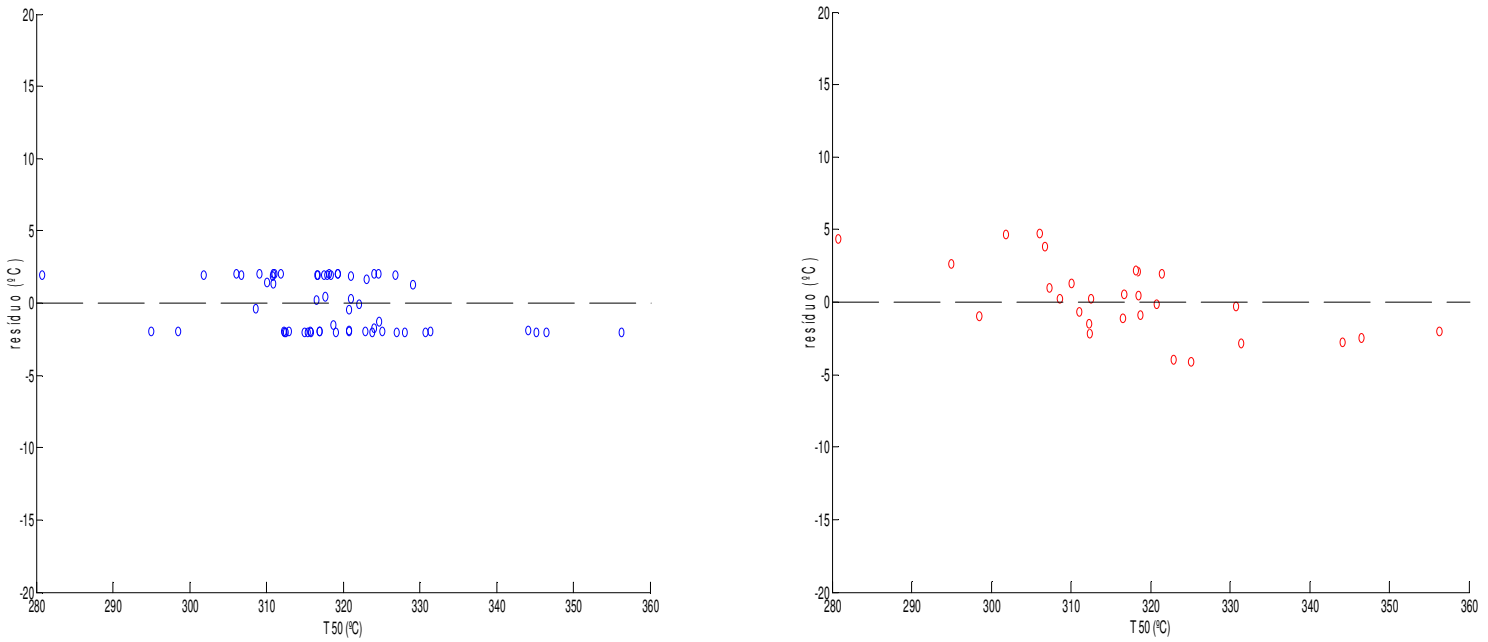
Tabela 8.11 – Modelos de calibração para a T50 obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento correção da linha base e centragem na média						
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM			
		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	nSV		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	nSV
RBF	Correção de linha base e centragem na média	4900	0,0365	1,8243	2,4718	0,9778	42	Correção de linha base e centragem na média	4900	0,0365	1,8243	2,4718	0,9778	42
Polinomial	Correção de linha base e centragem na média	1300	0,009	2,2468	3,0054	0,9666	40		1300	0,009	2,2468	3,0054	0,9666	40
Sigmoidal	primeira derivada	10,75	0,5681	6,8203	8,3337	0,7028	37		2,44	0,6542	9,9965	13,3468	0,3506	42
Linear	SNV	15,00	0,2989	3,8779	3,8149	0,9156	44		12,41	0,0329	3,1103	4,5583	0,9258	30



(a) (b)

Figura 8.17 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS para o T50



(a) (b)

Figura 8.18 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM para o T50

Tabela 8.12 – resultados de previsão dos modelos PLS e SVM para o T50

	PLS			SVM		
	Medido (ºC)	Previsto (ºC)	erro relativo (%)	Medido (ºC)	Previsto (ºC)	erro relativo (%)
1.	280,80	292,45	4,15	280,80	285,18	1,56
2.	295,00	304,21	3,12	295,00	297,66	0,90
3.	298,50	295,06	-1,15	298,50	297,53	-0,33
4.	301,80	313,93	4,02	301,80	306,49	1,55
5.	306,10	313,92	2,55	306,10	310,81	1,54
6.	306,70	304,95	-0,57	306,70	310,55	1,26
7.	307,30	304,98	-0,76	307,30	308,26	0,31
8.	308,60	310,19	0,52	308,60	308,84	0,08
9.	310,10	313,04	0,95	310,10	311,39	0,42
10.	311,10	312,51	0,45	311,10	310,44	-0,21
11.	312,30	314,33	0,65	312,30	310,77	-0,49
12.	312,40	317,42	1,61	312,40	310,21	-0,70
13.	312,50	312,67	0,05	312,50	312,76	0,08
14.	316,50	313,97	-0,80	316,50	315,34	-0,37
15.	316,70	316,04	-0,21	316,70	317,25	0,17
16.	318,20	321,87	1,15	318,20	320,39	0,69
17.	318,40	323,18	1,50	318,40	320,50	0,66
18.	318,50	319,42	0,29	318,50	318,95	0,14
19.	318,70	318,41	-0,09	318,70	317,79	-0,29
20.	320,80	315,52	-1,64	320,80	320,67	-0,04
21.	321,40	321,30	-0,03	321,40	323,32	0,60
22.	322,90	327,11	1,31	322,90	318,96	-1,22
23.	325,10	321,77	-1,02	325,10	320,97	-1,27
24.	330,70	329,10	-0,48	330,70	330,41	-0,09
25.	331,40	317,30	-4,25	331,40	328,54	-0,86
26.	344,10	340,98	-0,91	344,10	341,29	-0,82
27.	346,50	343,69	-0,81	346,50	344,01	-0,72
28.	356,30	353,68	-0,74	356,30	354,31	-0,56

8.2.5 – Modelos de regressão para a temperatura de destilação de 85% recuperados

Os modelos de regressão para a T85 foram feitos a partir da realização de ensaios pelo método de referência ASTM D86-09 para determinação desse parâmetro e com a obtenção de espectros na região do infravermelho próximo para 88 amostras.

8.2.5.1 – Modelo PLS

Obteve-se o modelo que fornece melhores resultados utilizando os dados com correção da linha base WLS e centrados na média. Utilizou-se 7 variáveis latentes, que explicam 99,8 % da variância dos dados. Os resultados são mostrados na tabela 8.13.

Tabela 8.13 – Resultados do melhor modelo de calibração para a T85 obtido com PLS.

Pré-processamento	RMSEC (ºC)	RMSEP (ºC)	R^2
Correção da linha base e centragem na média	4,63	5,40	0,749

A figura 8.19 ilustra o resultado do modelo ajustado com o algoritmo PLS e com as 60 amostras de calibração e as 28 amostras de validação.

A reprodutibilidade especificada pela norma ASTM D 86 é de 5,2 ºC, o que não possibilita o uso desse modelo de calibração tomando a comparação com o RMSEP obtido como parâmetro de decisão.

Assim, buscou-se a obtenção de um modelo de calibração com melhor poder de ajuste aos dados, visando diminuir o erro de previsão.

8.2.5.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a função kernel RBF e os dados com correção de linha base WLS e centrados na média. A figura 8.20 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 8.14. Também são mostrados os resultados obtidos para os dados com correção de linha base WLS e centrados na média, pré-processamento que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

Todas as funções kernel utilizadas, com exceção da função sigmoidal, proporcionaram resultados com RMSEP melhor do que o obtido com o modelo PLS. A função kernel sigmoidal proporciona modelos com valor de R^2 abaixo de 0,6 e seus resultados não são considerados na tabela 8.14.

Aqui mais uma vez verifica-se a consistência dos modelos obtidos, que utilizam um número não excessivo de vetores de suporte e valores próximos de RMSEC e RMSEP, indicando que não há sobreajuste dos modelos.

O melhor modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 28 % melhor em relação ao valor obtido com o modelo PLS. Esse valor, de 3,87 ºC, torna o modelo construído muito útil para utilização, uma vez que está bem abaixo do valor especificado para a reprodutibilidade da análise da T85 pelo método de referência.

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 8.19 e 8.20, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 8.21 e 8.22, respectivamente. Verifica-se que o modelo SVM possibilita uma modelagem sensivelmente melhor nesse espaço amostral, com a redução dos resíduos de previsão.

A tabela 8.15 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 8.22 mostra uma síntese dos resultados obtidos com SVM e PLS e dos valores de referência estabelecidos pelos métodos ASTM.

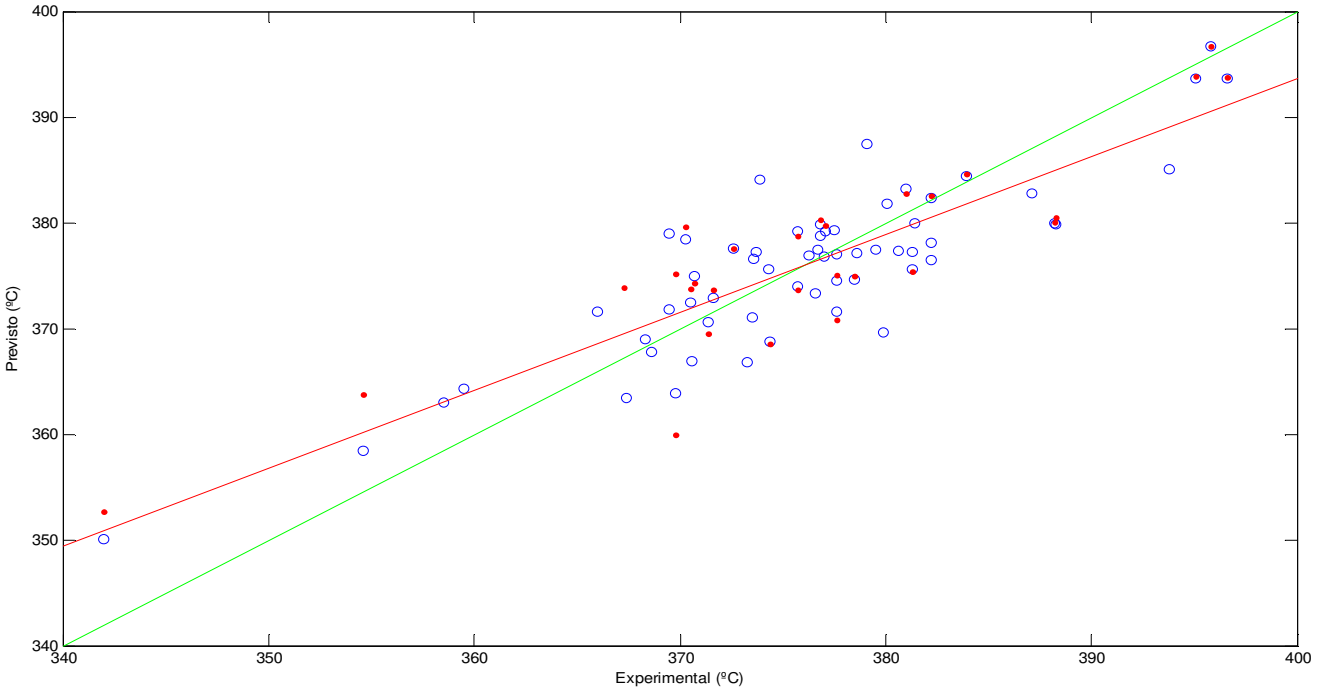


Figura 8.19 – Valores experimentais contra previstos para o modelo PLS para a T85 com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

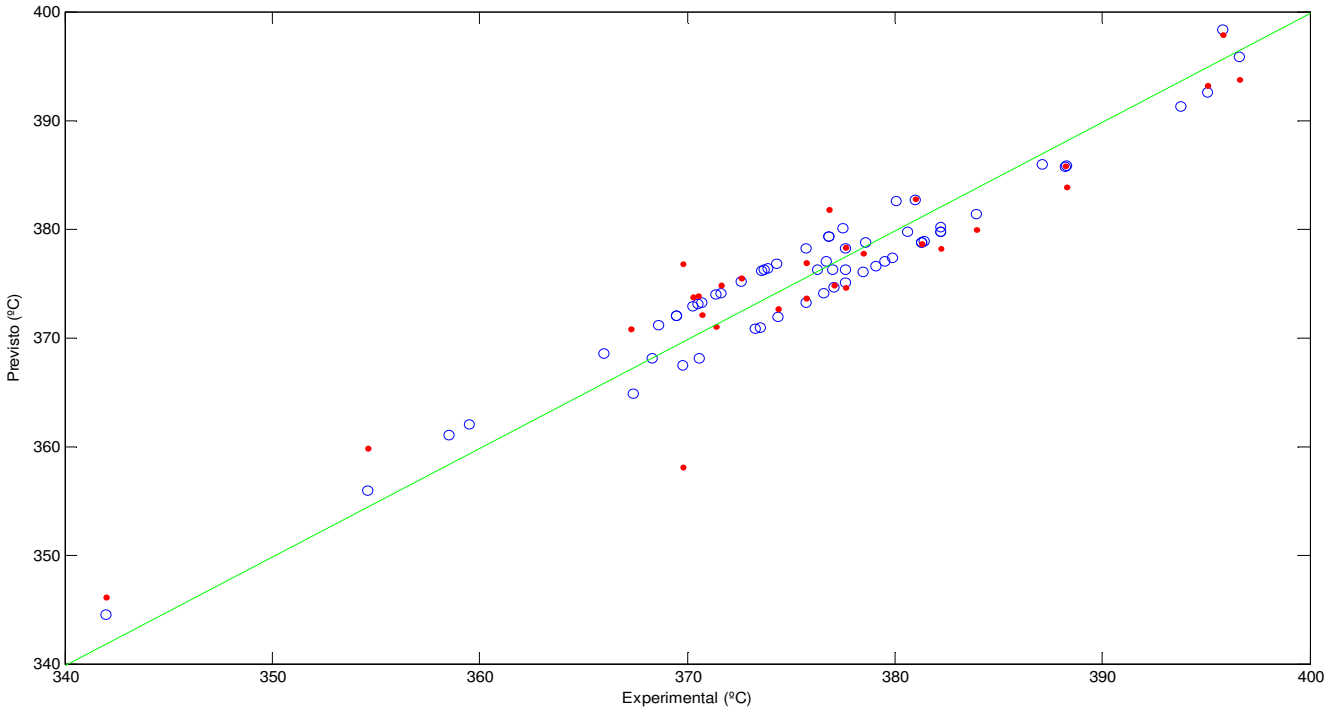
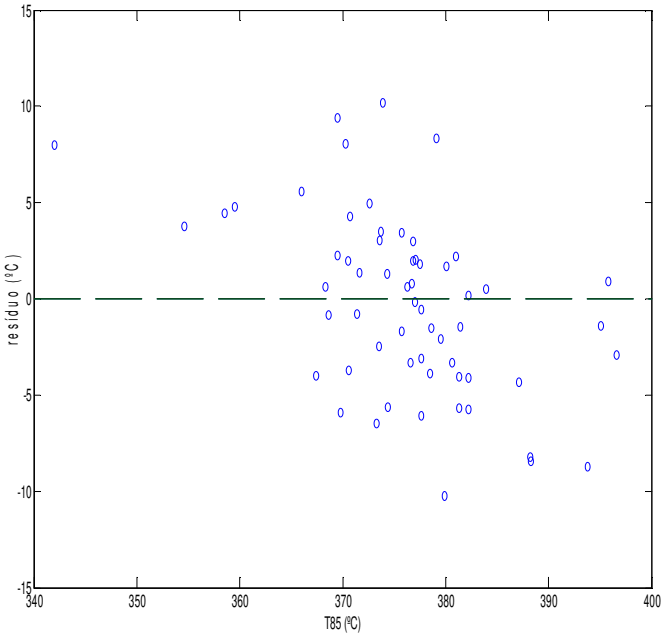


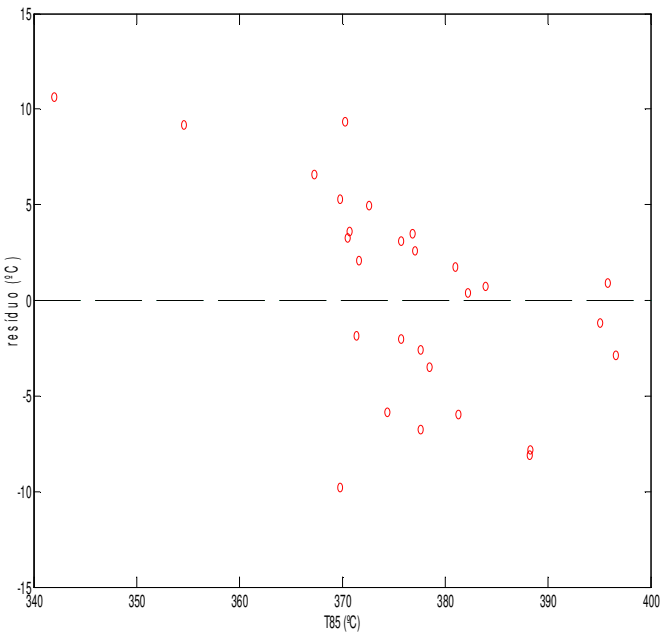
Figura 8.20 – Valores experimentais contra previstos para o modelo SVM para a T85 com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

Tabela 8.14 – Modelos de calibração para a T85 obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento correção da linha base e centragem na média							
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				
		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	nSV		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	nSV	
RBF	correção de linha base e centragem na média	5050	0,0450	2,2814	3,8739	0,9177	46	Correção de linha base e centragem na média	5050	0,0450	2,2814	3,8739	0,9177	46	
Polinomial	primeira derivada	600	0,14	3,1007	4,7697	0,8725	37		2200	0,50	3,1752	4,8272	0,8618	48	
Sigmoidal	-	-	-	-	-	-	-		-	-	-	-	-	-	
Linear	SNV	10,00	0,1604	3,8621	4,2575	0,8430	23		15,00	0,0233	3,1403	4,5807	0,8669	31	

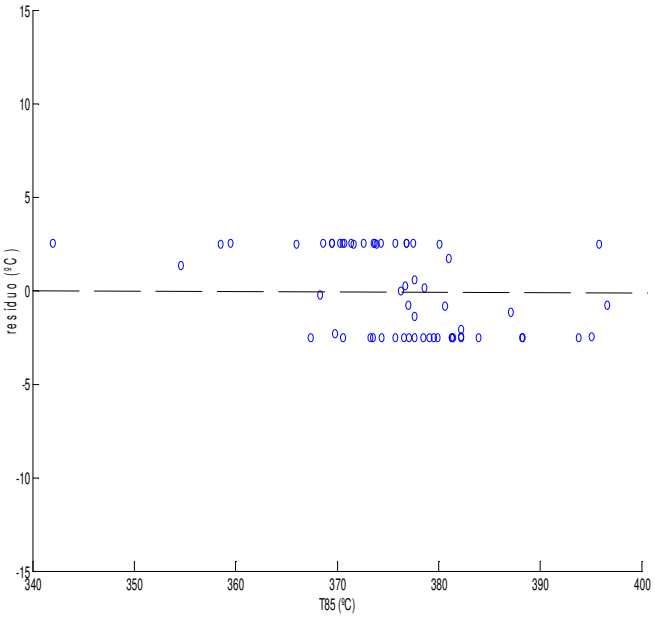


(a)

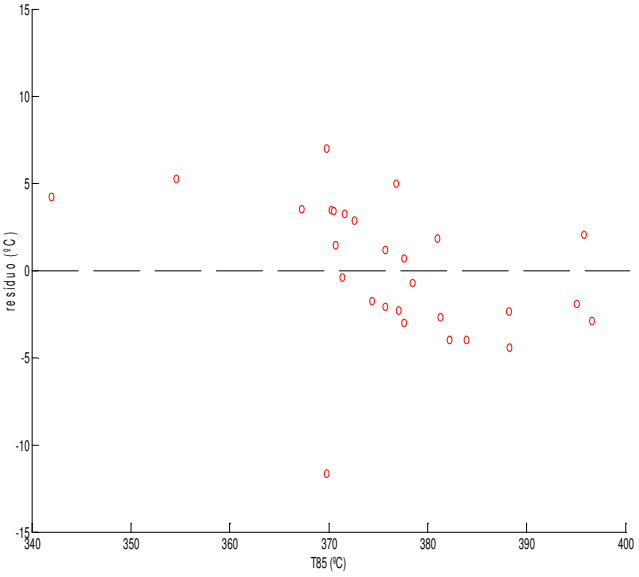


(b)

Figura 8.21 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS para o T85



(a)



(b)

Figura 8.22 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM para o T85

Tabela 8.15 – resultados de previsão dos modelos PLS e SVM para o T85

	PLS			SVM		
	Medido (ºC)	Previsto (ºC)	erro relativo (%)	Medido (ºC)	Previsto (ºC)	erro relativo (%)
1.	342,00	352,66	3,12	342,00	346,19	1,22
2.	354,60	363,80	2,59	354,60	359,86	1,48
3.	367,30	373,88	1,79	367,30	370,83	0,96
4.	369,80	359,99	-2,65	369,80	358,13	-3,16
5.	369,80	375,11	1,44	369,80	376,78	1,89
6.	370,30	379,66	2,53	370,30	373,76	0,93
7.	370,50	373,78	0,89	370,50	373,89	0,92
8.	370,70	374,28	0,97	370,70	372,16	0,39
9.	371,40	369,56	-0,49	371,40	371,01	-0,11
10.	371,60	373,67	0,56	371,60	374,85	0,88
11.	372,60	377,53	1,32	372,60	375,48	0,77
12.	374,40	368,55	-1,56	374,40	372,61	-0,48
13.	375,70	378,79	0,82	375,70	376,87	0,31
14.	375,70	373,67	-0,54	375,70	373,63	-0,55
15.	376,80	380,31	0,93	376,80	381,76	1,32
16.	377,10	379,70	0,69	377,10	374,79	-0,61
17.	377,60	375,04	-0,68	377,60	378,27	0,18
18.	377,60	370,82	-1,80	377,60	374,57	-0,80
19.	378,50	374,98	-0,93	378,50	377,77	-0,19
20.	381,00	382,73	0,45	381,00	382,82	0,48
21.	381,30	375,36	-1,56	381,30	378,62	-0,70
22.	382,20	382,59	0,10	382,20	378,21	-1,04
23.	383,90	384,64	0,19	383,90	379,93	-1,04
24.	388,20	380,10	-2,09	388,20	385,81	-0,62
25.	388,30	380,46	-2,02	388,30	383,85	-1,14
26.	395,10	393,89	-0,31	395,10	393,17	-0,49
27.	395,80	396,67	0,22	395,80	397,85	0,52
28.	396,60	393,71	-0,73	396,60	393,70	-0,73

8.2.6 – Modelos de regressão para a temperatura de destilação de 90% recuperados

Os modelos de regressão para a T90 foram feitos a partir da realização de ensaios pelo método de referência ASTM D86-09 para determinação desse parâmetro e com a obtenção de espectros na região do infravermelho próximo para 88 amostras.

8.2.6.1 – Modelo PLS

Obteve-se o modelo que fornece melhores resultados utilizando os dados com pré-processamento primeira derivada. Utilizou-se 9 variáveis latentes, que explicam 100 % da variância dos dados. Os resultados são mostrados na tabela 8.16.

Tabela 8.16 – Resultados do melhor modelo de calibração para a T90 obtido com PLS.

Pré-processamento	RMSEC (ºC)	RMSEP (ºC)	R^2
Primeira derivada	4,04	4,61	0,748

A figura 8.23 ilustra o resultado do modelo ajustado com o algoritmo PLS e com as 60 amostras de calibração e as 28 amostras de validação.

A reprodutibilidade especificada pela norma ASTM D 86 é de 5,4 ºC, o que torna possível o uso desse modelo de calibração tomando a comparação com o RMSEP obtido como parâmetro de decisão.

No entanto, mais uma vez devemos considerar a importância desse parâmetro de qualidade do óleo diesel na especificação da carga do processo de hidrotratamento. Assim, buscou-se a obtenção de um modelo de calibração com melhor poder de ajuste aos dados, visando diminuir o erro de previsão.

8.2.6.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a função kernel RBF e os dados com correção de linha base WLS e centrados na média. A figura 8.24 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 8.17. Também são mostrados os resultados obtidos para os dados com pré-processamento primeira derivada, pré-processamento que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

Todas as funções kernel utilizadas, com exceção da função sigmoidal, proporcionaram resultados com RMSEP melhor do que o obtido com o modelo PLS. A função kernel sigmoidal proporciona modelos com valor de R^2 abaixo de 0,6 e seus resultados não são considerados na tabela 8.17.

Aqui mais uma vez verifica-se a consistência dos modelos obtidos, que utilizam um número não excessivo de vetores de suporte e valores próximos de RMSEC e RMSEP, indicando que não há sobreajuste dos modelos.

O melhor modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 29 % melhor em relação ao valor obtido com o modelo PLS. Esse valor, de 3,26 ºC, torna o modelo construído muito útil para utilização, uma vez que está bem abaixo do valor especificado para a reprodutibilidade da análise da T90 pelo método de referência.

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 8.23 e 8.24, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 8.25 e 8.26, respectivamente. Verifica-se que o modelo SVM possibilita uma modelagem sensivelmente melhor nesse espaço amostral, com a redução dos resíduos de previsão.

A tabela 8.18 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 8.22 mostra uma síntese dos resultados obtidos com SVM e PLS e dos valores de referência estabelecidos pelos métodos ASTM.

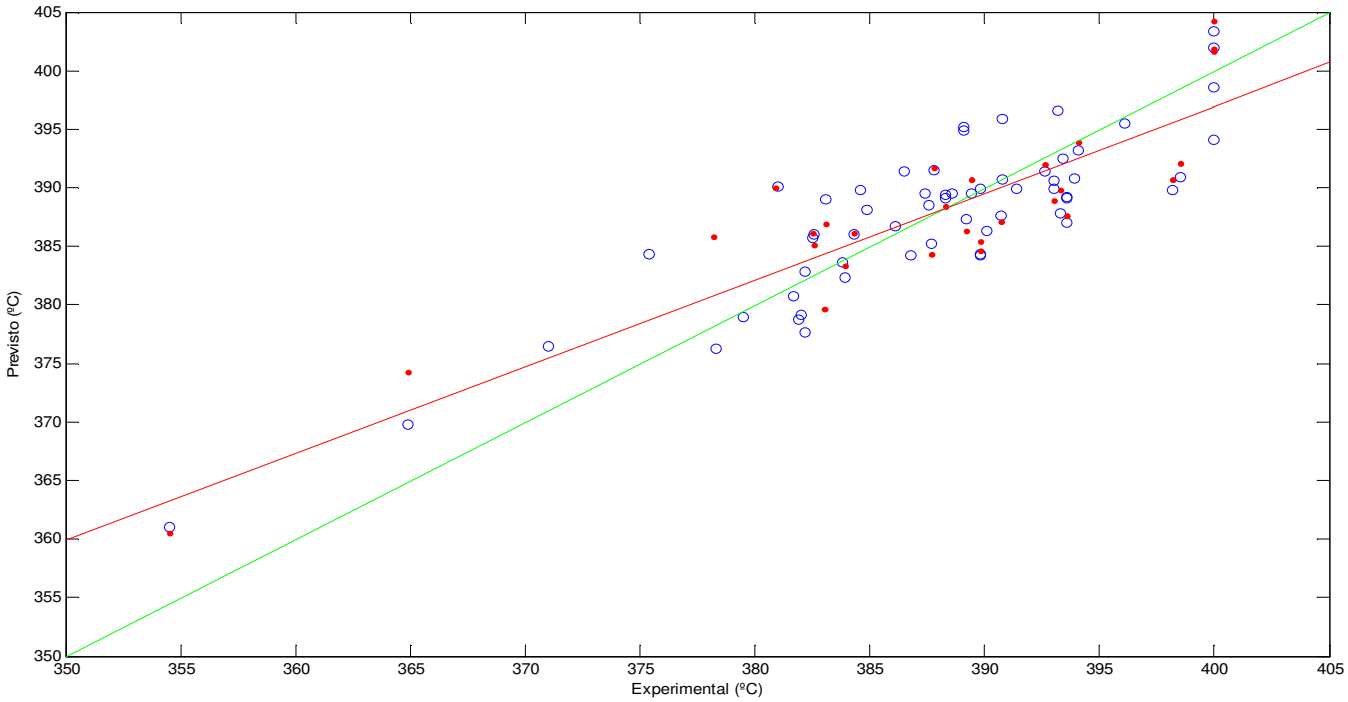


Figura 8.23 – Valores experimentais contra previstos para o modelo PLS para a T90 com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

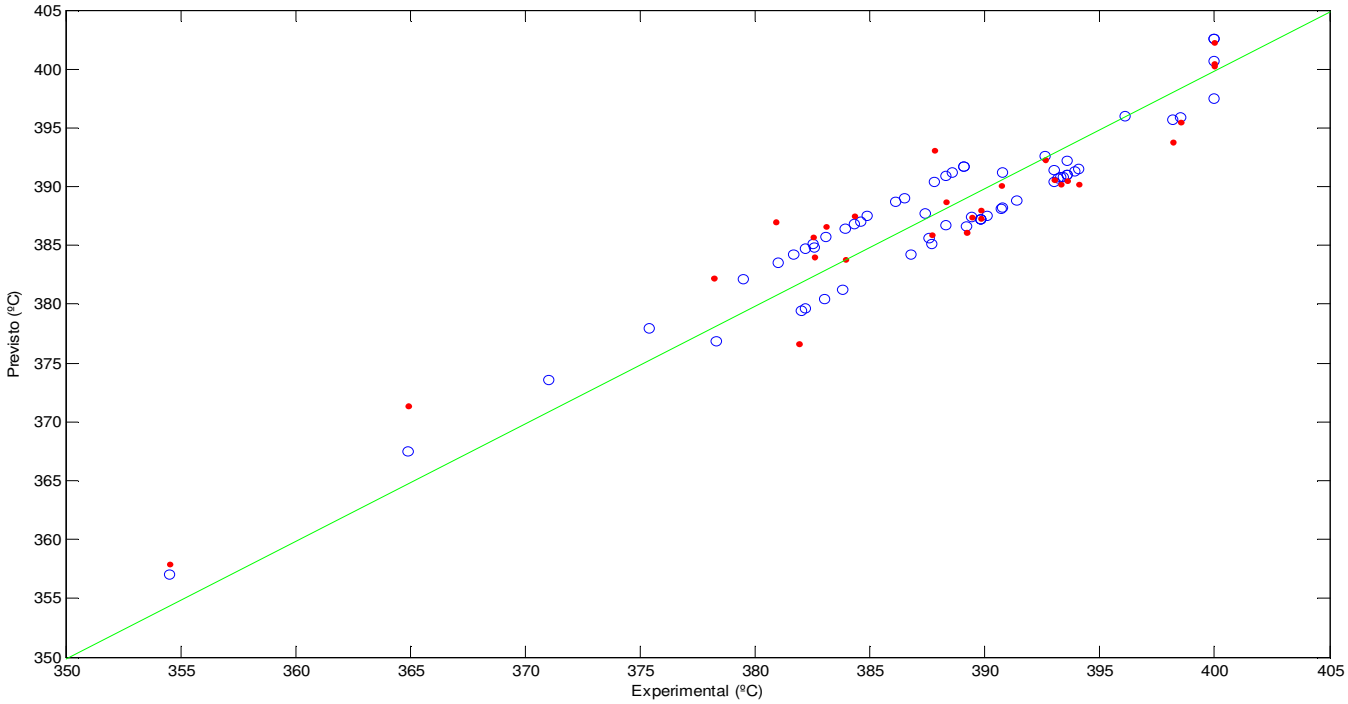
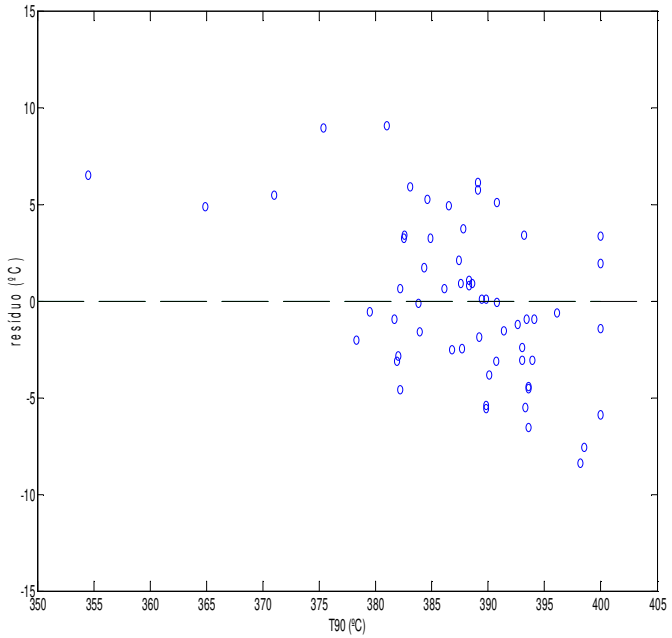


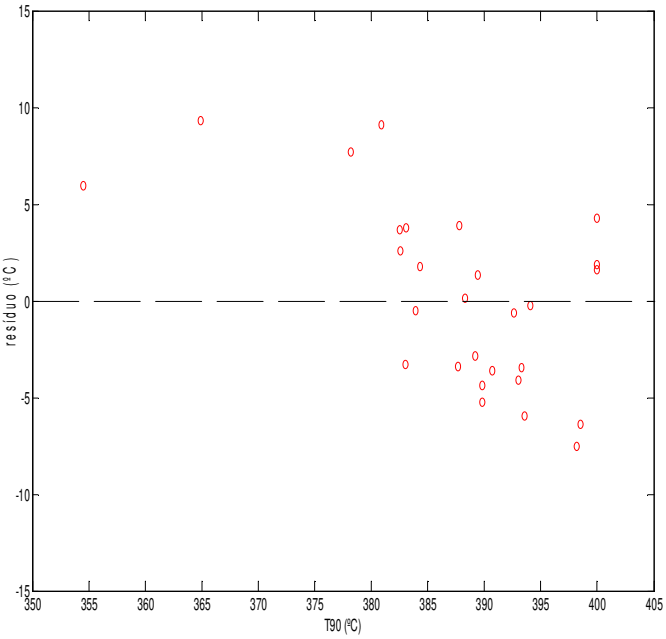
Figura 8.24 – Valores experimentais contra previstos para o modelo SVM para a T90 com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

Tabela 8.17 – Modelos de calibração para a T90 obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento primeira derivada SavGol							
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				
		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	nSV		C	ν	RMSEC (ºC)	RMSEP (ºC)	R^2	nSV	
RBF	Correção de linha base e centragem na média	4200	0,0650	2,3748	3,2659	0,9076	46	Primeira derivada	4500	0,0025	3,0024	4,3284	0,8414	33	
Polinomial	Correção de linha base e centragem na média	7000	0,3000	2,7934	3,9484	0,8690	46		5000	0,0200	2,9605	4,5685	0,8340	32	
Sigmoidal	-	-	-	-	-	-	-		-	-	-	-	-	-	
Linear	SNV	90,00	0,0550	2,9042	3,7872	0,8644	37		-	-	-	-	-	-	

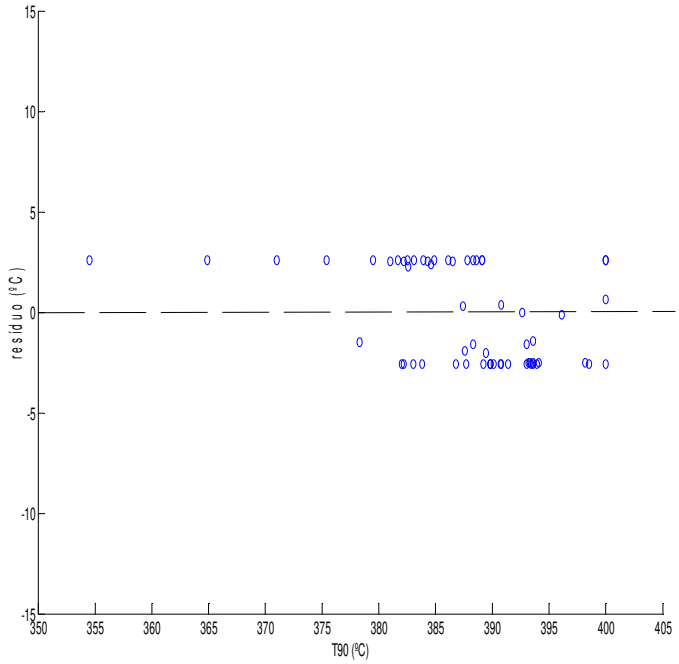


(a)

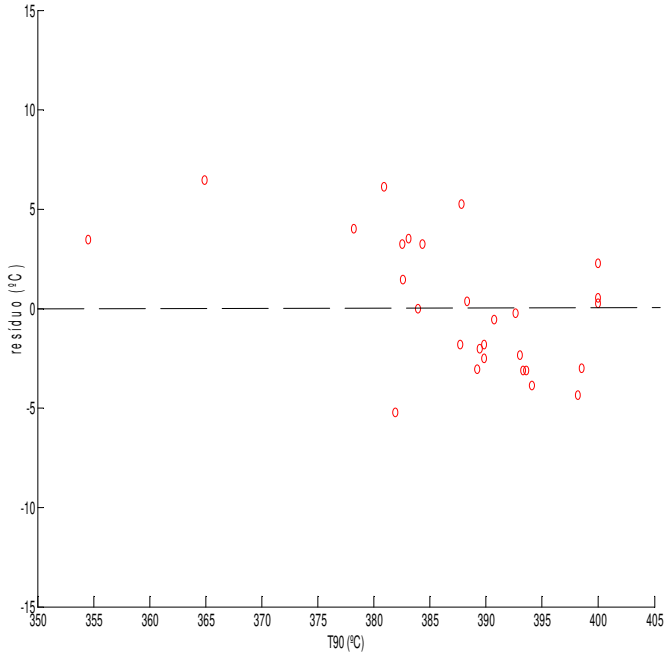


(b)

Figura 8.25 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS para o T90



(a)



(b)

Figura 8.26 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM para o T90

Tabela 8.18 – resultados de previsão dos modelos PLS e SVM para o T90

	PLS			SVM		
	Medido (ºC)	Previsto (ºC)	erro relativo (%)	Medido (ºC)	Previsto (ºC)	erro relativo (%)
1.	354,50	360,47	1,68	354,50	357,94	0,97
2.	364,90	374,22	2,56	364,90	371,36	1,77
3.	378,20	385,87	2,03	378,20	382,20	1,06
4.	380,90	389,98	2,39	380,90	387,00	1,60
5.	382,50	386,16	0,96	381,90	376,64	-1,38
6.	382,60	385,16	0,67	382,50	385,76	0,85
7.	383,00	379,69	-0,86	382,60	384,05	0,38
8.	383,10	386,88	0,99	383,10	386,63	0,92
9.	383,90	383,38	-0,14	383,90	383,86	-0,01
10.	384,30	386,09	0,47	384,30	387,53	0,84
11.	387,70	384,28	-0,88	387,70	385,89	-0,47
12.	387,80	391,69	1,00	387,80	393,04	1,35
13.	388,30	388,45	0,04	388,30	388,68	0,10
14.	389,20	386,36	-0,73	389,20	386,14	-0,79
15.	389,40	390,74	0,34	389,40	387,38	-0,52
16.	389,80	385,41	-1,13	389,80	387,99	-0,46
17.	389,80	384,58	-1,34	389,80	387,29	-0,64
18.	390,70	387,10	-0,92	390,70	390,11	-0,15
19.	392,60	391,98	-0,16	392,60	392,34	-0,07
20.	393,00	388,90	-1,04	393,00	390,62	-0,61
21.	393,30	389,85	-0,88	393,30	390,19	-0,79
22.	393,60	387,63	-1,52	393,60	390,47	-0,80
23.	394,10	393,88	-0,06	394,10	390,23	-0,98
24.	398,20	390,68	-1,89	398,20	393,81	-1,10
25.	398,50	392,14	-1,60	398,50	395,46	-0,76
26.	400,00	401,62	0,41	400,00	400,25	0,06
27.	400,00	404,25	1,06	400,00	402,27	0,57
28.	400,00	401,85	0,46	400,00	400,51	0,13

8.2.7 – Modelos de regressão para a densidade

Os modelos de regressão para a densidade foram feitos a partir da realização de ensaios pelo método de referência ASTM D4052-09 para determinação desse parâmetro e com a obtenção de espectros na região do infravermelho próximo para 88 amostras.

8.2.7.1 – Modelo PLS

Obteve-se o modelo que fornece melhores resultados utilizando os dados pré-processados com primeira derivada. Utilizou-se 9 variáveis latentes, que explicam 100 % da variância dos dados. Os resultados são mostrados na tabela 8.19.

Tabela 8.19 – Resultados do melhor modelo de calibração para a densidade.

Pré-processamento	RMSEC (g/mL)	RMSEP (g/mL)	R^2
Primeira derivada	0,0021	0,0022	0,894

A figura 8.27 ilustra o resultado do modelo ajustado com o algoritmo PLS e com as 60 amostras de calibração e as 28 amostras de validação.

A reprodutibilidade especificada pela norma ASTM D 4052 é de 0,0005 g/mL, o que torna impossível o uso desse modelo de calibração tomando a comparação com o RMSEP obtido como parâmetro de decisão.

Considerando a importância desse parâmetro de qualidade do óleo diesel na especificação da carga do processo de hidrotratamento buscou-se a obtenção de um modelo de calibração com melhor poder de ajuste aos dados, visando diminuir o erro de previsão.

8.2.7.2 – Modelos SVM

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a função kernel polinomial e os dados com correção de linha base WLS e centrados na média. A figura 8.28 ilustra o resultado obtido. Os parâmetros selecionados e o pré-processamento utilizado para o melhor resultado obtido para cada função kernel testada são mostrados na tabela 8.20. Também são mostrados os resultados obtidos para os dados com pré-processamento primeira derivada, pré-processamento que proporcionou o melhor modelo com PLS, apenas para fins de comparação.

Todas as funções kernel utilizadas, com exceção da função sigmoidal, proporcionaram resultados com RMSEP melhor do que o obtido com o modelo PLS. Aqui mais uma vez verifica-se a consistência dos modelos obtidos, que utilizam um número não excessivo de vetores de suporte e valores próximos de RMSEC e RMSEP, indicando que não há sobreajuste dos modelos.

O melhor modelo SVM obtido fornece um RMSEP de 0,0016 g/mL, que é aproximadamente 28 % melhor em relação ao valor obtido com o modelo PLS, não atingindo, porém, a reprodutibilidade especificada pelo método de referência.

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 8.27 e 8.28, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 8.29 e 8.30, respectivamente. Verifica-se que o modelo SVM possibilita uma modelagem sensivelmente melhor nesse espaço amostral, com a redução dos resíduos de previsão.

A tabela 8.21 mostra os valores de previsão obtidos para os conjuntos de validação com os modelos PLS e SVM e a tabela 8.22 mostra uma síntese dos resultados obtidos com SVM e PLS e dos valores de referência estabelecidos pelos métodos ASTM.

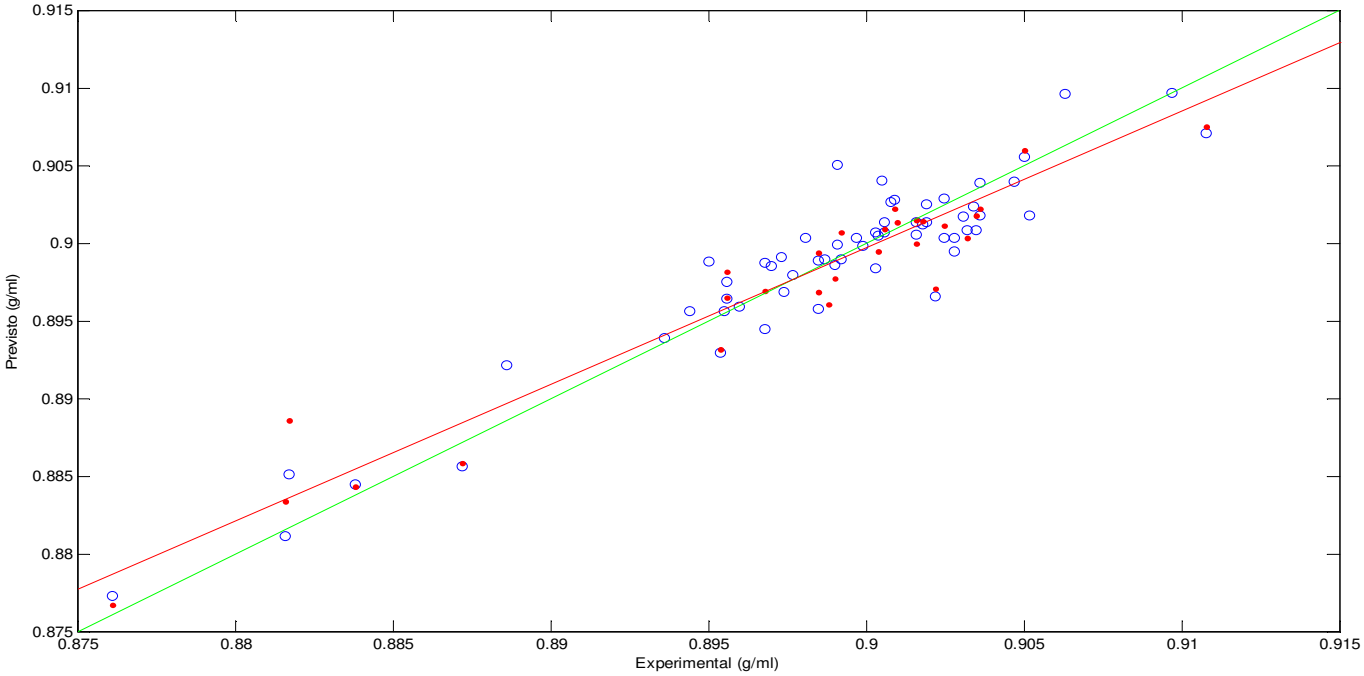


Figura 8.27 – Valores experimentais contra previstos para o modelo PLS para a densidade com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

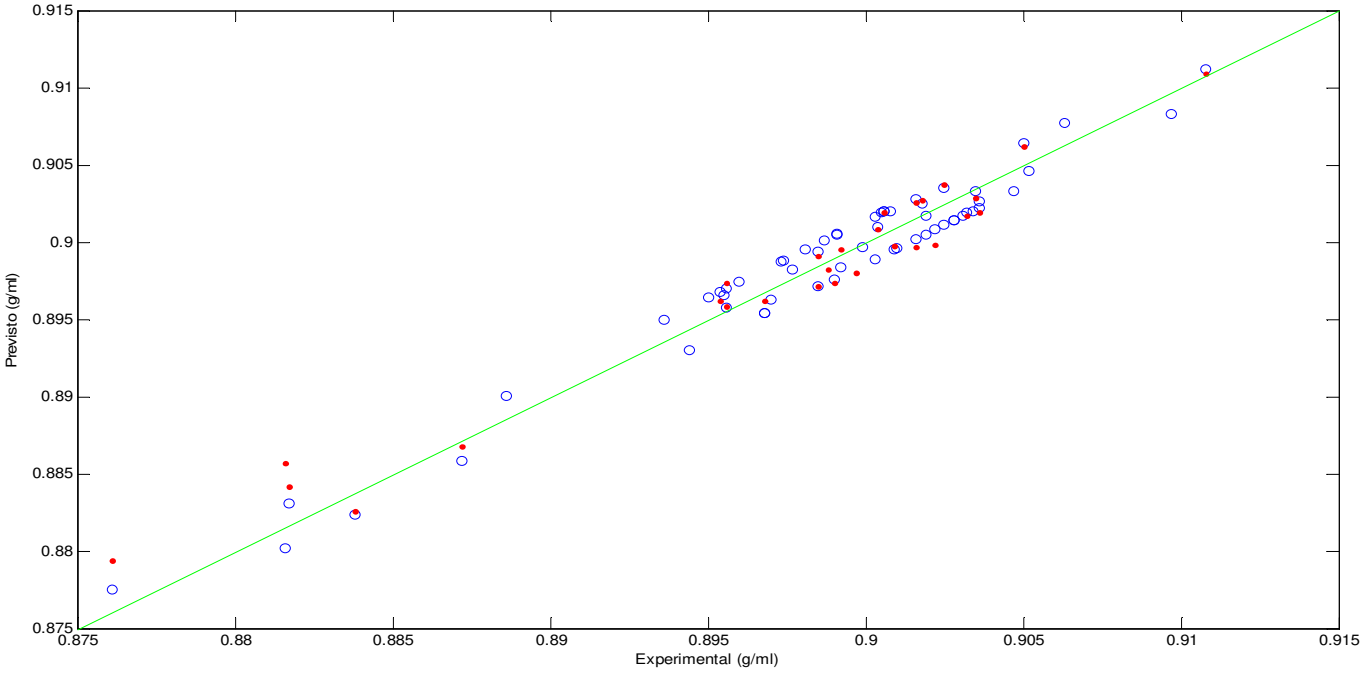
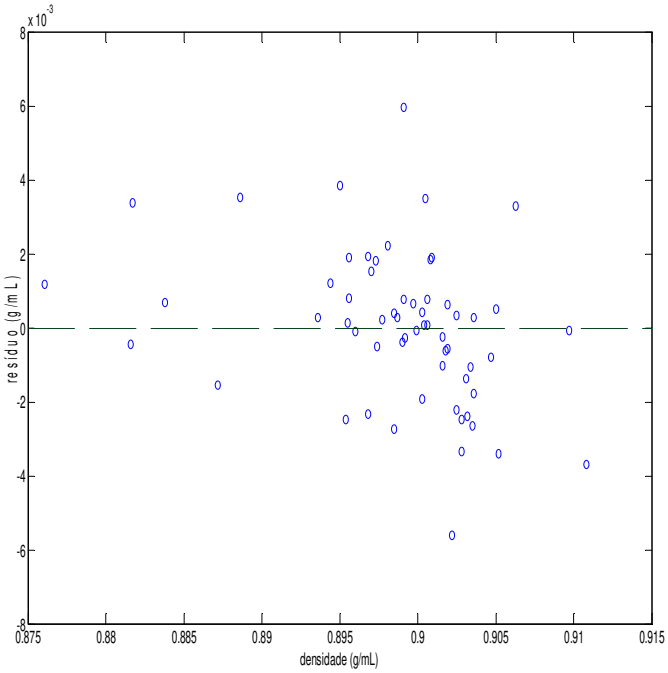


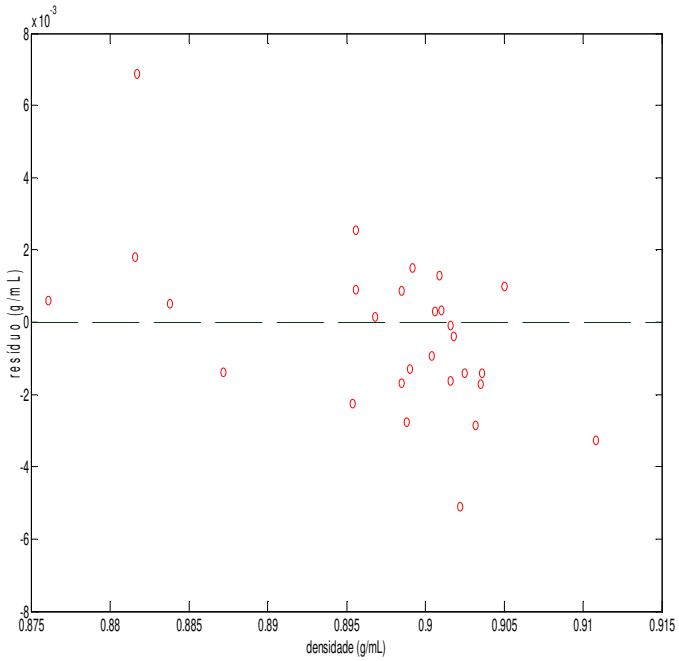
Figura 8.28 – Valores experimentais contra previstos para o modelo SVM para a densidade com as 60 amostras de calibração (○) e as 28 amostras de validação (●).

Tabela 8.20 – Modelos de calibração para a densidade obtidos com SVM

Melhores modelos SVM								Modelos SVM com pré-processamento primeira derivada							
Função kernel	Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				Pré-processamento	Parâmetros selecionados		Resultados do modelo SVM				
		C	ν	RMSEC (g/mL)	RMSEP (g/mL)	R^2	nSV		C	ν	RMSEC (g/mL)	RMSEP (g/mL)	R^2	nSV	
RBF	Correção de linha base e centragem na média	1900	0,075	0,0014	0,0017	0,9510	39	Primeira derivada	11,15	0,3816	0,0023	0,0028	0,8808	39	
Polinomial	Correção de linha base e centragem na média	1500	0,010	0,0013	0,0016	0,9569	40		-	-	-	-	-	-	
Sigmoidal	Primeira derivada SavGol	4,178	0,8285	0,0035	0,0043	0,7161	55		4,178	0,8285	0,0035	0,0043	0,7161	55	
Linear	Primeira derivada SavGol	23,20	0,0010	0,0017	0,0022	0,9330	33		23,20	0,0010	0,0017	0,0022	0,9330	33	

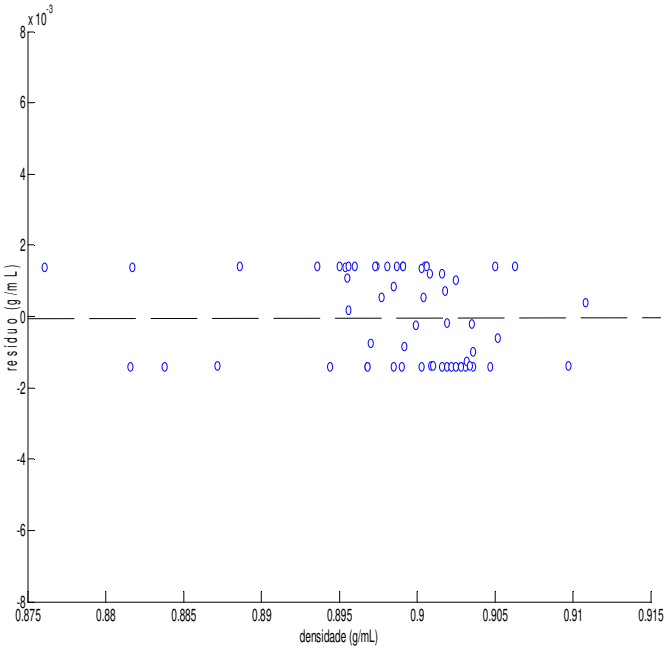


(a)

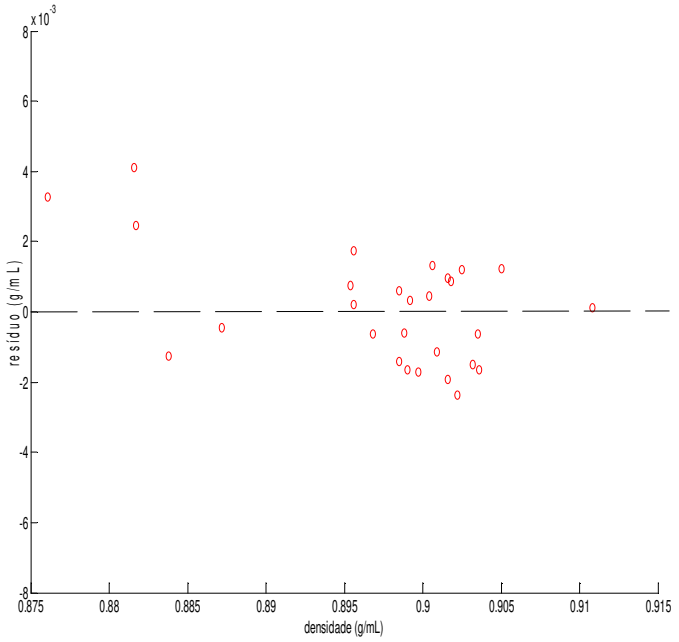


(b)

Figura 8.29 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS para a densidade



(a)



(b)

Figura 8.30 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM para a densidade

Tabela 8.21 – resultados de previsão dos modelos PLS e SVM para a densidade

	PLS			SVM		
	Medido (ºC)	Previsto (ºC)	Erro relativo (%)	Medido (ºC)	Previsto (ºC)	erro relativo (%)
1.	0,8761	0,8767	0,068	0,8761	0,8794	0,372
2.	0,8816	0,8834	0,205	0,8816	0,8857	0,467
3.	0,8817	0,8886	0,780	0,8817	0,8842	0,280
4.	0,8838	0,8843	0,057	0,8838	0,8825	-0,142
5.	0,8872	0,8858	-0,155	0,8872	0,8868	-0,049
6.	0,8954	0,8931	-0,252	0,8954	0,8962	0,085
7.	0,8956	0,8965	0,101	0,8956	0,8958	0,023
8.	0,8956	0,8981	0,285	0,8956	0,8973	0,194
9.	0,8968	0,8969	0,016	0,8968	0,8962	-0,069
10.	0,8985	0,8994	0,098	0,8985	0,8991	0,067
11.	0,8985	0,8968	-0,188	0,8985	0,8971	-0,155
12.	0,8988	0,8961	-0,306	0,8988	0,8982	-0,067
13.	0,8990	0,8977	-0,142	0,8990	0,8974	-0,182
14.	0,8992	0,9007	0,168	0,8992	0,8995	0,035
15.	0,9004	0,8995	-0,102	0,8997	0,8980	-0,189
16.	0,9006	0,9009	0,034	0,9004	0,9009	0,052
17.	0,9009	0,9022	0,145	0,9006	0,9019	0,145
18.	0,9010	0,9013	0,036	0,9009	0,8998	-0,125
19.	0,9016	0,9015	-0,012	0,9016	0,8997	-0,212
20.	0,9016	0,9000	-0,181	0,9016	0,9026	0,107
21.	0,9018	0,9014	-0,045	0,9018	0,9027	0,098
22.	0,9022	0,8971	-0,567	0,9022	0,8998	-0,263
23.	0,9025	0,9011	-0,156	0,9025	0,9037	0,134
24.	0,9032	0,9004	-0,315	0,9032	0,9017	-0,167
25.	0,9035	0,9018	-0,188	0,9035	0,9029	-0,069
26.	0,9036	0,9022	-0,155	0,9036	0,9019	-0,184
27.	0,9050	0,9060	0,109	0,9050	0,9062	0,136
28.	0,9108	0,9075	-0,359	0,9108	0,9109	0,013

Tabela 8.22 – Resultados dos modelos PLS e SVM e valores de referência dos métodos ASTM

Conjunto de dados	Região espectral (cm^{-1})	Parâmetro	Modelo PLS RMSEP R^2	Modelo SVM RMSEP R^2	Melhora do RMSEP (%)	ASTM Repetibilidade Reprodutibilidade	Espaço amostral
C	6444-8936	Ponto de anilina	1,02 ℃ 0,758	0,54 ℃ 0,935	53	D611-07 0,16 ℃ 0,5 ℃	57,6 ℃ a 71,6 ℃
		Indice de cetano	1,14 0,774	0,90 0,884	21	D976-06 precisão ± 2 CN	35,2 a 47,7
		PIE	8,5580 0,771	5,7851 0,891	32	D86-09 3,5 ℃ 8,5 ℃	140,7 ℃ a 257,6 ℃
		T50	5,40 ℃ 0,810	2,47 ℃ 0,978	54	D86-09 1,3 ℃ 3,0 ℃	280,8 ℃ a 356,3 ℃
		T85	5,40℃ 0,749	3,87 ℃ 0,917	28	D86-09 2,4 ℃ 5,2 ℃	342,0 ℃ a 396,6 ℃
		T90	4,61 ℃ 0,748	3,26 ℃ 0,908	29	D86-09 2,6 ℃ 5,4 ℃	354,5 ℃ a 400,0 ℃
		densidade	0,0022 g/mL 0,894	0,0016 g/mL 0,957	28	D4052-09 0,0001 g/mL 0,0005 g/mL	0,8761 g/mL a 0,9108 g/mL

8.2.8 – Comparação dos resultados dos modelos PLS e SVM – teste F

O procedimento citado no item 7.2.3 foi aqui repetido. Os valores críticos de $F_{27,27}$ ao nível de significância de 95 % são mostrados na tabela 8.23 juntamente com os valores calculados de F.

Tabela 8.23 – Resultados do teste-F na comparação dos modelos PLS e SVM

Parâmetro	F calculado	F crítico 95 %
Ponto de anilina	3,56	1,90
Índice de cetano	1,60	
PIE	2,19	
T50	4,78	
T85	1,94	
T90	1,99	
densidade	1,89	

Para os parâmetros ponto de anilina, PIE, T50, T85 e T90 há evidências, ao nível de significância de 95%, de que o modelo SVM fornece melhores resultados do que o modelo PLS. Para o parâmetro densidade o valor de F calculado está muito próximo do valor de F crítico com nível de confiança de 95 %. Para o parâmetro índice de cetano os modelos com PLS e SVM fornecem resultados semelhantes ao nível de significância de 95 %.

8.2.9 - Comparação dos resultados de referência com os dos modelos PLS e SVM

O procedimento citado no item 7.2.4 foi aqui repetido e os resultados são mostrados na tabela 8.24.

Tabela 8.24 – Percentual dos valores de referência que estão no intervalo estabelecido pelo método ASTM E 1655-05 para os modelos PLS e SVM

Parâmetro	Modelo PLS (%)	Modelo SVM (%)
Ponto de anilina	43	82
Indice de cetano	-	-
PIE	75	86
T50	54	79
T85	61	89
T90	75	93
densidade	21	25

Verifica-se que os resultados de previsão dos modelos SVM são consideravelmente superiores aos obtidos com os modelos PLS, embora não se tenha atingido a completa concordância com os métodos de referência segundo o procedimento descrito no método ASTM E1655-05.

8.3 – Conclusões

Os modelos SVM proporcionaram melhores resultados de previsão em relação aos modelos PLS para os sete parâmetros estudados. Os melhores resultados obtidos com SVM e PLS são mostrados na tabela 8.22. Com os modelos SVM desenvolvidos, os valores de RMSEP obtiveram uma melhora entre 21% e 54%, em relação aos modelos desenvolvidos com PLS, e todos os modelos de calibração têm valores de RMSEP menores do que os valores de reprodutibilidade estabelecidos pelos métodos de referência, com exceção do modelo de calibração para a densidade.

Com a utilização do Teste-F verifica-se que 5 dos 7 modelos desenvolvidos com SVM são estatisticamente melhores que os modelos PLS.

Entre os modelos SVM construídos para um mesmo parâmetro utilizando-se as quatro diferentes funções kernel, verificou-se que as funções kernel RBF e polinomial sempre proporcionam melhores resultados que o modelo PLS e o kernel linear proporcionou melhores resultados para 5 dos 7 parâmetros estudados. Entre os sete melhores modelos de regressão obtidos com SVM, quatro utilizam a função kernel polinomial e três utilizam a função kernel RBF. Por outro lado, o kernel sigmoidal proporcionou resultados inferiores aos obtidos com PLS para todos os parâmetros. Os melhores resultados obtidos com SVM para cada função kernel e os resultados obtidos com PLS para cada parâmetro estudado são mostrados na tabela 9.1, onde os modelos obtidos com R^2 inferior a 0,6 não foram considerados.

O pré-processamento dos dados com correção de linha base e centragem na média proporcionou os melhores resultados com SVM em seis dos sete melhores modelos.

O desenvolvimento de modelos de calibração com SVM proporcionou resultados que possibilitam a aplicação dos mesmos como ferramenta em Tecnologia Analítica de Processos – PAT e controle de qualidade. Com o desenvolvimento de modelos de calibração mais eficazes utilizando o algoritmo SVM e dados de espectroscopia NIR, para determinação *on line* de parâmetros para utilização no monitoramento dos parâmetros de qualidade da carga do processo de hidrotratamento torna-se possível obter maior otimização no controle da unidade.

9 – Conclusão geral - Modelos de regressão obtidos para determinação de parâmetros utilizados no controle de processos

Com a realização de modelos de regressão utilizando PLS e SVM para nove diferentes parâmetros, mostrados nos capítulos 7 e 8, utilizando três conjuntos de dados e três diferentes intervalos espectrais na região do infravermelho próximo obteve-se os melhores resultados com a utilização do algoritmo SVM.

Entre os nove melhores modelos de regressão obtidos com SVM, cinco utilizam a função kernel RBF e quatro utilizam a função kernel polinomial. Para os nove parâmetros de qualidade estudados, os modelos SVM com utilização do kernel RBF e polinomial têm melhores resultados que os modelos PLS. Para seis dos parâmetros estudados os modelos SVM com kernel linear possibilitam melhores resultados que os modelos PLS. Por outro lado, os modelos SVM com kernel sigmoidal proporcionam melhor resultado em relação aos modelos PLS em apenas um dos parâmetros estudados. A tabela 9.1 mostra esses resultados.

O pré-processamento dos dados com correção de linha base e centragem na média proporcionou os melhores resultados com SVM em oito entre os nove melhores modelos.

A comparação entre os modelos PLS e SVM utilizando o teste-F e valores de RMSEP demonstra que para os parâmetros ponto de fulgor, ponto de anilina, PIE, T50, T85 e T90 o modelo SVM é superior, com 95 % de confiança. O parâmetro densidade tem o valor de F calculado muito próximo do valor de F crítico com nível de confiança de 95 %.

Quanto a concordância dos resultados de previsão dos modelos PLS e SVM com os valores obtidos pelo método de referência, através da utilização do procedimento descrito pelo método ASTM E-1655-05, os modelos SVM mostraram melhores resultados, sendo que os resultados dos modelos de calibração para o ponto de fulgor e número de cetano concordam com os resultados obtidos com o método de referência.

A utilização do algoritmo SVM para resolver problemas de regressão com dados de espectroscopia NIR de amostras de óleo diesel mostrou resultados bastante superiores aos obtidos com o PLS. A escolha da função kernel mais adequada e a possibilidade de otimização de dois parâmetros no modelo SVM permitem um

refinamento no ajuste do modelo não possibilitado pela simples escolha do número de variáveis latentes com PLS.

A natureza do algoritmo que permite a modelagem de não linearidades presentes nas correlações estudadas e com um bom poder de generalização possibilitou uma performance superior do SVM com todos os conjuntos de dados e parâmetros estudados, sendo muito adequado para solução dos problemas abordados. Verifica-se que a utilização da função de perda ε-insensível pelo algoritmo SVM (ν-SVR) possibilita a obtenção de modelos com melhor capacidade de ajuste ao longo de todo espaço analítico utilizado, permitindo reduzidos erros de previsão inclusive para as amostras mais extremas, conforme as necessidades e condições que podem ser encontradas durante as análises em linha de produção, enquanto os modelos desenvolvidos com PLS não ajustam bem amostras em todo o intervalo analítico.

Tabela 9.1 – Resultados obtidos com os modelos SVM com as diferentes funções kernel e com os modelos PLS

Conjunto de dados	Parâmetro	**RBF** RMSEP R^2	**Polinomial** RMSEP R^2	**Sigmoidal** RMSEP R^2	**Linear** RMSEP R^2	**PLS** RMSEP R^2	ASTM Repetibilidade Reprodutibilidade
A	Ponto de fulgor	**1,98** **0,9357** **(1)**	2,03 0,8948 (1)	4,80 0,6166 (1)	2,20 0,8910 (1)	3,770 0,698 (2)	D56-05 1,2 ℃ 4,3 ℃
B	Número de cetano	**0,4535** **0,8946** **(1)**	0,4902 0,8923 (4)	0,5025 0,8767 (4)	0,5689 0,9053 (1)	0,5564 0,894 (1)	D613-08 0,8 2,8
C	Ponto de anilina	0,88 0,8814 (3)	**0,54** **0,9346** **(1)**	1,02 0,7913 (3)	-	1,02 0,758 (3)	D611-07 0,16 ℃ 0,5 ℃
	Indice de cetano	0,9864 0,8596 (1)	**0,9066** **0,8840** **(1)**	1,5105 0,6592 (1)	0,9944 0,8846 (1)	1,148 0,774 (1)	D976-06 precisão ± 2 CN
	PIE	5,8950 0,8825 (2)	**5,7851** **0,8909** **(3)**	-	8,6892 0,7770 (1)	8,5580 0,771 (1)	D86-09 3,5 ℃ 8,5 ℃
	T50	**2,4718** **0,9778** **(1)**	3,0054 0,9666 (1)	8,3337 0,7028 (3)	3,8149 0,9156 (2)	5,4058 0,810 (1)	D86-09 1,3 ℃ 3,0 ℃
	T85	**3,8739** **0,9177** **(1)**	4,7697 0,8725 (3)	-	4,2575 0,8430 (2)	5,4033 0,749 (1)	D86-09 2,4 ℃ 5,2 ℃
	T90	**3,2659** **0,9076** **(1)**	3,9484 0,8690 (1)	-	3,7872 0,8644 (2)	4,6090 0,748 (3)	D86-09 2,6 ℃ 5,4 ℃
	densidade	0,0017 0,9510 (1)	**0,0016** **0,9569** **(1)**	0,0043 0,7161 (3)	0,0022 0,9330 (3)	0,0022 0,894 (3)	D4052-09 0,0001 g/mL 0,0005 g/mL

Pré-processamento: (1) correção de linha base e centragem na média
(2) SNV
(3) primeira derivada SavGol
(4) correção de linha base

Obs. 1: O melhor resultado para cada parâmetro está em negrito

Obs. 2: Modelos com R^2 menor que 0,6 não foram considerados

10 – Determinação do teor de biodiesel em óleo diesel através de espectroscopia NIR e SVM

O monitoramento da qualidade dos combustíveis quanto ao teor de biodiesel no combustível óleo diesel é realizado pela Agência Nacional do Petróleo – ANP através do seu Programa de Monitoramento da Qualidade de Combustíveis líquidos - PMQC, visando assegurar a utilização do biocombustível conforme estabelecido na legislação.

Devido ao contínuo incremento da utilização do biodiesel como combustível de motores ciclo diesel torna-se importante o desenvolvimento de métodos mais simples e exatos para o controle de qualidade na produção e distribuição desse combustível.

Nesse trabalho pretende-se verificar o potencial do algoritmo SVM aplicado a dados de espectroscopia NIR para o desenvolvimento de modelos de calibração para determinação do teor de biodiesel em óleo diesel que se sejam mais eficazes e adequados às necessidades analíticas para esse tipo de mistura atualmente, evitando a construção de dois ou três modelos de calibração para determinação do biodiesel em óleo diesel conforme as normas ABNT e ASTM, respectivamente, sugerem.. Considerando-se que na prática atualmente a utilização de biodiesel em misturas com óleo diesel ocorre mais frequentemente entre 0 e 30 % (v/v), sugere-se um modelo de calibração para o intervalo de 0 a 35 % (v/v) de biodiesel em óleo diesel. Sugere-se também a utilização de um único modelo de calibração para o intervalo de 0 a 100 % (v/v) de biodiesel em óleo diesel. Demonstra-se a obtenção de modelos com SVM e compara-se os resultados com os obtidos com PLS.

10.1 – Parte experimental

Foram obtidos espectros na região do infravermelho próximo no intervalo de 4000 cm^{-1} a 9000 cm^{-1} para 81 amostras com teores entre 0% e 100% (v/v) de biodiesel na mistura de óleo diesel A S50 e biodiesel (B100) metílico de óleo soja, ambos fornecidos pela Petrobras Distribuidora S.A.. O B100 utilizado é produzido pela Camera S.A. e possui 97,2 % (m/m) de éster metílico (FAME), determinado pelo método EN14103.

Para obtenção dos espectros de transflectância na região do infravermelho próximo foi utilizado um espectrômetro Perkin Elmer Spectrum 100 MID/NIR, com fonte halógena e detector de sulfato de triglicina deuterada (DTGS). Utilizou-se uma placa de Petri de vidro como recipiente de amostra e um refletor de alumínio com caminho ótico de 0,5 mm como cela de transflectância, conforme ilustra a figura 10.1. Cada espectro foi obtido como uma média de 32 varreduras, com resolução de 4 cm^{-1}.

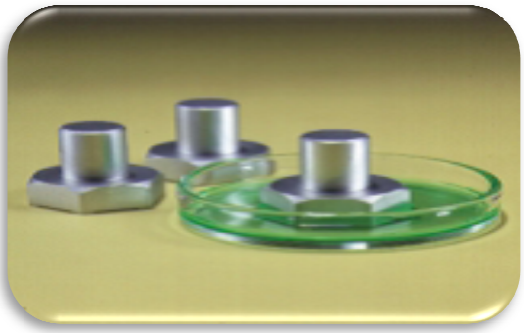


Figura 10.1 – Recipiente para amostra líquida e análise por transflectância

Foram realizados diferentes pré-processamentos dos dados para verificar qual proporciona a construção do melhor modelo utilizando os algoritmos PLS e SVM. Os pré-processamentos testados foram: correção de linha base e centragem na média; SNV; SNV e centragem na média; primeira derivada (janela com 15 pontos); primeira derivada SavGol (janela com 15 pontos) e centragem na média.

O pacote LIBSVM[85] versão 2.88 foi utilizado para o desenvolvimento dos modelos com SVM e o algoritmo genético foi aplicado para realização da otimização

paramétrica. Todos os programas são adequados para utilização com Matlab 7.7 da Mathworks.

Para obtenção dos modelos com SVM foram testadas diferentes funções kernel, tais como: linear, RBF, polinomial e sigmoidal.

Para construção dos modelos SVM os blocos de dados **X** e **y** dos conjuntos de calibração e validação foram previamente escalonados entre [0,1]. Foi utilizado como parâmetro γ do kernel RBF o valor *default* do pacote LIBSVM ($\gamma = 1/k$, onde k é o número de atributos ou variáveis nos dados de entrada) e o grau do polinômio no kernel polinomial igual a 3. Os parâmetros C e ν foram selecionados entre os intervalos de 0 a 10^4 e 10^{-4} a 1, respectivamente, utilizando-se algoritmo genético (GA). Ainda, para otimização paramétrica com GA estipulou-se a utilização de uma população com 20 indivíduos e um máximo de 15 gerações, pois observou-se que com essa configuração o valor do erro de validação cruzada se estabilizava, não havendo melhora com o aumento do número de gerações. A função objetivo a ser otimizada pelo GA, e definido inicialmente, consistiu simplesmente na utilização dos valores obtidos pela validação cruzada com 3 subconjuntos do conjunto de treinamento, buscando-se o menor valor do erro de validação cruzada. Os parâmetros a serem otimizados foram tratados como genes no GA.

Como a minimização do erro de validação cruzada no conjunto de treinamento não garante a obtenção dos parâmetros ótimos, eventualmente um *grid search* manual pode ser necessário, a partir dos valores previamente selecionados pelo GA, para refinamento do resultado, sempre considerando a utilização do adequado número de vetores de suporte e valores próximos de RMSEC e RMSEP de modo a evitar um sobreajuste do modelo.

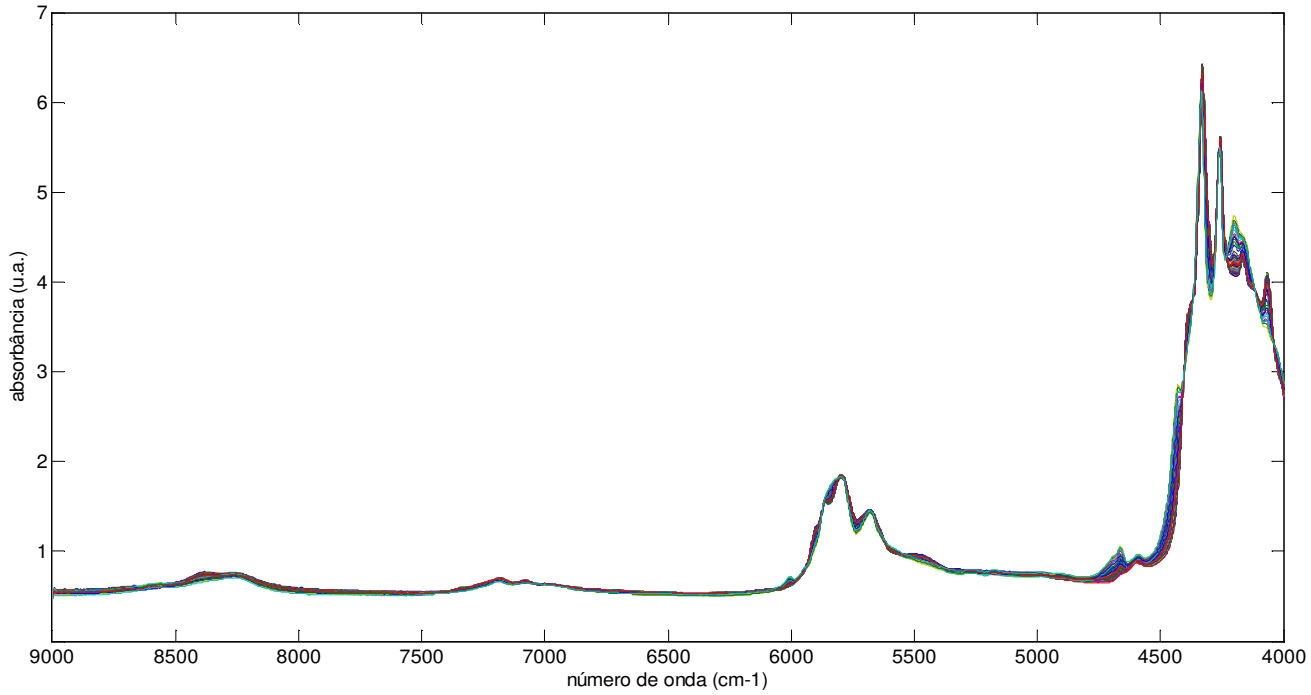
10.2 – Resultados e discussão

Na construção dos modelos de calibração para teores de 0 a 100 % (v/v) de biodiesel utilizaram-se 50 amostras de calibração e 31 amostras de validação. Para os modelos de calibração para teores de 0 a 35 % (v/v) de biodiesel utilizaram-se 41 amostras de calibração e 25 amostras de validação.

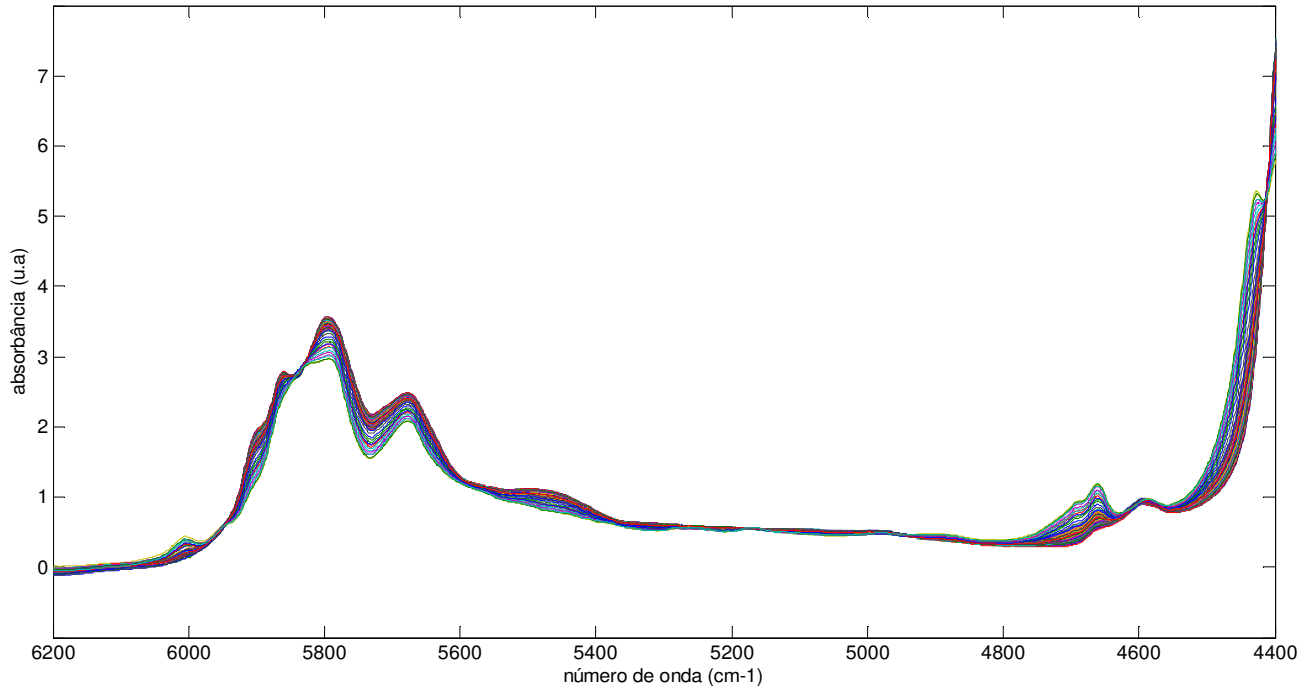
O intervalo espectral utilizado foi de 4000 cm^{-1} a 9000 cm^{-1}, que permite a calibração do teor de biodiesel em óleo diesel devido a ocorrência das bandas de combinação, primeiro e segundo sobreton de estiramentos C-H e C=C. Também, próximo a 4650 cm^{-1} ocorre uma banda de combinação de estiramento C-H e de estiramento da ligação C=O de éster, além disso a diferença nos espectros de NIR nas regiões de 4425 cm^{-1} e 6005 cm^{-1} onde os ésteres metílicos apresentam picos, enquanto os triacilglicerídeos exibem apenas ombros permitem a quantificação do biodiesel, conforme citado no item 2.1.

Foi testada a obtenção de modelos de calibração utilizando as regiões: (i) 4000 cm^{-1} a 9000 cm^{-1}, (ii) 4400 cm^{-1} a 6200 cm^{-1}, (iii) 4400 cm^{-1} a 4600 cm^{-1}, (iv) 4600 cm^{-1} a 4800 cm^{-1} e (v) 5950 cm^{-1} a 6100 cm^{-1}. Os espectros das 81 amostras utilizadas, com pré-processamento SNV, são mostrados na figura 10.2. Os melhores resultados de previsão dos modelos de calibração para 0-100 (%) e 0-35 (%) de biodiesel, utilizando o PLS ou o SVM, foram obtidos utilizando a região (iii). As demais regiões testadas também permitem a calibração do teor de biodiesel em óleo diesel.

Para comparar se há diferença significante entre os modelos SVM e PLS foi realizado um teste-F.



(a)



(b)

Figura 10.2 – Espectros das 81 amostras utilizadas nos conjuntos de calibração e validação. (a) região (i) e (b) região (ii)

10.2.1 – Modelo PLS – teor de biodiesel 0 – 100 % (v/v)

O melhor resultado foi obtido utilizando-se a região espectral (iii) e o pré-processamento SNV. Utilizou-se 3 variáveis latentes, que explicam 99,9% da variância dos dados. Os resultados são mostrados nas tabelas 10.1 e 10.2 e a figura 10.3 ilustra o resultado do modelo ajustado com o algoritmo PLS. A figura 10.5 ilustra a distribuição dos erros de calibração e previsão obtidos

Nesse modelo, embora se tenha obtido o melhor valor de RMSEP entre os modelos PLS desenvolvidos com as regiões espectrais e os diferentes pré-processamentos testados, observa-se uma tendência para desvios negativos nos valores de previsão. A utilização da região espectral (ii), que abrange as bandas de absorção em 4425 cm^{-1}, 4650 cm^{-1} e 6005 cm^{-1} e o pré-processamento SNV permite a obtenção de um modelo em que observa-se a distribuição aleatória dos erros de calibração e previsão, porém com valores de RMSEC e RMSEP muito superiores. Nesse modelo utilizou-se 3 variáveis latentes, que explicam 99,9% da variância dos dados. A figura 10.7 ilustra a distribuição dos erros de calibração e previsão do modelo ajustado com o algoritmo PLS utilizando a região espectral (ii) e os resultados são mostrados na tabela 10.1.

10.2.2 – Modelo SVM – teor de biodiesel 0 – 100 % (v/v)

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a região espectral (iii), a função kernel linear e os dados pré-processados com SNV. Os resultados são mostrados nas tabelas 10.1 e 10.2. A figura 10.4 ilustra o resultado obtido. Os parâmetros selecionados foram: C = 2 e ν = 0,1463. Nesse modelo foram utilizados 14 vetores de suporte. O adequado número de vetores de suporte utilizados evidencia o bom ajuste do modelo, sem ocorrência de sobreajuste.

O modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 10% melhor em relação ao valor obtido com o modelo PLS.

O valor do RMSEP obtido para ambos os modelos fica abaixo do exigido pela norma ABNT para determinação de teores de biodiesel na faixa de 8-30 % (v/v) de

biodiesel. Também, tomando o valor da reprodutibilidade estabelecida pelo método ASTM como comparação os valores de RMSEP obtidos para ambos os modelo ficam abaixo da reprodutibilidade mínima exigida para determinação de biodiesel a partir de 0% (v/v) na mistura com óleo diesel.

Tabela 10.1 – Resultados dos modelos PLS e SVM para 0-100 % (v/v) de biodiesel

Modelo	Região espectral (cm^{-1})	RMSEC (%) R^2	RMSEP (%)	Melhora do RMSEP (%)	ASTM D7371-07 reprodutibilidade (%)	ABNT-NBR 15568 RMSEP (%)
PLS	(ii) 4400 – 6200	0,6157 0,999	0,8302	10	0,76	0 – 8 % (v/v) = 0,1 8 – 30 % (v/v) = 1
	(iii) 4400 - 4600	0,1795 0,999	0,3172			
SVM	(ii) 4400 – 6200	0,0530 0,999	0,3050			
	(iii) 4400 - 4600	0,1698 0,999	0,2859			

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 10.3 e 10.4, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 10.5 e 10.6, respectivamente. Verifica-se que o modelo SVM possibilita um melhor ajuste ao longo de todo o espaço amostral, com a redução dos resíduos de previsão.

A utilização da região espectral (ii) também proporciona bons resultados utilizando o SVM. Nesse modelo utilizou-se a função kernel linear e os dados com pré-processamento SNV e centrados na média. Foram selecionados os parâmetros C= 2500 e ν=0,005 e utilizados 37 vetores de suporte. Os resultados são mostrados na tabela 10.1 e a distribuição dos erros dos conjunto de calibração e validação são mostrados na figura 10.8.

Entre os modelos citados com SVM utilizando as regiões espectrais (ii) e (iii), a utilização da região (iii) permite não apenas um melhor valor de RMSEP como também utiliza menos vetores de suporte, o que evidencia um modelo melhor ajustado.

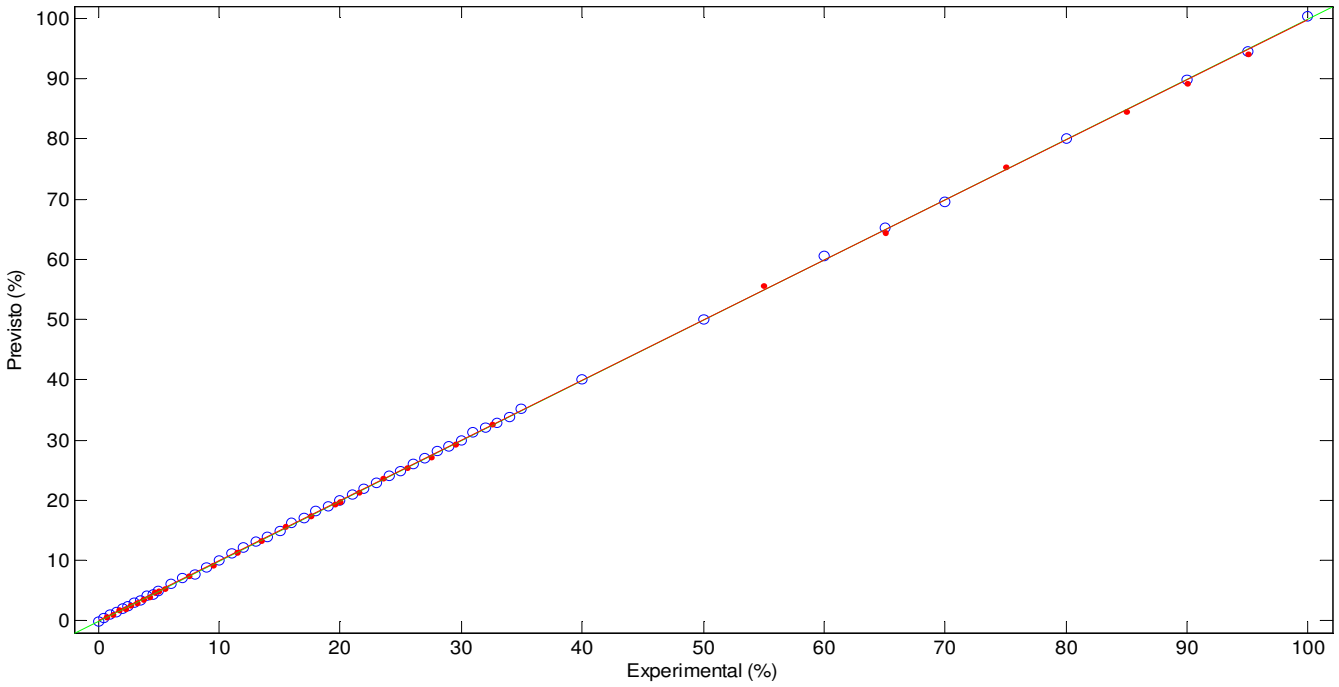


Figura 10.3 – Valores experimentais contra previstos para o modelo PLS com as 50 amostras de calibração (○) e as 31 amostras de validação (●).

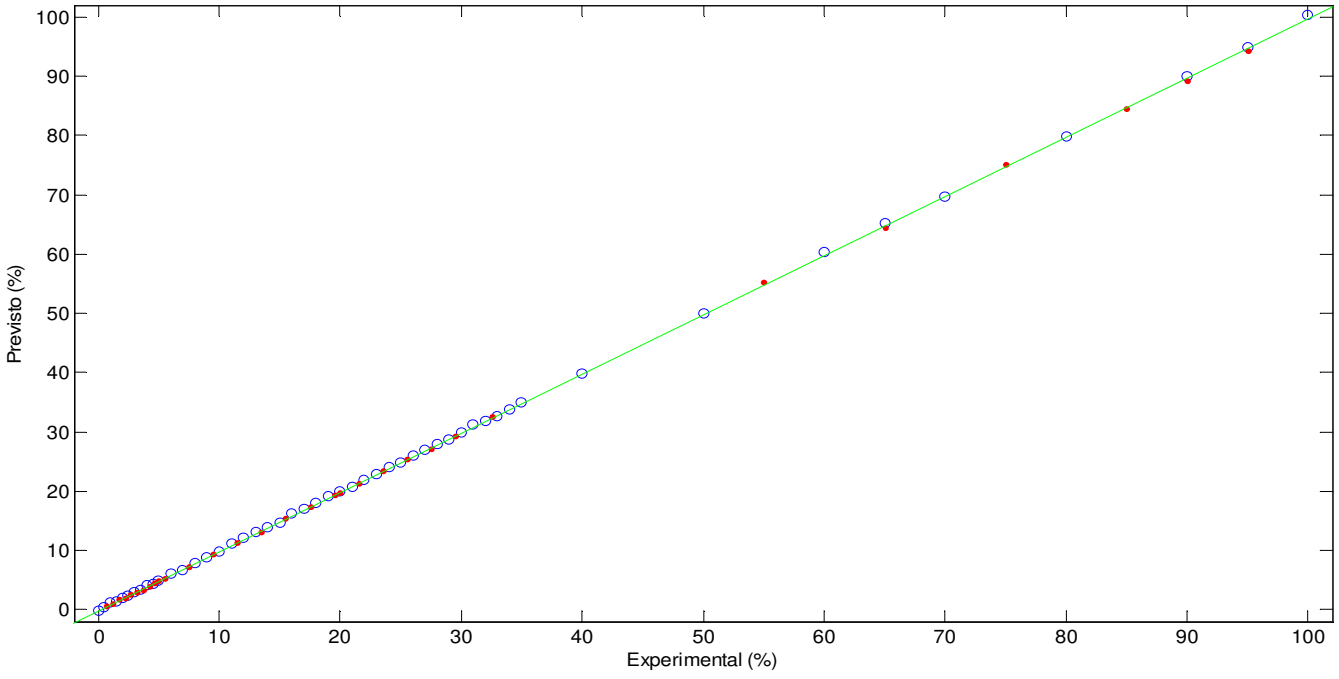
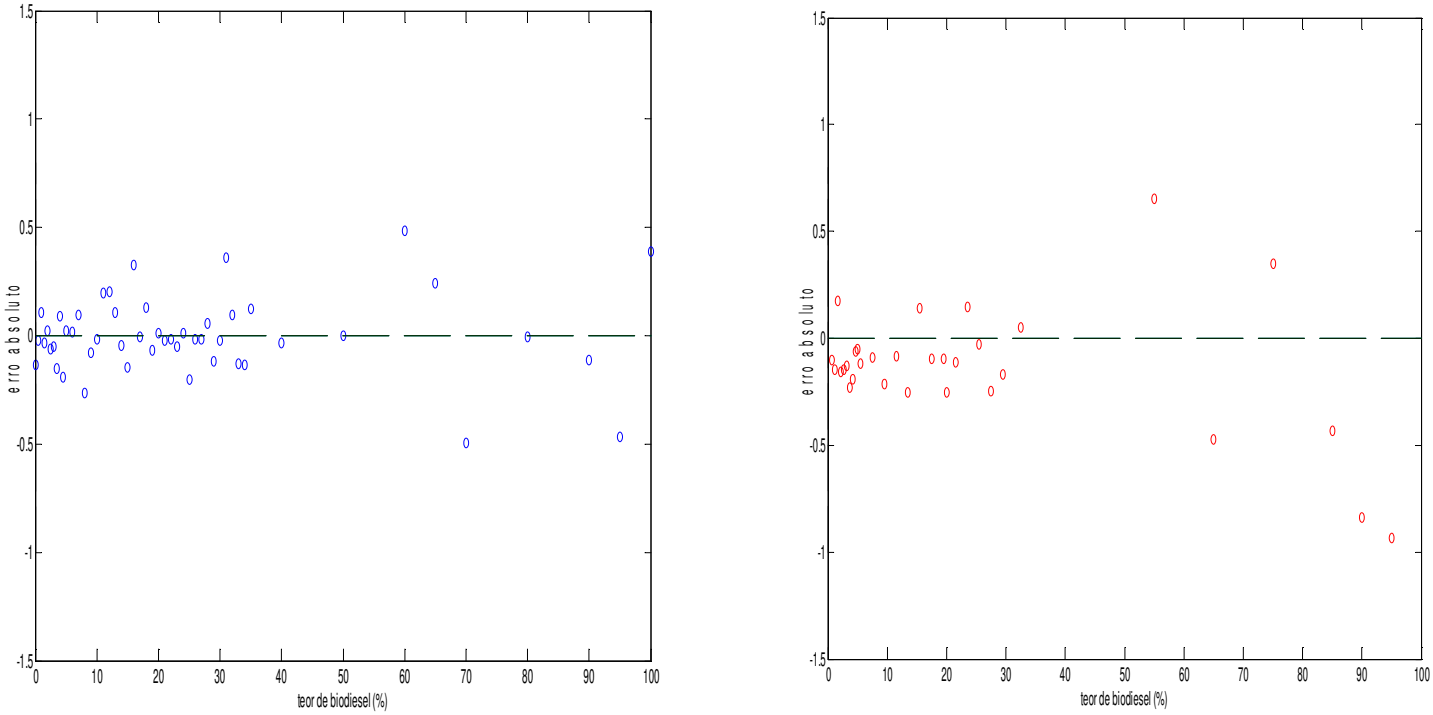
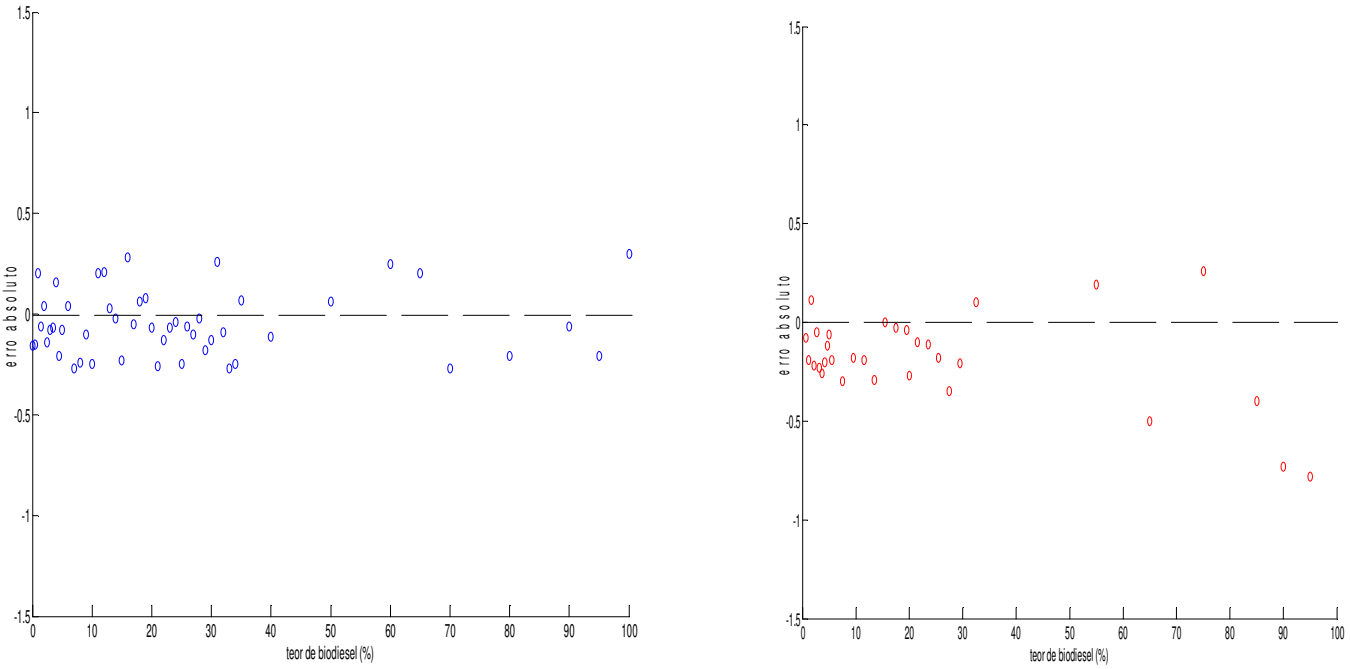


Figura 10.4 – Valores experimentais contra previstos para o modelo SVM com as 50 amostras de calibração (○) e as 31 amostras de validação (●).



(a) (b)

Figura 10.5 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS.

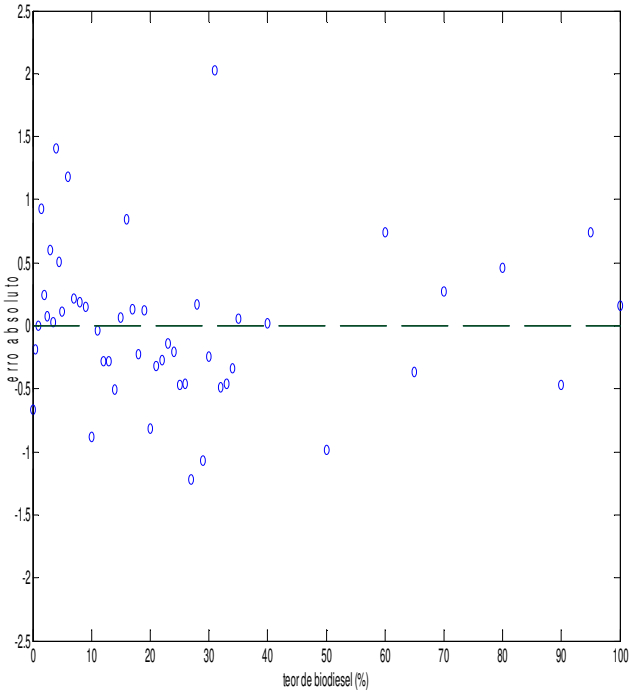


(a) (b)

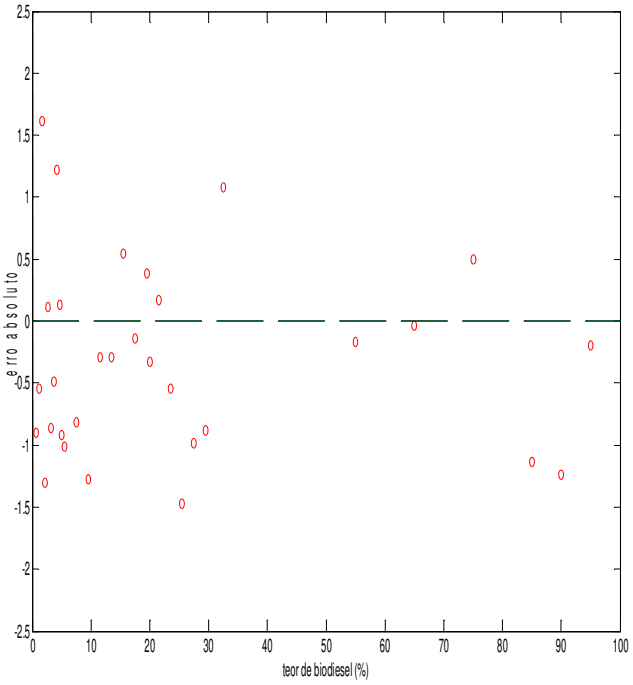
Figura 10.6 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM.

Tabela 10.2 – resultados de previsão dos modelos PLS e SVM para o teor de biodiesel 0 -100 % (v/v)

	PLS			SVM		
	nominal (%)	previsto (%)	erro relativo (%)	nominal (%)	previsto (%)	erro relativo (%)
1	0,70	0,60	-14,84	0,70	0,62	-12,01
2	1,20	1,05	-12,28	1,20	1,01	-16,14
3	1,70	1,87	10,13	1,70	1,81	6,41
4	2,20	2,04	-7,15	2,20	1,98	-10,21
5	2,70	2,55	-5,47	2,70	2,65	-1,82
6	3,20	3,07	-3,98	3,20	2,97	-7,16
7	3,70	3,47	-6,16	3,70	3,44	-7,14
8	4,20	4,01	-4,55	4,20	4,00	-4,79
9	4,70	4,64	-1,27	4,70	4,58	-2,45
10	5,00	4,95	-1,07	5,00	4,94	-1,10
11	5,50	5,38	-2,16	5,50	5,31	-3,51
12	7,50	7,41	-1,17	7,50	7,20	-3,97
13	9,50	9,28	-2,27	9,50	9,32	-1,91
14	11,50	11,42	-0,73	11,50	11,31	-1,62
15	13,50	13,24	-1,89	13,50	13,21	-2,14
16	15,50	15,64	0,91	15,50	15,50	-0,02
17	17,50	17,41	-0,53	17,50	17,47	-0,15
18	19,50	19,40	-0,50	19,50	19,46	-0,23
19	20,00	19,75	-1,26	20,00	19,73	-1,36
20	21,50	21,39	-0,52	21,50	21,40	-0,47
21	23,50	23,65	0,63	23,50	23,39	-0,45
22	25,50	25,47	-0,11	25,50	25,32	-0,70
23	27,50	27,25	-0,90	27,50	27,15	-1,26
24	29,50	29,33	-0,58	29,50	29,29	-0,73
25	32,50	32,55	0,16	32,50	32,60	0,30
26	55,00	55,65	1,19	55,00	55,19	0,34
27	65,00	64,53	-0,73	65,00	64,50	-0,77
28	75,00	75,35	0,46	75,00	75,26	0,35
29	85,00	84,57	-0,51	85,00	84,60	-0,48
30	90,00	89,16	-0,93	90,00	89,27	-0,81
31	95,00	94,07	-0,98	95,00	94,22	-0,82

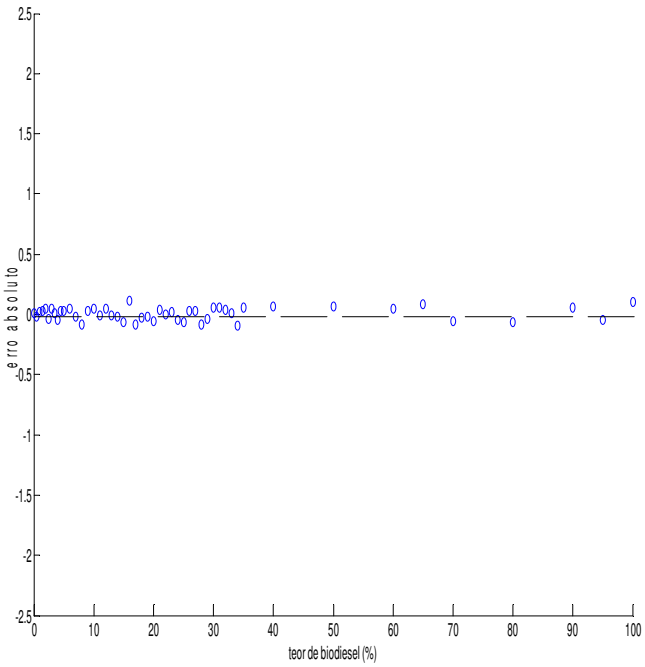


(a)



(b)

Figura 10.7 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS utilizando a região espectral (ii).



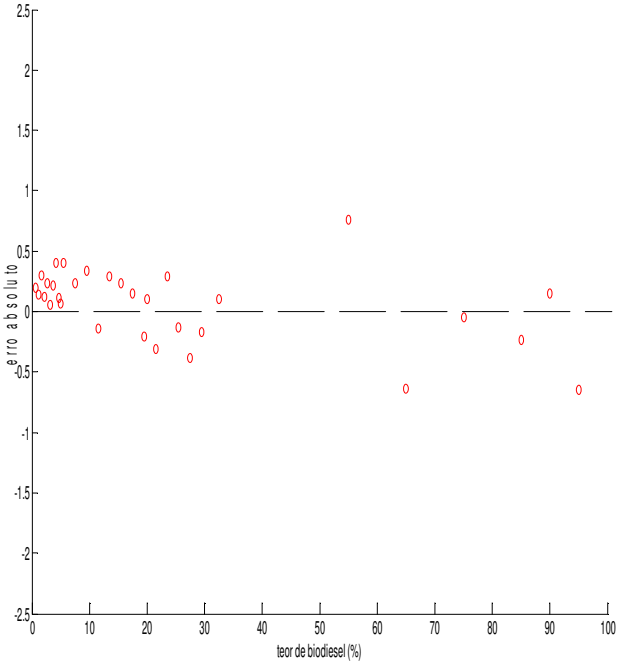


Figura 10.8 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM utilizando a região espectral (ii).

10.2.3 – Modelo PLS – teor de biodiesel 0 – 35 % (v/v)

O melhor resultado foi obtido utilizando-se a região espectral (iii) e o pré-processamento SNV. Utilizou-se 3 variáveis latentes, que explicam 99,9% da variância dos dados. Os resultados são mostrados nas tabela 10.3 e 10.4. A figura 10.9 ilustra o resultado do modelo ajustado com o algoritmo PLS. A figura 10.14 ilustra a distribuição dos erros de calibração e previsão obtidos.

Nesse modelo, embora se tenha obtido o melhor valor de RMSEP entre os modelos PLS desenvolvidos com as regiões espectrais e os diferentes pré-processamentos testados, observa-se uma tendência para desvios negativos nos valores de previsão. A utilização da região espectral (ii), que abrange as bandas de absorção em 4425 cm^{-1}, 4650 cm^{-1} e 6005 cm^{-1} e o pré-processamento primeira derivada permite a obtenção de um modelo em que se observa a distribuição aleatória dos erros de calibração e previsão, porém com valores de RMSEC e RMSEP muito superiores. Nesse modelo utilizou-se 2 variáveis latentes, que explicam 99,7% da variância dos dados. A figura 10.13 ilustra a distribuição dos erros de calibração e previsão do modelo ajustado com o algoritmo PLS utilizando a região espectral (ii) e os resultados são mostrados na tabela 10.3.

10.2.4 – Modelo SVM – teor de biodiesel 0 – 35 % (v/v)

O melhor modelo utilizando o algoritmo SVM foi obtido utilizando a região espectral (iii), a função kernel linear e os dados pré-processados com SNV. Os resultados são mostrados nas tabelas 10.3 e 10.4. A figura 10.10 ilustra o resultado obtido. Os parâmetros selecionados foram: C = 20 e ν = 0,007. Nesse modelo foram utilizados 12 vetores de suporte. O adequado número de vetores de suporte utilizados evidencia o bom ajuste do modelo, sem ocorrência de sobreajuste.

O modelo SVM obtido fornece um valor de RMSEP que é aproximadamente 13% melhor em relação ao valor obtido com o modelo PLS.

O valor do RMSEP obtido para os modelos com PLS e SVM atendem ao exigido pela norma ABNT para determinação de teores de biodiesel nas faixas de 0-8% e

8-30% (v/v) de biodiesel. Também, tomando o valor da reprodutibilidade estabelecida pelo método ASTM como comparação os valores de RMSEP obtidos para ambos os modelos ficam abaixo da reprodutibilidade mínima exigida para determinação de biodiesel a partir de 0% (v/v) na mistura com óleo diesel.

Tabela 10.3 – Resultados dos modelos PLS e SVM para 0-35 % (v/v) de biodiesel

Modelo	Região espectral (cm^{-1})	RMSEC (%) R^2	RMSEP (%)	Melhora do RMSEP (%)	ASTM D7371-07 reprodutibilidade (%)	ABNT-NBR 15568 RMSEP (%)
PLS	(ii) 4400 – 6200	0,5776 0,997	0,7636	13	0,76	0 – 8 % (v/v) = 0,1 8 – 30 % (v/v) = 1
	(iii) 4400 - 4600	0,1230 0,999	0,1298			
SVM	(ii) 4400 – 6200	0,0501 0,999	0,1208			
	(iii) 4400 - 4600	0,1178 0,999	0,1126			

Além dos valores de RMSEP e dos gráficos de valores medidos contra previstos das figuras 10.9 e 10.10, também pode-se verificar o melhor ajuste do modelo SVM, em relação ao PLS, através dos gráficos de resíduos para os conjuntos de calibração e validação dos modelos utilizando PLS e SVM mostrados nas figuras 10.11 e 10.12, respectivamente. Verifica-se que o modelo SVM possibilita um melhor ajuste ao longo de todo o espaço amostral, com a melhor distribuição e a redução dos resíduos de previsão.

A utilização da região espectral (ii) proporciona também bons resultados utilizando o SVM. Nesse modelo utilizou-se a função kernel linear e os dados com pré-processamento SNV. Foram selecionados os parâmetros C= 2 e ν=0,005 e utilizados 25 vetores de suporte. Os resultados são mostrados na tabela 10.3 e a distribuição dos erros dos conjunto de calibração e previsão são mostrados na figura 10.14.

Entre os citados modelos com SVM utilizando as regiões espectrais (ii) e (iii), a utilização da região (iii) permite não apenas um melhor valor de RMSEP como também utiliza menos vetores de suporte, o que evidencia um modelo melhor ajustado.

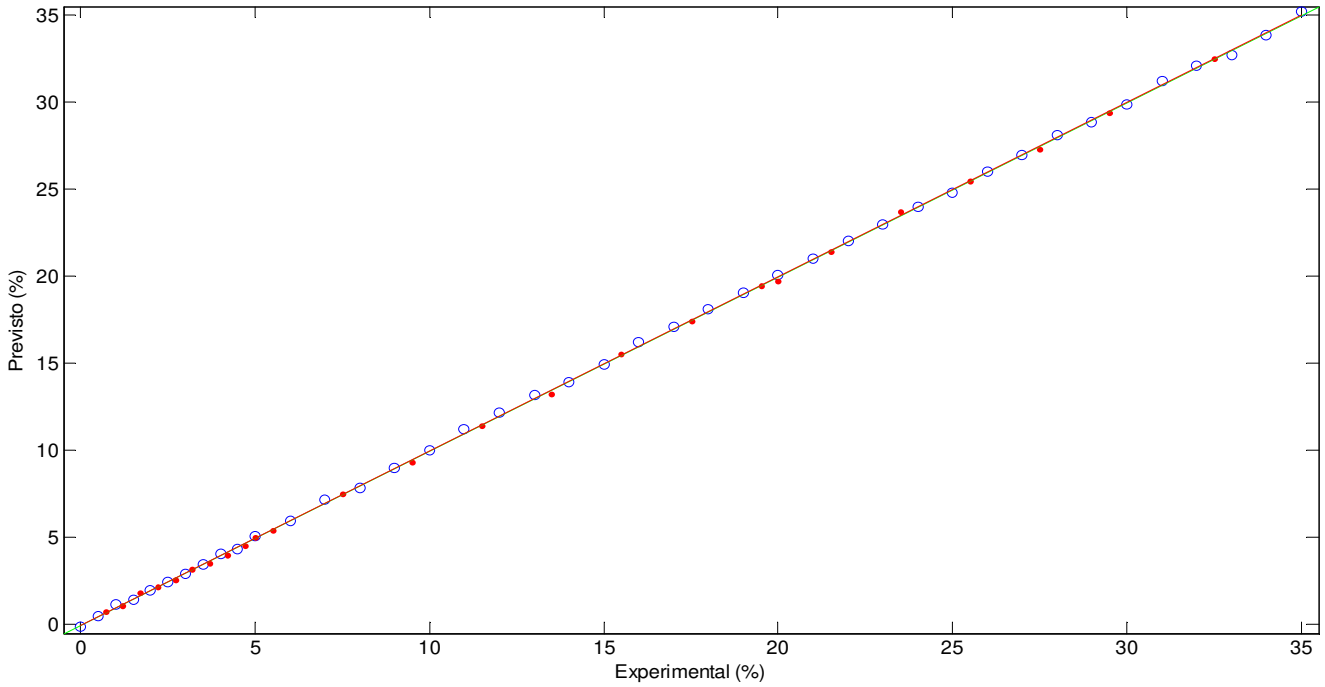


Figura 10.9 – Valores experimentais contra previstos para o modelo PLS com as 41 amostras de calibração (○) e as 25 amostras de validação (●).

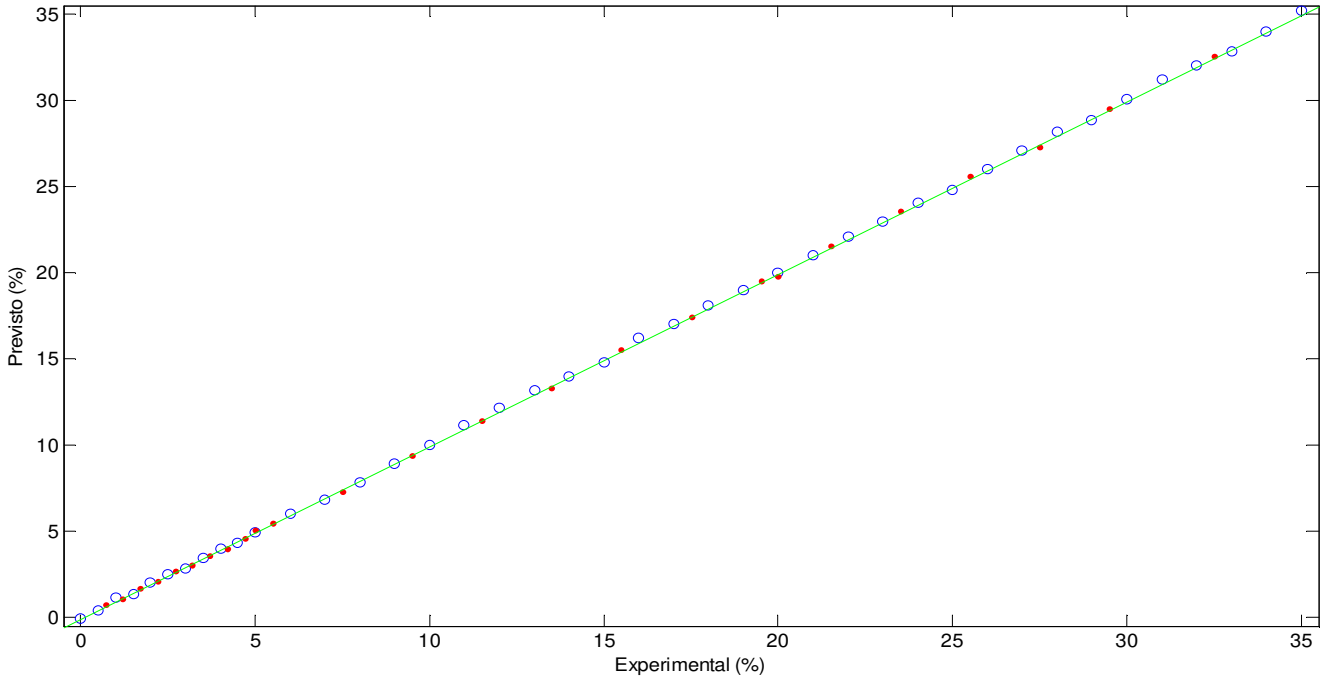
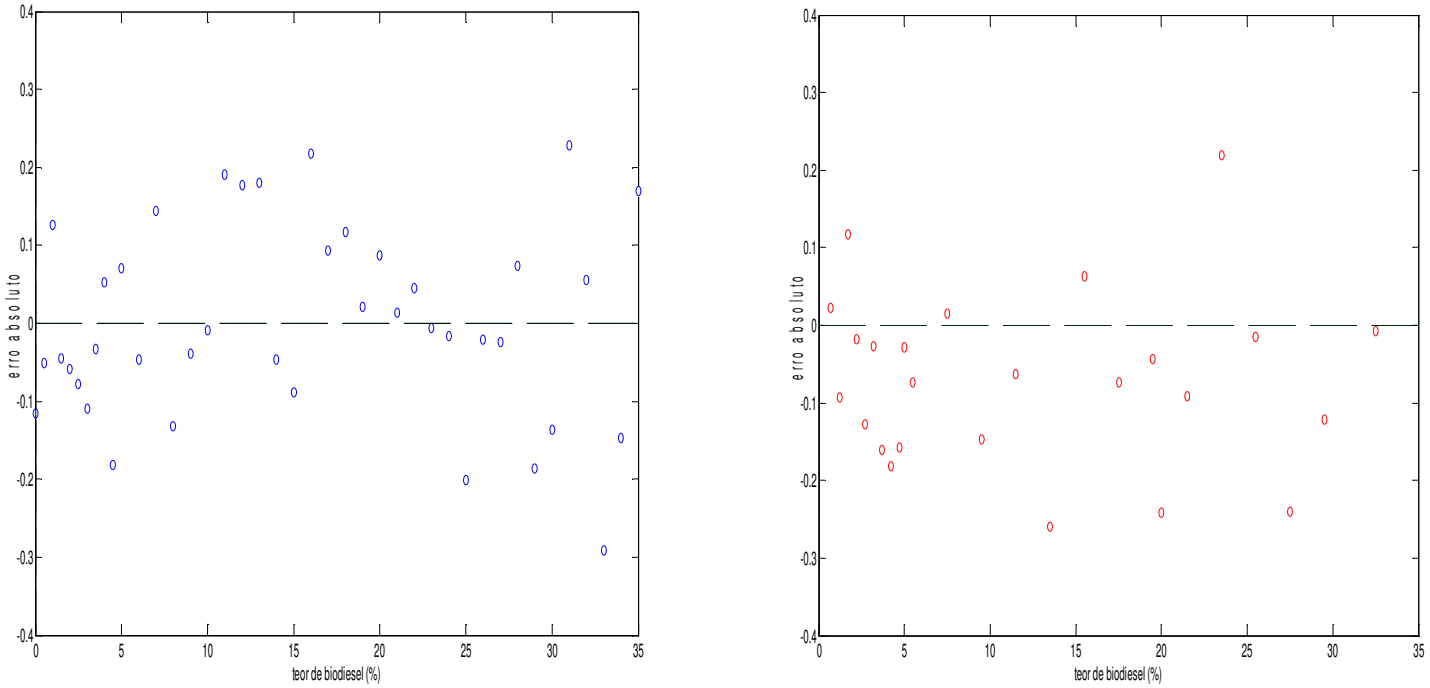
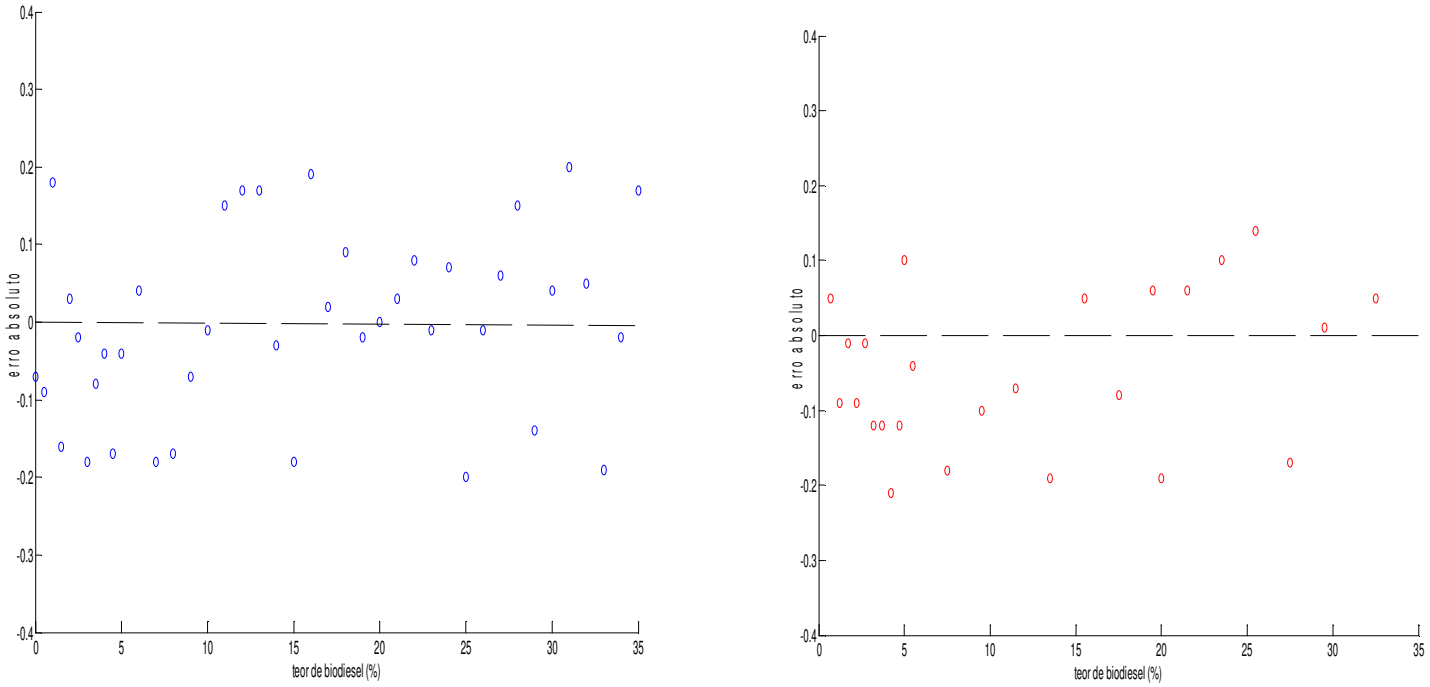


Figura 10.10 – Valores experimentais contra previstos para o modelo SVM com as 41 amostras de calibração (○) e as 25 amostras de validação (●).



(a) (b)

Figura 10.11 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS.

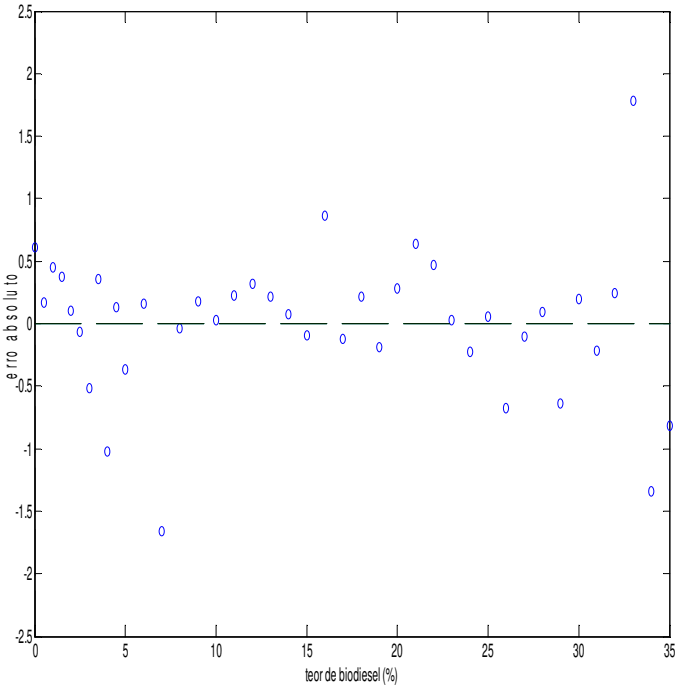


(a) (b)

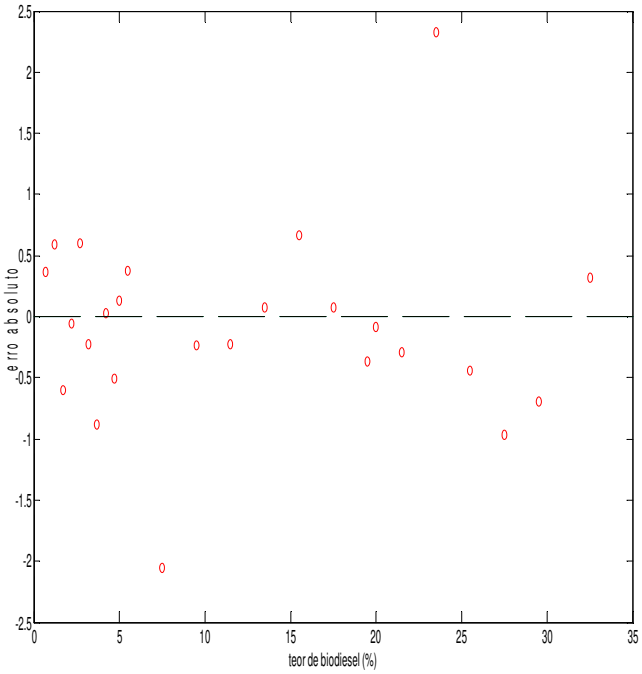
Figura 10.12 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM.

Tabela 10.4 – resultados de previsão dos modelos PLS e SVM para o teor de biodiesel 0 -35 % (v/v)

	PLS			SVM		
	nominal (%)	previsto (%)	erro relativo (%)	nominal (%)	previsto (%)	erro relativo (%)
1.	0,70	0,72	3,12	0,70	0,75	7,10
2.	1,20	1,11	-7,72	1,20	1,11	-7,53
3.	1,70	1,82	6,86	1,70	1,69	-0,63
4.	2,20	2,18	-0,83	2,20	2,11	-3,96
5.	2,70	2,57	-4,71	2,70	2,69	-0,21
6.	3,20	3,17	-0,85	3,20	3,08	-3,79
7.	3,70	3,54	-4,32	3,70	3,58	-3,13
8.	4,20	4,02	-4,32	4,20	3,99	-5,11
9.	4,70	4,54	-3,35	4,70	4,58	-2,49
10.	5,00	4,97	-0,56	5,00	5,10	1,95
11.	5,50	5,43	-1,34	5,50	5,46	-0,75
12.	7,50	7,51	0,20	7,50	7,32	-2,47
13.	9,50	9,35	-1,55	9,50	9,40	-1,08
14.	11,50	11,44	-0,54	11,50	11,43	-0,60
15.	13,50	13,24	-1,92	13,50	13,31	-1,43
16.	15,50	15,56	0,41	15,50	15,55	0,31
17.	17,50	17,43	-0,42	17,50	17,42	-0,45
18.	19,50	19,46	-0,22	19,50	19,56	0,29
19.	20,00	19,76	-1,21	20,00	19,81	-0,95
20.	21,50	21,41	-0,43	21,50	21,56	0,27
21.	23,50	23,72	0,93	23,50	23,60	0,43
22.	25,50	25,48	-0,06	25,50	25,64	0,54
23.	27,50	27,26	-0,87	27,50	27,33	-0,63
24.	29,50	29,38	-0,41	29,50	29,51	0,04
25.	32,50	32,49	-0,02	32,50	32,55	0,16

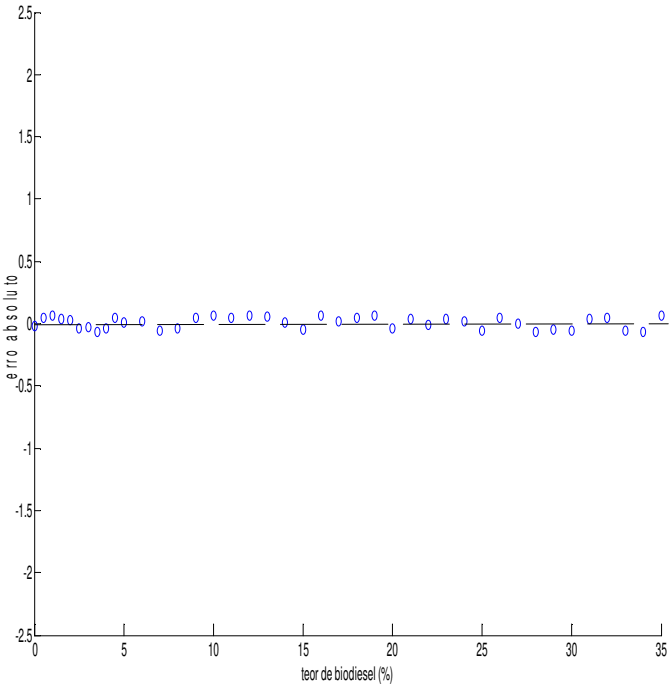


(a)

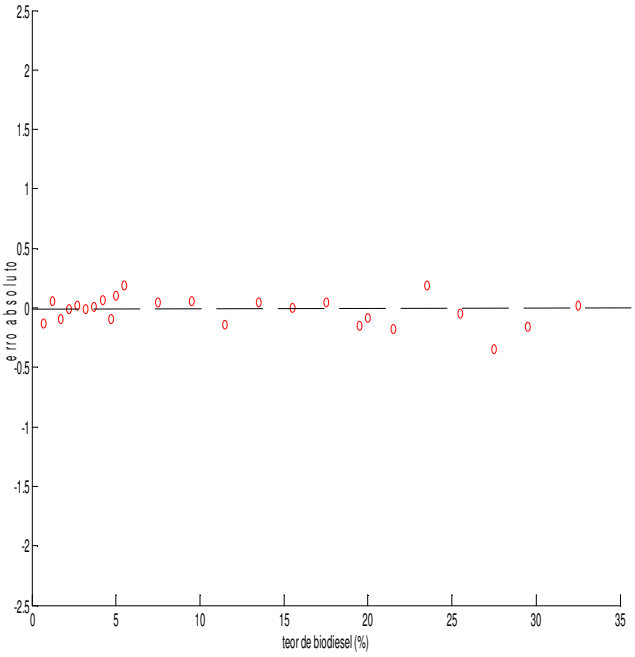


(b)

Figura 10.13 – Distribuição dos erros de calibração (a) e validação (b) do modelo PLS utilizando a região espectral (ii).



(a)



(b)

Figura 10.14 – Distribuição dos erros de calibração (a) e validação (b) do modelo SVM utilizando a região espectral (ii).

10.2.5 – Comparação dos resultados dos modelos SVM e PLS – teste F

O procedimento citado no item 7.2.3 foi aqui repetido para comparação entre os melhores modelos obtidos com PLS e SVM. O valores críticos de $F_{49,49}$ (teor de biodiesel de 0-100 % (v/v)) e $F_{40,40}$ (teor de biodiesel de 0-35 % (v/v)) ao nível de significância de 95 % são mostrados na tabela 10.5 juntamente com os valores calculados de F.

Tabela 10.5 – Resultados do teste F na comparação dos modelos PLS e SVM

modelo	F calculado	F crítico 95 %
0 – 100 % (v/v)	1,23	1,61
0 – 35 % (v/v)	1,33	1,69

Ao nível de significância de 95%, os modelos desenvolvidos com PLS e SVM fornecem resultados semelhantes. Essa constatação informa que os erros médios na estimativa do teor de biodiesel dos dois métodos são equivalentes. Porém observa-se que os modelos com SVM proporcionam uma melhor distribuição dos resíduos de calibração e validação bem como menores erros de previsão para as amostras mais extremas, evidenciando o melhor ajuste ao longo de todo o intervalo analítico.

10.3 – Conclusões

A utilização da região espectral (iii), que inclui a banda de absorção em 4425 cm^{-1}, relacionada ao grupo metila terminal de éster, proporciona a obtenção de modelos muito semelhantes com PLS e SVM, no entanto os modelos SVM em vista de sua melhor capacidade de ajuste ao longo de todo o espaço analítico fornece melhoras nos valores de RMSEP de 10% e 13%, para os modelos de 0-100% e 0-35% de biodiesel, respectivamente.

Nesse estudo obteve-se os melhores resultados utilizando o SVM e a função kernel linear. Tanto nos modelos PLS como nos modelos SVM obtiveram-se valores de coeficiente de correlação linear muito próximos ou iguais a 1. A utilização da função kernel linear nos modelos SVM demonstra que o mapeamento não linear dos dados do espaço de entrada para um espaço de elevada dimensão não possibilita os melhores resultados nesse caso. No entanto o algoritmo SVM utiliza a função de custo ε-insensível e permite a otimização de dois parâmetros do modelo, o que proporcionou o melhor ajuste dos modelos para o teor de biodiesel em relação ao PLS.

A utilização de reduzido número de variáveis (região espectral (iii) = 200 variáveis) em lugar da utilização de maior número de variáveis (região espectral (ii) = 1800 variáveis) também não proporciona um melhor desempenho da função kernel RBF em relação a função kernel linear, sugerindo que não é necessário o mapeamento não linear dos dados.

O melhor resultado fornecido pela função kernel linear não surpreende, uma vez que em estudos mostrados e discutidos nos capítulos 7 a 9 verifica-se que para seis dos nove parâmetros estudados obtiveram-se melhores resultados utilizando o SVM e o kernel linear em relação aos resultados obtidos com PLS, embora os resultados com kernel RBF e polinomial sejam ainda melhores do que os resultados com o kernel linear devido as peculiaridades dessas correlações que sugerem certo grau de não linearidade.

Embora todas as regiões espectrais testadas possibilitem a calibração do teor de biodiesel na mistura com óleo diesel, a região espectral (ii) proporciona solucionar a tendência de erros de previsão negativos dos modelos PLS, embora com sensível piora nos valores de RMSEP. Utilizando o SVM a região espectral (ii) fornece resultados muito semelhantes aos obtidos com a região espectral (iii) em termos de valores de RMSEP, evidenciando a melhor capacidade de ajuste do algoritmo SVM.

Os valores de RMSEP obtidos com os modelos PLS e SVM para determinação do teor de biodiesel entre 0-35% estão de acordo com o estabelecido pelo método de referência ABNT para determinação do teor de biodiesel nos modelos sugeridos entre 0-8% e entre 8-30% de biodiesel, com a vantagem de que os modelos ora desenvolvidos proporcionam a previsão em um intervalo amostral bem maior, possibilitando a utilização de apenas um modelo de calibração. Também, o valores de

RMSEP obtidos com os modelos PLS e SVM para determinação do teor de biodiesel entre 0-100% estão de acordo com o estabelecido pelo método de referência ABNT para determinação do teor de biodiesel a partir de 8%.

Comparando-se os valores de RMSEP obtidos com os modelos PLS e SVM com os valores de reprodutibilidade estabelecidos pelo método ASTM de referência, esses modelos também se mostram adequados e novamente têm a vantagem de possibilitar previsões em intervalos amostrais maiores do que o citado no método de referência, permitido a utilização de apenas um modelo de calibração.

Seção III – Modelos de classificação

11 – Classificação das frações que compõem o *pool* de óleo diesel através de dados de espectroscopia NIR e SVM

O desenvolvimento de eficazes modelos de classificação de múltiplas classes pode ser útil para o controle de qualidade e identificação de adulteração em óleo diesel, podendo servir como uma análise preliminar, para se necessário, posterior procedimento de análises mais específicas.

Nesse trabalho realizou-se um estudo sobre a performance do algoritmo SVM para o desenvolvimento de modelos de classificação com múltiplas classes, comparando-se os resultados com os obtidos com o algoritmo SIMCA, considerado de referência.

Obtiveram-se modelos de classificação de múltiplas classes com SVM e dados de espectroscopia NIR dos diferentes tipos de correntes que fazem parte do *pool* de óleo diesel na refinaria de Paulinia – Replan: diesel leve, diesel pesado, diesel HDT, querosene (QAV), nafta pesada e diesel externo. Além dessas seis correntes acrescentou-se o diesel produto final da refinaria, proveniente da mistura das mesmas.

11.1 – Parte experimental

Foram obtidos espectros na região do infravermelho próximo para seis diferentes correntes que compõem o *pool* de óleo diesel da refinaria de Paulínia – Replan e para o diesel produto final, proveniente da mistura dessas correntes.

Para obtenção dos espectros na região do infravermelho foi utilizado um espectrômetro ABB/Bomen MID/NIR com fonte glowbar (carbeto de silício) e detector de sulfato de triglicina deuterada (DTGS), usando uma cubeta de transmitância de CaF_2 de caminho óptico igual a 0,5 mm. Cada espectro foi obtido como uma média de 32 varreduras, com resolução de 4 cm^{-1}.

Para obtenção dos modelos foi utilizada a região de 3504cm^{-1} a 4466 cm^{-1}, que corresponde a região das bandas de combinação de modos vibracionais do grupo C-H.

A figura 11.1 mostra os espectros das 322 amostras utilizadas.

Foram realizados diferentes pré-processamentos dos dados para verificar qual proporciona a construção do melhor modelo utilizando os algoritmos SIMCA e SVM. Os pré-processamentos testados foram: correção de linha base WLS e centragem na média; SNV e centragem na média; e primeira derivada (janela com 15 pontos) e centragem na média.

O pacote LIBSVM[85] versão 2.88 foi utilizado para o desenvolvimento dos modelos com SVM e é adequado para utilização com Matlab 7.7 da Mathworks.

Para obtenção dos modelos de classificação com SVM foi utilizada a função kernel RBF. Testou-se a utilização dos algoritmos C-SVC e ν-SVC.

O parâmetro γ do kernel RBF foi selecionado juntamente com o parâmetro C ou ν do SVM. Os parâmetros C e γ (para o C-SVC) foram selecionados entre os intervalos de 0 a 1200 e 0,01 a 100, respectivamente. Os parâmetros ν e γ (para o ν-SVC) foram selecionados entre os intervalos 0,001 a 1 e 0,01 a 100, respectivamente. A seleção dos parâmetros foi realizada através do método *grid search*, visando a minimização do erro de validação cruzada no conjunto de treinamento.

Como a minimização do erro de validação cruzada no conjunto de treinamento não garante a obtenção da melhor exatidão para o conjunto de validação, eventualmente um refinamento manual dessa seleção pode ser necessário, sempre considerando a utilização do adequado número de vetores de suporte de modo a evitar um sobreajuste do modelo.

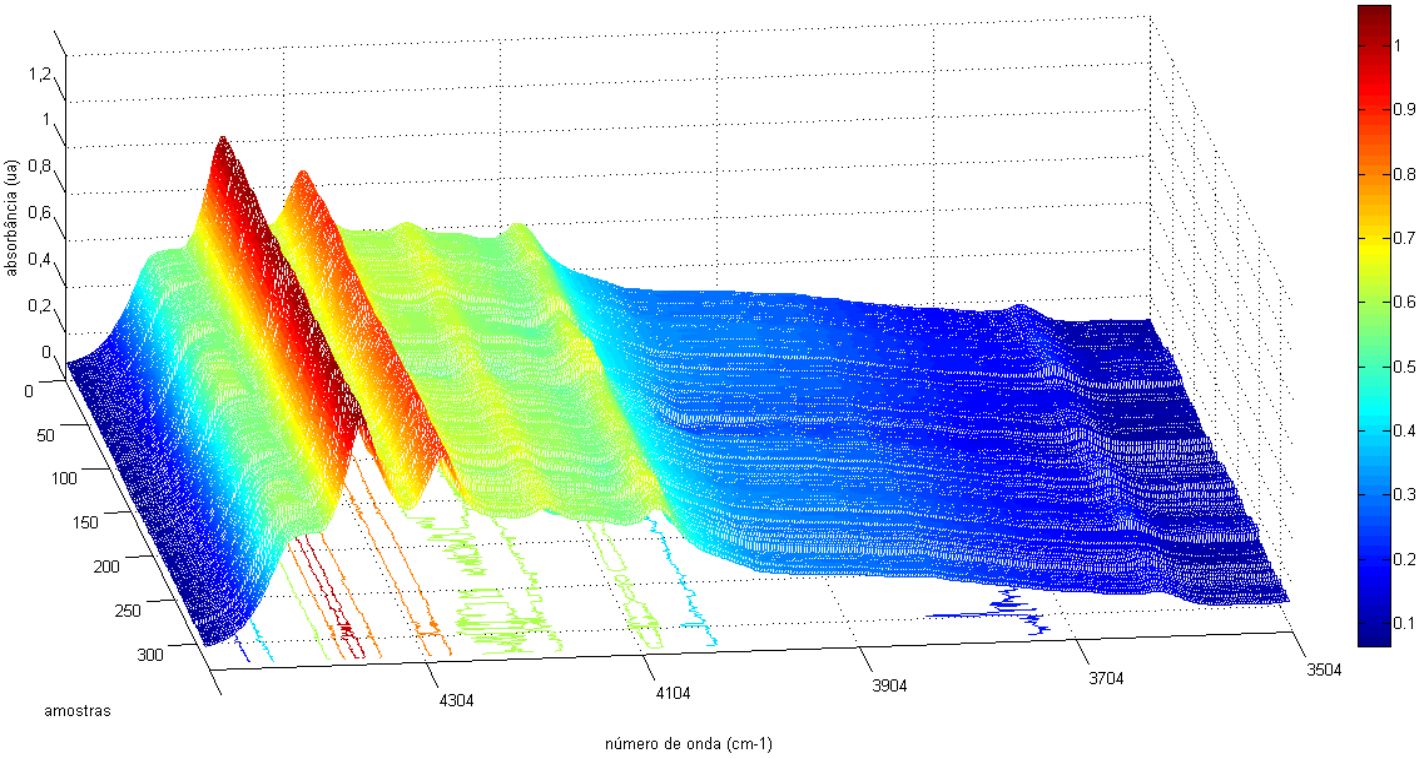


Figura 11.1 – Espectros das 322 amostras utilizadas no conjunto de dados E, que inclui as 7 classes estudadas.

11.2 – Resultados e discussão

Foram desenvolvidos cinco diferentes modelos de classificação com 2, 4, 6, 6 e 7 classes. Os diferentes conjuntos de dados dos modelos de classificação desenvolvidos são mostrados na tabela 11.1.

A região espectral utilizada, de 3504 cm^{-1} a 4466 cm^{-1}, corresponde a região das bandas de combinação de modos vibracionais do grupo C-H. Essa região possui diversas bandas de absorção atribuídas a combinação de modos vibracionais da ligação C-H de grupos metil e metileno em alcanos, ligação C-H em cicloalcanos e ligação C-H de anéis aromáticos. As atribuições de algumas bandas nessa região espectral podem ser verificadas na tabela 2.1.

Essa região espectral permite a classificação das correntes componentes do *pool* de óleo diesel porque cada corrente está associada a um maior ou menor teor de compostos hidrocarbonetos de cadeia carbonica alifática ou parafínica, cíclica ou

naftênica e aromática conforme a faixa de temperatura de destilação da correspondente fração do petróleo.

Nas frações leves, predominam hidrocarbonetos normais parafínicos, até cerca de 120 ºC, quando começa a ocorrer o equilibro com os ramificados. Em cerca de 150 ºC, em geral, passa a ocorrer predomínio dos parafínicos ramificados sobre os normais. Para um mesmo petróleo a ocorrência dos hidrocarbonetos parafínicos é maior para as frações mais leves, naftas, onde são majoritários, e bem menor para as frações mais pesadas, resíduos, onde em geral são minoritários.

Os cicloalcanos ou naftênicos ocorrem de forma majoritária nas frações médias, querosene, gasóleos atmosférico e de vácuo.

Os aromáticos são em geral minoritários nas frações leves e médias, ocorrendo em maior quantidade nas frações pesadas e residuais, onde, dependendo do tipo de petróleo podem ser as majoritárias. Entre os compostos mono-aromáticos, os alquil benzênicos são os maiores constituintes, entre estes pode-se citar o tolueno, etilbenzeno e xilenos, com pontos de ebulição entre 80 e 140 ºC. Os compostos com dois anéis aromáticos, como naftaleno, estão mais presentes nas frações médias e os compostos com três e quatro anéis aromáticos se concentram nas frações mais pesadas.

A tabela 4.1 mostra as faixas de destilação das frações do petróleo e a tabela 4.3 fornece uma referência quanto a ocorrência dos diferentes tipos de hidrocarbonetos em função da faixa de destilação.

Assim, com a variação dos teores de diferentes tipos de hidrocarbonetos presentes nas diferentes frações, tem-se a variação da quantidade dos grupos metil e metileno de cadeias normal, ramificada, cíclica e aromática e suas respectivas intensidades de sinais no espectro NIR permitem a classificação das diferentes classes.

Tabela 11.1 - Conjuntos de classificação para as diferentes correntes do pool de óleo diesel e óleo diesel produto final

	Conjunto A	Conjunto B	Conjunto C	Conjunto D	Conjunto E
Número de classes	2	4	6	6	7
Número de amostras: total conj. de calibração conj. de validação	 56 38 18	 191 138 53	 205 195 10	 294 193 101	 322 213 109
Classes	1 - diesel interno (REPLAN) 2- diesel externo (REVAP)	1 – diesel leve 2 – diesel pesado 3 – diesel HDT 4 – querosene (QAV)	1 – diesel leve 2 – diesel pesado 3 – diesel HDT 4 – querosene (QAV) 5 – nafta pesada 6 – diesel interno	1 – diesel leve 2 – diesel pesado 3 – diesel HDT 4 – querosene (QAV) 5 – nafta pesada 6 – diesel externo	1 – diesel leve 2 – diesel pesado 3 – diesel HDT 4 – querosene (QAV) 5 – nafta pesada 6 – diesel externo 7 – diesel interno
Distribuição das amostras por classe	1 – 28 (cal.: 19; val.: 9) 2 – 28 (cal.: 19; val.: 9)	1 – 59 (cal.: 40; val.: 19) 2 – 46 (cal.: 35; val.: 11) 3 – 54 (cal.: 40; val.: 14) 4 – 32 (cal.: 23; val.: 9)	1 – 40 (cal.: 40; val.: 0) 2 – 35 (cal.: 35; val.: 0) 3 – 45 (cal.: 40; val.: 5) 4 – 28 (cal.: 26; val.: 2) 5 – 26 (cal.: 26; val.: 0) 6 – 31 (cal.: 28; val.: 3)	1 – 60 (cal.: 40; val.:20) 2 – 53 (cal.: 35; val.: 18) 3 – 60 (cal.: 40; val.: 20) 4 – 39 (cal.: 26; val.: 13) 5 – 41 (cal.: 26; val.: 15) 6 – 41 (cal.: 26; val.: 15)	1 – 60 (cal.: 40; val.:20) 2 – 53 (cal.: 35; val.: 18) 3 – 60 (cal.: 40; val.: 20) 4 – 39 (cal.: 26; val.: 13) 5 – 41 (cal.: 26; val.: 15) 6 – 41 (cal.: 26; val.: 15) 7 – 28 (cal.: 20; val.: 8)

11.2.1 – Conjunto A - Modelo de classificação para 2 classes

Esse conjunto de dados foi composto de espectros de amostras de óleo diesel produto final da mistura das correntes que fazem parte do *pool* de óleo diesel da Replan e de amostras de óleo diesel externo, produzido na refinaria Henrique Lage -Revap em São José dos Campos/SP.

O número de amostras utilizado é mostrado na tabela 11.1 e os resultados obtidos para os modelos utilizando SIMCA e SVM são mostrados na tabela 11.2.

O melhor modelo obtido com SIMCA utilizou os dados pré-processados com correção de linha base e centrados na média.

O melhor modelo obtido com SVM utilizou os dados pré-processados com correção de linha e centrados na média. A utilização do SVM através do C-SVC proporcionou o melhor resultado. Utilizando os parâmetros selecionados C = 32 e $\gamma = 16$ obteve-se um modelo que utiliza 23 vetores de suporte, sendo 12 e 11 vetores de suporte para as classes 1 e 2, respectivamente.

Tabela 11.2 – Classificação do óleo diesel no Conjunto A

	SIMCA		SVM	
	$N_{acertos}/N_{total}$	Erros (%)	$N_{acertos}/N_{total}$	Erros (%)
Diesel interno	7/9	22	8/9	11
Diesel externo	8/9	11	8/9	11
total	15/18	17	16/18	11

Para um problema de classificação de apenas duas classes, verificou-se que o modelo SVM tem resultados um pouco melhores, embora os resultados obtidos com os algoritmos SIMCA e SVM sejam muito semelhantes.

Testou-se a utilização de um conjunto de calibração com menor número de amostras, sendo cinco amostras a menos para cada classe. Os dois algoritmos tiveram performance idêntica e com 14 % de erros.

11.2.2 – Conjunto B - Modelo de classificação para 4 classes

Esse conjunto de dados foi composto de espectros de amostras de quatro diferentes correntes de frações do petróleo que fazem parte do *pool* de óleo diesel da Replan e que compõem o óleo diesel produto final.

O número de amostras utilizado para cada diferente corrente é mostrado na tabela 11.1 e os resultados obtidos para os modelos utilizando SIMCA e SVM são mostrados na tabela 11.3.

Tabela 11.3 – Classificação das correntes do *pool* de óleo diesel no Conjunto B

	SIMCA		SVM	
	$N_{acertos}/N_{total}$	Erros (%)	$N_{acertos}/N_{total}$	Erros (%)
Diesel leve	19/19	0	19/19	0
Diesel pesado	10/11	9	10/11	9
Diesel HDT	12/14	14	14/14	0
Querosene (QAV)	8/9	11	8/9	11
total	49/53	8	51/53	4

O melhor modelo obtido com SIMCA utilizou os dados pré-processados com primeira derivada SavGol e centrados na média.

O melhor modelo obtido com SVM utilizou os dados pré-processados com correção de linha e centrados na média. A utilização do SVM através do C-SVC proporcionou o melhor resultado. Utilizando os parâmetros selecionados C = 8 e $\gamma = 32$ obteve-se um modelo que utiliza 51 vetores de suporte, sendo 17;5;17 e 12 vetores de suporte para as classes 1, 2, 3 e 4, respectivamente.

Para um problema de classificação com quatro classes verifica-se que o algoritmo SVM proporciona resultados melhores que os obtidos com SIMCA.

Com esse conjunto de dados, obteve-se para os três pré-processamentos testados e para o C-SVC e ν-SVC o mesmo número de acertos. O modelo citado foi considerado o melhor devido a utilização do menor número de vetores de suporte, considerado o mais adequado por ser mais robusto e evitar a possibilidade de sobreajuste do modelo.

11.2.3 – Conjunto C - Modelo de classificação para 6 classes

Esse conjunto de dados foi composto de espectros de amostras de cinco diferentes correntes de frações do petróleo que fazem parte do *pool* de óleo diesel da Replan e ainda de espectros de amostras de óleo diesel oriundas do produto do misturador de processo.

O número de amostras utilizado para cada diferente corrente é mostrado na tabela 11.1 e os resultados obtidos para os modelos utilizando SIMCA e SVM são mostrados na tabela 11.4.

Tabela 11.4 – Classificação das correntes do *pool* de óleo diesel e do óleo diesel produto final no Conjunto C

	SIMCA		SVM	
	$N_{acertos}/N_{total}$	Erros (%)	$N_{acertos}/N_{total}$	Erros (%)
Diesel leve	-	-	-	-
Diesel pesado	-	-	-	-
Diesel HDT	4/5	20	5/5	0
Querosene (QAV)	2/2	0	2/2	0
Nafta pesada	-	-	-	-
Diesel interno	0/3	100	3/3	0
total	6/10	40	10/10	0

Nesse conjunto de dados o conjunto de validação contém misturas simuladas preparadas em laboratório conforme mostra a tabela 11.5 e amostras oriundas do produto do misturador de processo. Tais amostras podem ser classificadas em três diferentes classes pertencentes às seis classes incluídas no conjunto de calibração.

Tabela 11.5 – Composição das amostras de validação com misturas simuladas

amostra	Frações das correntes nas amostras			classe
	Diesel HDT	Querosene (QAV)	Nafta pesada	
1	1	0	0	3
2	0,7	0	0,3	3
3	0,2	0,5	0,3	4
4	0,75	0,25	0	3
5	0,6	0,25	0,15	3
6	0,6	0,25	0,15	3
7	0,3	0,6	0,1	4
8	misturador			6
9	misturador			6
10	misturador			6

O melhor modelo obtido com SIMCA utilizou os dados pré-processados com SNV e centrados na média.

O melhor modelo obtido com SVM utilizou os dados pré-processados com primeira derivada SavGol e centrados na média. A utilização do SVM através do ν-SVC proporcionou o melhor resultado. Utilizando os parâmetros selecionados ν=0,285 e γ=19 obteve-se um modelo que utiliza 124 vetores de suporte, sendo 28;15;28;19;10 e 24 vetores de suporte para as classes 1, 2, 3, 4, 5 e 6, respectivamente.

Para esse complexo problema de classificação com seis classes verifica-se que o algoritmo SVM proporciona resultados muito melhores que os obtidos com SIMCA.

11.2.4 – Conjunto D - Modelo de classificação para 6 classes

Esse conjunto de dados foi composto de espectros de amostras de cinco diferentes correntes de frações do petróleo que fazem parte do *pool* de óleo diesel da Replan e de amostras de óleo diesel externo produzido na Revap e que também compõe o óleo diesel produto final.

O número de amostras utilizado para cada diferente corrente é mostrado na tabela 11.1 e os resultados obtidos para os modelos utilizando SIMCA e SVM são mostrados na tabela 11.6.

Tabela 11.6 – Classificação das correntes do *pool* de óleo diesel no Conjunto D

	SIMCA		SVM	
	$N_{acertos}/N_{total}$	Erros (%)	$N_{acertos}/N_{total}$	Erros (%)
Diesel leve	18/20	10	20/20	0
Diesel pesado	8/18	56	17/18	6
Diesel HDT	13/20	35	20/20	0
Querosene (QAV)	12/13	8	12/13	8
Nafta pesada	15/15	0	10/15	33
Diesel externo	0/15	100	11/15	27
total	66/101	35	90/101	11

O melhor modelo obtido com SIMCA utilizou os dados pré-processados com primeira derivada e centrados na média.

A figura 11.2 ilustra a Análise de Componentes Principais do conjunto de calibração do conjunto de dados D (pré-processamento primeira derivada e centrados na média).

O melhor modelo obtido com SVM utilizou os dados pré-processados com primeira derivada e centrados na média. A utilização do SVM através do ν-SVC proporcionou o melhor resultado. Utilizando os parâmetros selecionados ν=0,0625 e γ=16 obteve-se um modelo que utiliza 52 vetores de suporte, sendo 12;6;10;9;6 e 9 vetores de suporte para as classes 1, 2, 3, 4, 5 e 6, respectivamente.

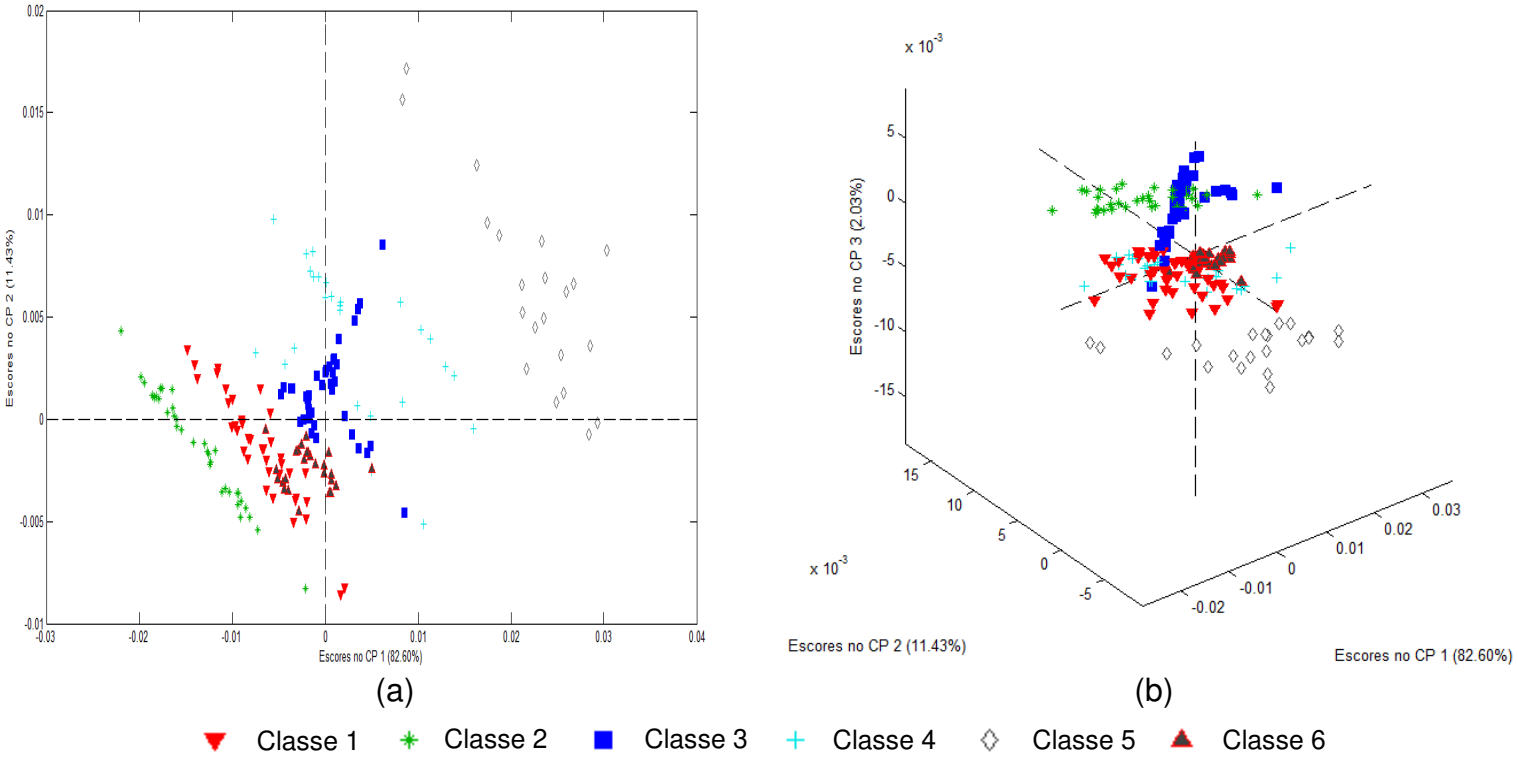


Figura 11.2 – PCA para o conjunto de dados D. (a) gráfico dos escores dos 2 primeiros componentes principais e (b) gráficos dos escores dos 3 primeiros componentes principais.

Para esse complexo problema de classificação com seis classes verifica-se que o algoritmo SVM proporciona um número de acertos muito superior em relação ao resultado obtido com SIMCA, fornecendo 24 % a mais de previsões corretas.

11.2.5 – Conjunto E - Modelo de classificação para 7 classes

Esse conjunto de dados foi composto de espectros de amostras de cinco diferentes correntes de frações do petróleo que fazem parte do *pool* de óleo diesel da Replan, de óleo diesel externo produzido na Revap e que também compõe o óleo diesel produto final e de amostras do produto do misturador de processo.

O número de amostras utilizado para cada diferente corrente é mostrado na tabela 11.1 e os resultados obtidos para os modelos utilizando SIMCA e SVM são mostrados na tabela 11.7.

Tabela 11.7 – Classificação das correntes do *pool* de óleo diesel e do óleo diesel produto final no Conjunto E

	SIMCA		SVM	
	$N_{acertos}/N_{total}$	Erros (%)	$N_{acertos}/N_{total}$	Erros (%)
Diesel leve	18/20	20	20/20	0
Diesel pesado	7/18	61	14/18	22
Diesel HDT	13/20	35	16/20	20
Querosene (QAV)	12/13	8	12/13	8
Nafta pesada	14/15	7	10/15	33
Diesel externo	2/15	87	8/15	47
Diesel interno	4/8	50	6/8	25
total	70/109	36	86/109	21

O melhor modelo obtido com SIMCA utilizou os dados pré-processados com correção de linha base e centrados na média.

A figura 11.3 ilustra a Análise de Componentes Principais do conjunto de calibração do conjunto de dados E (pré-processamento correção de linha base e centragem na média).

O melhor modelo obtido com SVM utilizou os dados pré-processados com primeira derivada SavGol e centrados na média. A utilização do SVM através do ν-SVC proporcionou o melhor resultado. Utilizando os parâmetros selecionados $\nu = 0{,}18$ e $\gamma = 60$ obteve-se um modelo que utiliza 106 vetores de suporte, sendo 22;9;17;14;10; 19 e 15 vetores de suporte para as classes 1, 2, 3, 4, 5, 6, e 7, respectivamente.

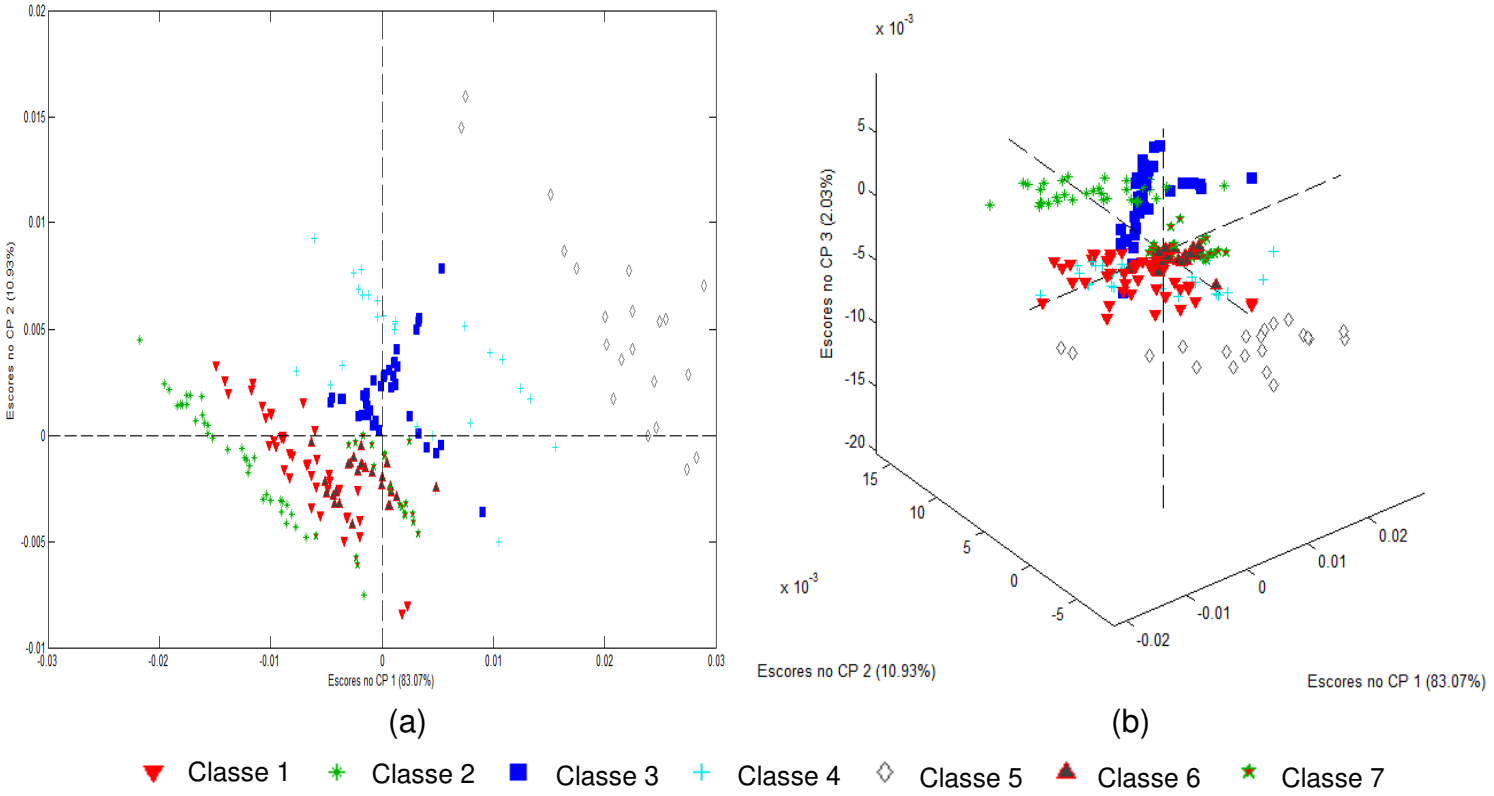


Figura 11.3 – PCA para o conjunto de dados E. (a) gráfico dos escores dos 2 primeiros componentes principais e (b) gráficos dos escores dos 3 primeiros componentes principais.

Para esse complexo problema de classificação com sete classes verifica-se que o algoritmo SVM proporciona um número de acertos muito superior em relação ao obtido com SIMCA, fornecendo 15 % a mais de previsões corretas.

11.3 – Conclusão

A utilização do algoritmo SVM para classificação de múltiplas classes de dados mostrou uma performance muito superior a obtida com SIMCA para problemas de classificação com seis e sete classes.

Com duas e quatro classes, embora os resultados com SVM sejam melhores, são próximos dos resultados obtidos com SIMCA.

Nesse trabalho, os modelos de classificação com SVM utilizaram um adequado número de vetores de suporte, de modo a evitar um sobreajuste dos modelos. Esse número é próximo de no máximo dois terços das amostras utilizadas no conjunto de calibração.

Verificou-se que em relação ao algoritmo C-SVC o algoritmo ν-SVC proporcionou o desenvolvimento de modelos de classificação mais eficazes com seis e sete classes. A excelente performance do ν-SVC relaciona-se a sua característica de controle mais eficaz do número de vetores de suporte utilizados, permitindo um melhor ajuste do modelo.

Com o desenvolvimento de modelos de classificação de múltiplas classes de dados mais eficazes eles podem ser utilizados em controle de qualidade na refinaria, onde busca-se a melhor relação da mistura de óleo diesel que proporciona a produção do derivado especificado utilizando na mistura as correntes de frações do petróleo disponíveis mais adequadas, com base nas diversas variáveis do processo.

Esse tipo de modelo também é útil para identificação de adulterações no óleo diesel, que podem ocorrer antes ou depois da adição de uma fração de biodiesel nas bases de distribuição das companhias distribuidoras.

12 – Classificação de óleos lubrificantes de motor quanto a presença de óleo naftênico e óleo vegetal

A qualidade de óleos lubrificantes para motores automotivos disponíveis no mercado é frequentemente monitorada pela ANP em seu Programa de Monitoramento da Qualidade de Lubrificantes – PMQL, em que eventualmente se constata a presença de óleo básico naftênico e/ou de óleo vegetal nos produtos comercializados, o que compromete a qualidade do óleo lubrificante para esse tipo de aplicação podendo ocasionar danos no motor.

A determinação do teor de carbono parafínico (C_P), carbono naftênico (C_N) e carbono aromático (C_A) em óleos básicos pode ser feito pelo método de referência ASTM D3238-95 através de cálculo que envolve quatro parâmetros: densidade, massa molecular, índice de refração e teor de enxofre ou através da calibração com dados de espectroscopia NIR, entre outras técnicas.

Nesse trabalho realizou-se um estudo sobre a performance do algoritmo SVM para o desenvolvimento de modelos de classificação com múltiplas classes visando identificar a mistura de óleo naftênico e/ou óleo vegetal em óleo parafínico (óleo básico e óleo lubrificante de motor). Tal desenvolvimento proporciona a obtenção de um método mais eficiente em relação ao citado método ASTM de referência para identificação de óleo naftênico em óleo básico além de proporcionar a identificação simultânea da presença de óleo vegetal e poder ser utilizado também para análise de amostras de óleo lubrificante de motor automotivo.

Obtiveram-se modelos de classificação de múltiplas classes utilizando SIMCA e SVM aplicados a dados de espectroscopia NIR de óleos básicos parafínicos e da mistura destes com óleo básico naftênico e/ou óleo vegetal. Foram feitas previsões de misturas preparadas com óleos básicos e óleo vegetal; misturas preparas com óleos lubrificantes de motor comerciais, óleos básicos e óleo vegetal; e também de óleos lubrificantes de motor puros (tais quais comercializados).

O desenvolvimento de eficazes modelos de classificação de múltiplas classes pode facilitar a análise para identificação de adulterações e ser útil para o controle de qualidade de óleos lubrificantes, podendo servir também como uma análise preliminar, para se necessário, posterior procedimento de análise com o método de referência.

12.1 – Parte experimental

Foram utilizados para compor as misturas nesse estudo oito diferentes tipos de óleos básicos minerais, fornecidos pela Petrobras Distribuidora S.A. e pela refinaria de Duque de Caxias/RJ, comumente utilizados em misturas lubrificantes comerciais, sendo sete óleos parafínicos e um óleo naftênico, além do óleo vegetal de soja refinado produzido pela Cargill: óleo parafínico spindle (PSP), óleo parafínico neutro leve (PNL), óleo parafínico neutro médio (PNM), óleo parafínico neutro pesado (PNP), óleo parafínico turbina leve (PTL), óleo parafínico turbina pesado (PTP), óleo parafínico bright stock (PBS), óleo naftênico hidrogenado (NH) e óleo vegetal de soja (VEG).

Com a utilização das misturas em proporções adequadas dos sete diferentes tipos de óleos parafínicos, que possuem diferentes viscosidades, conforme mostra a tabela 6.2, pode-se garantir um representativo espaço amostral quanto as viscosidades dos diferentes óleos lubrificantes de motor automotivo que utilizam óleos básicos do grupo I presentes no mercado. Nesse estudo incluiu-se misturas de óleos básicos que proporcionam viscosidades que variam entre os graus SAE 5W a 50.

Utilizaram-se os óleos básicos parafínicos puros e em mistura com o óleo básico naftênico e/ou com óleo vegetal para compor as quatro diferentes classes mostradas na tabela 12.1. Nos conjuntos de calibração, validação e previsão as amostras de misturas foram acrescidas de óleo naftênico em teores de 10 % (v/v) e 15 % (v/v) e o óleo vegetal foi acrescido em teor de 5 % (v/v). As amostras do conjunto de previsão foram preparadas substituindo o óleo básico parafínico por óleo lubrificante de motor automotivo mineral disponível no mercado. As composições das diferentes misturas utilizadas nos conjuntos de calibração, validação e previsão são mostradas nas tabelas 12.2, 12.6 e 12.7, respectivamente. O conjunto de calibração completo foi composto de 67 amostras com quatro classes, conforme mostra a tabela 12.2.

Para o conjunto de validação foram preparadas amostras de misturas idênticas as utilizadas no conjunto de calibração e também amostras de misturas que não estão no conjunto de calibração a fim de testar a capacidade de generalização dos modelos propostos. Essas misturas não presentes no conjunto de calibração diferem quanto aos óleos básicos componentes das misturas ou quanto ao teor de óleo naftênico

adicionado a mistura, sendo que utilizou-se nesse caso misturas com apenas 7 % (v/v) de óleo básico naftênico. Os conjuntos de validação são mostrados na tabela 12.6.

Para o conjunto de previsão utilizaram-se oito amostras (C1, C2, ..., C8) de óleos lubrificantes para motores automotivos regularmente comercializados no mercado nacional e produzidos por diferentes fabricantes (atuantes apenas no mercado regional ou que atuam também no mercado exterior), sendo produtos com diferentes classificações API de desempenho e com diferentes classificações SAE de viscosidade, conforme mostrado na tabela 12.7. Realizou-se a previsão dessas amostras tais quais comercializadas e também utilizando-as em misturas em lugar do óleo básico parafínico e adicionando-se teores de óleo básico naftênico e/ou óleo vegetal idênticos aos utilizados nos conjunto de calibração e validação. As amostras comerciais tais quais comercializadas (sem adulteração) devem em tese pertencer a classe 1 (óleo parafínico).

Foram obtidos espectros na região de 4000 cm^{-1} a 6100 cm^{-1} para os óleos parafínico, naftênico e vegetal puros, para óleos lubrificantes de motor automotivo e também para as misturas citadas nas tabelas 12.2, 12.6 e 12.7 cujos espectros são mostrados na figura 12.1.

Para obtenção dos espectros de transflectância na região do infravermelho próximo foi utilizado um espectrômetro Perkin Elmer Spectrum 100 MID/NIR, com fonte halógena e detector de sulfato de triglicina deuterada (DTGS). Utilizou-se uma placa de Petri de vidro como recipiente de amostra e um refletor de alumínio com caminho ótico de 0,5 mm como cela de transflectância. Cada espectro foi obtido como uma média de 32 varreduras, com resolução de 4 cm^{-1}.

Foram realizados diferentes pré-processamentos dos dados para verificar qual proporciona a construção do melhor modelo utilizando os algoritmos SIMCA e SVM. Os pré-processamentos testados foram: correção de linha base e centragem na média; SNV e centragem na média; e primeira derivada (janela com 15 pontos).

O pacote LIBSVM[85] versão 2.88 foi utilizado para o desenvolvimento dos modelos com SVM e é adequado para utilização com Matlab 7.7 da Mathworks.

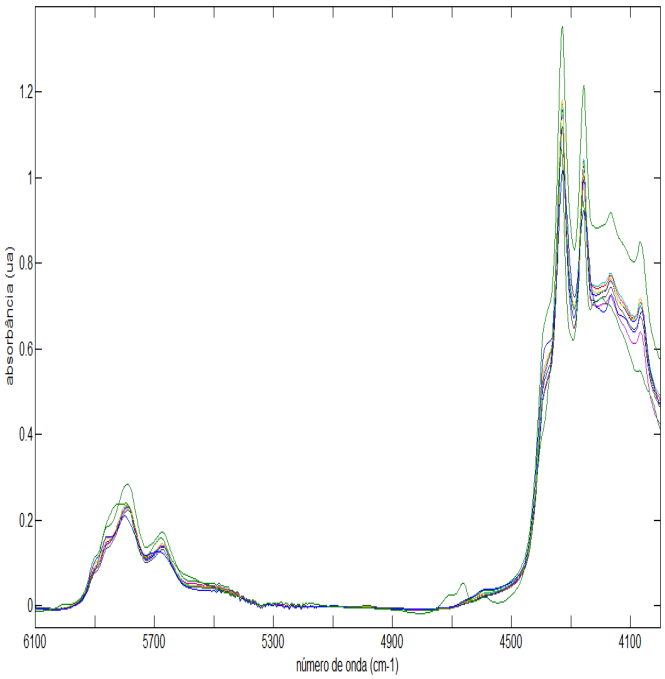
Para obtenção dos modelos de classificação com SVM foi utilizada a função kernel RBF. Testou-se a utilização dos algoritmos C-SVC e ν-SVC.

Tabela 12.1 - Conjuntos de amostras para os modelos de classificação com três e quatro classes

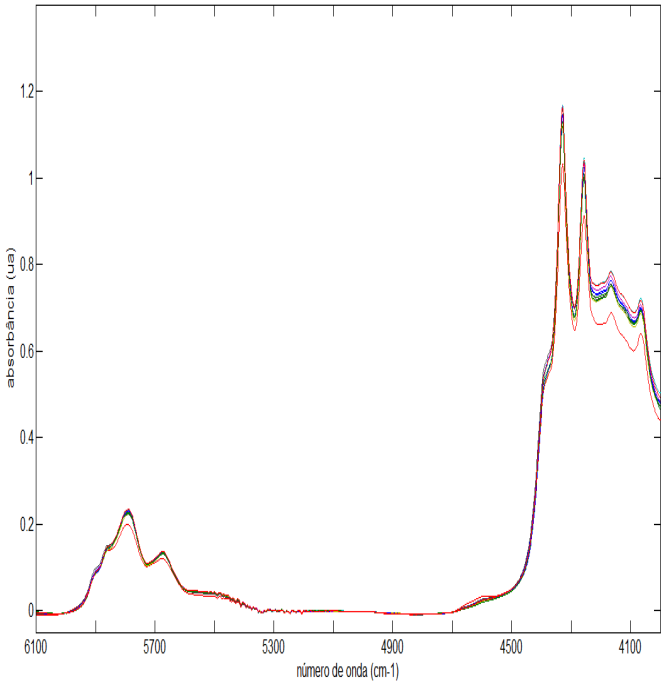
	Conjunto A	Conjunto B	Conjunto C
Número de classes	3	3	4
Número de amostras: total conj. de calibração conj. de validação conj. de previsão	 118 51 43 24	 125 56 45 24	 154 67 55 32
Classes	1 – óleo parafínico 2 – óleo parafínico + óleo naftênico 4 – óleo parafínico + óleo vegetal	1 – óleo parafínico 2 – óleo parafínico + óleo naftênico 3 – óleo parafínico + óleo naftênico + óleo vegetal	1 – óleo parafínico 2 – óleo parafínico + óleo naftênico 3 – óleo parafínico + óleo naftênico + óleo vegetal 4 – óleo parafínico + óleo vegetal
Distribuição das amostras de calibração e validação por classe	1 – 33 (cal.: 20; val.: 13) 2 – 40 (cal.: 20; val.: 20) 4 – 21 (cal.: 11; val.: 10)	1 – 33 (cal.: 20; val.: 13) 2 – 40 (cal.: 20; val.: 20) 3 – 28 (cal.: 16; val.: 12)	1 – 33 (cal.: 20; val.: 13) 2 – 40 (cal.: 20; val.: 20) 3 – 28 (cal.: 16; val.: 12) 4 – 21 (cal.: 11; val.: 10)

Tabela 12.2 – Amostras do conjunto de calibração

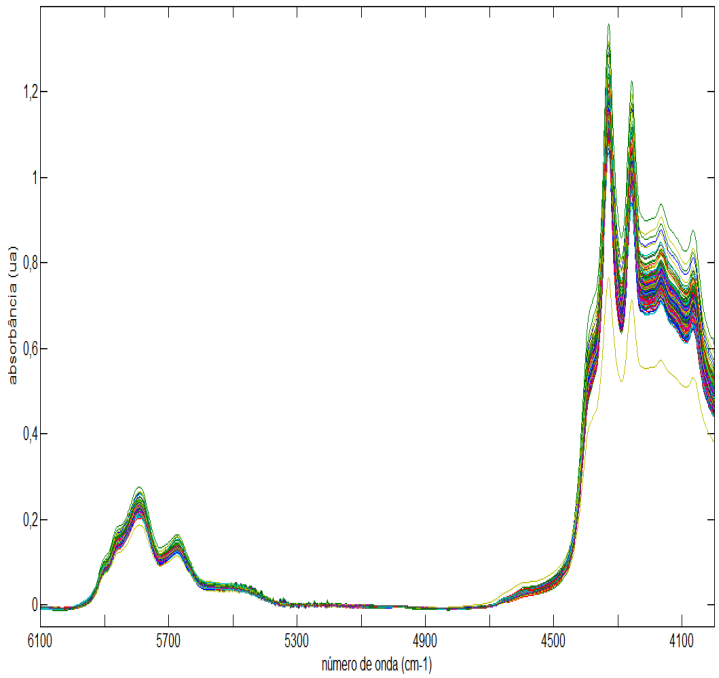
sequência	classe	Composições das amostras (frações em volume)
1	1	PNL
2		PNM
3		PNP
4		PSP
5		PTL
6		PTP
7		0,25 PBS + 0,75 PNP
8		0,50 PNL + 0,50 PNM
9		0,50 PNM + 0,50 PNP
10		0,90 PNM + 0,10 PNL
11		0,90 PNM + 0,10 PNP
12		0,90 PNP + 0,10 PNM
13		0,75 PNM + 0,25 PNP
14		0,75 PNP + 0,25 PNM
15		0,34 PBS + 0,66 PNP
16		0,15 PBS + 0,85 PNP
17		0,90 PNP + 0,10 PTP
18		0,50 PTL + 0,50 PTP
19		0,33 PNL + 0,33 PNM + 0,33 PNP
20		0,15 PNM + 0,55 PNP + 0,30 PBS
21	2	0,90 PTL + 0,10 NH
22		0,90 PNL + 0,10 NH
23		0,90 PNM + 0,10 NH
24		0,90 PNP + 0,10 NH
25		0,90 (0,25 PBS + 0,75 PNP) + 0,10 NH
26		0,90 PTP + 0,10 NH
27		0,90 PSP + 0,10 NH
28		0,85 PTL + 0,15 NH
29		0,85 PNL + 0,15 NH
30		0,85 PNM + 0,15 NH
31		0,85 PNP + 0,15 NH
32		0,85 PTP + 0,15 NH
33		0,85 PSP + 0,15 NH
34		0,85 (0,25 PBS + 0,75 PNP) + 0,15 NH
35		0,85 (0,50 PNL + 0,50 PNM) + 0,15 NH
36		0,85 (0,50 PNM + 0,50 PNP) + 0,15 NH
37		0,85 (0,75 PNM + 0,25 PNP) + 0,15 NH
38		0,85 (0,75 PNP + 0,25 PNM) + 0,15 NH
39		0,85 (0,15 PNM + 0,55 PNP + 0,30 PBS) + 0,15 NH
40		0,85 (0,33 PNL + 0,33 PNM + 0,33 PNP) + 0,15 NH
41	3	0,85 PTL + 0,10 NH + 0,05 VEG
42		0,85 PNL + 0,10 NH + 0,05 VEG
43		0,85 PNM + 0,10 NH + 0,05 VEG
44		0,85 PNP + 0,10 NH + 0,05 VEG
45		0,85 (0,25 PBS + 0,75 PNP) + 0,10 NH + 0,05 VEG
46		0,80 PTL + 0,15 NH + 0,05 VEG
47		0,80 PNL + 0,15 NH + 0,05 VEG
48		0,80 PNM + 0,15 NH + 0,05 VEG
49		0,80 PNP+ 0,15 NH + 0,05 VEG
50		0,80 (0,25 PBS + 0,75 PNP) + 0,15 NH + 0,05 VEG
51		0,80 (0,5 PNL + 0,50 PNM) + 0,15 NH + 0,05 VEG
52		0,80 (0, 5 PNM + 0,50 PNP) + 0,15 NH + 0,05 VEG
53		0,80 (0,75 PNM + 0,25 PNP) + 0,15 NH + 0,05 VEG
54		0,80 (0,75 PNP + 0,25 PNM) + 0,15 NH + 0,05 VEG
55		0,80 (0,15 PNM + 0,55 PNP + 0,30 PBS) + 0,15 NH + 0,05 VEG
56		0,80 (0,33 PNL + 0,33 PNM + 0,33 PNP) + 0,15 NH + 0,05 VEG
57	4	0,95 PTL + 0,05 VEG
58		0,95 PNL + 0,05 VEG
59		0,95 PNM + 0,05 VEG
60		0,95 PNP+ 0,05 VEG
61		0,95 (0,25 PBS + 0,75 PNP) + 0,05 VEG
62		0,95 (0,50 PNL + 0,50 PNM) + 0,05 VEG
63		0,95 (0,50 PNM + 0,50 PNP) + 0,05 VEG
64		0,95 (0,75 PNM + 0,25 PNP) + 0,05 VEG
65		0,95 (0,75 PNP + 0,25 PNM) + 0,05 VEG
66		0,95 (0,15 PNM + 0,55 PNP + 0,30 PBS) + 0,05 VEG
67		0,95 (0,33 PNL + 0,33 PNM + 0,33 PNP) + 0,05 VEG



(a)



(b)



(c)

Figura 12.1 – (a) espectros dos óleos básicos parafínicos, do óleo básico naftênico e do óleo vegetal, (b) espectros dos óleos lubrificantes comerciais, e (c) espectros das misturas preparadas em laboratório

O parâmetro γ do kernel RBF foi selecionado juntamente com o parâmetro C ou ν do SVM. Os parâmetros C e γ (para o C-SVC) foram selecionados entre os intervalos de 0 a 50000 e 0,01 a 100, respectivamente. Os parâmetros ν e γ (para o ν-SVC) foram selecionados entre os intervalos 0,001 a 1 e 0,01 a 100, respectivamente. A seleção dos parâmetros foi realizada através do método *grid search*, visando a minimização do erro de validação cruzada no conjunto de treinamento.

Como a minimização do erro de validação cruzada no conjunto de treinamento não garante a obtenção da melhor exatidão para o conjunto de validação, eventualmente um refinamento manual dessa seleção pode ser necessário, sempre considerando a utilização do adequado número de vetores de suporte de modo a evitar um sobreajuste do modelo.

12.2 – Resultados e discussão

Foram desenvolvidos três modelos de classificação, sendo dois modelos com três classes e um modelo com 4 classes. A tabela 12.1 mostra o rótulo de cada classe.

A região espectral utilizada, de 4000 cm^{-1} a 6100 cm^{-1}, corresponde a região das bandas de combinação e das bandas de primeiro sobreton de modos vibracionais do grupo C-H. Essa região possui diversas bandas de absorção atribuídas a combinação de modos vibracionais da ligação C-H de grupos metil e metileno em alcanos, ligação C-H em cicloalcanos, ligação C-H de anéis aromáticos e ligação C=C de compostos insaturados. Também há ocorrência da banda de combinação de estiramento do grupo C=O de ésteres. As atribuições de algumas bandas nessa região espectral podem ser verificadas no item 2.1 e na tabela 2.1.

Essa região espectral permite a classificação dos óleos básicos parafínicos puros e acrescidos de óleo naftênico e/ou óleo vegetal porque a cada óleo componente da mistura está associado um maior ou menor teor de hidrocarbonetos parafínicos, naftênicos, aromáticos e insaturados, além da presença do grupo carbonila no óleo vegetal.

A tabela 6.2, mostra as composições modais dos óleos básicos minerais do grupo I quanto as proporções de carbono parafínico (C_P), carbono naftênico (C_N) e

carbono aromático (C_A). De uma forma geral, observam-se aproximadamente as seguintes proporções:

- óleo parafínico (PNM): 66,2 % C_P, 26 % C_N, 7,8 % C_A
- óleo naftênico (NH20): 44,8 % C_P, 42,8 % C_N, 12,4 % C_A

Assim, com a variação dos teores de diferentes tipos de hidrocarbonetos presentes nas diferentes misturas, tem-se a variação da quantidade dos grupos metil e metileno de cadeias normal, ramificada, cíclica e aromática, a presença ou não do grupo carbonila e a variação do teor de compostos insaturados e suas respectivas intensidades de sinais no espectro NIR permitem a classificação das diferentes classes.

Inicialmente foram testados modelos utilizando toda a região espectral citada ou apenas as regiões de 4000-4800 cm^{-1} e 5450-6100 cm^{-1}. Como os resultados foram semelhantes optou-se por utilizar apenas as últimas regiões (1450 variáveis) visando obter um menor tempo de processamento dos dados.

12.2.1 – Conjunto A - Modelo de classificação para 3 classes

Esse modelo utilizou o conjunto A, composto de espectros de amostras de três diferentes classes: classe 1: óleo parafínico; classe 2: óleo parafinico + óleo naftênico; classe 4: óleo parafinico + óleo vegetal.

O número de amostras utilizado, a composição das misturas e os resultados obtidos para os conjuntos de validação e previsão para os modelos utilizando SIMCA e SVM são mostrados nas tabelas 12.1, 12.2, 12.6.e 12.7, respectivamente. Os resultados de previsão obtidos para os conjuntos de validação e previsão são mostrados de forma concisa na tabela 12.3.

Tabela 12.3 – Resultados de classificação dos modelos para o conjunto A

	Classe	SIMCA		SVM	
		$N_{acertos}/N_{total}$	Erros (%)	$N_{acertos}/N_{total}$	Erros (%)
Conjunto de validação	1	11/13	15	13/13	0
	2	19/20	5	20/20	0
	3	-	-	-	-
	4	10/10	0	10/10	0
	total	40/43	7	43/43	0
Conjunto de previsão	1	-	-	-	-
	2	8/8	0	8/8	0
	3	-	-	-	-
	4	3/8	63	4/8	50
	total	11/16	31	12/16	25

O melhores modelos obtidos com SIMCA utilizaram os dados pré-processados com primeira derivada.

O melhores modelos obtidos com SVM utilizaram os dados pré-processados com correção de linha base WLS e centrados na média. A utilização do SVM através do C-SVC proporcionou o melhor resultado. Utilizando os parâmetros selecionados C = 1024 e γ = 0,125 obteve-se um modelo que utiliza 26 vetores de suporte, sendo 12; 6; e 8 vetores de suporte para as classes 1, 2, e 4, respectivamente. O bom ajuste do modelo é evidenciado através da utilização de um adequado número de vetores de suporte.

Para as amostras de validação verifica-se que os modelos SIMCA e SVM têm bom desempenho, com uma pequena superioridade do modelo SVM que proporciona boa previsão inclusive para as 12 amostras não presentes no conjunto de calibração.

Com base no melhor desempenho do modelo SVM para as amostras do conjunto de validação considera-se esse modelo o mais adequado para realizar a previsão das amostras do conjunto de previsão.

Para esse problema de classificação com três classes verifica-se que o algoritmo SVM proporciona os melhores resultados:

- para as amostras do conjunto de validação há 100 % de acertos;
- para o conjunto de previsão as amostras comerciais classificadas como classe 2 sugere um teor de C_N e C_A além do comum para esse tipo de produto. A previsão das misturas utilizando os óleos comerciais mostram boa previsão para as amostras da classe 2, com 100% de acertos, porém, para as amostras da classe 4 com uma proporção de apenas 5% de óleo vegetal na mistura não houve boa previsão, com apenas 50% de acertos.

12.2.2 – Conjunto B - Modelo de classificação para 3 classes

Esse modelo utilizou o conjunto B, composto de espectros de amostras de três diferentes classes: classe 1: óleo parafínico; classe 2: óleo parafinico + óleo naftênico; classe 3: óleo parafinico + óleo naftênico + óleo vegetal.

O número de amostras utilizado, a composição das misturas e os resultados obtidos para os conjuntos de validação e previsão para os modelos utilizando SIMCA e SVM são mostrados nas tabelas 12.1, 12.2, 12.6 e 12.7, respectivamente. Os resultados de previsão obtidos para os conjuntos de validação e previsão são mostrados de forma concisa na tabela 12.4.

O melhores modelos obtidos com SIMCA utilizaram os dados pré-processados com primeira derivada.

O melhores modelos obtidos com SVM utilizaram os dados pré-processados com correção de linha base WLS e centrados na média. A utilização do SVM através do C-SVC proporcionou o melhor resultado. Utilizando os parâmetros selecionados C = 1024 e γ = 0,0625 obteve-se um modelo que utiliza 33 vetores de suporte, sendo 8; 13; e 12 vetores de suporte para as classes 1, 2, e 3, respectivamente. O bom ajuste do modelo é evidenciado através da utilização de um adequado número de vetores de suporte.

Para as amostras de validação verifica-se que os modelos SIMCA e SVM têm bom desempenho, com uma pequena superioridade do modelo SVM que proporciona boa previsão inclusive para as 12 amostras não presentes no conjunto de calibração.

Tabela 12.4 – Resultados de classificação dos modelos para o conjunto B

	Classe	SIMCA		SVM	
		$N_{acertos}/N_{total}$	Erros (%)	$N_{acertos}/N_{total}$	Erros (%)
Conjunto de validação	1	11/13	15	13/13	0
	2	19/20	5	19/20	5
	3	12/12	0	12/12	0
	4	-	-	-	-
	total	42/45	7	44/45	2
Conjunto de previsão	1	-	-	-	-
	2	7/8	13	7/8	13
	3	1/8	87	6/8	25
	4	-	-	-	-
	total	8/16	50	13/16	19

Com base no melhor desempenho do modelo SVM para as amostras do conjunto de validação considera-se esse modelo o mais adequado para realizar a previsão das amostras do conjunto de previsão.

Para esse problema de classificação com três classes verifica-se que o algoritmo SVM proporciona os melhores resultados principalmente para as amostras do conjunto de previsão:

- para as amostras de validação há 98 % de acertos;
- para o conjunto de previsão entre as oito amostras analisadas tais quais comercializadas obteve-se a classificação de seis delas na classe 1 e duas na classe 2, o que sugere um teor de C_N e C_A além do comum para esse tipo de produto nessas 2 amostras. A previsão das misturas utilizando os óleos comerciais mostram boa previsão obtendo 87% e 75% de acertos para as amostras das classes 2 e 3, respectivamente, sendo possível identificar uma proporção de apenas 5% de óleo vegetal na mistura.

12.2.3 – Conjunto C - Modelo de classificação para 4 classes

Esse modelo utilizou o conjunto C, composto de espectros de amostras de quatro diferentes classes: classe 1: óleo parafínico; classe 2: óleo parafinico + óleo naftênico; classe 3: óleo parafinico + óleo naftênico + óleo vegetal; e classe 4: óleo parafínico + óleo vegetal.

O número de amostras utilizado, a composição das misturas e os resultados obtidos para os conjuntos de validação e previsão para os modelos utilizando SIMCA e SVM são mostrados nas tabelas 12.1, 12.2, 12.6 e 12.7, respectivamente. Os resultados de previsão obtidos para os conjuntos de validação e previsão são mostrados de forma concisa na tabela 12.5.

A figura 12.2 ilustra a Análise de Componentes Principais do conjunto de calibração, onde é possível verificar o complicado problema de classificação das quatro classes estudadas.

O melhores modelos obtidos com SIMCA utilizaram os dados pré-processados com primeira derivada.

Os melhores modelos obtidos com SVM utilizaram os dados pré-processados com correção de linha base WLS e centrados na média. A utilização do SVM através do C-SVC proporcionou o melhor resultado. Utilizando os parâmetros selecionados C = 1024 e γ = 0,0625 obteve-se um modelo que utiliza 49 vetores de suporte, sendo 14; 13; 13 e 9 vetores de suporte para as classes 1, 2, 3 e 4, respectivamente. O bom ajuste do modelo é evidenciado através da utilização de um adequado número de vetores de suporte.

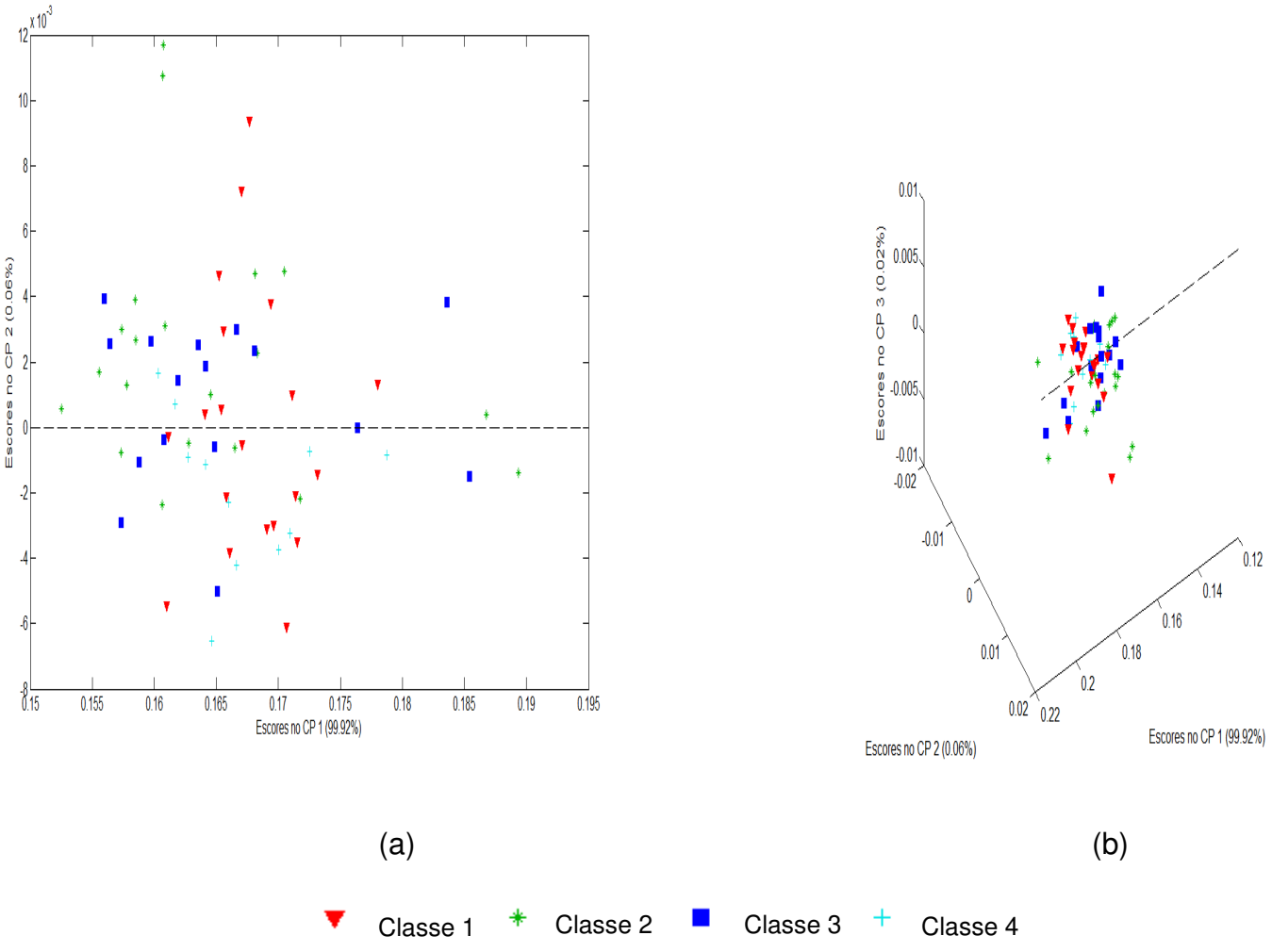


Figura 12.2 – PCA para o conjunto de dados C. (a) gráfico dos escores dos 2 primeiros componentes principais e (b) gráfico dos escores dos 3 primeiros componentes principais.

Para as amostras de validação verifica-se que os modelos com SIMCA e SVM têm bom desempenho, com uma pequena superioridade do modelo SVM pois embora com o mesmo número total de acertos este proporciona melhor previsão inclusive para todas as 12 amostras não presentes no conjunto de calibração.

Com base no melhor desempenho do modelo SVM para as amostras do conjunto de validação considera-se esse modelo o mais adequado para realizar a previsão das amostras do conjunto de previsão.

Tabela 12.5 – Resultados de classificação dos modelos para o conjunto C

	Classe	SIMCA		SVM	
		$N_{acertos}/N_{total}$	Erros (%)	$N_{acertos}/N_{total}$	Erros (%)
Conjunto de validação	1	11/13	15	12/13	8
	2	19/20	5	19/20	5
	3	12/12	0	12/12	0
	4	10/10	0	9/10	10
	total	52/55	5	52/55	5
Conjunto de previsão	1	-	-	-	-
	2	7/8	13	7/8	13
	3	2/8	75	6/8	25
	4	3/8	63	3/8	63
	total	12/24	50	16/24	33

Para esse problema de classificação com quatro classes verifica-se que o algoritmo SVM proporciona os melhores resultados, basicamente para as amostras do conjunto de previsão:

- para as amostras do conjunto de validação há 95 % de acertos;

- para o conjunto de previsão entre as oito amostras analisadas tais quais comercializadas obteve-se a classificação de seis delas na classe 1 e duas na classe 2 o que sugere um teor de C_N e C_A além do comum para esse tipo de produto. A previsão das misturas utilizando os óleos comerciais mostram boa previsão obtendo 87% e 75% de acertos para as classes 2 e 3, respectivamente, sendo possível identificar uma proporção de apenas 5% de óleo vegetal na mistura ternária, porém, para as amostras da classe 4 há elevado erro de previsão.

Tabela 12.6 – Previsões das amostras do conjunto de validação

sequência	Classe Nominal	Composições das amostras (frações em volume)	Conjunto A		Conjunto B		Conjunto C	
			SIMCA	SVM	SIMCA	SVM	SIMCA	SVM
1	1	PNP	2	1	2	1	2	1
2		0,25 PBS + 0,75 PNP	1	1	1	1	1	1
3		PSP	1	1	1	1	1	1
4		PTP	1	1	1	1	1	1
5		0, 5 PNL + 0,5 PNM	2	1	2	1	2	4
6		0, 5 PNM + 0,5 PNP	1	1	1	1	1	1
7		0,75 PNM + 0,25 PNP	1	1	1	1	1	1
8		0,75 PNP + 0,25 PNM	1	1	1	1	1	1
9		0,33 PNL + 0,33 PNM + 0,33 PNP	1	1	1	1	1	1
10		0,15 PNM + 0,55 PNP + 0,30 PBS	1	1	1	1	1	1
11		0,5 PTL + 0,5 PTP	1	1	1	1	1	1
12		0,5 PTL + 0,5 PNM *	1	1	1	1	1	1
13		0,5 PTP + 0,5 PNM *	1	1	1	1	1	1
14	2	0,9 PNP + 0,1 NH	2	2	2	2	2	2
15		0,9 (0,25 PBS + 0,75 PNP) + 0,1 NH	2	2	2	2	2	2
16		0,9 PTP + 0,1 NH	2	2	2	2	2	2
17		0,9 PSP + 0,10 NH	2	2	2	2	2	2
18		0,85 PNP + 0,15 NH	2	2	2	2	2	2
19		0,85 (0,25 PBS + 0,75 PNP) + 0,15 NH	2	2	2	2	2	2
20		0,85 PTP + 0,15 NH	2	2	2	2	2	2
21		0,85 PSP + 0,15 NH	2	2	2	2	2	2
22		0,85 (0,5 PNL + 0,5 PNM) + 0,15 NH	2	2	2	2	2	2
23		0,85 (0, 5 PNM + 0,5 PNP) + 0,15 NH	2	2	2	2	2	2
24		0,85 (0,75 PNM + 0,25 PNP) + 0,15 NH	2	2	2	2	2	2
25		0,85 (0,75 PNP + 0,25 PNM) + 0,15 NH	2	2	2	2	2	2
26		0,85 (0,33 PNL + 0,33 PNM + 0,33 PNP) + 0,15 NH	2	2	2	3	2	3
27		0,85 (0,15 PNM + 0,55 PNP + 0,30 PBS) + 0,15 NH	2	2	2	2	2	2
28		0,85 (0,5 PTL + 0,5 PNM) + 0,15 NH *	2	2	2	2	2	2
29		0,85 (0,5 PTP + 0,5 PNM) + 0,15 NH *	2	2	2	2	2	2
30		0,93 PTL + 0,07 NH *	1	2	1	2	1	2
31		0,93 PNL + 0,07 NH *	2	2	2	2	2	2
32		0,93 PNM + 0,07 NH *	2	2	2	2	2	2
33		0,93 (0,25 PBS + 0,75 PNP) + 0,07 NH *	2	2	2	2	2	2
34	3	0,85 PNP + 0,1 NH + 0,05 VEG	-	-	3	3	3	3
35		0,85 (0,25 PBS + 0,75 PNP) + 0,1 NH + 0,05 VEG	-	-	3	3	3	3
36		0,80 PNP + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
37		0,80 (0,25 PBS + 0,75 PNP) + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
38		0,80 (0,5 PNL + 0,5 PNM) + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
39		0,80 (0, 5 PNM + 0,5 PNP) + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
40		0,80 (0,75 PNM + 0,25 PNP) + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
41		0,80 (0,75 PNP + 0,25 PNM) + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
42		0,80 (0,15 PNM + 0,55 PNP + 0,30 PBS) + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
43		0,80 (0,33 PNL + 0,33 PNM + 0,33 PNP) + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
44		0,80 (0,5 PTL + 0,5 PNM) + 0,15 NH + 0,05 VEG *	-	-	3	3	3	3
45		0,80 (0,5 PTP + 0,5 PNM) + 0,15 NH + 0,05 VEG *	-	-	3	3	3	3
46	4	0,95 PNP + 0,05 VEG	4	4	-	-	4	3
47		0,95 (0,25 PBS + 0,75 PNP) + 0,05 VEG	4	4	-	-	4	4
48		0,95 (0, 5 PNL + 0,5 PNM) + 0,05 VEG	4	4	-	-	4	4
49		0,95 (0, 5 PNM + 0,5 PNP) + 0,05 VEG	4	4	-	-	4	4
50		0,95 (0,75 PNM + 0,25 PNP) + 0,05 VEG	4	4	-	-	4	4
51		0,95 (0,75 PNP + 0,25 PNM) + 0,05 VEG	4	4	-	-	4	4
52		0,95 (0,15 PNM + 0,55 PNP + 0,30 PBS) + 0,05 VEG	4	4	-	-	4	4
53		0,95 (0,33 PNL + 0,33 PNM + 0,33 PNP) + 0,05 VEG	4	4	-	-	4	4
54		0,95 (0,5 PTL + 0,5 PNM) + 0,05 VEG *	4	4	-	-	4	4
55		0,95 (0,5 PTP + 0,5 PNM) + 0,05 VEG *	4	4	-	-	4	4

* não tem réplica no conjunto de calibração

classe previsão não confere com a classe nominal

Tabela 12.7 – Previsões das amostras do conjunto de previsão

Sequência	Classe nominal	Composições das amostras (frações em volume)	Conjunto A		Conjunto B		Conjunto C	
			SIMCA	SVM	SIMCA	SVM	SIMCA	SVM
1	-	**C1 – API CF - SAE 15W40**	2	1	2	1	2	1
2	2	0,85 C1 + 0,15 NH	2	2	2	2	2	2
3	3	0,80 C1 + 0,15 NH + 0,05 VEG	-	-	2	3	3	3
4	4	0,95 C1 + 0,05 VEG	4	4	-	-	4	4
5	-	**C2 – API SL - SAE 15W40**	2	2	2	1	2	1
6	2	0,85 C2 + 0,15 NH	2	2	2	2	2	2
7	3	0,80 C2 + 0,15 NH + 0,05 VEG	-	-	2	3	2	3
8	4	0,95 C2 + 0,05 VEG	2	2	-	-	2	3
9	-	**C3 – API CI/SL - SAE 15W40**	2	1	2	1	2	1
10	2	0,85 C3 + 0,15 NH	2	2	2	3	2	3
11	3	0,80 C3 + 0,15 NH + 0,05 VEG	-	-	2	3	2	3
12	4	0,95 C3 + 0,05 VEG	2	4	-	-	2	4
13	-	**C4 - API SJ - SAE 20W50**	2	1	3	1	3	1
14	2	0,85 C4 + 0,15 NH	2	2	3	2	3	2
15	3	0,80 C4 + 0,15 NH + 0,05 VEG	-	-	3	3	3	3
16	4	0,95 C4 + 0,05 VEG	4	4	-	-	4	3
17	-	**C5 – API SL - SAE 20W50**	2	2	2	2	2	2
18	2	0,85 C5 + 0,15 NH	2	2	2	2	2	2
19	3	0,80 C5 + 0,15 NH + 0,05 VEG	-	-	2	3	2	3
20	4	0,95 C5 + 0,05 VEG	2	2	-	-	2	3
21	-	**C6 – API SL - SAE 20W50**	2	2	2	2	2	2
22	2	0,85 C6 + 0,15 NH	2	2	2	2	2	2
23	3	0,80 C6 + 0,15 NH + 0,05 VEG	-	-	2	3	2	3
24	4	0,95 C6 + 0,05 VEG	2	2	-	-	3	1
25	-	**C7 – API CF - SAE 40**	2	1	2	1	2	1
26	2	0,85 C7 + 0,15 NH	2	2	2	2	2	2
27	3	0,80 C7 + 0,15 NH + 0,05 VEG	-	-	2	1	2	1
28	4	0,95 C7 + 0,05 VEG	4	4	-	-	4	4
29	-	**C8 - API SF - SAE 50**	2	2	2	1	2	1
30	2	0,85 C8 + 0,15 NH	2	2	2	2	2	2
31	3	0,80 C8 + 0,15 NH + 0,05 VEG	-	-	2	2	2	2
32	4	0,95 C8 + 0,05 VEG	2	1	-	-	3	1

classe previsão não confere com a classe nominal

12.3 – Conclusão

A utilização do algoritmo SVM para classificação de múltiplas classes de dados de espectroscopia NIR de amostras de óleos lubrificantes mostrou uma performance superior a obtida com SIMCA para problemas de classificação com três e quatro classes. Demonstrou-se a possibilidade de realizar a identificação de teores de óleo naftênico além do comum (elevados teores de compostos naftênicos e aromáticos) em óleos básicos minerais e óleos lubrificantes de motor do grupo I. Os modelos SVM desenvolvidos proporcionaram boa previsão para teores de óleo naftênico superiores ao

comum para óleos minerais parafínicos, identificando acréscimos de óleo naftênico a partir de 7 % (v/v) em mistura com óleo parafínico. Bons resultados foram também obtidos para a identificação simultânea de teores superiores ao comum de óleo naftênico e a presença de óleo vegetal a partir do teor de 5% (v/v).

O menor número relativo de amostras presentes no conjunto de calibração e a baixa concentração do óleo vegetal na mistura com óleo parafínico podem ser o motivo da dificuldade encontrada na previsão da classe contendo apenas o óleo vegetal em mistura com óleo parafínico.

Utilizando o SVM verificou-se que os modelos para os conjuntos B e C têm desempenho muito bom e bastante parecido quanto a previsão das classes 1, 2 e 3 nos conjunto de validação e previsão. O modelo para o conjunto A tem desempenho pouco superior ao modelo para o conjunto C quanto a previsão da classe 4 nos conjunto de validação e previsão. No entanto, o modelo para o conjunto C pode ser considerado o mais adequado uma vez que possibilita a classificação das 4 classes com desempenho muito parecido aos dos modelos com três classes.

O superior poder de generalização do algoritmo SVM é demonstrado ao proporcionar uma melhor previsão de amostras do conjunto de validação e do conjunto de previsão que apresentam composição distinta das amostras utilizadas no conjunto de calibração.

O desenvolvimento dos modelos ora propostos caracterizam um método mais simples e eficiente para determinação da presença de óleo básico naftênico em óleo básico parafínico em relação ao método de referência ASTM D3238-95, além de poder ser utilizado também para determinação da presença de óleo básico naftênico em óleo lubrificante de motor. Além disso o desenvolvimento desses modelos de classificação dispensam a realização do método de referência também na construção do modelo. Soma-se a essas vantagens a capacidade de identificação simultânea de óleo básico naftênico e óleo vegetal na mistura com óleo básico parafínico ou óleo lubrificante de motor.

Com o desenvolvimento de modelos mais eficazes de classificação de múltiplas classes de dados eles podem ser utilizados para identificação de adulterações e em controle de qualidade de óleos básicos parafínicos e óleo lubrificante de motor automotivo, como uma alternativa ao método de referência ASTM (no caso dos óleos

básicos) e a métodos espectroscópicos e cromatográficos, mais trabalhosos ou mais caros, ou também como uma análise preliminar para seleção de amostras para análise pelo método de referência.

13 – Conclusão geral

Os resultados obtidos demonstram que o SVM proporciona a obtenção de modelos de regressão melhor ajustados e a obtenção de melhores resultados de previsão, pois o SVM pode não apenas obter a relação linear principal existente entre os dados espectroscópicos e as propriedades de interesse, mas, também pode modelar a não linearidade existente na relação entre os conjuntos de dados. Por outro lado, em relação aos modelos desenvolvidos com PLS, verificou-se que a obtenção dos modelos com SVM demandam um maior tempo de trabalho, para obtenção da otimização paramétrica, o que é plenamente justificado pela obtenção de modelos mais eficazes. A possibilidade de otimização de dois parâmetros no modelo SVM permite um refinamento no ajuste do modelo não possibilitado pela simples escolha do número de variáveis latentes com PLS.

A natureza do algoritmo que permite a modelagem de não linearidades presentes nas correlações estudadas e com um bom poder de generalização possibilita uma performance superior do SVM em todos os conjuntos de dados e modelos de regressão e classificação, sendo muito adequado para solução dos problemas estudados.

O melhor desempenho do algoritmo SVM em modelos de regressão e classificação, em relação ao PLS e ao SIMCA, respectivamente, pode ser atribuído a uma combinação de características próprias do SVM, tais como: (i) utilização da função kernel para elevação da dimensionalidade do espaço dos dados e cálculo do produto interno e (ii) possibilidade de otimizar ao menos dois parâmetros durante o ajuste do modelo.

Com o desenvolvimento de modelos de regressão para 11 diferentes parâmetros de qualidade de combustíveis e o desenvolvimento de 8 modelos de classificação para combustíveis ou lubrificantes verificou-se quais as melhores condições quanto ao pré-processamento dos dados e função kernel a utilizar para tratamento de dados de espectroscopia NIR desses tipos de amostras.

O desenvolvimento de modelos de calibração com SVM proporcionou resultados que possibilitam a aplicação dos mesmos como ferramenta em Tecnologia Analítica de Processos – PAT e controle de qualidade de produtos acabados. Com o desenvolvimento de modelos de calibração mais eficazes utilizando o algoritmo SVM e

dados de espectroscopia NIR, para determinação *on line* de parâmetros utilizados em controle de processos torna-se possível obter um melhor aproveitamento das correntes disponíveis na refinaria de modo a contribuir para a otimização da produção bem como otimizar a produção de unidades de processamento como o HDT. Esse desenvolvimento é importante porque proporciona uma melhora na produtividade de óleo diesel, reduzindo a importação desse derivado e aumentando o lucro da refinaria.

Além disso, proporciona simples e eficazes alternativas aos métodos de referência utilizados também em análises de rotina no controle de qualidade de produtos acabados.

14 - Referências bibliográficas

[1]P. Geladi; B. R. Kowalski; Partial least squares regression: a tutorial; Anal. Chim. Acta, 185 (1986) 1
[2]C. Cortes, V. Vapnik, Support Vector Networks, Machine Learning 20 (1995) 273
[3] J. B. Callis, D. L. Illman, B. R. Kowalski, Process analytical chemisty, Anal. Chem. 59 (1987) 624A
[4] H. Yang; P. R. Griffiths; J. D. Tate; Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase Fourier transform infrared spectra; Anal. Chim. Acta, 489 (2003) 125
[5] R. M. Balabin; R. Z. Safieva; E. I. Lomakina; Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction; Chemom. Intell. Lab. Syst., 88, (2007), 183
[6] D. Svozil; Introduction to multilayer feed-forward neural networks; Chemom. Intell. Lab. Syst., 39 (1997) 43
[7] U. Thissen, M. Pepers, B. Ustun, W. J. Melssen, L. M. C. Buydens, Comparing support vector machines to PLS for spectral regression applications, Chemom. Intel. Lab. Syst., 73 (2004) 169.
[8] A. F. Bueno, Caracterização de petróleo por espectroscopia no infravermelho próximo, dissertação de mestrado, Instituto de química, Universidade Estadual de Campinas, Campinas, 2004
[9] E. Bertran; M. Blanco; S. Maspoch; M. C. Ortiz; M. S. Sanchez; L. A. Sarabia; Handling intrinsic non-linearity in near-infrared reflectance spectroscopy; Chemom. Intell. Lab. Syst., 49 (1999) 215
[10] H. Li; Y. Liang; Q. Xu; Support vector machines and its application in chemistry; Chemom. Intell. Lab. Syst., 95 (2009) 188
[11] H. Zou; H. Wu; R. Yu; Variable weighted least-squares support vector machine for multivariate spectral analysis; Talanta, 80 (2010) 1698
[12] V. O. Santos Jr.; F. C. C. Oliveira; D. G. Lima; A. C. Petry; E. Garcia; P. A. Z. Suarez; J. C.Rubim; A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis; Anal. Chim. Acta, 547 (2005) 188
[13] N. Pasadakis; S. Sourligas; Ch. Foteinopoulos; Prediction of the distillation profile and cold properties of diesel fuels using mid-IR spectroscopy and neural networks; Fuel, 85 (2006) 1131
[14] R. M. Balabin; R. Z. Safieva; E. I. Lomakina; Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques; Anal. Chim. Acta, 671 (2010) 27
[15] K. Brudzewski; A. Kesik; K. Kolodziejczyk; U. Zborowska; J. Ulaczyk; Gasoline quality prediction using gas chromatography and FTIR spectroscopy: an artificial intelligence approach; Fuel, 85 (2006) 553
[16] F. S. Falla; C. Larini; G. A. C. Le Roux; F. H. Quina; L. F. L. Moro; C. A. O. Nascimento; Characterization of crude petroleum by NIR; J. Petroleum Science and Eng., 51 (2006) 127
[17] C. Vendeuvre; R. Guerrero; F. Bertoncini; M. Hennion; Characterisation of middle distillates by comprehensive two dimensional gas chromatography (GCxGC): a powerful

alternative for performing various standard analysis of middle distillates; J. Chromatogr. A, 1086 (2005) 21
[18] J. Workman Jr., L. Weyer, Practical guide to interpretive near infrared spectroscopy, CRC Press, Boca Raton, 2008
[19] C. Pasquini, Near infrared spectroscopy: fundamentals, practical aspects and analytical applications, J. Braz. Chem. Soc., 14 (2003) 198
[20] L. Bokobza, Near infrared spectroscopy, J. Near Infrared Spectrosc., 6 (1998) 3
[21] A. D. Skoog, F. J. Holler, T. A. Nieman, Princípios de análise instrumental, 5º ed., Bookman, Porto Alegre, 2002.
[22] J. M. Andrade; M. V. Garcia; P. Lopez-Mahia; D. Prada; A review of the main factors influencing the FT-IR-PLS abilities exemplified with petrochemical qualimetric applications; Talanta, 44 (1997) 2167
[23] H. Swierenga; A. P. Weijer; R. J. van Wijk; L. M. C. Buydens; Strategy for constructing robust multivariate calibration models; Chemom. Intell. Lab. Syst., 49 (1999) 1
[24] L. A. Sacorague, Avaliação de diferentes regiões do espectro do infravermelho próximo na determinação de parâmetros de qualidade de combustíveis empregando ferramentas quimiométricas, tese de doutorado, Instituto de química, Universidade Estadual de Campinas, Campinas, 2004
[25] G. Knothe, J.V. Gerpen, J. Krahl, L. P. Ramos, Manual de biodiesel, Ed. Edgard Blucher ltda., São Paulo, (2008)
[26] G. Knothe, Rapid monitoring of transesterification and assessing biodiesel fuel quality by near-infrared spectroscopy using a fiber optic probe, J. Am. Oil Chem. Soc., 76 (1999) 795
[27] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, Chemom. Intell. Lab. Syst., 2 (1987) 37
[28] H. Martens, T. Næs, Multivariate Calibration. 1st ed., John Wiley & Sons, Chichester, (1989)
[29] M. M. de Sena, R.J. Poppi, R. T. Frighetto, P. J. Valarini, Avaliação do uso de métodos quimiométricos em análise de solos, Quím. Nova, 23 (2000) 547
[30] M. P. Derde, D. L. Massart, Comparison of the performance of the class modeling techniques UNEQ, SIMCA and PRIMA, Chemom. Intell. Lab. Syst., 4 (1988) 65
[31] M. P. Derde, D. L. Massart, Supervised pattern recognition: the ideal method?, Anal. Chim. Acta, 191 (1986) 1
[32] M. Sjostrom, B. R. Kowalski, A comparison of five pattern recognition methods based on the classification results from six real data bases, Anal. Chim. Acta, 112 (1979) 11
[33] F.S. Oliveira, L.S.G. Teixeira, M.C.U. Araujo, M. Korn, Screening analysis to detect adulteration in brazilian gasoline samples using distillation curves, Fuel, 83 (2004) 917
[34] R. De Maesschalck, A. Candolfi, D. L. Massart, S. Heuerding, Decision criteria for soft independent modelling of class analogy applied to near infrared data, Chemom. Intell. Lab. Syst., 47 (1999) 65
[35] V. Vapnik, A. Lerner, Pattern recognition using generalized portrait method, Automation and Remote Control, 24 (1963) 774
[36] V. Vapnik, Estimation of dependences based on empirical data, Springer, Berlin, (1982)
[37] V. Vapnik, The nature of statistical learning theory, Springer, New York, (1995)

[38] A. C. Lorena, A.C.P.L.F de Carvalho, Uma introdução às Support Vector Machines, RITA, 14 (2007) 43
[39] B. Scholkopf, A. J. Smola, Learning with kernels, MIT press, Cambridge, MA (2002)
[40] V. Vapnik, S. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, Advances in Neural Information Processing Systems 9, MIT Press, Cambridge (1997)
[41] A. I. Belousov, S. A. Verzakov, J. von Frese, A flexible classification approach with optimal generalization performance: support vector machines, Chemom. Intell. Lab. Syst., 64 (2002) 15
[42] R. Semolini, Support vector machines, inferência transdutiva e o problema de classificação, dissertação de mestrado, Faculdade de engenharia elétrica e de computação, Universidade Estadual de Campinas, Campinas, 2002
[43] B. Ustun, W. J. Melssen, M. Oudenhuijzen, L. M. C. Buydens, Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization, Anal. Chim. Acta. 544 (2005) 292
[44] A. J. Smola, B. Scholkopf, A tutorial on support vector regression, Statistics and computing 14 (2004) 199
[45] C. A. M. Lima, Comitê de máquinas: uma abordagem unificada empregando máquinas de vetores suporte, tese de doutorado, Faculdade de engenharia elétrica e de computação, Universidade Estadual de Campinas, Campinas, 2004
[46] N. Chen, W. Lu, J. Yang, G. Li, Support vector machine in chemistry, Word Scientific Publishing Co. Pte. Ltd., Singapore, (2004)
[47] T. Hofmann, B. Scholkopf, A. Smola, Kernel methods in machine learning, The annals of statistics, 36 (2008) 1171
[48] A. J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans, Introduction to large margin classifiers, In (Eds.) A. J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans, Advances in large margin classifiers, MIT press, Cambridge, MA, (2000)
[49] A. Chalimourda, B. Scholkopf, A. J. Smola, Experimentally optimal ν in support vector regression for different noise models and paremeters settings, Neural Networks 17 (2004) 127
[50] O. Ivancius, Applications of support vector machines in chemistry, In: Rewiews in Computational Chemistry, eds. K. B. Lipkowitz, T. R. Cundari, Wiley-VCH, Weinheim, 23 (2007) 291
[51] M. S. Bazaraa, H. D. Sherali, C. M. Shetty, Nonlinear Programming – Theory and algorithms, 2nd edition, John Wiley & Sons inc., Hoboken, 1993
[52] O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, J.-P. Huvenne, Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation, Chemom. Intell. Lab. Syst., 96 (2009) 27
[53] J. Luts, F. Ojeda, R. Van de Plas, B. de Moor, S. Van Huffel, J. A. K. Suykens, A tutorial on support vector machine based methods for classification problems in chemometrics, Anal. Chim. Acta, 665 (2010), 129
[54] B. Scholkopf, A. J. Smola, R. C. Williamson, P. L. Bartlett, New Support Vector Algorithms, Neural Computation, 12 (2000) 1207
[55] P. A. Costa Filho, R. J. Poppi, Algoritmo genético em química, Química Nova, 22 (1999) 405
[56] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A practical guide to support vector classification, 2009, http://www.csie.ntu.edu.tw/~cjlin

[57] A. C. Ferreira, Modelos de otimização na produção de óleo diesel: uma aplicação industrial, tese de doutorado, Faculdade de engenharia química, Universidade Estadual de Campinas, Campinas, 2008
[58] D.H.G. Bezerra, F. L. Borges, R.B. Mendonça, U. Kopcak, Avaliação e desenvolvimento de modelos para cálculo de misturas para produção de óleo diesel, monografia do curso de especialização em engenharia de processamento de petróleo, Instituto de química, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2008
[59] Anuário estatístico brasileiro do petróleo, gás natural e biocombustíveis – 2011, Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP, http://www.anp.gov.br
[60] M. L. M. Valle, Produtos do setor de combustíveis e de lubrificantes, Publit Soluções editoriais, Rio de Janeiro, 2007
[61] C. Song, Introduction to chemistry of diesel fuels, In: Chemistry of diesel fuels, eds.: C. Song, C.S. Hsu, I. Mochida, Taylor & Francis, New York, 2000
[62] A. Singh, J. F. Forbes, P. J. Vermeer, S. S. Woo, Model based real time optimization of automotive gasoline blending operations, Journal of process control, 10 (2000) 43
[63] M. Chèbre, Y. Creff, N. Petit, Feedback control and optimization for the production of commercial fuels by blending, Journal of process control, 20 (2010) 441
[64] R.M.C.F. da Silva, Cinética e modelagem do reator de HDT, dissertação de mestrado, Faculdade de engenharia química, Universidade Estadual de Campinas, Campinas, 1995
[65] M. A. Rude, A. Schirmer, New microbial fuels: a biotech perspective, Current opinion in microbiology, 12 (2009) 274
[66] E. J. Steen, Y. Kang, G. Bokinsky, Z. Hu, A. Schirmer, A. McClure, S. B. del Cardayre, J. D. Keasling, Microbial production of fatty-acid-derived fuels and chemicals from plant biomass, Nature, 463 (2010) 559
[67] T. L. Alleman, L. Fouts, R. L. McCormick, Quality analysis of wintertime B6-B20 biodiesel blend samples collected in the United States, Fuel Processing Technology, 92 (2011) 1297
[68] A. Macor, F. Avella, D. Faedo, Effects of 30 % v/v biodiesel/diesel fuel blend on regulated and unregulated pollutant emissions from diesel engines, Applied Energy, 88 (2011) 4989
[69] J. R. Garcia, O programa nacional de produção e uso de biodiesel brasileiro e a agricultura familiar na região nordeste, dissertação de mestrado, Instituto de economia, Universidade Estadual de Campinas, Campinas, 2007
[70] A comprehensive analysis of biodiesel impacts on exhaust emissions, Draft technical report, U. S. Environmental Protection Agency – EPA, 2002
[71] D. F. Amaral, Desmistificando o programa nacional de produção e uso do biodiesel. A visão da indústria brasileira de óleos vegetais, ABIOVE, São Paulo, 2009, http://abiove.com.br
[72] R. Rinaldi, C. Garcia, L. L. Marciniuk, A. V. Rossi, U. Schuchardt, Síntese de biodiesel: uma proposta contextualizada de experimento para laboratório de química geral, Química Nova, 30 (2007) 1374
[73] M. F. Pimentel, G. M. G. S. Ribeiro, R. S. Cruz, L. Stragevitch, J. G. A. Pacheco Filho, L. S. G. Teixeira, Determination of biodiesel content when blended with mineral diesel fuel using infrared spectroscopy and multivariate calibration, Microchemical Journal, 82 (2006) 201

[74] J. S. Oliveira, R. Montalvão, L. Daher, P. A. Z Suarez, J. C. Rubim, Determination of methyl ester contents in biodiesel blends by FTIR-ATR and FTNIR spectroscopies, Talanta, 69 (2006) 1278
[75] G. Knothe, Determining the blend level of mixtures of biodiesel with conventional diesel fuel by fiber-optic near infrared spectroscopy and ^{1}H nuclear magnetic resonance spectroscopy, J. Am. Oil Chem. Soc., 78 (2001) 1025
[76] M. A. Aliske, G. F. Zagonel, B. J. Costa, W. Veiga, C. K. Saul, Measurement of biodiesel concentration in a diesel oil mixture, Fuel, 86 (2007) 1461
[77] A. J. Caines, R. F. Haycock, Automotive lubricants reference book, Society of Automotive Engineers, Inc., Warrendale, 1996
[78] Standard test method for calculation of carbon distribution and structural group analysis of petroleum oils by the n-d-m method – ASTM D3238-95, ASTM international, West Conshohocken, 2010
[79] S. Z. Erhan, S. Asadauskas, Lubricant basestocks from vegetable oils, Industrial Crops and Products, 11 (2000) 277
[80] Boletim mensal do monitoramento dos lubrificantes – agosto 2010, Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP, 2010
[81] Boletim mensal do monitoramento dos lubrificantes – dezembro 2010, Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP, 2010
[82] R. M. Balabin, R. Z. Safieva, Motor oil classification by base stock and viscosity based on near infrared (NIR) spectroscopy data, Fuel, 87 (2008) 2745
[83] F. S. G. Lima, M. A. S. Araujo, L. E. P. Borges, Determination of lubricant base oil properties by near infrared spectroscopy using different sample and variable selection methods, J. Near Infrared Spectroscopy, 12 (2004) 159
[84] R. W. Kennard, L. A. Stone, Computer aided design of experiments, Technometrics, 11 (1969) 137
[85] C. C. Chang, C. J. Lin, LIBSVM: a library for support vector machines, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm
[86] M. M. de Sena, M. G. Trevisan, R. J. Poppi, Talanta, 68 (2006) 1707
[87] N. Baccan, J. C. de Andrade, O.E. Godinho, J.S. Barone, Quimica analítica quantitativa elementar, 3ª ed., Ed. Edgard Blucher Ltda, São Paulo, 2001
[88] J. W. B. Braga, Aplicação e validação de modelos de calibração de segunda ordem em química analítica, tese de doutorado, Instituto de química, Universidade Estadual de Campinas, Campinas, 2008