PAULO ROBERTO FILGUEIRAS

REGRESSÃO POR VETORES DE SUPORTE APLICADO NA DETERMINAÇÃO DE PROPRIEDADES FÍSICO-QUÍMICAS DE PETRÓLEO E BIOCOMBUSTÍVEIS

CAMPINAS

2014

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA

PAULO ROBERTO FILGUEIRAS

REGRESSÃO POR VETORES DE SUPORTE APLICADO NA DETERMINAÇÃO DE PROPRIEDADES FÍSICO-QUÍMICAS DE PETRÓLEO E BIOCOMBUSTÍVEIS

ORIENTADOR: PROF. DR. RONEI JESUS POPPI

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

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA POR PAULO ROBERTO FILGUEIRAS, E ORIENTADA PELO PROF.DR. RONEI JESUS POPPI.

_____
Assinatura do Orientador

CAMPINAS
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Química
Simone Lucas Gonçalves de Oliveira - CRB 8/8144

F475r Filgueiras, Paulo Roberto, 1982-
Regressão por vetores de suporte aplicado na determinação de propriedades físico-químicas de petróleo e biocombustíveis / Paulo Roberto Filgueiras. – Campinas, SP : [s.n.], 2014.

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

1. Quimiometria. 2. Petróleo. 3. Biodiesel. 4. Regressão por vetores de suporte. I. Poppi, Ronei Jesus. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Support vector regression applied to the determination of physicochemical properties of petroleum and biofuels
Palavras-chave em inglês:
Chemometrics
Crude oil
Biodiesel
Support vector regression
Área de concentração: Química Analítica
Titulação: Doutor em Ciências
Banca examinadora:
Ronei Jesus Poppi [Orientador]
Eustáquio Vinícius Ribeiro de Castro
Jez Willian Batista Braga
José Alberto Fracassi da Silva
Francisco Benedito Teixeira Pessine
Data de defesa: 27-08-2014
Programa de Pós-Graduação: Química

Agradecimentos

Este trabalho não foi construído sozinho, muitos contribuíram para eu poder chegar a estes resultados. Agradeço:

Aos meus irmãos: Anderson, Carlos e Laiz pela força familiar e em especial à minha mãe Teresinha pela confiança depositada durante estes anos de estudos;

Ao Prof. Dr. Ronei Jesus Poppi, pela orientação e oportunidade de realização deste trabalho;

Ao grupo LAQQA: Carlos D. L. Albuquerque, Guilherme Alexandrino, Mariana A. Ramos, Guto, Luciana A. Terra, Luciana F. Oliveira, Débora, Mônica, Márcia, André e Humberto pelo apoio e discussões sempre proveitosas;

Aos amigos da república D5: Bruno, André (Caiçara), Gabriel Chaves, Carlos (Bixo) e Mineiro pela confiança e estadia nestes 3 anos de Campinas. Em especial ao Wanderson Romão e ao Marcos F. Franco pela indicação e confiança depositada em mim no início do doutorado;

Aos meus amigos do jiu-jitsu: Sensei Carlinhos, Alex, Gafanhoto, Grilo, Tabajara, Zidane, Rafael, Piloto, Júlia, Tatão e Moicano pela ajuda nos treinamentos que sempre foram uma válvula de escape do estresse da rotina acadêmica;

Em especial à minha noiva Samantha Ribeiro Campos da Silva, pelos incentivos, companheirismo e compreensão para eu continuar trabalhando mesmo nos momentos difíceis.

Ao Instituto de Química na Universidade Estadual de Campinas pelas condições de trabalho.

Ao LabPetro pelas amostras concedidas e ao CNPq pelo apoio financeiro (Processo número 146807/2011-1).

Curriculum Vitae

Formação acadêmica

Doutorando em Ciências – área Química Analítica, 2014. Universidade Estadual de Campinas – UNICAMP, Campinas-SP.

Mestre em química – área Físico-Química, 2011. Universidade Federal do Espírito Santo – UFES, Vitória-ES.

Bacharel e licenciado em Química, 2008. Universidade Federal do Espírito Santo – UFES, Vitória-ES.

Licenciado em Matemática, 2004. Faculdade de Filosofia, Ciências e Letras de Alegre, Alegre-ES.

Atividades acadêmicas

Artigos publicados durante o doutorado

Paulo R. Filgueiras, Cristina M. S. Sad, Alexandre R. Loureiro, Maria F. P. Santos, Eustáquio V. R. Castro, Júlio C. M. Dias, Ronei J. Poppi. Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. Fuel 2014, 116, 123-130.

Paulo R. Filgueiras, Júlio C. L. Alves, Cristina M. S. Sad, Eustáquio V. R. Castro, Júlio C. M. Dias, Ronei J. Poppi. Evaluation of trends in residuals of multivariate calibration models by permutation test. Chemometrics and Intelligent Laboratory Systems 2014, 133, 33-41.

Paulo R. Filgueiras, Júlio C. L. Alves, Ronei J. Poppi. Quantification of animal fat biodiesel in soybean biodiesel and B20 diesel blends using near infrared spectroscopy and synergy interval support vector regression. Talanta 2014, 119, 582-589.

Luciana A. Terra, Paulo R. Filgueiras, Lílian V. Tose, Wanderson Romão, Douglas D. de Souza, Eustáquio V. R. de Castro, Mirela S. L. de Oliveira, Júlio C. M. Dias,

Ronei J. Poppi. Petroleomics by electrospray ionization FT-ICR mass spectrometry coupled to partial least squares with variable selection methods: prediction of the total acid number of crude oils. Analyst 2014, 139, 4908–4916.

Luciana F. Oliveira, Soraia C. G. N. Braga, Paulo R. Filgueiras, Fábio Augusto, Ronei J. Poppi. Assessment of robustness on analysis using headspace solid-phase microextraction and comprehensive two-dimensional gas chromatography through experimental designs. Talanta 2014, 129, 303-308.

Álvaro C. Neto, Emanuele C. S. Oliveira, Valdemar Lacerda Jr., Eustáquio V.R. Castro, Wanderson Romão, Renzo C. Silva, Roni G. Pereira, Tiago Sten, Paulo R. Filgueiras, Ronei J. Poppi. Quality control of ethanol fuel: Assessment of adulteration with methanol using ^{1}H NMR. Fuel 2014, 135, 387-392.

Thieres M. C. Pereira, Josué A. Q. júnior, Rafael S. Ortiz, Werickson F. C. Rocha, Denise C. Endringer, Paulo R. Filgueiras, Ronei J. Poppi, Wanderson Romão. Viagra and Cialis blister packaging fingerprinting using Fourier transform infrared spectroscopy (FTIR) allied with chemometric methods. Analytical Methods 2014, 6, 2722-2728.

André M. Souza, Márcia C. Breitkreitz, Paulo R. Filgueiras, Jarbas J. R. Rohwedder, Ronei J. Poppi. Experimento didático de quimiometria para calibração multivariada na determinação de paracetamol em comprimidos comerciais utilizando espectroscopia no infravermelho próximo: um tutorial, parte II. Química Nova 2013, 36, 1057-1065.

Paulo R. Filgueiras, Luciana A. Terra, Eustáquio V. R. Castro, Lize M. S. L. Oliveira, Júlio C. M. Dias, Ronei J. Poppi. Prediction of the distillation temperatures of crude oils using 1H NMR and support vector regression analyses with estimated confidence intervals. Energy & Fuels (artigo submetido).

RESUMO

REGRESSÃO POR VETORES DE SUPORTE APLICADO NA DETERMINAÇÃO DE PROPRIEDADES FÍSICO-QUÍMICAS DE PETRÓLEO E BIOCOMBUSTÍVEIS

O petróleo é constituído por uma mistura complexa de composição química heterogênea. Sua completa avaliação envolve cerca de 700 ensaios físico-químicos, consumindo de 10 a 70 litros de amostra em aproximadamente 1 ano de análises. Visando reduzir tempo e quantidade de amostra, nesta Tese, métodos espectroscópicos aliados à Regressão por Vetores de Suporte (SVR) foram aplicados na determinação de algumas propriedades físico-químicas de petróleos e biocombustíveis. Diferentes abordagens para otimização e interpretação dos modelos SVR foram desenvolvidas: técnica para determinar as variáveis mais importantes na construção dos modelos, estimativa de intervalos de confiança nas previsões e avaliação de tendências nos resíduos. Foram realizadas quatro aplicações com diferentes técnicas instrumentais. A primeira aplicação foi direcionada a interpretação dos modelos SVR construídos a partir de espectros de infravermelho médio (MIR) na determinação da gravidade API, viscosidade cinemática e teor de água em petróleos. Na segunda aplicação foi desenvolvido um método para estimar o intervalo de confiança de modelos SVR aplicados a espectros de Ressonância Magnética Nuclear de próton (RMN de ^{1}H) na determinação das temperaturas equivalentes a 10%, 50% e 90% de volume destilado de petróleo. Na terceira aplicação foi desenvolvido um método para selecionar variáveis espectrais e otimizar os parâmetros do modelo SVR simultaneamente por algoritmo genético, aplicado a espectros de Ressonância Magnética Nuclear de carbono 13 (RMN de ^{13}C) na determinação de saturados, aromáticos, resinas e asfaltenos (SARA) em petróleos. Na última aplicação, procurou-se selecionar variáveis espectrais utilizando o método de sinergismo de intervalos, aplicado a espectros de infravermelho próximo (NIR) para quantificar biodiesel de gordura animal em mistura com biodiesel de soja e diesel B20. Os resultados apontam o SVR como excelente ferramenta para calibração multivariada aplicada a dados complexos como petróleo e biocombustíveis.

ABSTRACT

SUPPORT VECTOR REGRESSION APPLIED TO THE DETERMINATION OF PHYSICOCHEMICAL PROPERTIES OF PETROLEUM AND BIOFUELS

Crude oil is composed by a complex mixture of heterogeneous chemical composition. Its full evaluation involves about 700 physicochemical experiments, consuming about 10-70 liters of sample in there about 1 year of analysis. In order to reduce time and amount of sample, in this Thesis, spectroscopic methods combined with Support Vector Regression (SVR) were applied in determination of physicochemical properties of petroleum and biofuels. Different approaches for optimization and interpretation of SVR models were developed: techniques to determine the most important variables in the model development, determination of confidence intervals in predictions and assessment of trends in residuals. Four applications with different instrumental techniques were performed. The first application was directed to interpretation of SVR models built from mid infrared (MIR) spectra to determination of the API gravity, kinematic viscosity and water content in petroleum. In the second application, it was developed a method to estimation of the confidence interval of SVR models applied in spectra of proton nuclear magnetic resonance (^{1}H NMR) for the determination of equivalent temperatures to 10%, 50% and 90% of distillate volume in petroleum. In the third application it was developed a method for spectral variables selection and optimization the SVR model parameters simultaneously by genetic algorithm, applied to nuclear magnetic resonance spectra of carbon 13 (^{13}C NMR) in determination of saturates, aromatics, resins and asphaltenes (SARA) in petroleum. In the last application, it was proposed a method for spectral variables selection using the synergism of intervals, applied to near-infrared (NIR) spectra to quantify biodiesel from animal fat in blend with biodiesel from soybean and diesel B20. The results indicate the SVR as an excellent tool for multivariate calibration applied to complex dataset such as petroleum and biofuels.

Sumário

Lista de abreviaturas..... xvii

Lista de tabelas..... xix

Lista de figuras..... xxi

1 Introdução..... 1
1.1 Objetivos..... 3
1.2 Estrutura da tese..... 3

2 Métodos quimiométricos..... 5
2.1 Regressão por Mínimos Quadrados Parciais – PLS..... 6
2.2 Máquina de Vetores de Suporte – SVM..... 7
2.2.1 Margem suave..... 11
2.2.2 Problema linearmente não separável..... 12
2.2.3 Regressão por vetores de suporte..... 19
2.3 Otimização dos modelos de calibração multivariada..... 18
2.4 Métodos de avaliação de modelos de calibração..... 22
2.4.1 Avaliação de erros sistemáticos..... 22
2.4.2 Avaliação de erro de tendência..... 23
2.4.3 Comparação da exatidão entre modelos de calibração..... 26

3 Aplicação 1: Determinação da gravidade API, viscosidade cinemática e teor de água em petróleo por espectroscopia ATR-FTIR e calibração multivariada..... 29
3.1 Introdução..... 29
3.1.2 Objetivos..... 30
3.2 Metodologia..... 31
3.2.1 Construção dos modelos de calibração multivariada..... 32
3.3 Resultados e discussões..... 33
3.3.1 Análise da importância das variáveis no modelo..... 40
3.4 Conclusões..... 48

4 Aplicação 2: Cálculo do intervalo de confiança para modelos de regressão SVR por ensemble tipo *boosting* na determinação de temperaturas equivalentes a 10%, 50% e 90% de volume de destilado em petróleos por espectroscopia de RMN de ^{1}H..... 49
4.1 Introdução..... 49
4.1.2 Objetivos..... 50
4.2 *Boosting* ensemble..... 51
4.3 Metodologia..... 53
4.3.1 Valores de Referência..... 53
4.3.2 Medidas de RMN de ^{1}H..... 54

4.3.3 modelos de calibração..... 54
4.3.4 Determinação dos intervalos de confiança..... 55
4.4 Resultados e discussões..... 55
4.5 Conclusões..... 67

5 Aplicação 3: Determinação de saturados, aromáticos, resinas e asfaltenos em petróleo por RMN de ^{13}C e regressão por vetores de suporte com seleção de variáveis por algoritmo genético..... 69
5.1 Introdução..... 69
5.1.2 Objetivo..... 70
5.2 Algoritmo genético..... 71
5.3 Metodologia..... 75
5.4 Resultados e discussões..... 77
5.5 Conclusões..... 86

6 Aplicação 4: Quantificação de biodiesel de gordura animal em biodiesel de soja e diesel B20 usando espectroscopia no infravermelho próximo e regressão por vetores de suporte com sinergismo de intervalos..... 87
6.1 Introdução..... 87
6.1.2 Objetivo..... 89
6.2 Metodologia..... 89
6.3 Resultados e discussões..... 91
6.3.1 Quantificação de biodiesel de gordura animal em misturas com biodiesel de soja..... 92
6.3.2 Quantificação de biodiesel de gordura animal no diesel B20..... 95
6.4 Conclusões..... 99

7 Conclusões gerais..... 101

8 Referências bibliográficas..... 103

Apêndice A – Formulação de um problema de classificação binária por um separador linear..... 113
Apêndice B – Otimização dos parâmetros do modelo SVR por grade de pesquisa..... 115

Lista de abreviaturas

ANN	a*rtificial neural networks*
API	*american petroleum Institute*
ASTM	*american society for testing and materials*
ATR	*attended total reflectance*
FTIR	*Fourier transformed infrared*
GA	*genetic algorithm*
GA-SVR	*genetic algorithm - support vector regression*
GC × GC-FID	*comprehensive two-dimensional gas chromatography with flame ionization detection*
ISO	*international organization for Standardization*
LabPetro	laboratório de pesquisa e desenvolvimento de metodologias para análise de óleos pesados
MIR	*middle infrared spectroscopy*
MSC	*multiplicative signal correction*
OSH	*optimal separating hyperplane*
PLS	*partial least squares*
R^2cv	coeficiente de determinação para validação cruzada
R^2p	coeficiente de determinação para previsão
RBF	*radial basis function*
RMN de ^{13}C	ressonância magnética nuclear de carbono 13.
RMN de ^{1}H	ressonância magnética nuclear de próton.
RMSECV	*root mean square error of cross-validation*
RMSEP	*root mean square error of prediction*
SARA	saturados, aromáticos, resinas e asfaltenos
SEC	*standard error of calibration*
SIMDIS	*simulated distillation*
SNV	*standard normal variate*
SDV	*standard deviation of the validation errors*
SVM	*support vector machines*
SVR	*support vector regression*

T10%	temperatura equivalente a 10% de volume de evaporados
T50%	temperatura equivalente a 50% de volume de evaporados
T90%	temperatura equivalente a 90% de volume de evaporados
VL	variáveis latentes

Lista de tabelas

Tabela 1.1. Classificação de petróleos de acordo com a gravidade API..... 4

Tabela 3.1. Resultados dos modelos PLS e SVR na determinação de propriedades físico-químicas de petróleo..... 34

Tabela 4.1. Resultados de previsão dos modelos de calibração..... 58

Tabela 4.2. Valores de t-calculado para o teste para viés e p-valores do teste de permutação para erros de tendência..... 62

Tabela 5.1. Resultados dos parâmetros estatísticos dos modelos GA-SVR..... 84

Tabela 6.1. Resultados dos modelos de calibração para quantificação do teor de biodiesel de gordura animal em mistura com biodiesel de soja..... 92

Tabela 6.2. Resultados dos modelos de calibração para quantificação do teor de biodiesel de gordura animal no diesel B20..... 96

Lista de figuras

Figura 2.1. Representação esquemática da construção da matriz de dados $\mathbf{X}$ 6

Figura 2.2. Gráfico ilustrativo de um problema de classificação binária linearmente separável no espaço original..... 8

Figura 2.3. Separação possível das classes por hiperplanos (a); separação das classes pelo hiperplano de separação ótima (b); soluções do problema de classificação binária com mínimos locais (c); solução do problema de classificação binária com mínimo global (d)..... 9

Figura 2.4. Hiperplano de separação ótima com margens e os vetores de suporte (a). Hiperplanos de separação e margem (b)..... 10

Figura 2.5. Gráfico ilustrativo de um problema de classificação binária linearmente inseparável no espaço original..... 11

Figura 2.6. Gráfico ilustrativo de um problema de classificação binária não linearmente separável no espaço original (a). Projeção das amostras de entrada no espaço de características de alta dimensão (b). Representação ilustrativa de uma função não linear de classificação no espaço original (c)..... 14

Figura 2.7. Exemplo gráfico das funções kernel mais utilizadas para mapeamento em máquinas de vetores de suporte..... 15

Figura 2.8. Ilustração gráfica da transformação de um problema de regressão em classificação binária. Problema de regressão (a); adição e subtração de uma constante aos dados de entrada (b); transformação do problema de regressão original em um problema de classificação binária (c); hiperplano que define a função de regressão (d)..... 16

Figura 2.9. Expressão gráfica do limite de tolerância ε-insensivel..... 17

Figura 2.10. Resíduos simulados normalmente distribuídos com média zero e variância unitária: sem presença de tendência (a); histograma dos resíduos sem tendência (b); resíduos com tendência quadrática (c); histograma dos resíduos com tendência quadrática (d)..... 24

Figura 2.11. Histograma ilustrativo da distribuição dos coeficientes polinomiais de maior ordem (b_n) para os ajustes permutados. A linha vertical vermelha representa o valor de b_n^* (neste caso positivo) para os dados originais. Em cinza são os coeficientes menores que b_n^* e em verde os coeficientes maiores que b_n^* 26

Figura 2.12. Histograma ilustrativo da distribuição das diferenças médias das permutações. A linha vertical vermelha representa o valor de $\bar{d}^*$ para os dados originais. Em cinza são as diferenças médias menores que $\bar{d}^*$ e em verde as diferenças médias maiores que $\bar{d}^*$ 27

Figura 3.1. Espectros de infravermelho dos petróleos usados no modelo de calibração, sem pré-processamento (a); após pré-processamento por MSC (b)..... 33

Figura 3.2. Gráfico da gravidade API prevista pelo modelo PLS pelos valores de referência (a); resíduos da modelagem PLS (b); Gráfico da gravidade API prevista pelo modelo SVR pelos valores de referência (c); resíduos da modelagem SVR (d). Amostras de calibração (o) e previsão (◊)..... 35

Figure 3.3. Gráfico da viscosidade cinemática prevista pelo modelo PLS pelos valores de referência (a); resíduos da modelagem PLS (b); Gráfico da viscosidade cinemática prevista pelo modelo SVR pelos valores de referência (c); resíduos da modelagem SVR (d). Amostras de calibração (o) e previsão (◊)..... 37

Figura 3.4. Resíduos da modelagem PLS na determinação da viscosidade cinemática (a); histograma dos coeficientes quadráticos do teste de permutação dos resíduos do modelo PLS (b); Resíduos da modelagem SVR na determinação da viscosidade cinemática (c); histograma dos coeficientes quadráticos do teste de permutação dos resíduos do modelo SVR (d); Amostras de calibração (o) e previsão (◊)..... 38

Figure 3.5. Gráfico do teor de água em petróleos previsto pelo modelo PLS pelos valores de referência (a); resíduos da modelagem PLS (b); Gráfico do teor de água em petróleos previsto pelo modelo SVR pelos valores de referência (c); resíduos da modelagem SVR (d). Amostras de calibração (o) e previsão (◊)..... 40

Figura 3.6. Espectro ATR-FTIR médio das amostras de calibração (a); coeficientes de regressão do modelo PLS, com 6 variáveis latentes, na determinação da gravidade API (b); p-vetor do modelo SVR na determinação da gravidade API..... 42

Figura 3.7. Figura 3.9. Espectros de ATR-FTIR das amostras de petróleo do conjunto de calibração com gradiente de cor em função do aumento da gravidade API da amostra. Azul: menor gravidade API; vermelho: maior gravidade API..... 43

Figura 3.8. Espectro ATR-FTIR médio das amostras de calibração (a); coeficientes de regressão do modelo PLS, com 6 variáveis latentes, na determinação da viscosidade cinemática (b); p-vetor do modelo SVR na determinação da viscosidade cinemática..... 44

Figura 3.9. Espectros de ATR-FTIR das amostras de petróleo do conjunto de calibração com gradiente de cor em função do aumento da viscosidade cinemática da amostra. Azul: menor viscosidade cinemática; vermelho: maior viscosidade cinemática..... 45

Figura 3.10. Espectro ATR-FTIR médio das amostras de calibração (a); coeficientes de regressão do modelo PLS, com 4 variáveis latentes, na determinação da teor de água (b); p-vetor do modelo SVR na determinação da teor de água..... 46

Figura 3.11. Espectros de ATR-FTIR das amostras de petróleo do conjunto de calibração com gradiente de cor em função do aumento do teor de água da amostra. Azul: menor teor de água; vermelho: maior teor de água..... 47

Figura 4.1. Histograma da gravidade API medida para as 35 amostras de petróleo utilizadas na modelagem..... 54

Figura 4.2. Espectros de RMN de ^{1}H para um petróleo extraleve com gravidade API de 54,0 (—) e um petróleo pesado com gravidade API de 17,0 (—)..... 56

Figura 4.3. Espectros de RMN de ^{1}H das 35 amostras de petróleos utilizadas na modelagem antes do alinhamento (a) e após alinhamento pelo programa icoshift (b)..... 57

Figura 4.4. Temperatura equivalente a 10% de volume destilado de petróleo medido pelo método ASTM D 7169 *versus* valor previsto pelo modelo PLS (a), ensemble PLS (b), SVR (c) e ensemble SVR (d). As barras verticais representam o limite de confiança de 95%..... 59

Figura 4.5. Temperatura equivalente a 50% de volume destilado de petróleo medido pelo método ASTM D 7169 *versus* valor previsto pelo modelo PLS (a), ensemble PLS (b), SVR (c) e ensemble SVR (d). As barras verticais representam o limite de confiança de 95%..... 60

Figura 4.6. Temperatura equivalente a 90% de volume destilado de petróleo medido pelo método ASTM D 7169 *versus* valor previsto pelo modelo PLS (a), ensemble PLS (b), SVR (c) e ensemble SVR (d). As barras verticais representam o limite de confiança de 95%..... 61

Figura 4.7. Gráfico de barras dos pesos do modelo ensemble SVR para as propriedades: T10% (a), T50% (b) e T90% (c); e frequência de amostragem de cada amostra de calibração para construção do modelo eSVR para as propriedades: T10% (d), T50% (e) e T90% (f)..... 63

Figura 4.8. RMSEP% das propriedades T10%, T50% e T90% dos modelos PLS, ensemble PLS, SVR e ensemble SVR..... 64

Figura 4.9. Gráfico do p-vetor médio do modelo ensemble SVR em função do deslocamento químico para as propriedades: T10% (a), T50% (b) e T90% (c)..... 66

Figura 5.1. Exemplo da estrutura de uma molécula de resina (a) e asfalteno (b)..... 70

Figura 5.2. Fluxograma das principais etapas de um algoritmo genético básico..... 72

Figura 5.3. Exemplo ilustrativo da otimização de parâmetros do modelo SVR simultaneamente a seleção de variáveis de espectros de RMN de ^{13}C por algoritmo genético..... 77

Figura 5.4. Histograma da gravidade API das 65 amostras de petróleos utilizadas..... 78

Figura 5.5. Espetros de RMN de ^{13}C de petróleos brutos originais (a) e após alinhamento utilizando o programa icoshift (b)..... 79

Figura 5.6. Gráfico do RMSECV em função da frequência de seleção das variáveis espectrais obtidas com execução de 100 modelos GA-SVR..... 80

Figura 5.7. Histograma dos parâmetros do modelo SVR otimizados pelo algoritmo genético na determinação do teor de saturados, aromáticos e resinas mais asfaltenos. A linha vertical vermelha é o valor mediano da distribuição..... 81

Figura 5.8. (a) espectro médio de RMN de C13 das amostras de calibração com região de predominância de carbonos aromáticos destacada à direita. Frequência das variáveis selecionadas para saturados (b), aromáticos (c) e resinas e asfaltenos (d). A linha horizontal em vermelho representa a frequência mínima de seleção para construção do modelo GA-SVR..... 83

Figura 5.9. Gráfico dos teores de SARA medidos segundo a norma ASTM D 7169 pelos valores de estimados pelo modelo GA-SVR. (a) Saturados, (b) Aromáticos, (c) Resinas e Asfaltenos. Amostras de calibração (o) e previsão (◊)..... 85

Figura 6.1. Espectro NIR de biodiesel de gordura animal com 31% m/m de biodiesel de soja (—), biodiesel de soja (—) e diesel B20 com 20% m/m de biodiesel de soja (—), dividido em 10 intervalos..... 91

Figura 6.2. Gráfico dos resíduos dos modelos de calibração em função do teor de biodiesel de gordura animal em biodiesel de soja com ajuste quadrático aplicado aos dados (—). Modelo PLS (a), siPLS (b), SVR (c) e siSVR (d). Amostras de calibração (o) e previsão (◊)..... 93

Figura 6.3. histograma dos coeficientes quadráticos do teste de permutação dos resíduos dos modelos: PLS (a), siPLS (b), SVR (c) e siSVR (d). A linha vermelha vertical sólida refere-se ao coeficiente quadrático ajustado aos dados..... 95

Figura 6.4. Gráfico dos resíduos dos modelos de calibração em função do teor de biodiesel de gordura animal no diesel B20 com ajuste quadrático aplicado aos dados (—). Modelo PLS (a), siPLS (b), SVR (c) e siSVR (d). Amostras de calibração (o) e previsão (◊)..... 97

Figura 6.5. histograma dos coeficientes quadráticos do teste de permutação dos resíduos dos modelos: PLS (a), siPLS (b), SVR (c) e siSVR (d). A linha vermelha vertical sólida refere-se ao coeficiente quadrático ajustado aos dados..... 98

Figura A.1. Gráfico ilustrativo da utilização de um hiperplano como classificador primário no método SVM..... 113

1 Introdução

O petróleo é constituído por uma mistura de compostos orgânicos altamente complexa de composição química muito heterogênea.[1] Devido a esta complexidade sua qualidade no setor de produção e exploração é avaliada por propriedades físico-químicas. Estas propriedades indicam mudanças que possam ocorrer no petróleo de um mesmo campo produtor e auxiliam na elaboração de estratégias de transporte e refino além de informarem sobre potenciais derivados esperados.[1-4]

A partir de dados de 2013, a Petrobras estima dobrar a produção de petróleo até 2020, devido à participação de campos produtores na camada do pré-sal. A avaliação do petróleo produzido envolve cerca de 700 ensaios físico-químicos, consumindo de 10 a 70 litros de petróleo, em não menos de 4 meses, ao custo estimado de mais de 80 mil dólares. Dos 700 ensaios, 26 são diretamente no petróleo bruto. Devido à grande perspectiva de aumento da produção aliado ao longo tempo para análises e grandes quantidades de amostras, o desenvolvimento de metodologias analíticas rápidas e em menores quantidades de amostras são estratégicas para o controle do processo de produção e refino.

Um dos principais derivados do petróleo é o óleo diesel, que é comercializado como uma mistura de diesel de petróleo e biodiesel. O biodiesel é atualmente o principal substituto do diesel de petróleo, devido suas características similares, e tem sido utilizado em muitos países como combustível para motores diesel, usualmente em mistura de até 20% v/v em diesel de petróleo, conhecido como B20.[5]

Estudos recentes envolvendo diesel, biodiesel, petróleo e derivados mostraram a potencialidade de técnicas espectroscópicas como o infravermelho[6-18] e a ressonância magnética nuclear[19-24] para o desenvolvimento de metodologias analíticas para estimar propriedades físico-químicas de óleos, apresentando como principais vantagens o uso de pequenas quantidades de amostra, procedimentos analíticos rápidos com mínimo pré-tratamento de amostra. Nestas metodologias, a

conversão da resposta instrumental na informação de interesse, requer a utilização de técnicas de calibração multivariada.[6-24]

A regressão por mínimos quadrados parciais (PLS, do inglês *Partial Least Squares)*[25] atualmente é o método de calibração multivariada mais utilizado em química analítica. Sua teoria é amplamente descrita na literatura[25-28] e disponível em pacotes de vários softwares estatísticos. O método é rápido computacionalmente além de ser eficaz quando os dados apresentam relação linear, normalmente obtido em matrizes mais simples. Em um conjunto de dados não lineares a técnica pode, em situações não muito severas, ser adaptada para o tratamento dos dados,[29,30] entretanto o melhor procedimento consiste em utilizar um método de calibração não linear, como as redes neurais artificiais (ANN, do inglês *Artificial Neural Networks*) ou máquina de vetores de suporte (SVM, do inglês *Support Vector Machines*).[31,32]

A SVM é um método de aprendizagem de máquina desenvolvido principalmente por Cortes & Vapnik,[33] originalmente para resolver problemas de classificação binária, entretanto, a técnica foi estendida para tratar problemas de classificação múlticlasses[34,35] e regressão (SVR, do inglês *Support Vector Regression*).[31,36]

Alguns fatores contribuem para a pouca utilização do método SVR, em relação ao PLS, em química analítica: menor disponibilidade do método em pacotes estatísticos devido a maior complexidade do algoritmo, maior dificuldade na otimização de parâmetros internos do modelo e obtenção de modelos de calibração não interpretáveis.

1.1 Objetivos

- Desenvolver metodologias analíticas para determinação de parâmetros físico-químicos de petróleos e biocombustíveis utilizando técnicas espectroscópicas associadas à regressão por vetores de suporte.
- Desenvolver metodologias estatísticas buscando torna os modelos de regressão por vetores de suporte mais interpretáveis possibilitando a identificação das variáveis mais importantes na construção do modelo e estimativa das figuras de mérito para o método não linear.
- Desenvolver metodologias para seleção de variáveis aplicada à regressão por vetores de suporte.

1.2 Estrutura da Tese

Nesta Tese são desenvolvidas aplicações do SVR para determinação de parâmetros de qualidade em petróleo e biocombustíveis, assim como o desenvolvimento de metodologias buscando a interpretação e validação de modelos SVR. Foram realizadas quatro aplicações com diferentes técnicas instrumentais em matrizes de petróleo biodiesel. Os petróleos utilizados possuem características físico-químicas muito diferentes, variando do extraleve a asfáltico, conforme a classificação de petróleo de acordo com a gravidade API (Tabela 1.1).

A seguir, no capítulo 2 é descrita a parte teórica dos algoritmos PLS e SVR assim como os métodos mais comuns de otimização e avaliação dos modelos construídos.

No capítulo 3 é mostrada uma aplicação direcionada a interpretação dos modelos SVR construídos. Foram determinados a gravidade API (escala de densidade da *American Petroleum Institute*), viscosidade cinemática e teor de água em petróleos a partir de espectros no infravermelho médio (MIR) buscando identificar as regiões espectrais com maior contribuição para construção do modelo.

No capítulo 4 é exposto um método para determinar o intervalo de confiança na estimativa do modelo SVR através da técnica *boosting* ensemble. Nesta aplicação foram determinadas as temperaturas equivalentes a 10%, 50% e 90% de volume de destilado de petróleo a partir de espectros de ressonância magnética nuclear de próton (RMN de ^{1}H).

No capítulo 5 é apresentado um método para selecionar variáveis espectrais e otimizar os parâmetros do modelo SVR simultaneamente por algoritmo genético. Foram determinados teores de: saturados, aromáticos, resinas e asfaltenos (SARA) em petróleos a partir de espectros de ressonância magnética nuclear de carbono 13 (RMN de ^{13}C).

No capítulo 6 é mostrado um procedimento para selecionar variáveis espectrais utilizando o método de sinergismo de intervalos. Nesta aplicação o teor de biodiesel proveniente de gordura animal foi quantificado em mistura com biodiesel de soja e no diesel B20 a partir de espectros no infravermelho próximo (NIR).

No capítulo 7 são apresentadas as conclusões gerais da Tese.

Tabela 1.1. Classificação de petróleos de acordo com a gravidade API.

Petróleo	Gravidade API	Densidade relativa a 20°C
Extraleve	> 40,0	< 0,821
Leve	40,0 – 33,0	0,821 – 0,857
Médio	33,0 – 27,0	0,857 – 0,889
Pesado	27,0 – 19,0	0,889 – 0,937
Extrapesado	19,0 – 15,0	0,937 – 0,962
Asfáltico	< 15,0	> 0,962

Fonte: GUIMARÃES, 2004.[37]

2 Métodos quimiométricos

Os métodos quimiométricos podem ser divididos em três grandes áreas: planejamento e otimização de experimentos, métodos qualitativos e de reconhecimento de padrões e calibração multivariada.[38] A química analítica quantitativa teve grande impacto com o desenvolvimento das técnicas de calibração multivariada. Nesta é possível estimar uma propriedade de interesse a partir de outras medições, que normalmente são espectros obtidos por procedimentos analíticos mais simples, rápidos, menos onerosos e que dependem de menor quantidade de amostra.

Em calibração multivariada os dados são organizados algebricamente em vetores e matrizes. Considere o conjunto de dados $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$, com $\mathbf{x}_i \in R^m$ sendo o vetor contendo o espectro e $y_i \in R$ o valor de referência para a amostra i. Em problemas quantitativos y_i representa o valor da propriedade de interesse e em problemas qualitativos a classe a qual a amostra pertence.

Os espectros são organizados na matriz de dados $\mathbf{X}$ de forma que cada amostra represente um vetor linha (Figura 2.1). Cada variável espectral (número de onda, deslocamento químico, etc.) é emparelhada na mesma coluna, assim todos os espectros devem ser medidos com mesma resolução espectral. Para n amostras com m variáveis espectrais, a matriz de dados terá dimensão $\mathbf{X}_{(n,m)}$ (n linhas por m colunas).

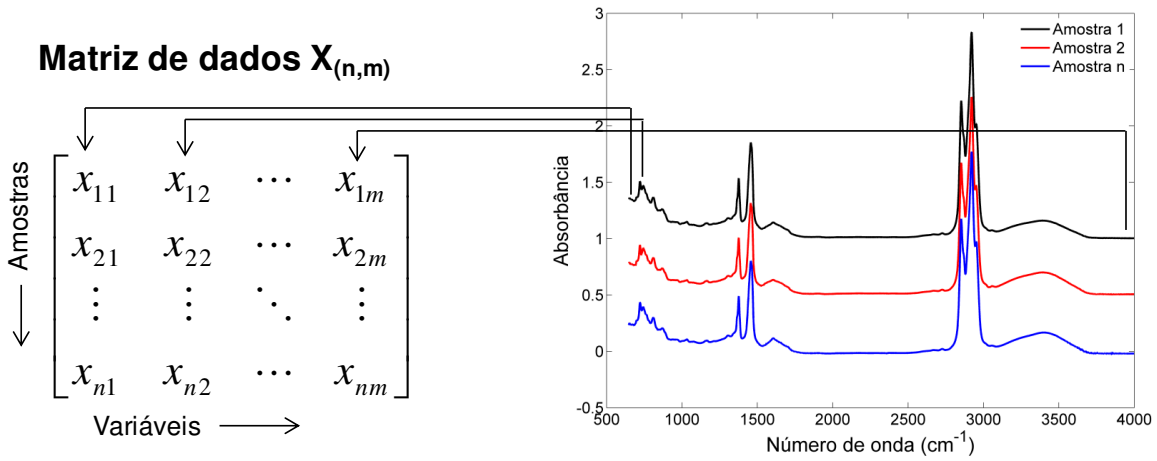

Figura 2.1. Representação esquemática da construção da matriz de dados $\mathbf{X}$.

2.1 Regressão por Mínimos Quadrados Parciais – PLS[25-28]

Atualmente a regressão por mínimos quadrados parciais[25-28] (PLS, do inglês *Partial Least Squares*) é o método de calibração multivariada mais utilizado em química analítica. Na regressão PLS a matriz de variáveis dependentes $\mathbf{X}$ e o vetor $\mathbf{y}$ contendo a propriedade de interesse são decompostos simultaneamente em uma soma de "a" variáveis latentes (VL), dada por:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \tag{2.1}$$

$$\mathbf{y} = \mathbf{Uq}^T + \mathbf{f} \tag{2.2}$$

onde **T** e **U** são os *scores*; **P** e **q** os *loadings* (pesos) e **E** e **f** os resíduos de **X** e **y** respectivamente. A matriz de *scores* **T** é estimada pela combinação linear de **X** com coeficientes ponderados por **W** (pesos):

$$\mathbf{T} = \mathbf{XW} \tag{2.3}$$

A partir de **W**, os coeficientes de regressão do modelo PLS podem ser estimados por:

$$\mathbf{b}_{PLS} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{q}^T \tag{2.4}$$

O modelo linear PLS pode ser representado por:

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{b}_{PLS} \tag{2.5}$$

onde valores da propriedade de interesse para amostras futuras podem ser estimados pelo modelo utilizando a combinação linear do vetor $\mathbf{x}_i$ amostral pelos coeficientes do modelo. Na construção do modelo de regressão PLS é necessário otimizar o número de variáveis latentes.

PLS é o método mais utilizado em calibração multivariada, quando os dados apresentam tendência linear, o método é rápido e computacionalmente eficiente. Mesmo com um conjunto de dados não linear a técnica pode, com certa dificuldade, ser adaptada. Entretanto, métodos de calibração não linear como redes neurais artificiais (ANN) ou Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machines*) são melhores recomentados. Problemas reais envolvendo matrizes complexas como petróleo, derivados e biocombustíveis, podem apresentar resultados insatisfatórios quando modelados simplesmente pelo método PLS. Seleção de variáveis ou aproximações não lineares[29] são alternativas normalmente utilizadas para contornar estes problemas, entretanto, uma excelente alternativa é aplicação de métodos não lineares de regressão como o SVR.

2.2 Máquina de Vetores de Suporte – SVM[33-35]

SVM é um método de aprendizagem de máquina desenvolvido por Vapnik nos anos 90[33,39] originalmente para tratar problemas de classificação binária.

Considere o caso de classificar duas classes de padrões linearmente separáveis em seu espaço original, como mostrado na Figura 2.2. Os dados de entrada são: ($\mathbf{x}_1$,y_1), ($\mathbf{x}_2$,y_2), … ($\mathbf{x}_n$,y_n) onde $x_i \in R^m$ é um vetor pertencente a uma das duas classes $y_i \in \{-1,+1\}$. O objetivo do SVM é encontrar um hiperplano que separe as classes sem erro, dado pela Equação:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{2.6}$$

Onde $\mathbf{w} \in R^m$ é um vetor de pesos, que multiplicado pelo vetor amostral $\mathbf{x}$ e somado à constante *b* define o hiperplano que separa as classes e permite

identificar a classe ao qual a amostra pertence. Se as classes são linearmente separáveis em seu espaço original, podem existir vários hiperplanos capazes de resolver o problema sem erros.

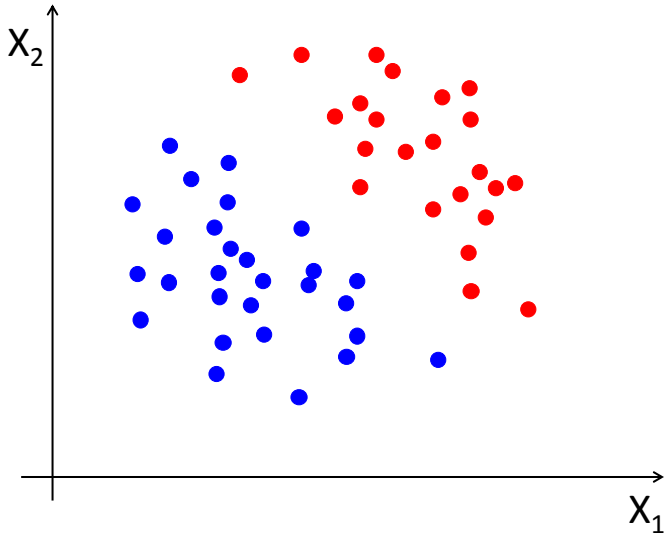

Figura 2.2. Gráfico ilustrativo de um problema de classificação binária linearmente separável no espaço original.

Na Figura 2.3a são mostrados vários hiperplanos que podem resolver o problema proposto. Quando várias soluções são possíveis, a cada treinamento, podemos obter qualquer uma das possíveis soluções (hiperplanos), chegando assim a vários mínimos locais (Figura 2.3c). O SVM utiliza o princípio da minimização do risco estrutural, no qual obtém o hiperplano de classificação com máxima distância das amostras de classificação (Figura 2.3b), chamado de hiperplano de separação ótima (OSH, do inglês *Optimal Separating Hyperplane*). Com este princípio, apenas uma solução é obtida para o problema proposto, fazendo as soluções convergirem para um mínimo global (Figura 2.3d).

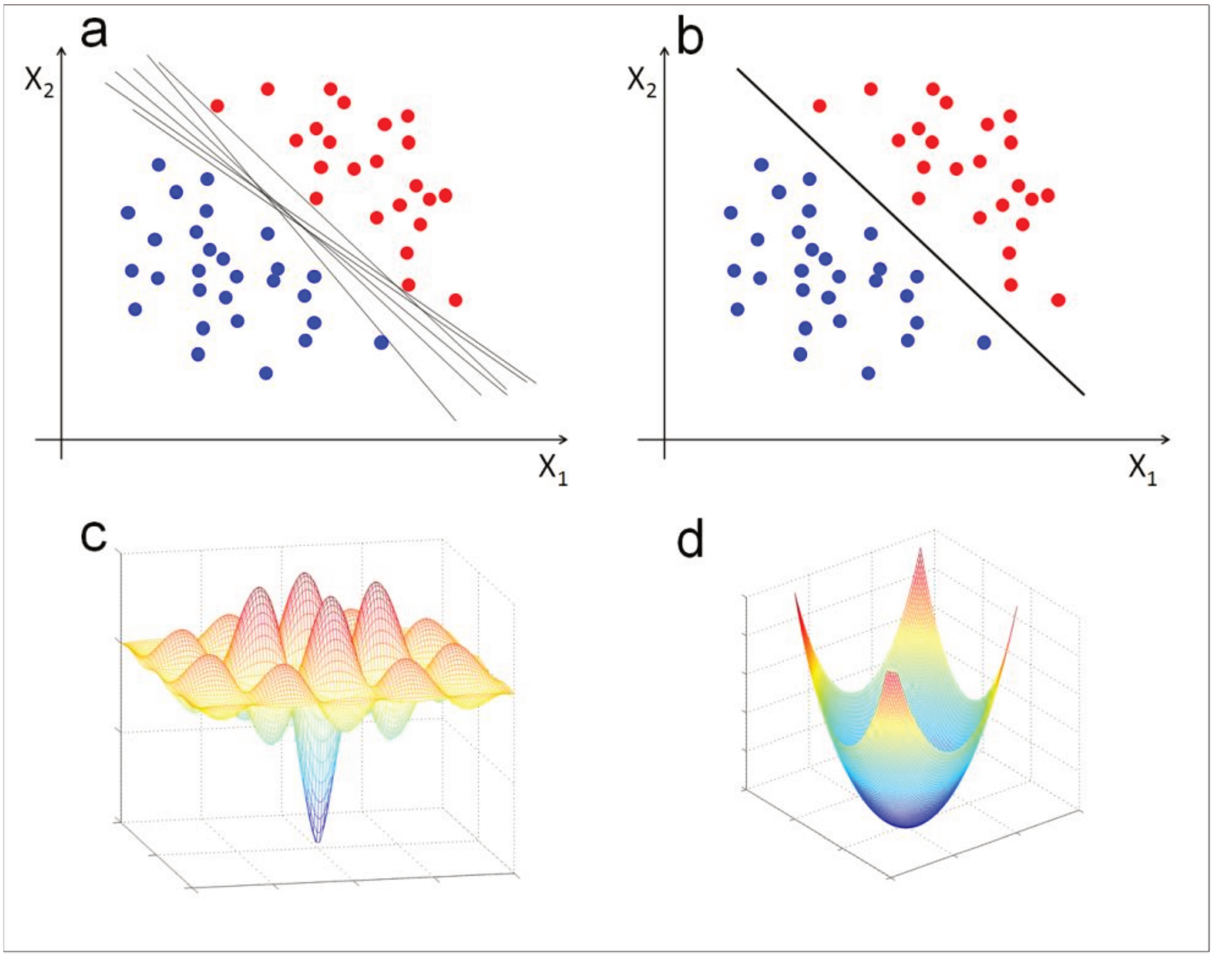

Figura 2.3. Separação possível das classes por hiperplanos (a); separação das classes pelo hiperplano de separação ótima (b); soluções do problema de classificação binária com mínimos locais (c); solução do problema de classificação binária com mínimo global (d).

Para uma única solução ótima, a função procurada apresentada na Equação 2.6 é descrita como um problema de otimização convexa:[33]

$$\text{minimize: } \frac{1}{2}\|\mathbf{w}\|^2 \tag{2.7}$$

$$\text{sujeito a : } \begin{cases} \mathbf{w}\cdot\mathbf{x}+b \geq +1 & \forall \quad y \in \{+1\} \\ \mathbf{w}\cdot\mathbf{x}+b \leq -1 & \forall \quad y \in \{-1\} \end{cases} \tag{2.8}$$

A dedução matemática que origina a Equação 2.7 é apresentada no Apêndice A. O OSH pode garantir ao SVM a melhor habilidade de previsão para

amostras futuras, ou seja, melhor capacidade em discriminar amostras de um conjunto de teste real. A máxima separação entre as classes é denominada margem e as amostras que limitam a separação entre as classes são chamadas de vetores de suporte (Figura 2.4a). Pelas restrições apresentadas na Equação 2.8 podemos observar que os vetores de suporte são as amostras nos quais $\mathbf{w}\cdot\mathbf{x}+b=+1$ para amostras pertencentes à classe +1 e $\mathbf{w}\cdot\mathbf{x}+b=-1$ para amostras pertencentes à classe -1 (Figura 2.4(b)).

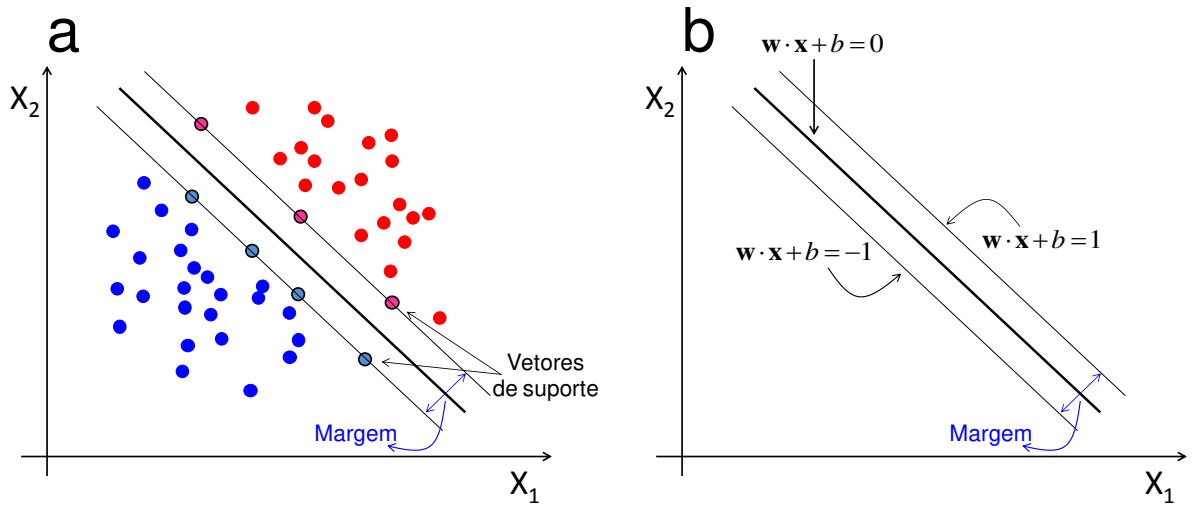

Figura 2.4. Hiperplano de separação ótima com margens e os vetores de suporte (a). Hiperplanos de separação e margem (b).

Os vetores de suporte do modelo SVM são as amostras que demarcam a fronteira entre as classes, sendo assim as mais importantes para a classificação. O modelo SVM é construído apenas com as amostras que são vetores de suporte. Na representação esquemática da Figura 2.4a, o modelo final seria construído apenas com 7 amostras.

A representação acima é válida para um conjunto de amostras linearmente separáveis no espaço original das amostras, porém na maioria das aplicações práticas isso não é possível.

2.2.1 Margem suave[33-35,39]

Na maior parte dos problemas reais não é possível separar duas classes de amostras por um hiperplano. Para contornar este problema, Corte e Vapnik (1995)[33] introduziram o conceito de variável de folga ξ_i, no qual admitem que algumas amostras possam ter um erro associado a sua classificação. Este erro é proporcional a distância da amostra ao hiperplano definido pelos vetores de suporte da classe à qual pertencem (Figura 2.5).

A margem para esta nova situação pode assumir um valor muito alto, assim, os erros de classificação são ponderados por uma constante C adicionada ao problema de otimização, tornando-o:

$$\text{minimize: } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{2.9}$$

$$\text{sujeito a : } \begin{cases} \mathbf{w}\cdot\mathbf{x} + b \geq +1 - \xi_i \\ \mathbf{w}\cdot\mathbf{x} + b \leq -1 + \xi_i \\ \xi_i \geq 0 \end{cases} \tag{2.10}$$

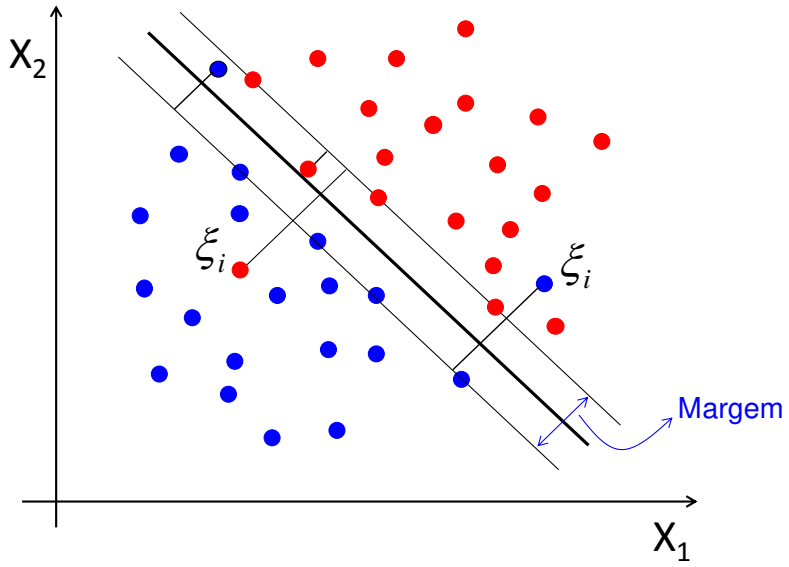

Figura 2.5. Gráfico ilustrativo de um problema de classificação binária linearmente inseparável no espaço original.

A constante C controla os erros de classificação. Para um baixo valor de C admite-se que muitas amostras possam ser classificadas com erros, gerando uma função de classificação mais suave. Caso o valor de C seja alto, a função de classificação torna-se menos suave para que as amostras não sejam erroneamente classificadas, podendo gerar um maior número de vetores de suporte. Assim o valor de C deve ser ajustado para cada problema.

Mesmo com o problema linear ajustado para o caso onde é admitido erros de classificação, problemas reais podem não ser lineares. Assim, uma transformação é aplicada aos dados de entrada do ajuste de uma função não linear.

2.2.2 Problema linearmente não separável[39-41]

O SVM[36] é popularmente conhecido como um método não linear. Mesmo em problema não linear, a função objetiva do SVM continua a mesma; encontrar um hiperplano de máxima separação entre as classes. Para problemas não lineares uma transformação dimensional prévia é aplicada aos dados de entrada.

Um problema de classificação binária linearmente não separável é mostrado na Figura 2.6a. Nela não é possível obter uma função linear que discrimine as duas classes de amostras, mesmo com aplicação de variável de folga. Neste caso, as amostras de entrada são mapeadas do espaço original para um espaço de alta dimensão chamado de espaço de características. Isto é feito pela aplicação de uma função kernel não linear (φ).

Por exemplo, cada amostra apresentada na Figura 2.6(a) possui um par de coordenadas definida por duas variáveis $\mathbf{x}_i = \left[x_i^{(1)}, x_i^{(2)}\right]$; $\mathbf{x}_j = \left[x_j^{(1)}, x_j^{(2)}\right]$. No espaço original estas classes são linearmente inseparáveis, porém este problema pode ser contornado aumentando a dimensão original dos dados pela aplicação de uma função kernel:[40]

$$\varphi(\mathbf{x}) = \left\{ \left(x^{(1)}\right)^2, \left(x^{(2)}\right)^2, \sqrt{2} \cdot x^{(1)} \cdot x^{(2)} \right\} \tag{2.11}$$

Assim, as amostras *i* e *j* terão como novas coordenadas:

$$\varphi(\mathbf{x}_i) = \left\{ \left(x_i^{(1)}\right)^2, \left(x_i^{(2)}\right)^2, \sqrt{2} \cdot x_i^{(1)} \cdot x_i^{(2)} \right\} \tag{2.12}$$

$$\varphi(\mathbf{x}_j) = \left\{ \left(x_j^{(1)}\right)^2, \left(x_j^{(2)}\right)^2, \sqrt{2} \cdot x_j^{(1)} \cdot x_j^{(2)} \right\} \tag{2.13}$$

Pelo aumento da dimensão dos dados realizado pelo mapeamento das amostras no espaço de características de alta dimensão as classes podem ser linearmente separáveis, como mostrado na Figura 2.6(b). Devido a aplicação da função kernel as amostras mapeadas possuem novas dimensões que são funções das variáveis originais, como pode ser visto nas Equações 2.12 e 2.13. Esta separação linear no espaço de características equivale a aplicação de uma função não linear no espaço original dos dados (Figura 2.6(c)).

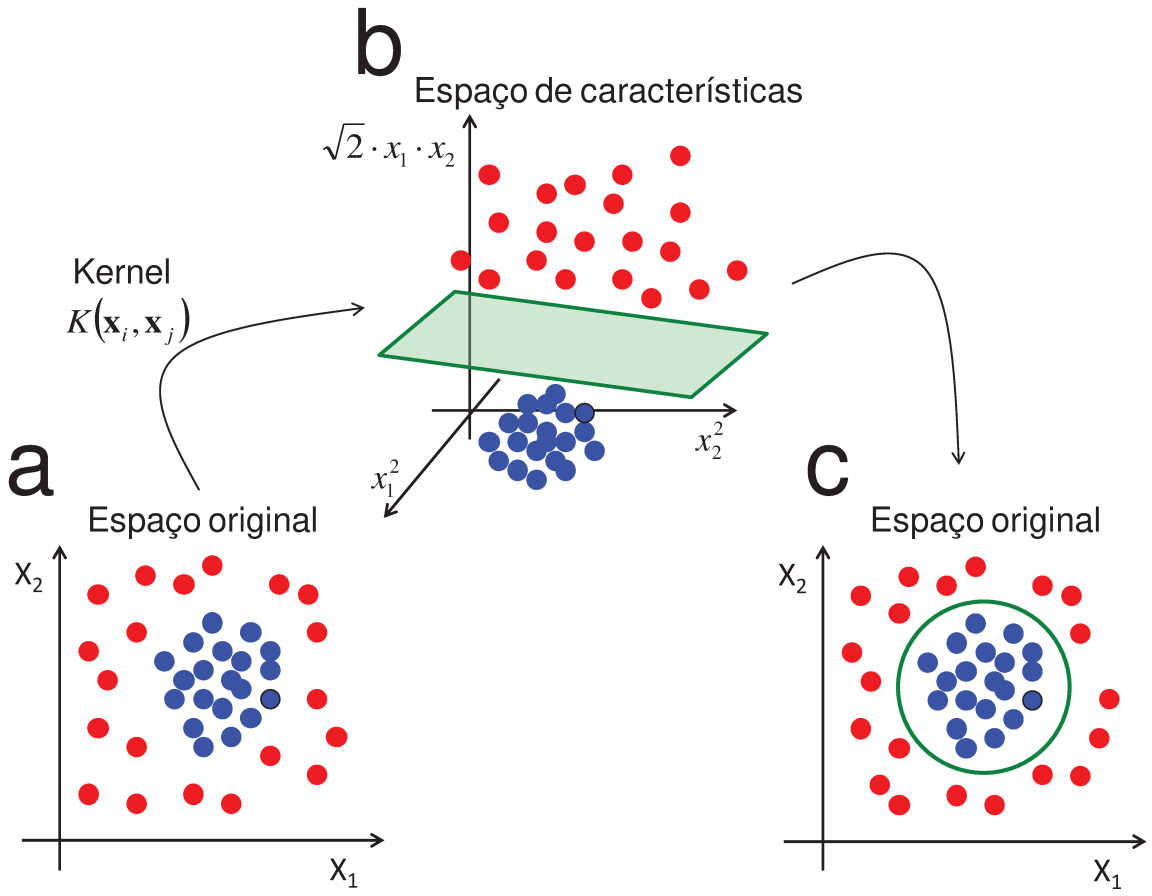

Figura 2.6. Gráfico ilustrativo de um problema de classificação binária não linearmente separável no espaço original (a). Projeção das amostras de entrada no espaço de características de alta dimensão (b). Representação ilustrativa de uma função não linear de classificação no espaço original (c).

As funções kernel mais conhecidas em vetores de suporte são: linear, polinomial, sigmoide e a função de base radial (RBF do inglês, *radial basis function*).[40] Esta última normalmente é a aplicada. Um esboço destas funções é mostrado abaixo.

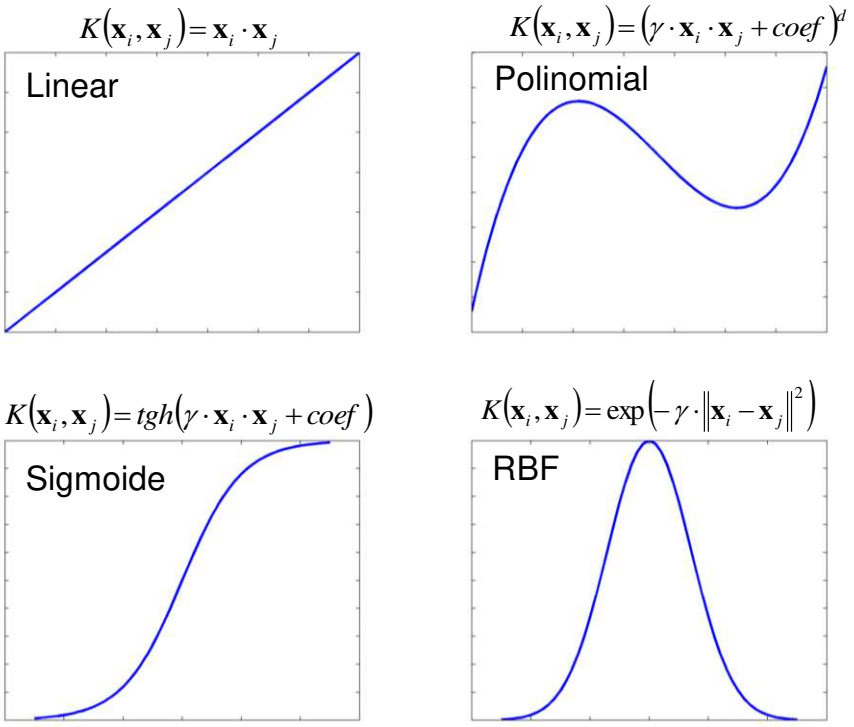

Figura 2.7. Exemplo gráfico das funções kernel mais utilizadas para mapeamento em máquinas de Vetores de Suporte.

2.2.3 Regressão por vetores de suporte[36,41]

O SVM foi originalmente desenvolvido para resolver problemas de classificação. No entanto, problemas de regressão podem ser resolvidos pelo método de classificação binária, através do procedimento: para cada amostra $\mathbf{x}_i$ da regressão (Figura 2.8a) um número positivo *d* é adicionado e subtraído do correspondente valor de interesse y_i (Figura 2.8b), que neste caso é uma variável quantitativa, formando assim duas classes, uma positiva e outra negativa.[41] Observando apenas os dois novos conjuntos formados (Figura 2.8c), podemos ver claramente a formação das duas classes. O hiperplano de máxima separação obtido para estas duas novas classes formadas passará justamente pelos valores de y_i, pois a máxima separação será $y_i + 0$ (Figura 2.8d). Assim o problema de

regressão é transformado em um problema de classificação binária e o algoritmo SVM pode ser aplicado normalmente.

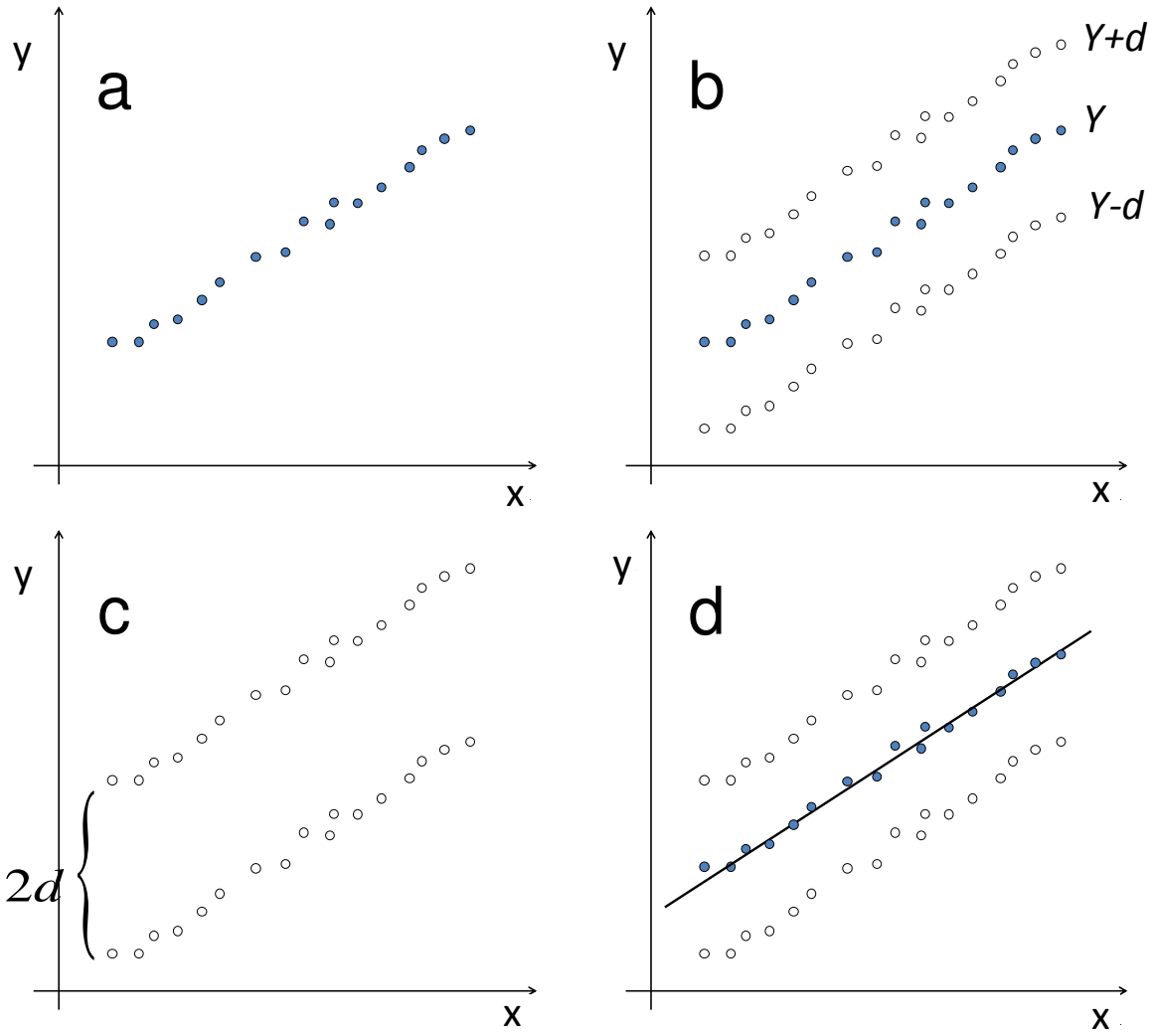

Figura 2.8. Ilustração gráfica da transformação de um problema de regressão em classificação binária. Problema de regressão (a); adição e subtração de uma constante aos dados de entrada (b); transformação do problema de regressão original em um problema de classificação binária (c); hiperplano que define a função de regressão (d).

No método de regressão por vetores de suporte (SVR) alguns ajustes são feitos na função objetiva:[36]

$$\text{Minimize: } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \tag{2.14}$$

$$\text{sujeito a: } \begin{cases} y_i - \mathbf{w}\cdot\phi(\mathbf{x}_i) - b \le \varepsilon + \xi_i \\ \mathbf{w}\cdot\phi(\mathbf{x}_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{2.15}$$

onde a constante $C > 0$ pondera os erros da função ($\xi_i, \xi_i^* \ge 0$) que são delimitados por uma tolerância "ε", descrito pela função de perda ε-insensivel, como pode ser observado graficamente na Figura 2.9. O limite de tolerância ε forma um tubo em torno da função de regressão onde é aceitável erros na exatidão. Nesta abordagem temos duas variáveis de folga: ξ_i e ξ_i^* que estão relacionadas aos erros positivos e negativos respectivamente.

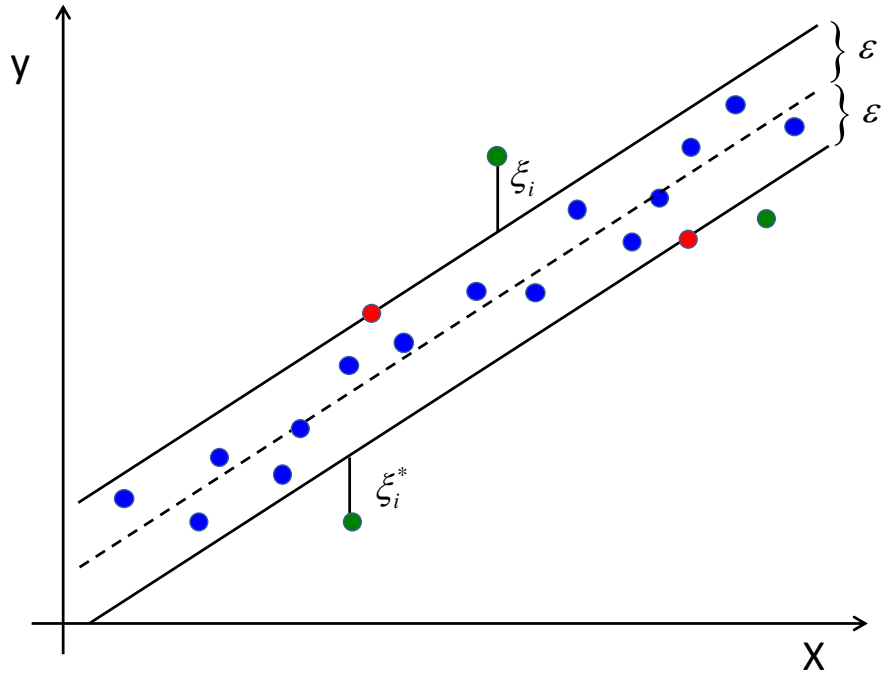

Figura 2.9. Expressão gráfica do limite de tolerância ε-insensivel. Adaptada do artigo de Fong *et al.*[31]

O problema descrito pelas Equações 2.14 e 2.15 é resolvido pela aplicação dos multiplicadores de Lagrange:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{N}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right) - \sum_{i=1}^{N}\alpha_i\left(\varepsilon + \xi_i - y_i + \mathbf{w}\phi(\mathbf{x}_i) + b\right) - \sum_{i=1}^{N}\alpha_i^*\left(\varepsilon + \xi_i^* + y_i - \mathbf{w}\phi(\mathbf{x}_i) - b\right) \quad (2.16)$$

onde L é a Lagrangiana, $\eta_i^{(*)}$ e $\alpha_i^{(*)}$ são os multiplicadores de Langrange que satisfazem a restrição: $\eta_i^{(*)}, \alpha_i^{(*)} \geq 0$. As derivadas parciais de L com relação às respectivas variáveis primárias ($\mathbf{w}, b, \xi_i, \xi_i^*$), tem como otimização:

$$\partial_b L = \sum_{i=1}^{N}\left(\alpha_i^* - \alpha_i\right) = 0 \quad (2.17)$$

$$\partial_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right)\cdot\mathbf{x}_i = 0 \quad (2.18)$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \quad (2.19)$$

A partir da Equação 2.18, obtemos o valor de $\mathbf{w}$:

$$\mathbf{w} = \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right)\cdot\mathbf{x}_i \quad (2.20)$$

Assim, a equação procurada, conforme apresentado na Equação 2.6 tem como solução:

$$f(\mathbf{x}) = \sum_{i=1}^{m}\left(\alpha_i - \alpha_i^*\right)K(\mathbf{x}_i, \mathbf{x}) + b \quad (2.21)$$

onde $K(\mathbf{x}_i, \mathbf{x})$ representa a função kernel aplicada aos dados de entrada, α e α^* podem ser representados como um vetor ($\boldsymbol{\alpha}$) no qual pondera as amostras utilizadas na modelagem. Apenas os vetores de suporte possuem valores de $\boldsymbol{\alpha}$ diferentes de zero, sendo as únicas que contribuem ara construção do modelo.

Durante a aplicação do SVR algumas constantes mencionadas acima como C e γ da função kernel devem ser otimizadas. As margens da função ε -insensivel também pode ser otimizada.

Atualmente, são utilizados dois tipos de algoritmos baseado em vetores de suporte para regressão: e -SVR e ν -SVR.[36] A função objetiva apresentada na

Equação 2.14 refere-se ao algoritmo e-SVR. Esta reformulação pode apresentar problemas na escolha do valor adequado para a margem, possibilitando reduzir a margem ao mesmo tempo em que aumenta de forma excessiva o número de vetores de suporte. No algoritmo ν -SVR, um parâmetro adicional ν é acrescentado na Equação 2.14 para controlar o número de vetores de suporte:

$$\text{Minimize: } \frac{1}{2}\|\mathbf{w}\|^2 + C\left[\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)\right] \tag{2.22}$$

Após a escolha de ν, a margem (ε) é automaticamente definida pelo algoritmo. Neste algoritmo, ν é o limite inferior de proporção entre vetores de suporte e o número total de amostras. Na utilização deste algoritmo, ν também é um parâmetro a ser otimizado.

2.3 Otimização dos modelos de calibração multivariada

A utilização dos modelos de calibração PLS ou SVR requerem a otimização de alguns parâmetros. Para o modelo PLS, apenas o número de variáveis latentes deve ser otimizado,[42,43] enquanto que no modelo SVR, pelo menos 2 parâmetros devem ser otimizados: as constantes C e γ.[41] Entretanto, o limite de tolerância (ε) e as constantes "e" e "ν" também podem ser otimizadas no SVR.

Quanto maior o número de parâmetros otimizados, maior o custo computacional e a possibilidade de superajuste aos dados. Para evitar este último problema, neste trabalho foi aplicado os métodos de validação:

- Validação externa: quando um conjunto de dados é separado do conjunto de calibração para ser utilizado apenas como validação. Assim, constrói-se um modelo com as amostras de calibração e o aplica nas amostras de validação.
- Validação interna: quando as próprias amostras do conjunto de calibração são utilizadas para validação do modelo, este procedimento é conhecido como validação cruzada ou *cross-validation*. Os métodos mais comuns de validação interna são:

1. *leave-one-out*: neste procedimento uma amostra é removida do conjunto de calibração para validação enquanto constrói-se o modelo com as $n-1$ amostras restantes. Neste caso n modelos são construídos até que todas as amostras de calibração sejam utilizadas para validação;
2. *k-fold*: este procedimento consiste em dividir as n amostras de calibração em *k* subconjuntos mutuamente exclusivos de mesmo tamanho. Um subconjunto é utilizado para validação enquanto os *k-1* restantes são utilizados para construção do modelo. O procedimento é repetido k vezes até que todos os subgrupos tenham sido utilizados como validação. Se os subconjuntos são escolhidos em blocos, o método é denominado *contiguos block*; caso sejam escolhidos aleatoriamente, *random block*; ou sejam ordenadas e retiradas de forma ordenada *venetian blinds*.

No procedimento de validação cruzada as amostras do conjunto de calibração são previstas por diferentes modelos construídos com as próprias amostras de calibração, mas nunca uma mesma amostra participando da calibração e validação simultaneamente. A partir dos valores previstos das amostras de calibração, a raiz quadrada dos erros médio de validação (RMSECV) é calculado:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{ncal}\left(y_{cal,i} - \hat{y}_{cal,i}\right)^2}{ncal}} \tag{2.23}$$

Onde $y_{ref,i}$ e $y_{est,i}$ são os valores de referência de calibração e estimado pelo procedimento de validação cruzada para $ncal$ amostras do conjunto de calibração.

No modelo PLS, o número ótimo de variáveis latentes é definido pelo gráfico do RMSECV em função do número de variáveis latentes. O número ótimo é definido pelo valor mínimo ou menor número de variáveis latentes no qual não se tem mais mudança significativa no valor de RMSECV.

No modelo SVR, por ter mais de um parâmetro a ser otimizado o procedimento mais simples a ser utilizado é uma grade de pesquisa. Neste caso, as constantes a serem otimizadas têm certos valores fixados dentro de seu campo de existência e uma combinação de todas as possibilidades são montadas para construção dos modelos. No final, a combinação de parâmetros com mínimo

RMSECV é escolhida como o melhor conjunto de valores. Na modelagem SVR a otimização dos parâmetros do modelo é uma etapa crítica para aplicação do método.

Após otimização do modelo, este está pronto para ser aplicado em um conjunto de dados que não fez parte de sua construção ou otimização, conjunto este denominado de previsão ou teste. A qualidade dos modelos de calibração multivariada gerados normalmente é avaliada pela a raiz quadrada dos erros médio de previsão, RMSEP:[44,45]

$$RMSEP = \sqrt{\frac{1}{nprev} \cdot \sum_{i=1}^{nprev} \left(y_{prev,i} - \hat{y}_{prev,i} \right)^2} \qquad (2.24)$$

onde $nprev$ é o número de amostras de previsão (ou teste), $y_{prev,i}$ e $\hat{y}_{prev,i}$ são respectivamente, os valores de referência e previsto pelo modelo para as amostras do conjunto de previsão.

O ajuste linear entre os valores de referência e previstos é verificada pelo coeficiente de determinação:

$$R^2 = 1 - \frac{\sum_i \left(y_{ref,i} - y_{est,i} \right)^2}{\sum_i \left(y_{ref,i} - \bar{y}_{médio} \right)^2} \qquad (2.25)$$

onde $\bar{y}_{médio}$ e $y_{est,i}$ são os valores médio e previsto pelo modelo, respectivamente, podendo ser aplicado tanto às amostras do conjunto de libração ou previsão. O valor de R^2 varia de 0 a 1, sendo quanto mais próximo de 1 melhor o ajuste do modelo. Quando os dados avaliados são resultados do procedimento de validação cruzada denominamos R^2cv, quando estes são do conjunto de previsão denomina-se R^2p.

2.4 Métodos de avaliação de modelos de calibração

Uma importante etapa durante a construção de modelos de calibração está na avaliação dos resíduos (e_i) do modelo, calculados por:

$$e_i = y_{ref,i} - y_{est,i} \tag{2.26}$$

onde $y_{ref,i}$ e $y_{est,i}$ são os valores de referência e estimado pelo modelo de calibração, respectivamente. Os resíduos devem ser avaliados quanto a presença de erros sistemáticos e tendência. Havendo a presença destes tipos de erros nos resíduos, o modelo gerado pode ser considerado insatisfatório.

2.4.1 Avaliação de erros sistemáticos

Os erros sistemáticos afetam a estimativa sempre no mesmo sentido, gerando resultados abaixo ou acima do valor esperado. O método mais comum para avaliar a presença deste tipo de erro é o teste para viés.[42] Estes representam o desvio médio de n medições:

$$viés = \frac{\sum_{i=1}^{n}(y_{ref,i} - y_{est,i})}{n} \tag{2.27}$$

A presença de erros sistemáticos é avaliada por um teste-t bicaudal no qual as hipóteses testadas são:

H_0: $viés = 0$;

H_1: $viés \neq 0$;

O desvio padrão dos erros (SVD, do inglês *Standard Deviation of the Validation Errors*) e a estatística de teste (t_{calc}) são determinados por:

$$SVD = \sqrt{\frac{\sum_{i=1}^{n}\left\{\left[\left(y_{ref,i} - y_{est,i}\right) \cdot viés\right]^2\right\}}{n-1}} \tag{2.28}$$

$$t_{calc} = \frac{|viés|\sqrt{n}}{SVD} \tag{2.29}$$

A estatística t_{calc} é comparada com o valor de *t* tabelado (t_{tab}) da distribuição t-*student* com $n-1$ graus de liberdade e nível de significância α (neste trabalho sempre foi adotado $\alpha = 0{,}05$). Se $t_{calc} \leq t_{tab}$, não há evidências estatísticas, ao nível de confiança $1-\alpha$, para rejeição de H_0. Neste caso não há evidências da presença de erros sistemáticos nos resíduos. Caso contrário, $t_{calc} > t_{tab}$ rejeita-se H_0 e admite-se a presença de erros sistemáticos nos resíduos.

2.4.2 Avaliação de erro de tendência

A presença de tendência em resíduos do modelo pode ser avaliada pelo grau de ajuste do modelo aos dados.[46,47] Entretanto, esta análise apenas pode ser realizada quando há replicas autênticas o que, em calibração multivariada, não é comum devido ao tempo computacional e às vezes a impossibilidade de réplicas autênticas para algumas amostras.

O procedimento normalmente adotado para avaliar tendência em resíduos de calibração multivariada é subjetivo, através da observação visual do gráfico dos resíduos em função dos valores de referência, como mostrado na Figura 2.10(a). Neste procedimento, diferentes analistas podem tomar decisões conflitantes quando a presença de resíduos não está muito clara. Para evitar conclusões subjetivas foi implementado um teste de permutação não paramétrico para avaliar erros de tendência em resíduos de calibração multivariada.

A presença de tendência não está associada à distribuição dos resíduos, como pode ser visto na Figura 2.10. Supondo o experimento de tomarmos um número aleatório com distribuição N(0,1), e repetir este procedimento 100 vezes, armazenando 100 números aleatórios. Um gráfico representativo destes 100 valores é mostrado na Figura 2.10(a) e seu histograma na Figura 2.10(b). Caso os mesmos números aleatórios fossem tomados em outra ordem, poderíamos ter diferentes formas de representação do gráfico de dispersão, porém mantendo constante o histograma de distribuição dos resíduos. Uma forma pouco provável, mas possível é apresentada na Figura 2.10(c) e seu histograma na Figura 2.10(d).

Na Figura 2.10(c) observa-se claramente uma tendência quadrática nos dados, apesar de estarem normalmente distribuídos com média zero e variância unitária.

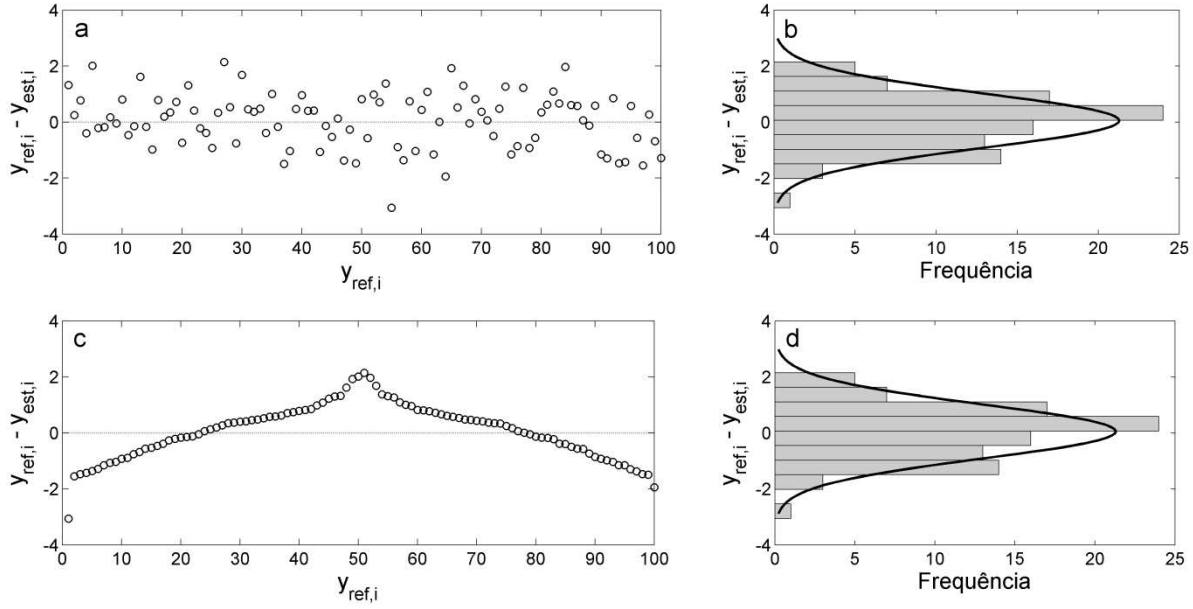

Figura 2.10. Resíduos simulados normalmente distribuídos com média zero e variância unitária: sem presença de tendência (a); histograma dos resíduos sem tendência (b); resíduos com tendência quadrática (c); histograma dos resíduos com tendência quadrática (d). Fonte: Filgueiras *et al.*[48]

O teste de permutação não paramétrico baseia-se na probabilidade da tendência observada nos dados ser devida ao acaso.[48] Na avaliação da presença de tendência nos resíduos de calibração, a primeira etapa do teste é definir as hipóteses a serem testadas:

H_0: os resíduos e_i são independentes de y_i.

H_1: os resíduos e_i são dependentes de y_i conforme a equação:

$$e_i = g(y_i) + \varepsilon_i \tag{2.30}$$

onde ε_i é um erro aleatório independente e $g(y_i)$ alguma função polinomial que pode modelar a relação entre os resíduos e os valores de referência. A presente equação segue a forma:

$$g(y_i) = b_n y_i^n + b_{n-1} y_i^{n-1} + \ldots + b_1 y_i + b_0 \quad (2.31)$$

Na hipótese alternativa, a dependência dos resíduos com os valores de referência é proposta, e presume-se que todo o efeito aleatoriamente presente em y_i é devido somente a variável ε_i.

Haverá evidências de tendência nos resíduos se o coeficiente polinomial de maior ordem (b_n) for estatisticamente significativo, ao nível de significância α adotado. Assim, o grau do polinômio deve ser definido previamente à aplicação do teste. O teste pode ser resumido nos seguintes passos:

(i) calcular o coeficiente polinomial b_n ajustado para os dados originais dos resíduos em função dos valores de referência. Assim, o coeficiente será denominado b_n^*;

(ii) Permutar randomicamente somente o vetor $\mathbf{y}$ contendo a propriedade de interesse;

(iii) Calcular o coeficiente b_n^i para o *i*-ésimo vetor $\mathbf{y}$ permutado;

(iv) Comparar b_n^* com b_n^i;

(v) Repetir as etapas (ii) a (iv) k-vezes.

Os coeficientes b_n^i dos ajustes permutados não tem sentido físico por serem aleatórios. Assim a distribuição destes coeficientes constituirá a distribuição para o teste. O p-valor do teste é determinado pela proporção do número de vezes em que $b_n^* > b_n^i$, como mostrado na Figura 2.11. Não há necessidade dos coeficientes polinomiais permutados b_n terem distribuição normal, entretanto, muitas vezes ela é obtida. Se o p-valor do teste for menor que o nível se significância α adotado, não existem evidências significativas para aceitar H_0, e a tendência observada é significativa. Caso contrário, não há evidências para rejeitar H_0 e aceita-se a hipótese dos resíduos serem aleatórios.

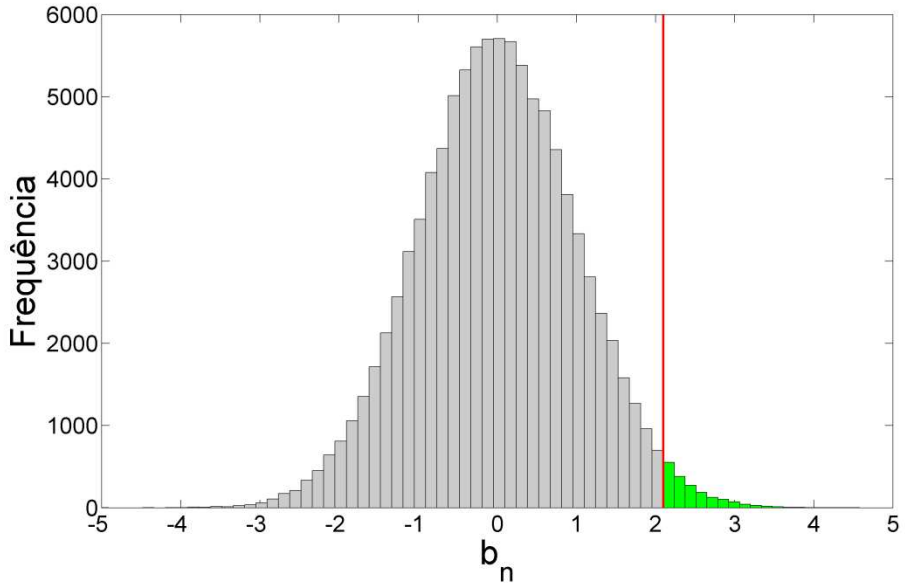

Figura 2.11. Histograma ilustrativo da distribuição dos coeficientes polinomiais de maior ordem (b_n) para os ajustes permutados. A linha vertical vermelha representa o valor de b_n^* (neste caso positivo) para os dados originais. Em cinza são os coeficientes menores que b_n^* e em verde os coeficientes maiores que b_n^*.

2.4.3 Comparação da exatidão entre modelos de calibração

Quando se utiliza mais de um método para construção de modelos de calibração surge uma etapa critica que é a comparação dos modelos. Neste trabalho foi adotado o teste randômico para exatidão.[49] O teste consiste em avaliar os resíduos de previsão dos dois modelos (A e B) a serem comparados, segundo as hipóteses:

H_0: $RMSEP_A^2 = RMSEP_B^2$;

H_1: $RMSEP_A^2 \neq RMSEP_B^2$.

O algoritmo de teste segue os passos:

(i) Calcular a diferença média quadrática real dos resíduos dos modelos A e B:

$$\bar{d}^{*} = \sum_{i=1}^{n}\left[\frac{1}{n}\left(e_{A,i}^{2} - e_{B,i}^{2}\right)\right] \tag{2.32}$$

(ii) Permutar randomicamente somente o vetor com os resíduos do modelo A mantendo a ordem dos resíduos de B;

(iii) Calcular as diferenças médias das permutações $\bar{d}_i$;

(iv) Comparar $\bar{d}^{*}$ com $\bar{d}_i$;

(v) Repetir as etapas (ii) a (iv) k-vezes.

O p-valor do teste é determinado pela proporção do número de vezes em que $\bar{d}^{*} > \bar{d}_i$, como mostrado na Figura 2.12. Se o p-valor do teste for menor que o nível se significância α adotado, não existem evidências significativas para aceitar H_0 e os modelos tem exatidões diferentes. Caso contrário, não há evidências para rejeitar H_0 e aceita-se a hipótese H_1, não havendo diferença entre a exatidão dos modelos.

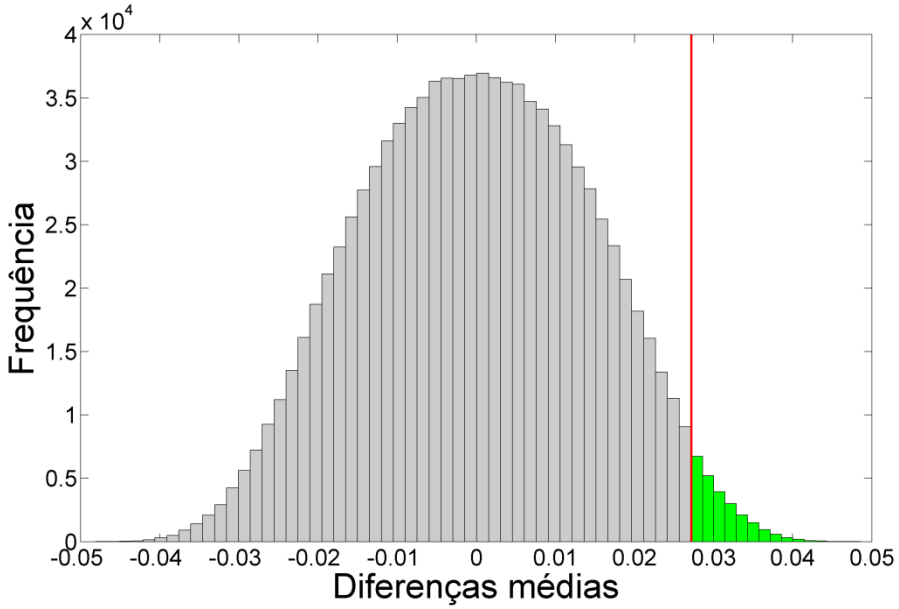

Figura 2.12. Histograma ilustrativo da distribuição das diferenças médias das permutações. A linha vertical vermelha representa o valor de $\bar{d}^{*}$ para os dados originais. Em cinza são as diferenças médias menores que $\bar{d}^{*}$ e em verde as diferenças médias maiores que $\bar{d}^{*}$.

3 Aplicação 1: Determinação da gravidade API, viscosidade cinemática e teor de água em petróleo por espectroscopia ATR-FTIR e calibração multivariada

3.1 Introdução

O petróleo é avaliado principalmente por medidas físico-químicas como: gravidade API (escala de densidade do *American Petroleum Institute*), viscosidade cinemática, teor de água e sedimentos, etc. A partir da grande variedade de petróleos existentes o *American Petroleum Institute* criou uma escala para densidade denominada gravidade API.[1] Esta é uma escala hidrométrica utilizada para expressar a densidade relativa de óleos e varia inversamente com a densidade relativa. A escala API em relação à densidade é dada pela equação:

$$API = \frac{141{,}5}{\rho(60^{o}/60^{o}F)} - 131{,}5 \qquad (3.1)$$

Onde $\rho(60^{o}/60^{o}F)$ é a densidade específica do óleo a 60ºF em relação à densidade da água na mesma temperatura. A gravidade API pode definir tipos de petróleos e direciona-los para correta unidade de destilação. Variação superior a 1 unidade na gravidade API pode modificar toda a logística de refino de um petróleo, uma vez que as unidades de destilação são construídas para destilar tipos específicos de petróleos.

A gravidade API e a viscosidade cinemática afetam a viabilidade econômica dos campos produtores, pois estas propriedades auxiliam no dimensionamento dos equipamentos utilizados na exploração e na estimativa do preço do barril do petróleo. Após iniciada a exploração, estas propriedades influenciam o processo de decisão da produção, pois controlam o intervalo de produção do reservatório e auxiliam na perfuração de novos poços.[1-4] Durante a exploração e produção, água e sedimentos são contaminantes indesejáveis que podem ocasionar problemas no transporte e na refinaria, como corrosão de equipamentos, acidentes durante o processo de destilação ou efeitos adversos na qualidade dos produtos destilados.[50,51] Sua determinação possibilita avaliar o preço de venda, custos de

produção e royalties.[1] As estimativas destes parâmetros em tempo real auxiliam rápida tomada de decisão sobre a produção contínua de um poço ou campo.

A espectroscopia no infravermelho tem se destacado, nos últimos anos, na análise quantitativa de petróleo,[10-13] derivados[9,15,16,18] e biodiesel,[6,7,14,17] apresentando como principais vantagens o uso de pequenas quantidades de amostra, procedimentos experimentais rápidos com mínimo pré-tratamento da amostra, redução na geração de resíduos e custos. Entretanto, a conversão do sinal instrumental na resposta de interesse, requer a utilização de técnicas de calibração multivariada. O método linear de regressão por mínimos quadrados parciais (PLS) é atualmente o mais aplicado em química analítica. Entretanto, devido à natureza complexa do petróleo, métodos de regressão não lineares como a Regressão por Vetores de Suporte (SVR) são mais indicados.

Devido à aplicação da função kernel em SVR, informações sobre as variáveis originais são perdidas tornando o modelo não interpretável. Isso é uma desvantagem na aplicação do método SVR em relação ao PLS. O mapeamento kernel resulta em uma matriz de similaridade entre as amostras do conjunto de calibração. Os vetores de suporte obtidos posteriormente durante a modelagem possuem informações a respeito das amostras mais importantes e sua relação com as demais amostras. Assim, os vetores de suporte contêm informações relevantes para interpretar quais variáveis de entrada são responsáveis pelo modelo de regressão construído.[52]

3.1.2 Objetivos

Determinar a gravidade API, viscosidade cinemática e teor de água em petróleos utilizando espectroscopia no infravermelho médio com reflectância total atenuada (ATR-FTIR) e métodos de regressão multivariada PLS e SVR.

Identificar as regiões espectrais com maior contribuição na construção do modelo de calibração SVR.

3.2 Metodologia

Neste estudo foram utilizadas 68 amostras de *blends* de petróleos bruto da bacia sedimentar da costa brasileira. Estas amostras foram analisadas no Laboratório de Pesquisa e Desenvolvimento de Metodologias para Análise de Óleos Pesados (LabPetro) do Departamento de Química da Universidade Federal do Espírito Santo, seguindo seus respectivos padrões técnicos de análises:

A gravidade API das amostras foi medida segundo a norma ASTM ISSO 12185.[53] A densidade foi determinada pela injeção da amostra no densímetro digital (Anton Paar modelo DMA 5000), medida a 50°C e posteriormente estimada a 20°C para o cálculo da gravidade API.

A viscosidade cinemática foi medida segundo a norma ASTM D 7042-04,[54] pela injeção da amostra no viscosímetro digital automático (Anton Paar Stabinger SVM 3000). A viscosidade cinemática das amostras foram medidas a 50°C e 60°C sendo posteriormente estimada a viscosidade cinemática a 40°C por regressão, segundo a equação que relaciona a viscosidade cinemática com a temperatura.[55] No setor de exploração e produção de petróleo, a viscosidade cinemática é analisada a 40°C mas, para óleos muito viscosos, medir diretamente nesta temperatura gera erros muito altos. Assim, para estes óleos, a viscosidade cinemática é medida a duas temperaturas superiores e extrapolada para 40°C.

O teor de água foi medido pelo método de reagente Karl Fischer, segundo a norma ASTM D 4377-06.[56] O solvente usado durante as análises foi uma mistura de metanol e clorofórmio (20% v/v). Para padronização do reagente de Karl Fischer, água destilada foi solubilizada nos solventes. Um titulador Metrohm Karl Fischer (modelo 836 Titrando) equipado com eletrodo duplo de platina foi utilizado para a determinação do teor de água. A norma ASTM D 4377-00 cobre resultados na faixa de 0,02 a 2% v/v de água em óleo. Amostras com resultados acima deste limite podem ser analisadas pela técnica, mas não são descritos pela referida norma.

Espectros de infravermelho na região do médio (ATR-FTIR) das amostras foram tomados em um espectrômetro, marca BOMEM com acessório horizontal para Refletância Total Atenuada de aproximadamente, modelo SPLA 2000-102,

com cristal de ZnSe. Espectros na região do infravermelho médio (4000,00–646,10 cm^{-1}) foram tomados, sendo cada espectro amostral uma média de 16 varreduras, empregando resolução de 4 cm^{-1}. O espectro do branco foi registrado com o ar ambiente e subtraído do espectro da amostra. A umidade relativa do ar e a temperatura ambiente da sala permaneceram sempre em torno de 36% e 24 ºC, respectivamente.

3.2.1 Construção dos modelos de calibração multivariada

Para construção dos modelos de calibração, as 68 amostras foram divididas em dois conjuntos disjuntos: 48 para construção dos modelos e 20 para previsão, utilizando o algoritmo Kennard-Stone.[57] Antes da construção dos modelos, a linha de base dos espectros foi corrigida pelo algoritmo de correção multiplicativa de sinal (MSC, do inglês *Multiplicative Signal Correction*).[58] O modelo PLS foi construído na plataforma PLS Toolbox[59] e o SVR com o pacote LIBSVM.[60] O número de variáveis latentes do modelo PLS e os parâmetros C e ε do modelo SVR foram otimizados por validação cruzada *leave-one-out*.

Para obter informação à respeito das variáveis mais importantes na construção do modelo SVR, foi utilizada a metodologia proposta por Üstun e colaboradores.[52] Este método busca informações contidas nos vetores de suporte do modelo para torná-lo interpretável como os coeficientes de regressão do modelo PLS. O método proposto relaciona a matriz de dados de entrada com os vetores de suporte ($\boldsymbol{\alpha}$) do modelo SVR:

$$\mathbf{p}vector^{T}_{(nx1)} = \mathbf{X}^{T}{}_{(nxm)} \cdot \boldsymbol{\alpha}_{(mx1)} \qquad (3.2)$$

O vetor $\boldsymbol{\alpha} \geq 0$ possui valor zero para amostras que não são vetores de suporte. No método proposto para identificação das variáveis mais importantes na construção do modelo, utiliza-se informação apenas das amostras vetores se suporte e pondera seus espectros de acordo com o valor de α_i obtido para cada amostra.

3.3 Resultados e discussões

As 68 amostras de óleo estudadas apresentam gravidade API entre 16 e 23, que corresponde a óleos pesados e extra pesados de acordo com a classificação de petróleos por densidade proposta por Guimarães *et al.*[37] A viscosidade cinemática a 40°C encontra-se na faixa de 57 a 250 mm^2s^{-1}. Os petróleos utilizados neste estudo são muito densos, o que dificulta a medida de viscosidade cinemática diretamente a 40ºC. O teor de água dos óleos variou de 0,1 a 6,1% v/v. Os espectros das amostras de calibração são mostrados na Figura 3.1. A variação na linha de base espectral foi corrigida pelo pré-processamento: correção multiplicativa de sinal.

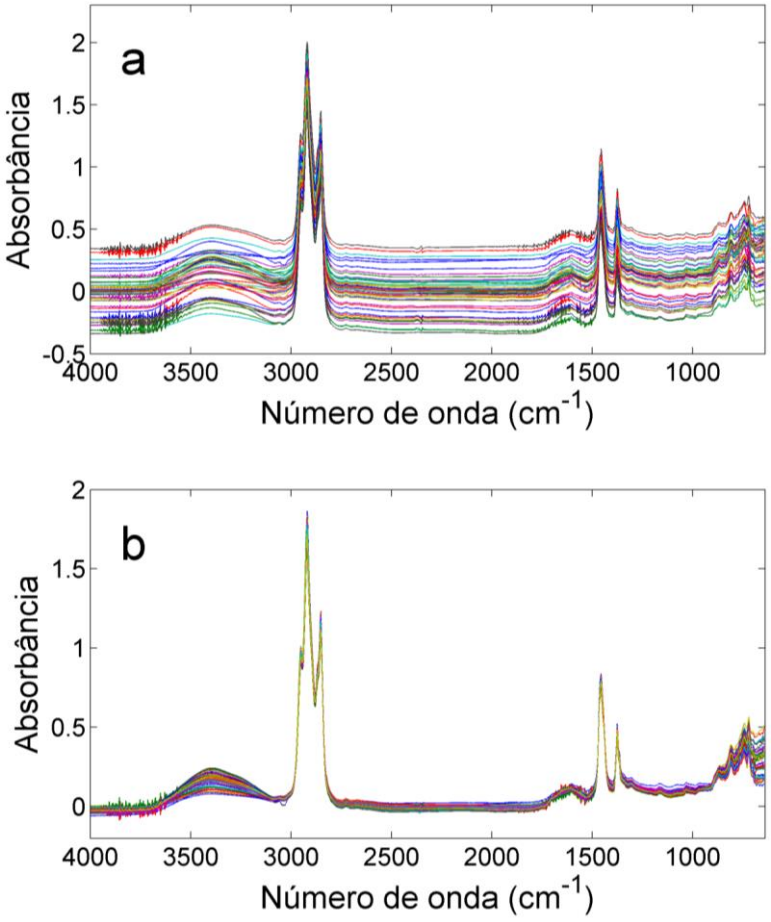

Figura 3.1. Espectros no infravermelho dos petróleos usados no modelo de calibração, sem pré-processamento (a); após pré-processamento por MSC (b).

Os resultados dos parâmetros estatísticos dos modelos PLS e SVR para as três propriedades estudadas são apresentados na Tabela 3.1. Os modelos SVR apresentaram menores erros de previsão (RMSEP) e maiores coeficientes de determinação (R^2p) que os modelos PLS, indicando que os sistemas em estudo parecem se comportar de forma não linear. Na modelagem da propriedade gravidade API, o modelo PLS foi construído com 6 variáveis latentes e o modelo SVR com os valores de C e ε de 134,3 e 0,0452 respectivamente. Os modelos PLS e SVR apresentaram RMSEP's de 0,38 e 0,25 respectivamente. A diferença apresentada é estatisticamente significativa pelo teste randômico para exatidão, com p-valor de 0,030. A boa relação entre os valores previstos pelo modelo SVR e medidos pelo método de referência é observado nos coeficientes de determinação: 0,9817 para validação cruzada, 0,9751 para previsão e na Figura 3.2, onde o melhor ajuste no do modelo SVR em relação ao PLS pode ser visualizado pela menor intensidade dos resíduos.

Tabela 3.1. Resultados dos modelos PLS e SVR na determinação de propriedades físico-químicas de petróleo.

Parâmetros		Gravidade API	Viscosidade (mm^2s^{-1})	Teor de água (% v/v)
PLS	R^2cv	0,9292	0,8183	0,9189
	R^2p	0,9461	0,7811	0,9670
	Viés-cv	-0,01	0	0,00
	Viés-p	-0,09	11	-0,14
	RMSECV	0,42	20	0,38
	RMSEP	0,38	27	0,34
SVR	R^2cv	0,9817	0,9661	0,9768
	R^2p	0,9751	0,8584	0,9767
	Viés-cv	0,01	1	0,03
	Viés-p	-0,05	1	-0,12
	RMSECV	0,22	8	0,20
	RMSEP	0,25	22	0,26

O teste-t para erros sistemáticos foi aplicado aos resíduos dos modelos construídos e não apresentaram resultados significativos ao nível de significância de 5%. Os resíduos dos modelos visualmente parecem se comportar de forma aleatória assim, o teste de permutação para avaliar tendências não foi aplicado a estes dados.

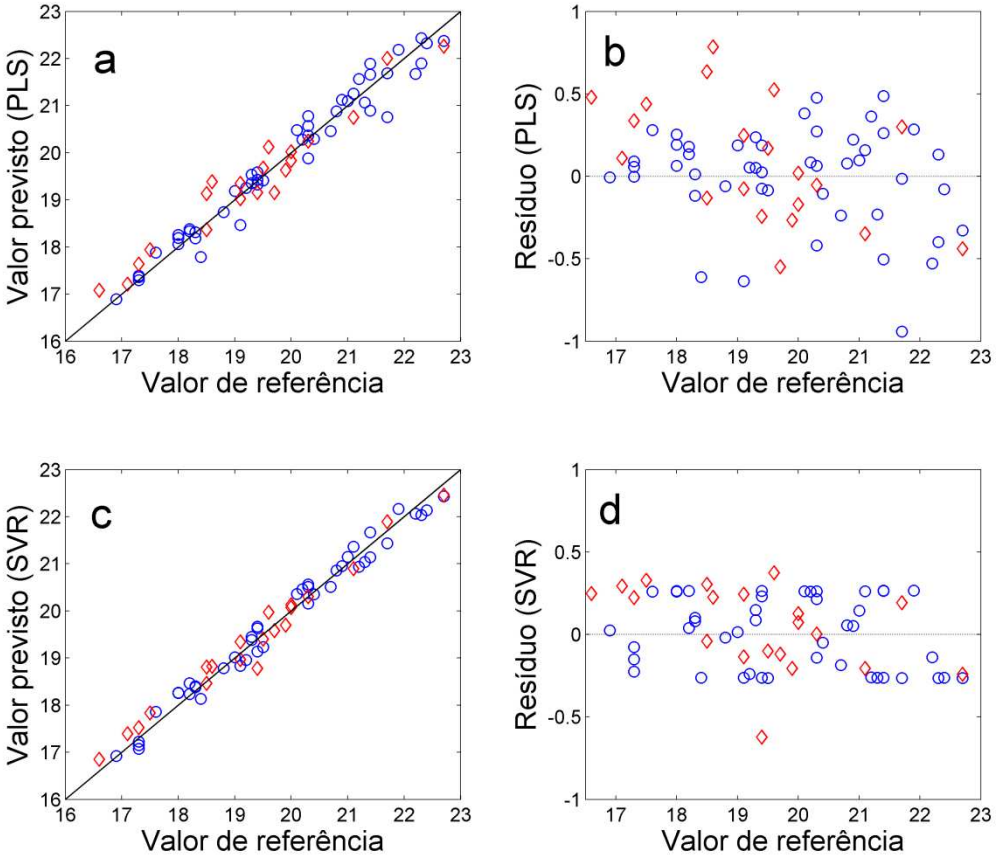

Figura 3.2. Gráfico da gravidade API prevista pelo modelo PLS pelos valores de referência (a); resíduos da modelagem PLS (b); Gráfico da gravidade API prevista pelo modelo SVR pelos valores de referência (c); resíduos da modelagem SVR (d). Amostras de calibração (o) e previsão (◊).

Abbas *et al.*[61] estimaram a gravidade API de petróleos por espectroscopia ATR-FTIR associado à calibração multivariada PLS. O resultado obtido foi um RMSEP de 1.66. Na indústria do petróleo é aceito variação máxima de 1 unidade

na gravidade API para óleos de mesma procedência. Assim, este resultado excede o limite de erro adotado na empresa brasileira sendo inviável de utilização prática. No trabalho, os autores utilizaram óleos de 7 bacias geográficas contendo ampla variação físico-química. Nesta tese, foram utilizados apenas óleos da bacia sedimentar da costa brasileira. Devido ao uso de amostras muito diferentes, uma melhor comparação da exatidão dos modelos dos dois trabalhos pode ser avaliada pelo RMSEP%:

$$RMSEP\% = 100 \cdot \frac{RMSEP}{\bar{y}_{previsão}} \tag{3.3}$$

Pela maior faixa de trabalho utilizado, RMSEP de 1,66 representa 4,61% do erro médio de previsão, enquanto que o RMSEP de 0,25 obtido pelo modelo SVR para esta tese, representa apenas 1,3%. Estes resultados reforçam o uso de um método de calibração não linear na modelagem da propriedade gravidade API em petróleos, obtendo valores próximos aos aceitos pela indústria do petróleo.

A viscosidade cinemática varia exponencialmente com a temperatura de medição do óleo. Para petróleos pesados e extrapesados, normalmente esta propriedade é medida em elevadas temperaturas (acima de 50°C) para posteriormente ser estimada a 40°C. Na modelagem da propriedade viscosidade cinemática a 40°C, o modelo PLS foi construído com 6 variáveis latentes e o modelo SVR com os valores de C e ε de 18 e 0,0001 respectivamente. No modelo PLS, os erros de validação cruzada e previsão (RMSECV de 20 mm^2s^{-1} e RMSEP de 27 mm^2s^{-1}) apresentam uma leve discordância, porém não significativa. Entretanto, para o modelo SVR a diferença é significativa, com RMSECV de 8 mm^2s^{-1} e RMSEP de 22 mm^2s^{-1}. Isto indica que pode ter ocorrido superajuste do modelo SVR na etapa de validação cruzada. A diferença dos RMSEP's dos modelos é significativa pelo teste randômico para exatidão, com significância de 5%, com p-valor de 0,038.

Na Figura 3.3, observa-se melhor ajuste do modelo SVR aos dados experimentais, com R^2p de 0,8584 contra 0,7811 para o modelo PLS (Tabela 3.1). Os resíduos da modelagem PLS (Figura 3.3b) parecem seguir um padrão quadrático. O teste-t para erros sistemáticos foi aplicado aos resíduos dos

modelos construídos e não apresentam resultados significativos ao nível de significância de 5%.

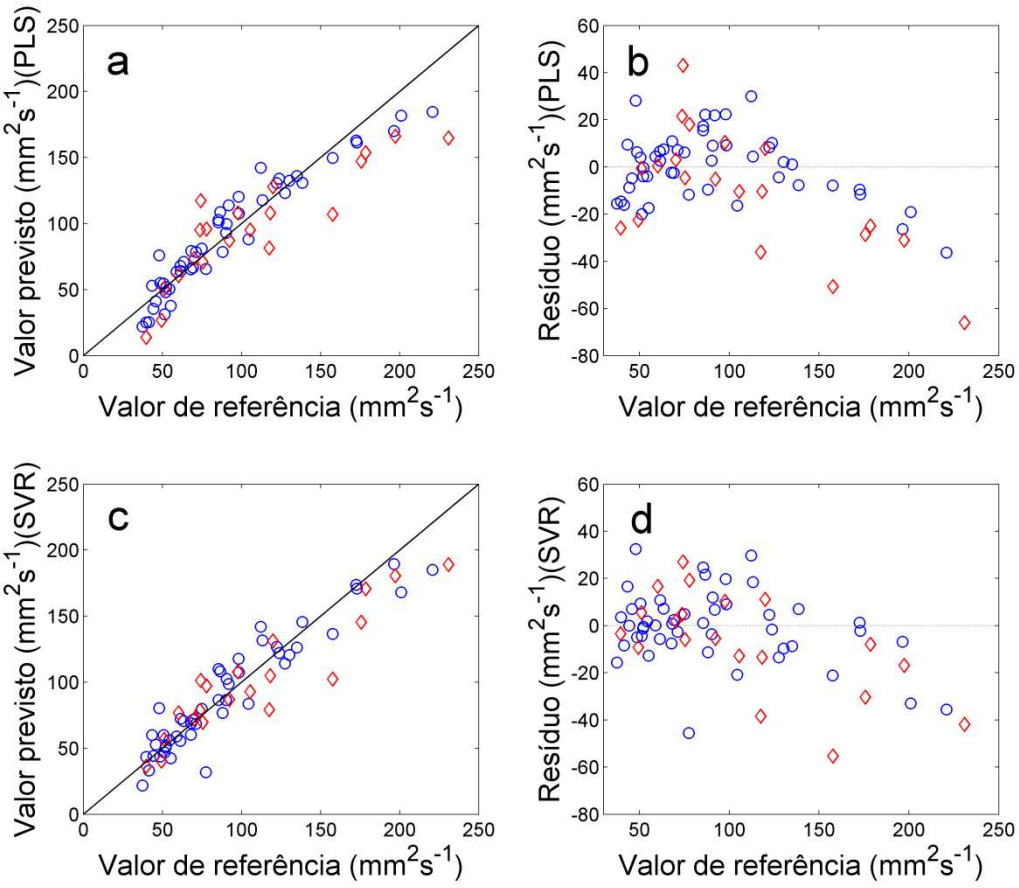

Figure 3.3. Gráfico da viscosidade cinemática prevista pelo modelo PLS pelos valores de referência (a); resíduos da modelagem PLS (b); Gráfico da viscosidade cinemática prevista pelo modelo SVR pelos valores de referência (c); resíduos da modelagem SVR (d). Amostras de calibração (o) e previsão (◊).

Para avaliar se os resíduos da modelagem PLS e SVR seguem tendência quadrática, como observado nas Figuras 3.3b e 3.3d, o ajuste quadrático aos resíduos precisa ter coeficiente quadrático estatisticamente significativo. Nas Figuras 3.4a e 3.4c, o ajuste quadrático aos resíduos é mostrado e parece ser mais significativo nos resíduos do modelo PLS. Nestes casos, o teste de permutação é utilizado para avaliar a probabilidade da curvatura observada

(coeficiente quadrático do polinômio ajustado aos dados) ser simplesmente devida ao acaso, ou seja, não haver relação matemática entre os resíduos e os valores de viscosidade medidos experimentalmente.

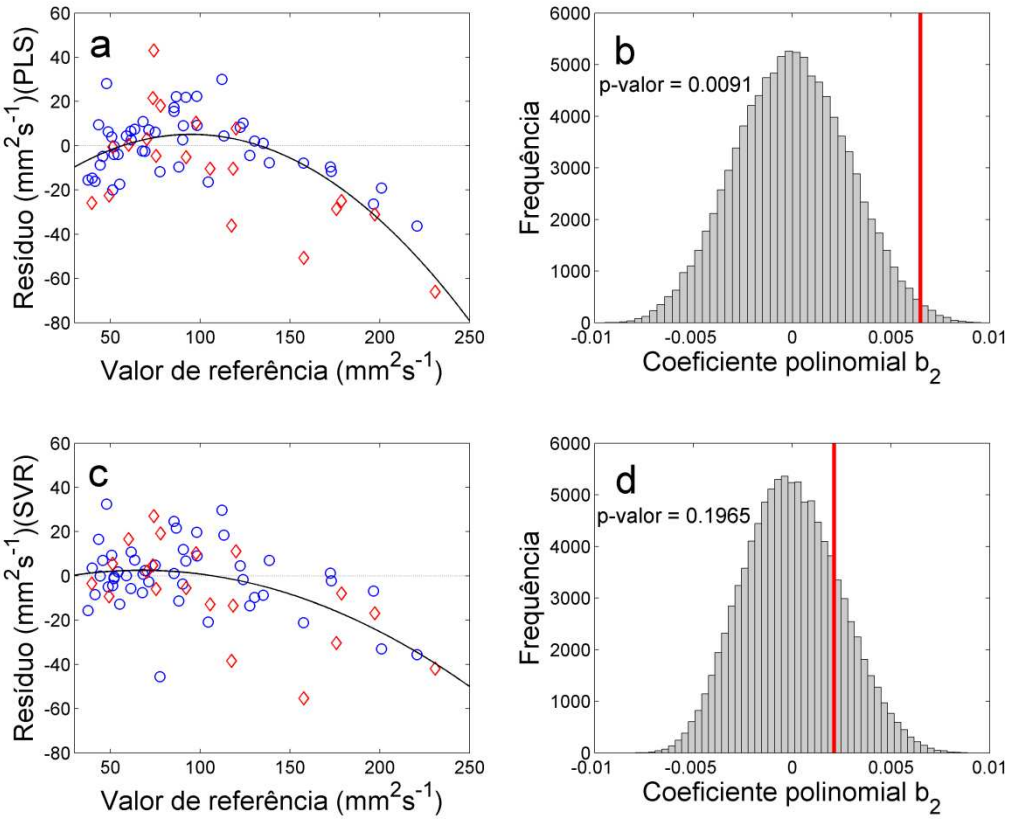

Figura 3.4. Resíduos da modelagem PLS na determinação da viscosidade cinemática (a); histograma dos coeficientes quadráticos do teste de permutação dos resíduos do modelo PLS (b); Resíduos da modelagem SVR na determinação da viscosidade cinemática (c); histograma dos coeficientes quadráticos do teste de permutação dos resíduos do modelo SVR (d); Amostras de calibração (o) e previsão (◊).

Nas Figuras 3.4b e 3.4d são mostradas as distribuições dos coeficientes quadráticos obtidos pela permutação dos valores de referência. Estes coeficientes não apresentam significado físico, mas servem para construção de uma população de coeficientes quadráticos a partir dos dados obtidos na modelagem. Testes não

paramétricos têm vantagem de não assumirem nenhuma hipótese sobre a distribuição de probabilidades da população da qual os dados foram retirados. No entanto, os coeficientes quadráticos parecem distribuir-se normalmente com média zero. A linha vertical vermelha refere-se ao coeficiente quadrático obtido com os dados reais, e o p-valor a área à direita da linha. Nestes casos, quanto mais à direita estiver a linha vermelha, menor o p-valor do teste, menor a probabilidade da tendência observada ser devido ao acaso. Pelos resultados obtidos no teste, a probabilidade da tendência quadrática observada nos resíduos do modelo PLS (Figura 3.4a) ser devida ao acaso é 0,91%; e para os resíduos do modelo SVR (Figura 3.4c) é 19,65%. Tomando um nível de significância de 5% para o teste, podemos concluir que o modelo PLS apresenta tendência quadrática em seus resíduos, enquanto que para o modelo SVR não há tendência.

Na modelagem da propriedade teor de água, o modelo PLS foi construído com 4 variáveis latentes e o modelo SVR com os valores de C e ε de 138,3 e 0,0141 respectivamente. Os modelos SVR e PLS apresentaram exatidões equivalentes, pelo teste randômico para exatidão ao nível de significância de 5%, com RMSEP de 0,26% v/v e 0,34% v/v, respectivamente. Os resultados de modelagem apresentam boa relação linear com os valores de referência, com coeficientes de determinação acima de 0,9 (Tabela 3.1). Na Figura 3.5, observa-se que várias amostras apresentam teor de água acima de 2% v/v, limite contemplado pela norma ASTM D 4377-06, mas seguem linearidade até 6,1% v/v.

O teste-t para erros sistemáticos foi aplicado aos resíduos dos modelos construídos e não apresentaram resultados significativos ao nível de significância de 5%. Os resíduos dos modelos visualmente parecem se comportar de forma aleatória assim, o teste de permutação para avaliar tendências não foi aplicado a estes dados.

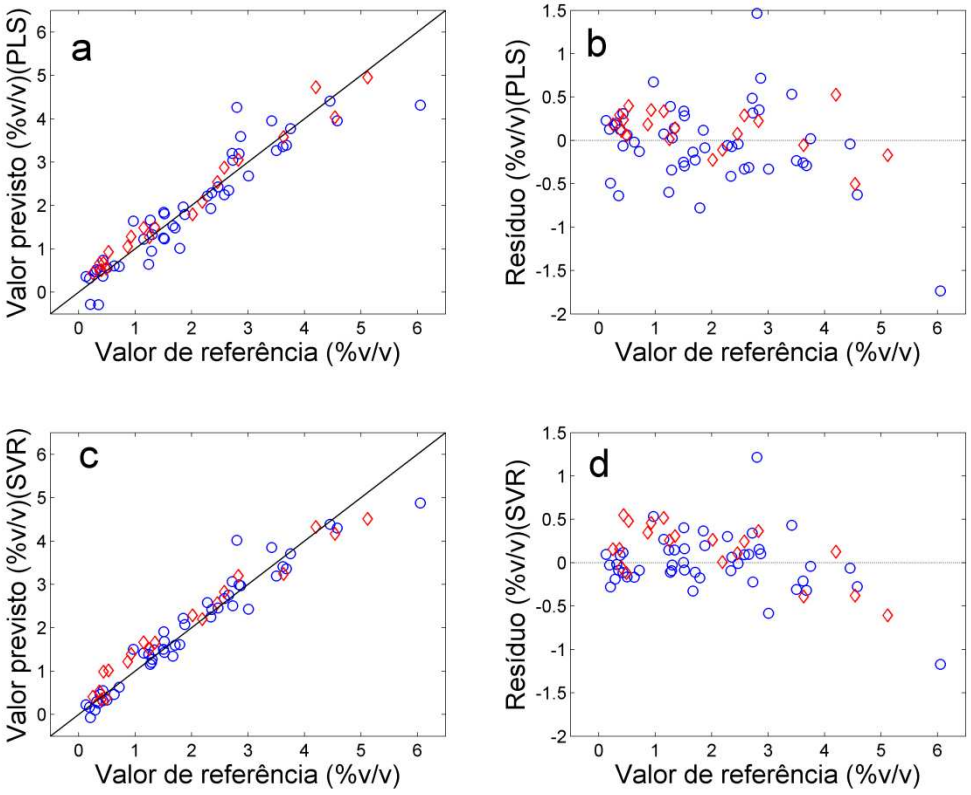

Figure 3.5. Gráfico do teor de água em petróleos previsto pelo modelo PLS pelos valores de referência (a); resíduos da modelagem PLS (b); Gráfico do teor de água em petróleos previsto pelo modelo SVR pelos valores de referência (c); resíduos da modelagem SVR (d). Amostras de calibração (o) e previsão (◊).

3.3.1 Análise da importância das variáveis no modelo

O modelo PLS está bem consolidado na área de calibração multivariada por apresentar bons resultados de calibração e fácil interpretação dos modelos construídos. As variáveis mais significativas para construção do modelo PLS podem ser identificadas pelos pesos ou pelos coeficientes de regressão do modelo. Utilizando o método desenvolvido por Üstun de colaboradores,[52] torna-se possível interpretar os modelos SVR pelo cálculo e análise do p-vetor.

A Figura 3.6a apresenta o espectro ATR-FTIR médio das 48 amostras de calibração. As Figuras 3.6b e 3.6c indicam as regiões espectrais com maior contribuição na construção dos modelos PLS e SVR utilizando os coeficientes de regressão do PLS e p-vetor do SVR, para determinação da gravidade API. A região abaixo de 900 cm^{-1}, referente à região de impressão digital apresenta grande importância nos modelos PLS e SVR. Ordenando os espectros ATR-FTIR com o aumento da gravidade API da amostra, podemos observar que a região de impressão digital realmente está relacionada com a gravidade API (Figura 3.7). O aumento da gravidade API na amostra resulta em redução na intensidade de absorbância na faixa da impressão digital. Esta relação é vista nos coeficientes de regressão do modelo PLS (Figura 3.6b) e nitidamente no gráfico do p-vetor do modelo SVR (Figura 3.6c).

A região em torno de 2900 cm^{-1}, referente à banda de estiramento C–H de carbonos primários e secundários, também apresenta grande importância nos modelos construídos (Figuras 3.6b e 3.6c). O aumento na intensidade da banda nesta região pode estar relacionado ao aumento da cadeia linear carbônica de compostos parafínicos. Na Figura 3.7 observa-se o aumento da absorbância nesta região relacionada com o aumento da gravidade API dos óleos.

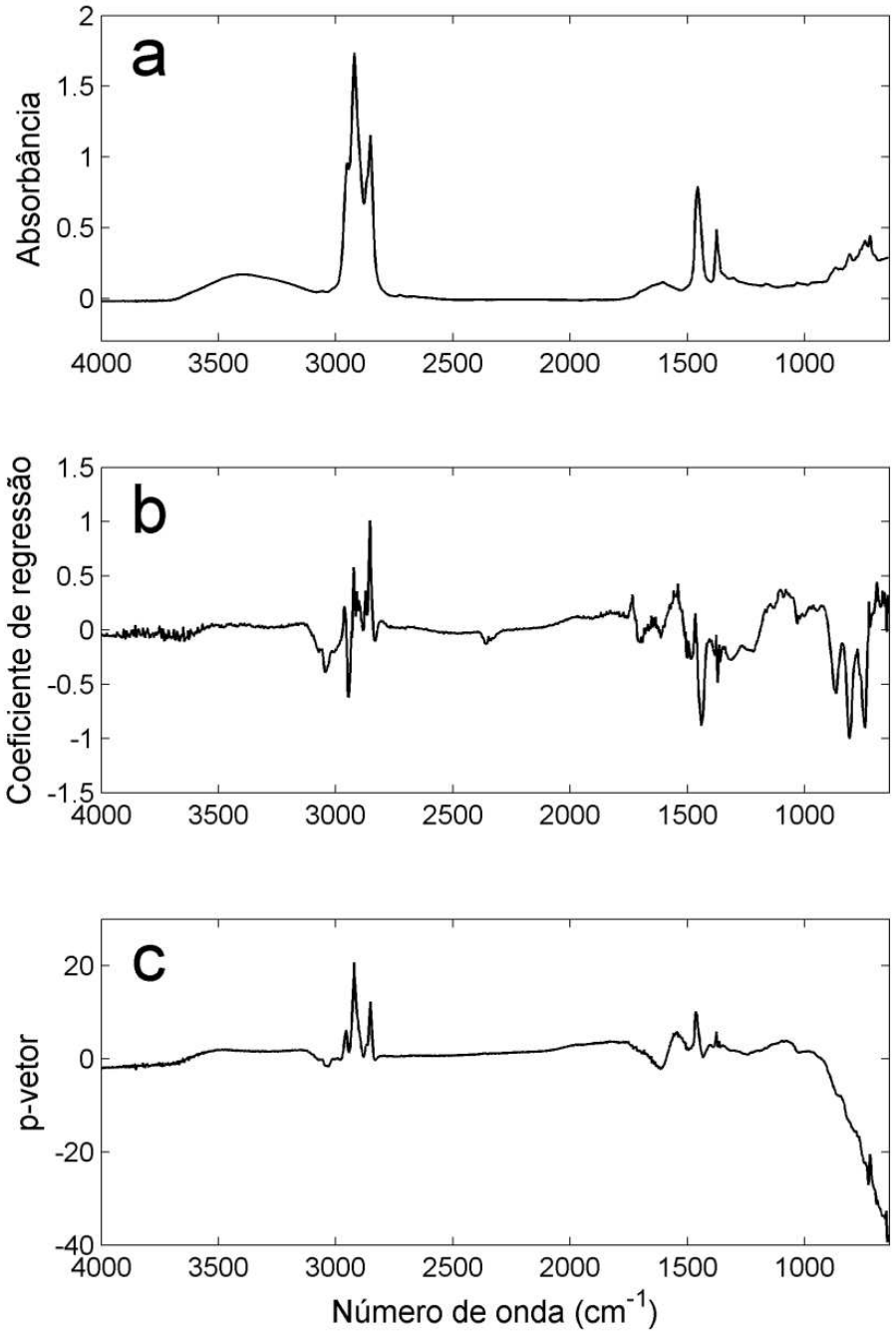

Figura 3.6. Espectro ATR-FTIR médio das amostras de calibração (a); coeficientes de regressão do modelo PLS, com 6 variáveis latentes, na determinação da gravidade API (b); p-vetor do modelo SVR na determinação da gravidade API.

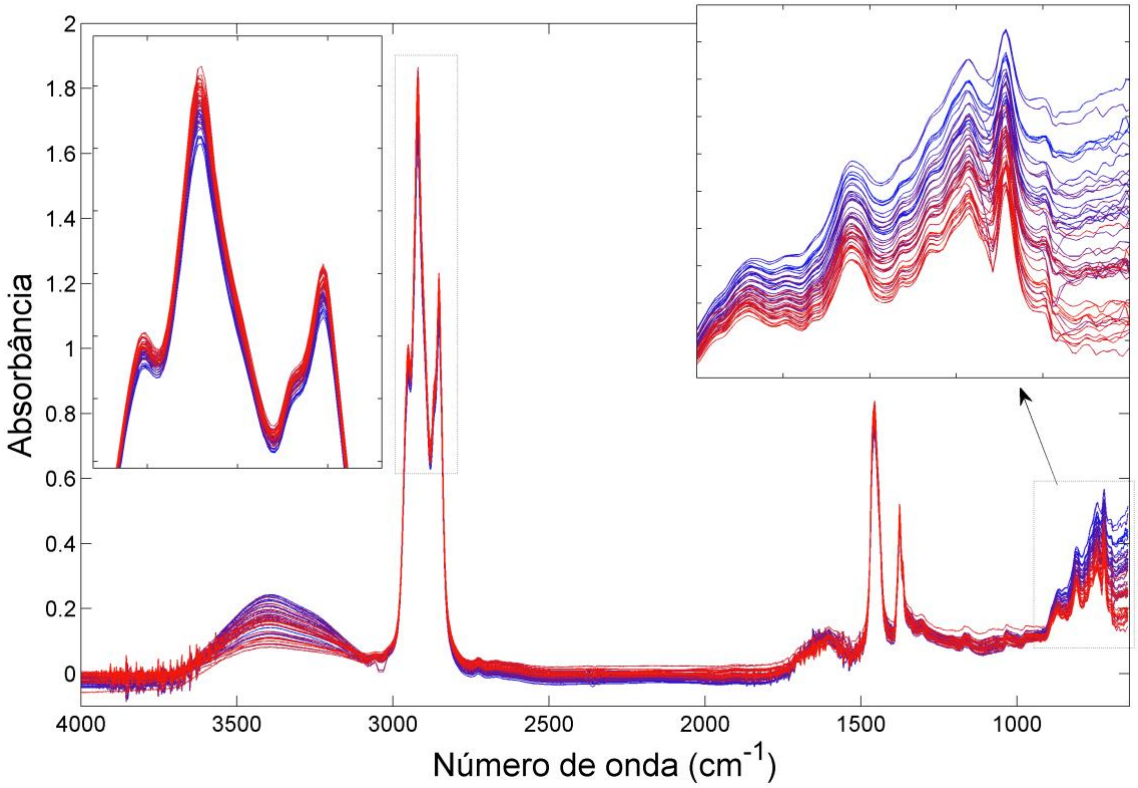

Figura 3.7. Espectros de ATR-FTIR das amostras de petróleo do conjunto de calibração com gradiente de cor em função do aumento da gravidade API da amostra. Azul: menor gravidade API; vermelho: maior gravidade API.

As regiões espectrais mais importantes na construção dos modelos para viscosidade cinemática são as mesmas que ao obtidas para a gravidade API, como esperado. Na Figura 3.8a está apresentado o espectro médio e na Figura 3.8b os coeficientes de regressão do modelo PLS parecem estar invertidos na região de impressão digital com relação aos coeficientes obtidos do modelo PLS para determinação da gravidade API (Figura 3.6). Esta relação inversa era esperada e pode ser visto na Figura 3.9 que o aumento da absorbância está diretamente relacionado com a elevação da viscosidade cinemática dos óleos. A mesma relação inversa não pôde ser visualizada no gráfico do p-vetor (modelo SVR para a viscosidade cinemática), este se parece mais com os p-vetores obtidos para a modelagem da gravidade API.

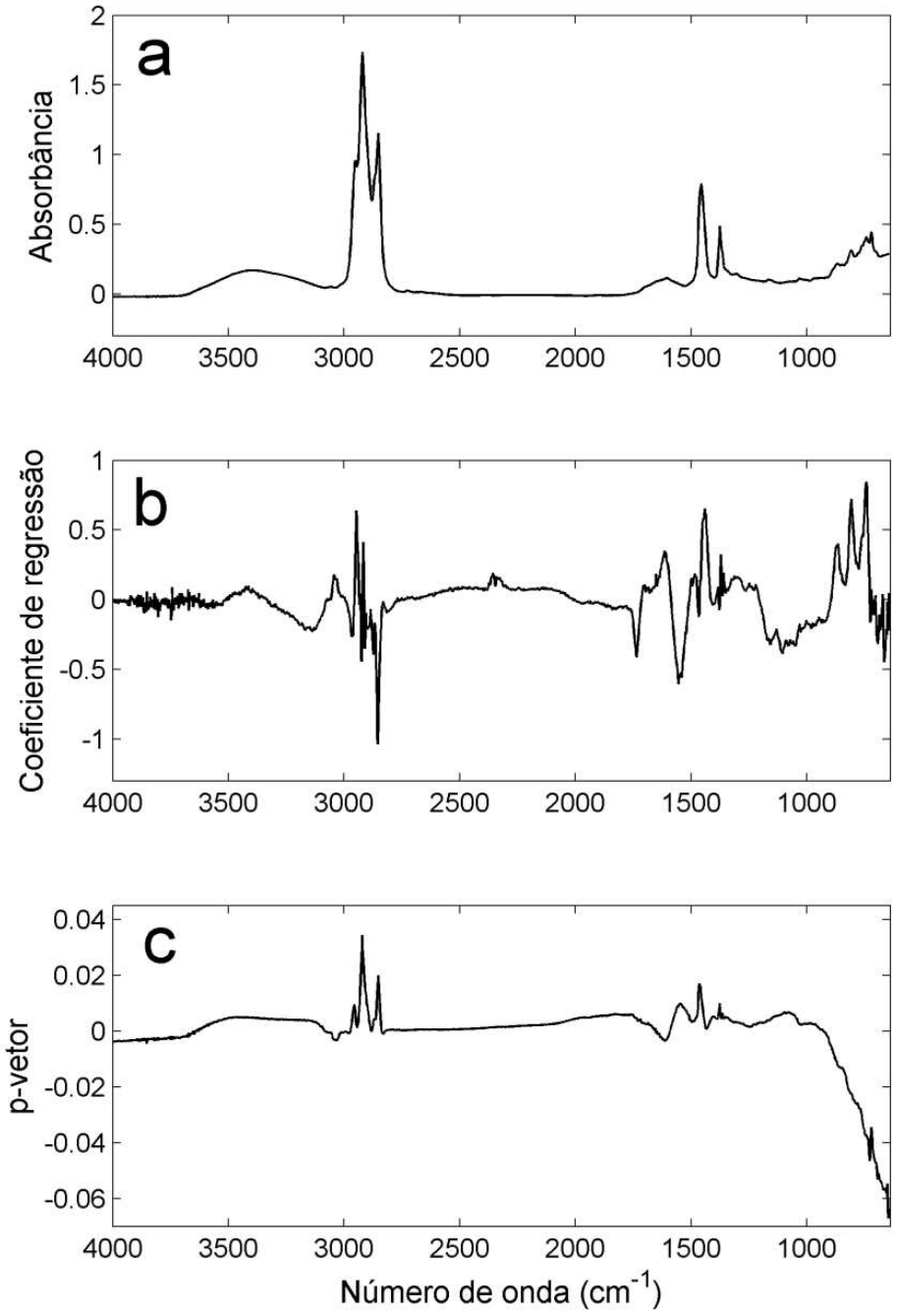

Figura 3.8. Espectro ATR-FTIR médio das amostras de calibração (a); coeficientes de regressão do modelo PLS, com 6 variáveis latentes, na determinação da viscosidade cinemática (b); p-vetor do modelo SVR na determinação da viscosidade cinemática.

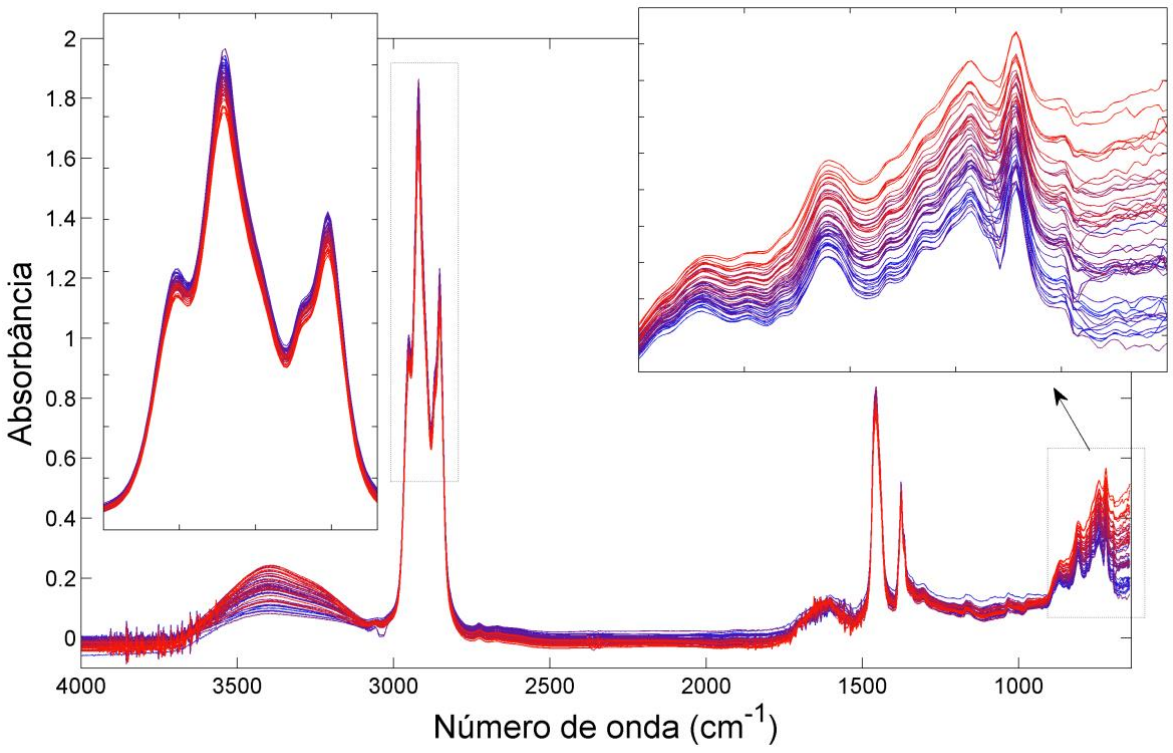

Figura 3.9. Espectros de ATR-FTIR das amostras de petróleo do conjunto de calibração com gradiente de cor em função do aumento da viscosidade cinemática da amostra. Azul: menor viscosidade cinemática; vermelho: maior viscosidade cinemática.

A elevação da absorbância na região em torno de 2900 cm^{-1} está relacionada a redução da viscosidade cinemática (Figura 3.9). Quanto maior a quantidade das cadeias carbônicas lineares presentes no petróleo, maior a interação entre elas e maior a resistência ao movimento, resultando num aumento da viscosidade cinemática do óleo.

Em relação à determinação do teor de água (Figura 3.10), a presença de água em amostras de óleos pode ser identificada pela larga banda de 3100-3600 cm^{-1} e associada à banda em torno de 1650 cm^{-1}. A região de 3100-3600 cm^{-1} característica do estiramento O–H, está muito bem definida e com grande importância no gráfico do p-vetor do SVR (Figura 3.10c) enquanto que praticamente não aparece no modelo PLS (Figura 3.10b).

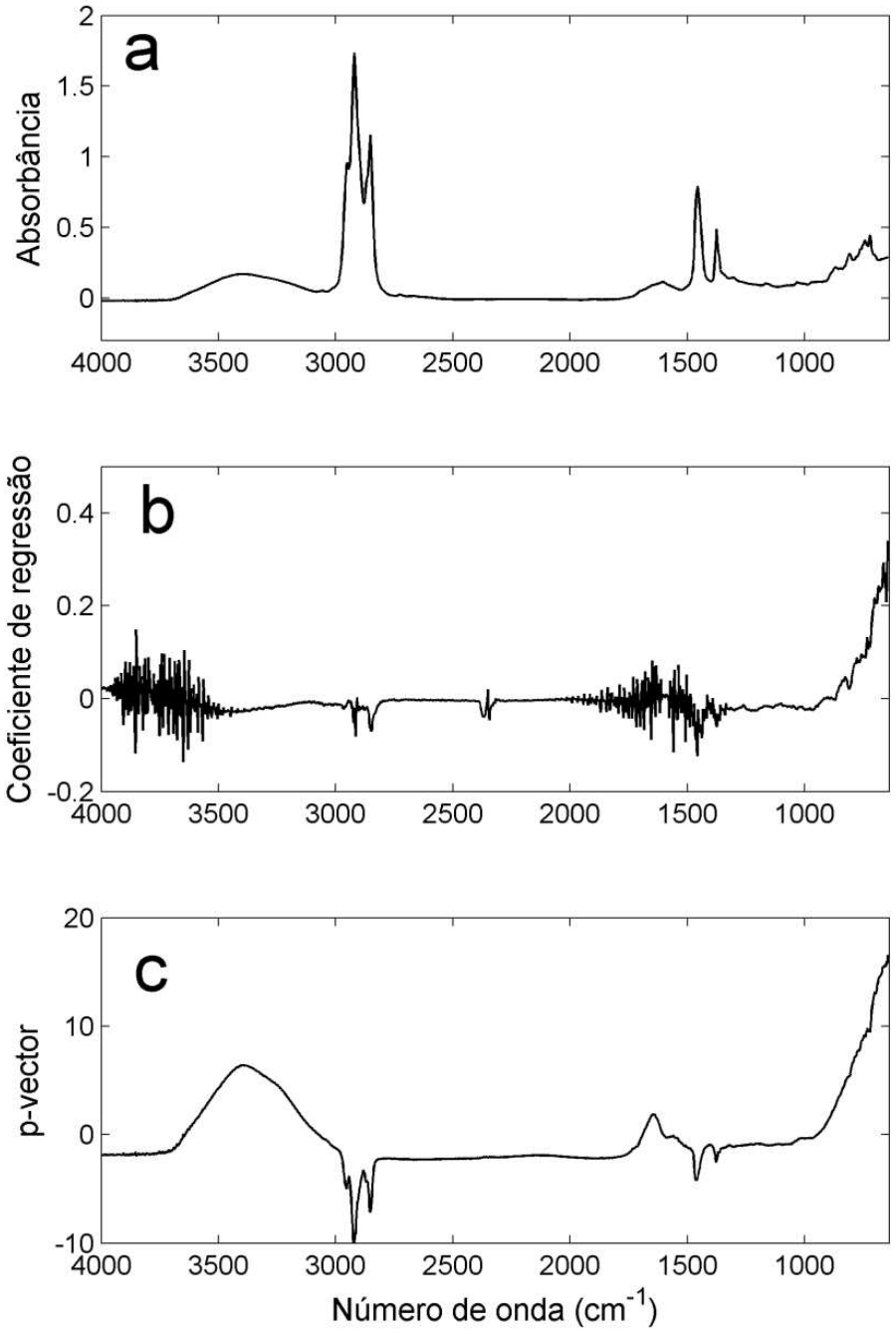

Figura 3.10. Espectro ATR-FTIR médio das amostras de calibração (a); coeficientes de regressão do modelo PLS, com 4 variáveis latentes, na determinação da teor de água (b); p-vetor do modelo SVR na determinação da teor de água.

A região em torno de 2900 cm^{-1}, referente à banda de estiramento C–H de carbonos primários e secundários, apresenta uma relação inversa com o teor de água no petróleo (Figura 3.11). Esta região espectral também apresenta importância nos modelos PLS e SVR construído (Figuras 3.10b e 3.10c). Observado na Figura 3.11 que sua relação é inversamente proporcional ao teor de água nas amostras. Para o modelo PLS esta região parece não ter importância (Figura 3.10b). Outra importante região a ser destacada nas Figuras 3.10b e 3.10c é abaixo de 900 cm^{-1}. Na Figura 3.11 observa-se relação direta do aumento do teor de água com aumento da absorbância na região de impressão digital, que corrobora os modelos PLS e SVR construídos.

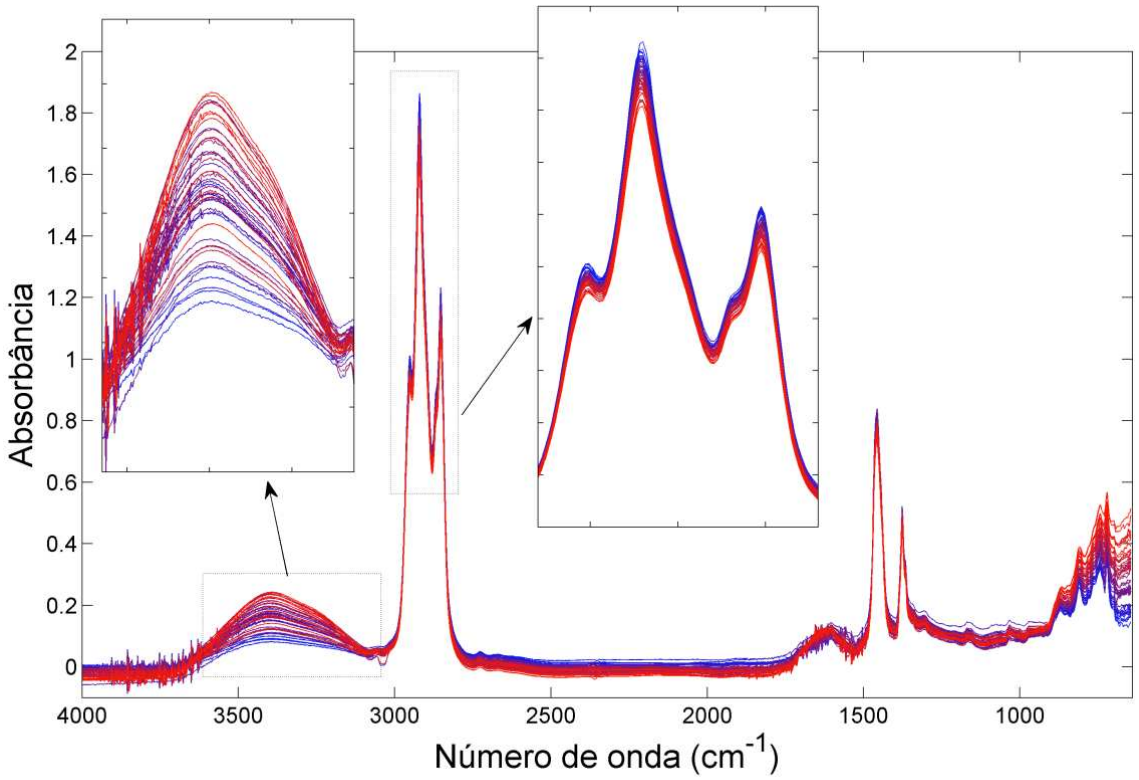

Figura 3.11. Espectros de ATR-FTIR das amostras de petróleo do conjunto de calibração com gradiente de cor em função do aumento do teor de água da amostra. Azul: menor teor de água; vermelho: maior teor de água.

3.4 Conclusões

A técnica de ATR-FTIR associada à metodologias de calibração multivariada é eficiente para determinar a gravidade API, viscosidade cinemática e teor de água em petróleos. A simplicidade na preparação da amostra, rapidez na análise, utilização de pequena quantidade de amostra e capacidade em determinar, simultaneamente, as três propriedades do petróleo com apenas um único espectro são as maiores vantagens no uso da metodologia proposta.

O modelo SVR é mais exato que PLS na determinação da gravidade API e viscosidade cinemática, enquanto que para o teor de água ambos os métodos são equivalentes. O produto da matriz espectral pelos vetores de suporte do modelo SVR (p-vetor) possibilita a interpretação das regiões espectrais com uma maior contribuição na construção do modelo SVR, levando a resultados tão interpretáveis quanto os coeficientes de regressão do modelo PLS. Um artigo com esses resultados foi publicado na revista Fuel.[62]

4 Aplicação 2: Cálculo do intervalo de confiança para modelos de regressão SVR por ensemble tipo *boosting* na determinação de temperaturas equivalentes a 10%, 50% e 90% de volume de destilado em petróleos por espectroscopia de RMN de ^{1}H

4.1 Introdução

O petróleo é constituído por uma mistura complexa de compostos que se evaporam a diferentes temperaturas. Assim, a qualidade dos óleos depende de várias propriedades físico-químicas medidas por métodos padrão que, na maioria das vezes, são demorados, custosos e requerem grandes quantidades de amostra e solvente. As temperaturas equivalentes a 10% (T10%), 50% (T50%) e 90% (T90%) de volume de evaporados são propriedades relevantes para estimar futuros cortes no petróleo bruto, condições ótimas de destilação e auxiliar na determinação de seu valor econômico.

A Ressonância Magnética Nuclear de próton (RMN de ^{1}H) vem sendo aplicada no controle de qualidade de óleos e combustíveis por apresentar vantagens em fornecer detalhes sobre os grupos funcionais moleculares permitindo a caracterização da amostra a nível molecular.[22] Contudo, para relacionar os espectros obtidos com a propriedade de interesse, em alguns casos é necessário aplicação de técnicas de calibração multivariada.

Pesquisas envolvendo métodos espectroscópicos combinados com quimiometria têm alcançado resultados promissores na estimativa de propriedades físico-químicas de petróleos, utilizando uma abordagem rápida e de baixo custo.[22-24,61-63] A regressão por mínimos quadrados parciais (PLS) pode ser considerada o método de regressão padrão mais utilizado em Química Analítica. O PLS gera bons resultados para a maioria dos problemas analíticos e a estimativa de suas figuras de mérito é descrita em alguns trabalhos científicos.[42,44,45,64,65] Contudo, em algumas aplicações práticas em matrizes complexas são indicados métodos de regressão não lineares, como redes neurais ou regressão de vetores de suporte (SVR).

O SVR é um método de regressão alternativo quando o problema em estudo envolve matrizes complexas, como óleos e combustíveis.[6,7,18,48,62] Uma importante informação a ser abordada na estimativa da propriedade de interesse é o intervalo de confiança. Em métodos não lineares onde é difícil de estabelecer uma equação analítica para o intervalo de confiança, métodos de reamostragem são procedimentos alternativos que podem fornecer boa estimativa do intervalo de confiança para previsão realizada.[66-69] Ensemble tipo *boosting* é um procedimento de reamostragem com ponderação iterativa aplicada ao algoritmo de aprendizagem enquanto os modelos de regressão são construídos.[70]

Estudos recentes apresentam resultados promissores na estimativa de temperaturas de destilação de derivados de petróleo em bruto utilizando os métodos multivariados. Godoy *et al.*[71] estimaram as temperaturas equivalentes a 10%, 50% e 90% de volume de destilado em gasolina por cromatografia gasosa bidimensional abrangente, com detecção por ionização de chama (GC × GC-FID). Lira *et al.*[15] estimaram as temperaturas equivalentes a 50% e 90% de volume destilado de misturas de diesel/biodiesel por espectroscopia de infravermelho na região do próximo (NIR) e médio (MIR). Galvão *et al.*[72] estimaram as temperaturas de destilação de 10% e 90% no diesel por espectroscopia NIR.

4.1.2 Objetivos

Estimar as temperaturas referentes a 10%, 50% e 90% de volume destilado de petróleo bruto utilizando espectroscopia de RMN de ^{1}H e regressão por vetores de suporte.

Utilizar a metodologia ensemble tipo *boosting* para estimar o intervalo de confiança na previsão obtida pelo modelo de regressão por vetores de suporte.

4.2 *Boosting* ensemble

Ensemble é uma técnica estatística que combina os resultados de vários modelos individuais para gerar uma estimativa única. Em problemas de regressão, cada modelo individual é chamado de regressor. As abordagens predominantes para geração dos regressores são *baggin* (*bootstrap agregation*) e *boosting*.

Bagging, proposto por Breiman,[73,74] é baseado em amostragem com reposição. Neste método, vários subconjuntos de calibração são gerados a partir de amostragem uniforme do conjunto original de dados (conjunto de treinamento). Em cada subconjunto amostrado um modelo de calibração é construído e utilizado na previsão das amostras do conjunto de teste.[72] Cada subconjunto tem o mesmo número de amostras; no entanto, algumas amostras do conjunto original podem ser escolhidas mais de uma vez.

No método *boosting*,[75,76] diferentes subconjuntos são gerados a partir das amostras do conjunto de calibração com probabilidades de seleção variável para cada amostra. A probabilidade de uma amostra ser selecionada para construção do submodelo de calibração depende de seu erro de calibração.[70] Assim, a reamostragem no método *boosting* normalmente é realizada sem reposição, ou seja, para um dado subconjunto não há repetição de uma mesma amostra como ocorre no método *bagging*. Neste trabalho, a estratégia *boosting* foi utilizada para determinar o intervalo de confiança em modelos PLS e SVR.

Inicialmente as amostras de trabalho são divididas em dois conjuntos: calibração e previsão. A partir das amostras de calibração, um subconjunto de amostras é selecionado para construção do primeiro regressor e utilizado para prever as amostras de calibração e teste. Inicialmente, cada amostra do conjunto de calibração tem a mesma probabilidade, $P(i)$, de ter ser escolhida para construção do primeiro regressor. Esta probabilidade é dada por:

$$P(i) = 1 - \frac{C_{ncal-1,nsubcal}}{C_{ncal,nsubcal}} \tag{4.1}$$

onde *ncal* é o número de amostras do conjunto de calibração, *nsubcal* é o número de amostras escolhidas como subconjunto para construção do regressor e C é o

operador de combinação. Neste trabalho, os subconjuntos foram formados com 70% das amostras do conjunto de calibração.[70]

A partir da previsão das amostras de calibração, o erro padrão de calibração (SEC) é calculado para o regressor:

$$SEC = \sqrt{\sum_{i=1}^{n} \frac{\left(y_{ref,i} - y_{est,i}\right)^2}{df}} \tag{4.2}$$

onde $y_{ref,i}$ é o valor da propriedade de interesse, $y_{est,i}$ o valor estimado pelo modelo e df é o número de graus de liberdade do modelo, que para modelos PLS é determinado pelo número de amostras do subconjunto de calibração menos o número de variáveis latentes menos 1 se os dados espectrais foram centrados na média. Para a calibração multivariada não linear, o número de graus de liberdade utilizado foi igual ao número de amostras do subconjunto de calibração.

Antes da subsequente reamostragem, a probabilidade de escolha das amostras de calibração é alterada. As amostras de calibração com erros menores que o SEC têm suas probabilidades de escolha aumentadas, ao passo que as amostras com erros elevados terão menor probabilidade de escolha. Este procedimento é realizado através da ponderação das amostras de calibração pelo peso w_i. Para cada amostra de calibração um valor z_i é determinado pela equação:

$$z_i = \frac{\left|y_{ref,i} - y_{est,i}\right|}{SEC} \tag{4.3}$$

O valor z_i fornece uma normalização do erro de calibração para cada amostra, como uma função do valor de SEC. Para suavizar elevados desvios de calibração, uma função logarítmica é aplicada aos valores de z_i. O novo peso para todas as amostras de calibração, escolhida no regressor, é calculado por:

$$w_i = \log\left(k + z_i\right) \tag{4.4}$$

onde k é uma constante utilizada para evitar que erros de amostras com elevados pesos sejam inflacionados após cada regressor. Quanto mais próximo de 10 for o valor de k, mais suavizada será o novo peso das amostras escolhidas. Neste

trabalho foi utilizado k igual a 9, para aumentar a suavização dos pesos w_i durante a construção dos 100 regressores individuais.

Cada regressor construído é utilizado na estimativa das amostras do conjunto de previsão. Ao final da modelagem, cada amostras de previsão terá tantas estimativas quanto o número de reamostragens. A estimativa final é representa pela média dos valores de previsão de todos os regressores construídos e a partir de sua dispersão é calculado o intervalo de confiança da estimativa realizada.

4.3 Metodologia

4.3.1 Valores de Referência

Neste estudo foram utilizados 35 amostras de petróleo da bacia sedimentar da costa brasileira. Os valores de referência da propriedade de interesse: T10%, T50% e T90% de petróleo bruto foram medidos por destilação simulada (SIMDIS, do inglês *Simulated Distillation*), utilizando o método de referência ASTM D 7169.[77] A medida dessas propriedades depende fortemente da gravidade API do petróleo bruto. As amostras usadas neste estudo apresentam gravidades API variando de 17,0 a 54,0, cuja distribuição é mostrada na Figura 4.1. Estes valores cobrem uma vasta gama de gravidades API para diferentes tipos de petróleos brutos pesados.

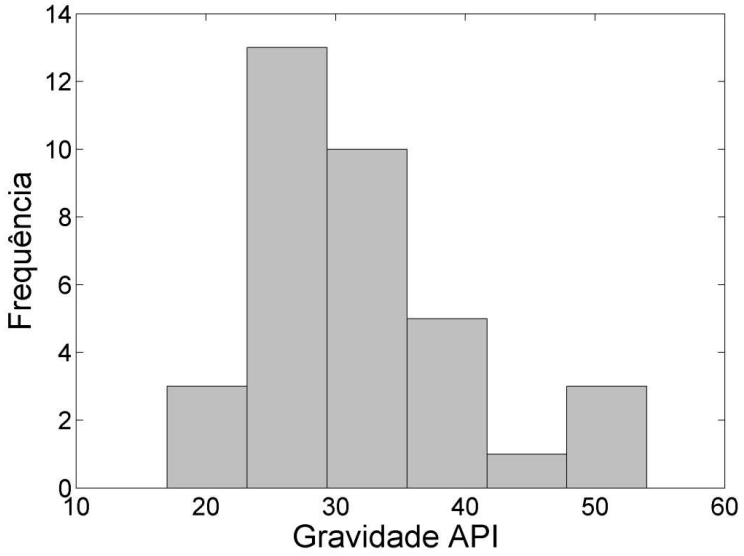

Figura 4.1. Histograma da gravidade API medida para as 35 amostras de petróleo utilizadas na modelagem.

4.3.2 Medidas de RMN de ^{1}H

Os espectros de RMN de ^{1}H foram medidos a 27°C utilizando diclorometano deuterado como solvente, em um equipamento Varian MR-400, com campo magnético de 399,8 MHz, utilizando as seguintes condições: largura espectral 6410,3 Hz, duração de impulso 10,0 µs, tempo de relaxação de 30 segundos e número de transientes 64.

4.3.3 modelos de calibração

As 35 amostras de petróleo foram divididas em 24 para calibração e 11 para previsão pelo algoritmo Kennard-Stone.[57] Antes da construção dos modelos de calibração, os espectros de RMN de ^{1}H foram alinhados utilizando o programa icoshift.[78] Os espectros foram pré-processados por variação normal padrão (SNV, do inglês *standard normal variate*).[58] Adicionalmente, os dados espectrais e valores da propriedade de interesse foram normalizados na faixa de 0 a 1, antes

da construção dos modelos SVR. Os modelos PLS e SVR foram desenvolvidos utilizando a metodologia do *boosting* ensemble, chamado ePLS (ensemble PLS) e eSVR (ensemble SVR). Nos modelos eSVR os p-vetores foram determinados. Os parâmetros dos modelos SVR e eSVR foram otimizados por uma grade de pesquisa com procedimento de validação cruzada "5-*fold*".

4.3.4 Determinação dos intervalos de confiança

O intervalo de confiança (*IC*) para o modelo PLS foi determinada de acordo com a norma ASTM E 1655-12:[42]

$$IC_i = y_{est,i} \pm t_{\alpha,df} SEC \sqrt{(1+h_i)} \quad (4.5)$$

onde t é o valor da distribuição t de *Student* com graus de liberdade df e nível de significância α, e h_i o *leverage* para cada amostra. O *leverage* representa a distância de cada amostra ao centro dos dados, sendo muito utilizada para a detecção de valores anômalos. O *leverage* foi determinado por:

$$h_i = \mathbf{t}_{nlv,i}^T \left(\mathbf{T}_{nlv}^T \mathbf{T}_{nlv}\right)^{-1} \mathbf{t}_{nlv,i} \quad (4.6)$$

onde **T** é a matriz de *scores* do modelo PLS para o conjunto de amostras de calibração, $\mathbf{t}_i$ é o vetor de *scores* para a amostra *i* com *nlv* variáveis latentes.

Para os intervalos de confiança estimados por uma metodologia *boosting* ensemble, 100 regressores foram construídas com k = 9.

4.4 Resultados e discussões

O petróleo bruto é composto predominantemente por hidrocarbonetos alifáticos, naftênicos e grupos moleculares de compostos aromáticos. Assim, a técnica de RMN de ^{1}H é adequada para análise do petróleo bruto ao nível molecular. Petróleos com diferentes gravidades API apresentam grandes diferenças em seus espectros de RMN de ^{1}H. Na Figura 4.2 são mostrados os

espectros de um petróleo extraleve com API de 54,0 e um óleo pesado com API de 17.

Nestes espectros as seguintes regiões de deslocamentos químicos podem ser destacadas: 0,5-1,0 ppm, CH_2 de cadeia alquílica longa relacionada a compostos alifáticos, com maior intensidade no petróleo extraleve; 1,2-1,3 ppm, CH_2 também de cadeia alquílica longa, mas presentes em ambos óleos; 1,3-1,6 ppm, CH_2 β-aromáticos identificado com maior intensidade no óleo extraleve; 2,1-2,4 ppm, CH_3 α-aromáticos identificado com maior intensidade no óleo extraleve; 5,3-5,4 ppm, olefinas, com diferentes características para o óleo extraleve e pesados; e 6,8-7,4 ppm, referente a mono-aromáticos com maior intensidade no espectro do óleo extraleve. Outro aspecto importante observado nos espectros é a ausência de água, notada na região de 4,6-4,9 ppm.

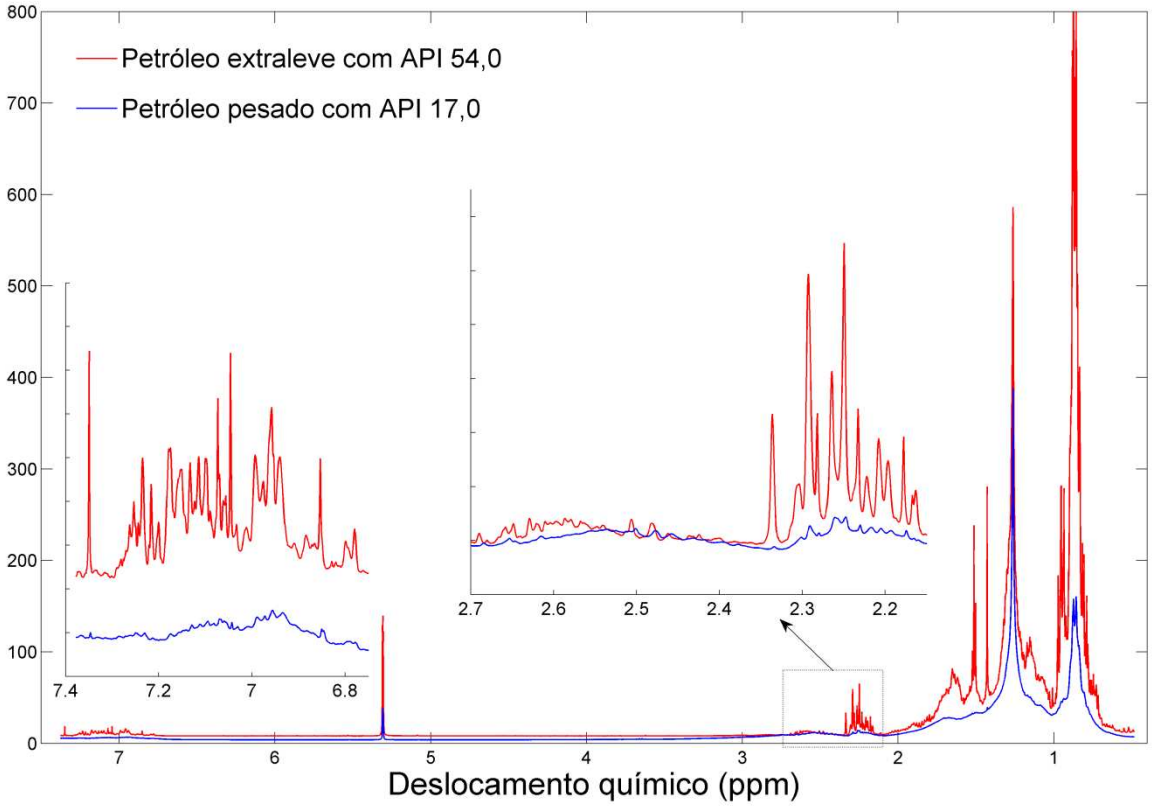

Figura 4.2. Espectros de RMN de ^1H para um petróleo extraleve com gravidade API 54,0 (—) e um petróleo pesado com gravidade API 17,0 (—).

Antes da construção dos modelos de calibração, os espectros de RMN de ^{1}H foram alinhados utilizando o software icoshift (Figura 4.3). A rotina icoshift para Matlab está disponível gratuitamente para download no site www.models.life.ku.dk. Na Figura 4.3 é observado o alinhamento dos espectros com destaque para duas regiões, antes do alinhamento (Figura 4.3a) e após alinhamento (Figura 4.3b).

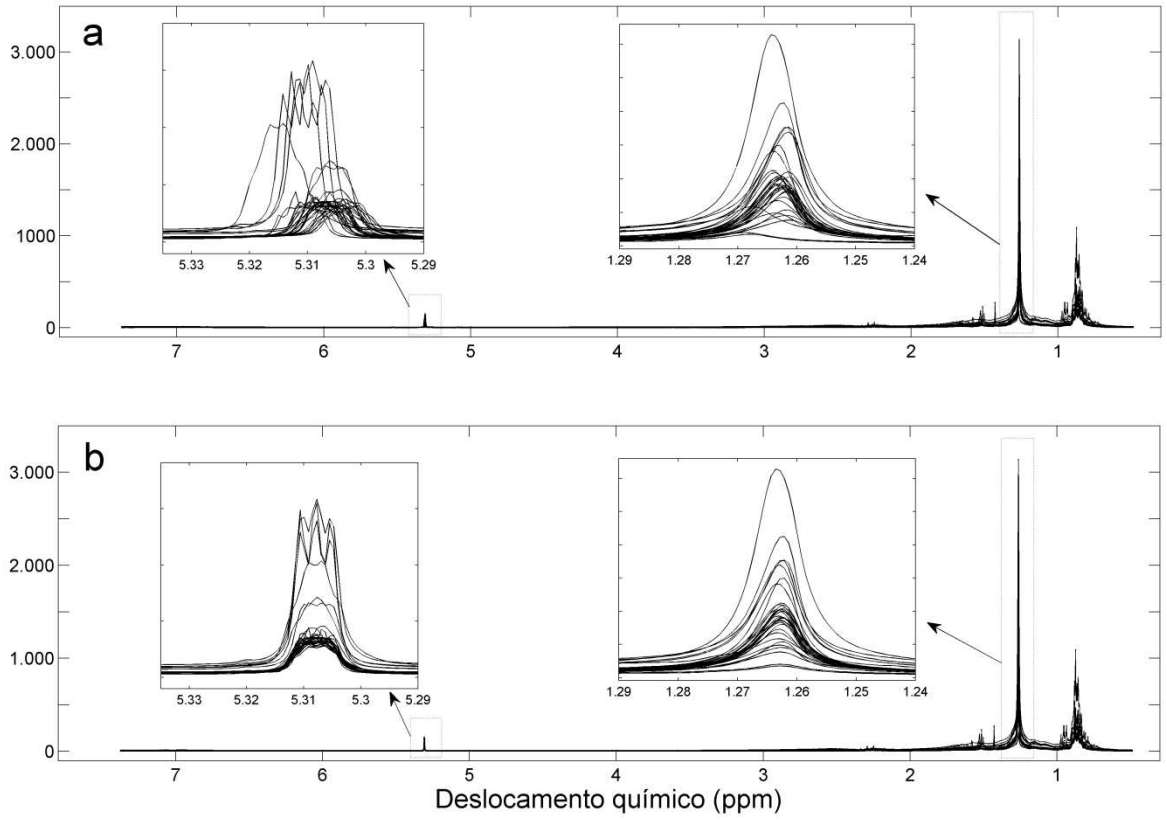

Figura 4.3. Espectros de RMN de ^{1}H das 35 amostras de petróleos utilizadas na modelagem antes do alinhamento (a) e após alinhamento pelo programa icoshift (b).

Neste trabalho, as 3 propriedades: T10%, T50% e T90% de volume destilado do petróleo são determinadas a partir do mesmo conjunto de espectros de RMN de ^{1}H, mas construindo um modelo para cada parâmetro. O método eSVR apresentou menor RMSEP para as 3 propriedades estudadas (Tabela 4.1).

Tabela 4.1. Resultados de previsão dos modelos de calibração.

Modelo	RMSEP (°C)	R^2p	Viés (°C)
	Propriedade T10%		
PLS	15,6	0,7972	0,6
ePLS	15,1	0,8047	0,2
SVR	28,4	0,6802	-0,1
eSVR	11,6	0,8949	3,6
	Propriedade T50%		
PLS	24,2	0,9223	2,0
ePLS	23,4	0,9265	1,2
SVR	22,8	0,9416	-7,9
eSVR	14,4	0,9784	0,8
	Propriedade T90%		
PLS	39,0	0,9127	-4,5
ePLS	39,9	0,9144	-6,5
SVR	37,0	0,9166	1,4
eSVR	23,3	0,9753	1,0

Na calibração da propriedade T10%, os modelos PLS, ePLS e eSVR apresentaram exatidão equivalentes e superior ao modelo SVR, segundo o teste randômico para exatidão. Para esta propriedade o método *boosting* ensemble aplicado a regressão por vetores de suporte mostrou ser muito vantajoso. Na Figura 4.4, é observado um melhor ajuste do modelo eSVR em relação ao modelo SVR. O modelo PLS apresentou intervalo de confiança de aproximadamente mesma magnitude e maiores que os modelos ePLS e eSVR. Isso se deve a pouca influência do *leverage* da amostra no cálculo do intervalo de confiança. Na Equação 4.5 pode-se observar que o peso do *leverage* no cálculo do IC é proporcional a sua raiz quarada. Assim, amostras muito distante da posição média dos dados podem apresentar intervalos de confiança semelhantes a amostras mais próximas da posição média.

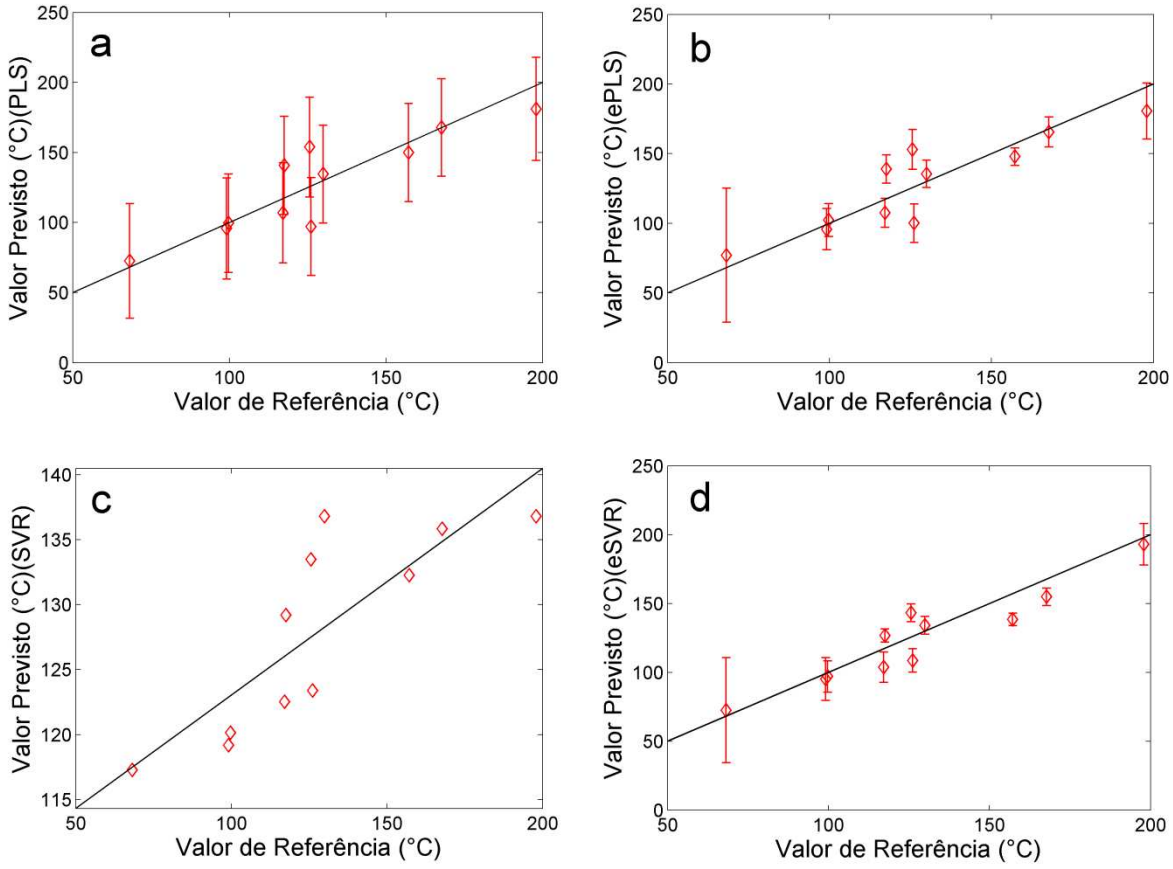

Figura 4.4. Temperatura equivalente a 10% de volume destilado de petróleo medido pelo método ASTM D 7169 *versus* valor previsto pelo modelo PLS (a), ensemble PLS (b), SVR (c) e ensemble SVR (d). As barras verticais representam o limite de confiança de 95%.

Na calibração da propriedade T50%, o modelo eSVR apresentou melhor exatidão que os modelo PLS e ePLS, segundo o teste randômico para exatidão. Na Figura 4.5 é observada grande diferença no intervalo de confiança obtido pelos modelos utilizando a metodologia *boosting* ensemble (ePLS e eSVR) e o método PLS. Os métodos ePLS e eSVR apresentaram maior intervalo de confiança para amostra com menor temperatura T50% enquanto que o modelo PLS apresentou mesma magnitude do intervalo de confiança para esta amostra.

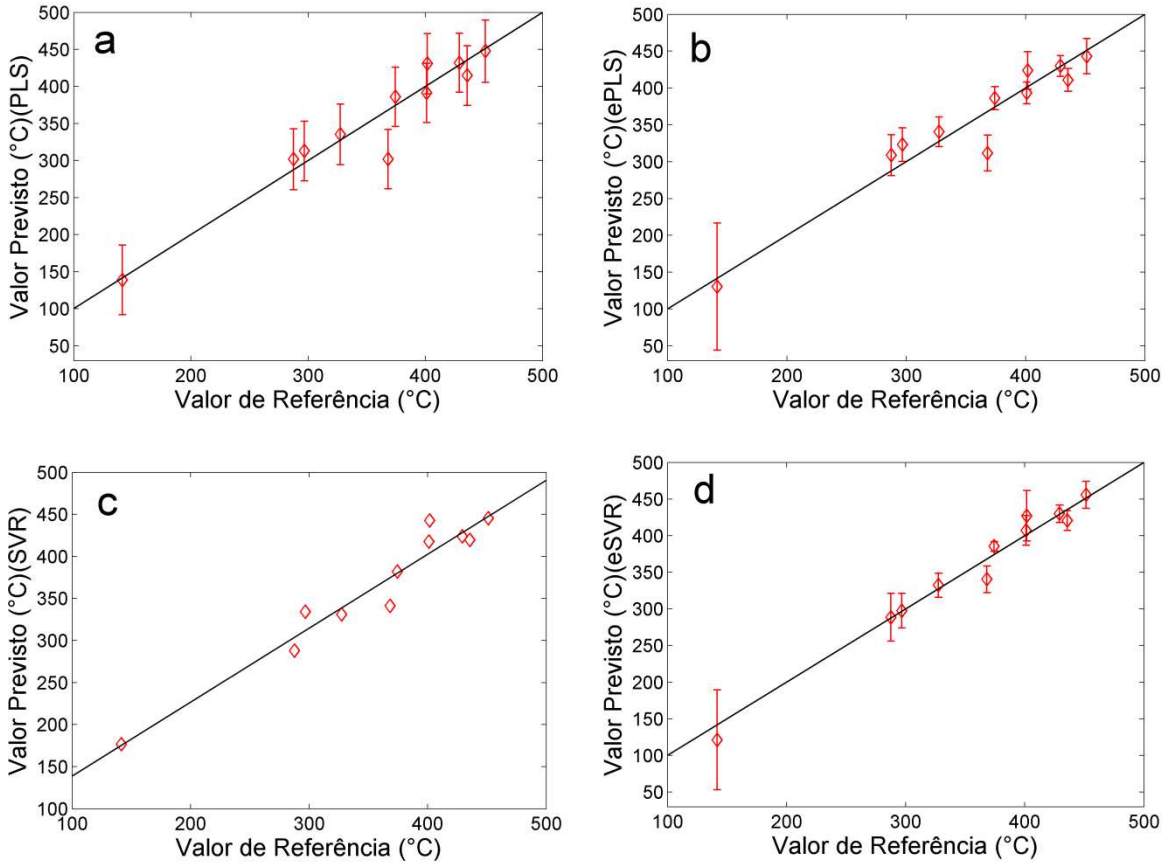

Figura 4.5. Temperatura equivalente a 50% de volume destilado de petróleo medido pelo método ASTM D 7169 *versus* valor previsto pelo modelo PLS (a), ensemble PLS (b), SVR (c) e ensemble SVR (d). As barras verticais representam o limite de confiança de 95%.

Na calibração da propriedade T90%, o modelo eSVR apresentou melhor exatidão que o modelo PLS, segundo o teste randômico para exatidão. Na Figura 4.6, foi observado o mesmo comportamento dos intervalos de confiança obtidos na modelagem da propriedade T50%.

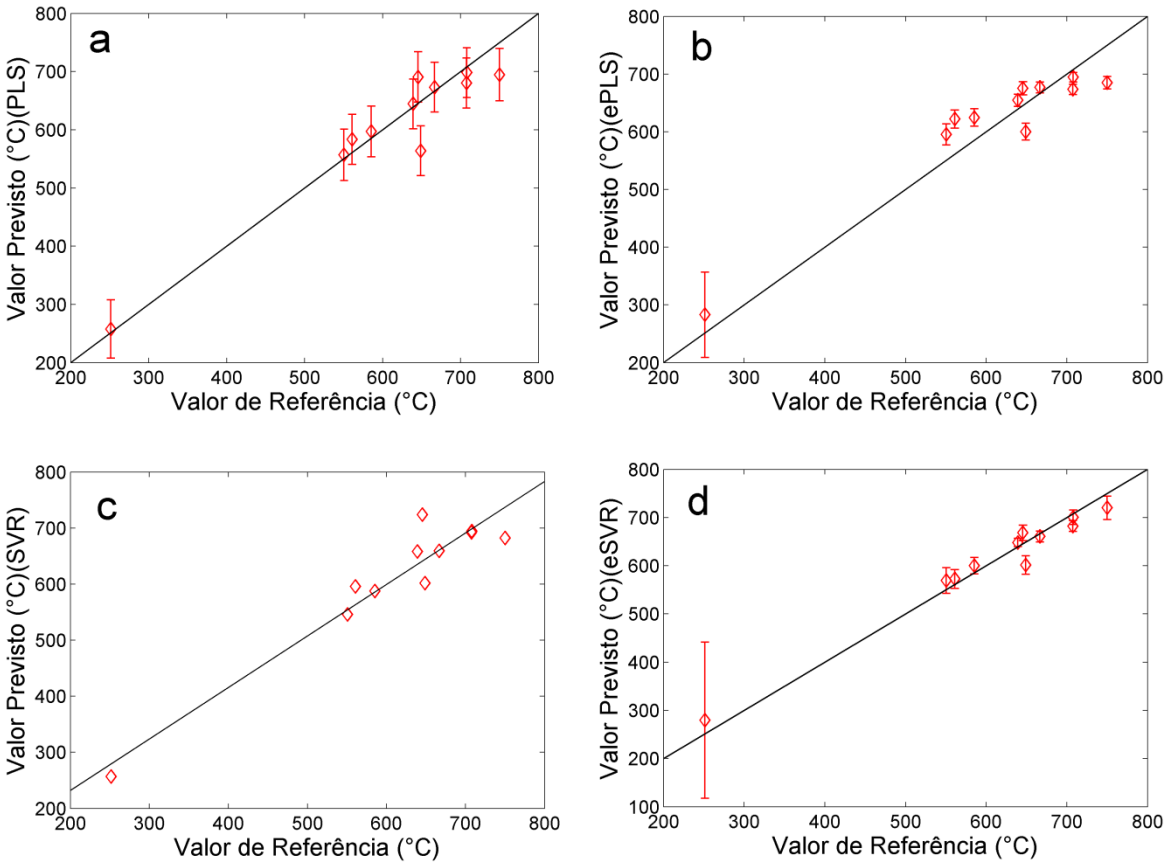

Figura 4.6. Temperatura equivalente a 90% de volume destilado de petróleo medido pelo método ASTM D 7169 *versus* valor previsto pelo modelo PLS (a), ensemble PLS (b), SVR (c) e ensemble SVR (d). As barras verticais representam o limite de confiança de 95%.

Os resíduos dos modelos de calibração foram avaliados para identificação de erros sistemáticos ou de tendência. Os testes para viés e teste de permutação para erros de tendência mostrou não haver estes tipos de erros nos resíduos dos modelos de calibração ao nível de significância de 5%. Na Tabela 4.2 são mostrados os valores de t-calculado para o teste para viés e os p-valores do teste de permutação para erros de tendência.

Tabela 4.2. Valores de t-calculado para o teste para viés e p-valores do teste de permutação para erros de tendência.

Viés teste: t-calculado *[1]				
Propriedade	PLS	ePLS	SVR	eSVR
T10%	0,11	0,04	0,00	1,01
T50%	0,26	0,16	1,16	0,18
T90%	0,60	0,52	0,12	0,14
Teste de permutação para tendência: p-valor *[2]				
Propriedade	PLS	ePLS	SVR	eSVR
T10%	0,225	0,235	0,428	0,455
T50%	0,367	0,383	0,158	0,458
T90%	0,379	0,168	0,271	0,334

*[1]: t_{tab} crítico de 2,23 ($\alpha = 0{,}05$ e $df = 10$).
*[2]: p-valor crítico de 0,05.

No procedimento proposto *boosting* ensemble, cada amostra de calibração tem sua probabilidade de seleção para a próxima amostragem alterada de acordo com o peso (w) após construção de cada regressor. Quanto maior o peso w da amostra, menor probabilidade de ser escolhida e, consequentemente, menor sua contribuição no modelo final construído. Os pesos finais das amostras de calibração obtidas para o modelo eSVR são mostrados na Figura 4.7. Adicionalmente ao número de vezes que cada amostra de calibração foi selecionada para construção dos regressores.

A amostra de número 6 apresentou maior peso, com valor próximo de 1, apenas para a modelagem de propriedade T10% (Figura 4.7a). Desta forma, esta amostra foi menos frequentemente selecionada para construção dos regressores, como mostrado na Figura 4.7d. Estes resultados indicam que a amostra de número 6 é anômala (*outlier*) apenas para modelagem da propriedade T10%, pois seu peso é de mesma magnitude na modelagem das propriedades T50% e T90%.

Na determinação da propriedade T10% o modelo SVR apresentou RMSEP muito maior que os outros modelos (Tabela 4.1). Este menor ajuste do modelo pode ser devido a uma amostra anômala no conjunto de calibração, tendo em

vista que a diferença no RMSEP dos modelos SVR e eSVR para as propriedades T50% e T90% é muito menor que a observada para a propriedade T10%. Isso pode ser devido ao fato que o modelo SVR procura uma função menos suave, com maior curvatura, na tentativa de explicar o comportamento não linear dos dados com uma amostra anômala.

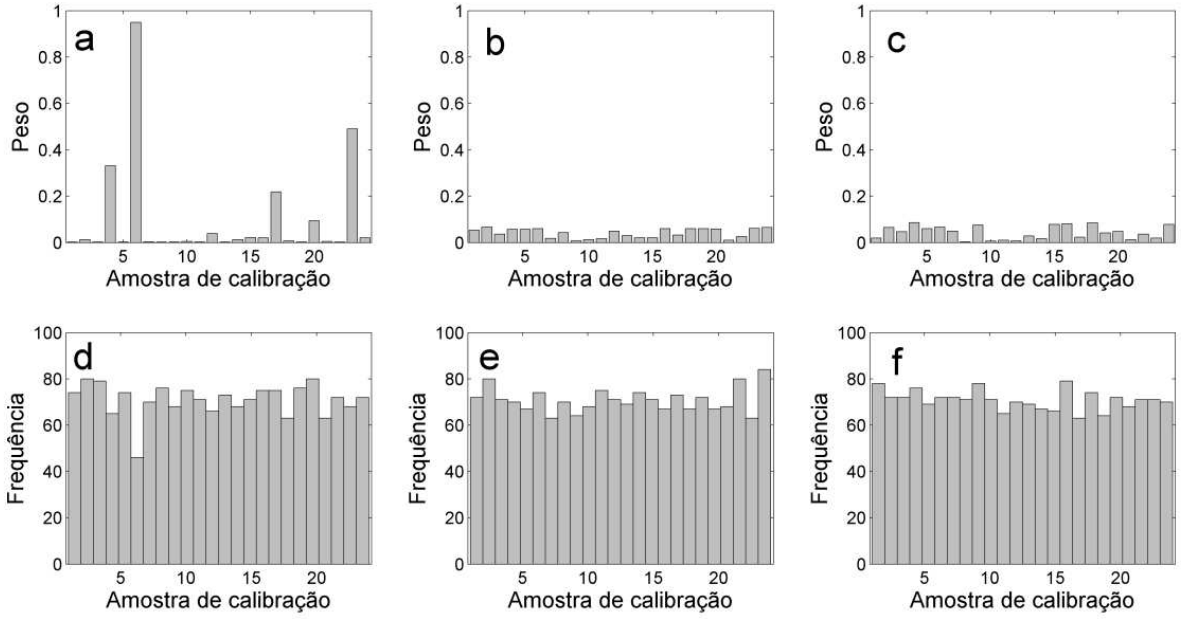

Figura 4.7. Gráfico de barras dos pesos do modelo ensemble SVR para as propriedades: T10% (a), T50% (b) e T90% (c); e frequência de amostragem de cada amostra de calibração para construção do modelo eSVR para as propriedades: T10% (d), T50% (e) e T90% (f).

Para melhor comparação da exatidão dos modelos, foi calculado o percentual do RMSEP em relação ao valor médio de previsão:

$$RMSEP\% = 100\frac{RMSEP}{\bar{y}_{previsão}} \tag{4.7}$$

onde que $\bar{y}_{previsão}$ é a média das amostras de conjunto de previsão. A partir destes valores é possível observar que o modelo eSVR gerou menores erros de previsão para todas as 3 propriedades (Figura 4.8). Em termos relativos, a menor exatidão foi obtida na calibração da propriedade T10%, o qual é a propriedade em que foi

observada um *outlier* no conjunto de calibração. O alto RMSEP% no modelo SVR para modelar a propriedade T10% mostra que o modelo não linear é mais significativamente influenciado pela presença de *outliers* que o modelo linear PLS. A utilização da metodologia *boosting* ensemble pode minimizar este efeito durante a modelagem.

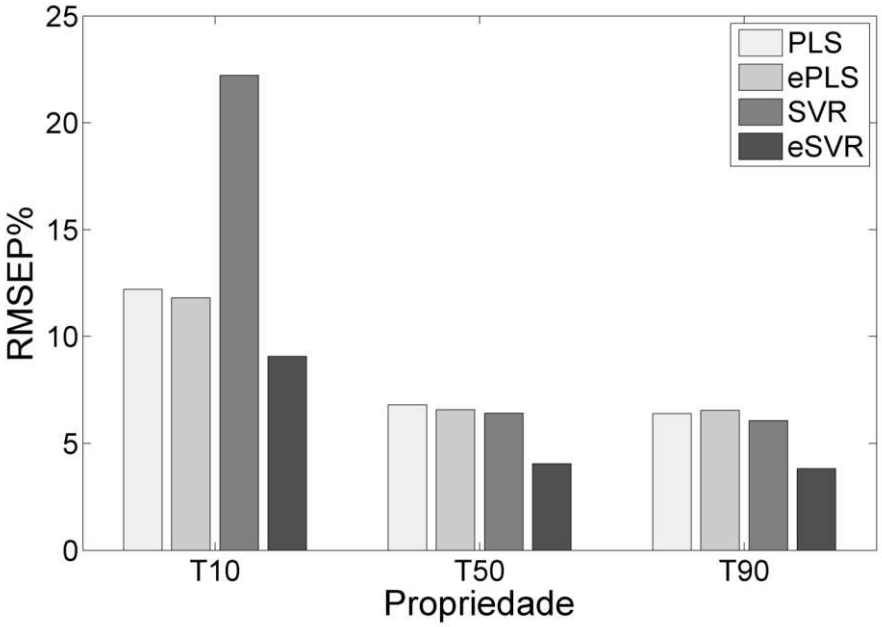

Figura 4.8. RMSEP% das propriedades T10%, T50% e T90% dos modelos PLS, ensemble PLS, SVR e ensemble SVR.

O p-vetor foi calculado em cada regressor do modelo eSVR. A Figura 4.9 apresenta o gráfico dos valores médios do p-vetor para as três propriedades. Os p-vetores médios são diretamente relacionados com os valores da propriedade de interesse (temperaturas relativas a 10%, 50% e 90% do volume destilado) e indicam regiões espectrais com maior importância na construção do modelo eSVR. Na Figura 4.9, observa-se que os deslocamentos químicos de prótons alfa aromáticos, região de 2,35-2,15 ppm são de grande importância para o modelo T10% e T90%. Entretanto, a região de 1,00-0,50 ppm, correspondente a prótons CH_2 de cadeias alquílicas longas de compostos alifáticos, é importante para a determinação da propriedade T50%.

Em T10%, frações de petróleo mais leves, constituídas principalmente por parafinas e naftenos de baixa massa molar (menos do que 8 átomos de carbono) já foram destiladas, restando as parafinas e naftenos de massa molar mais elevado (maior do que 8 átomos de carbono), compostos aromáticos, aromáticos nafteno e asfaltenos (que consiste principalmente de estruturas poliaromáticos). Neste ponto da destilação, a fração de compostos aromáticos em relação às parafinas e naftenos é superior àquela para o T50%, em que uma grande parte dos compostos aromáticos e aromáticos nafteno já foi destilada.

Em T50%, devido à menor proporção relativa de parafinas, aromáticos e naftênicos, a região de 2,35-2,15 ppm é menos importante na modelagem desta propriedade. Alternativamente, a região de 1,00-0,50 ppm, relativo a prótons CH_2 de cadeias alquílicas longas de compostos alifáticos, é mais importante para o modelo T50% em comparação com aos modelos T10 e T90% (Figura 4.9b). Este fato sugere que a destilação de compostos aromáticos e aromáticos nafteno, restando as estruturas poliaromáticas de asfaltenos.

Em T90%, restam apenas a fração asfáltica de petróleo, que consiste em maior proporção de compostos aromáticos policíclicos contendo de 6 a 20 anéis aromáticos. Neste ponto da destilação a região de 2,35-2,15 ppm torna-se novamente importante para o modelo.

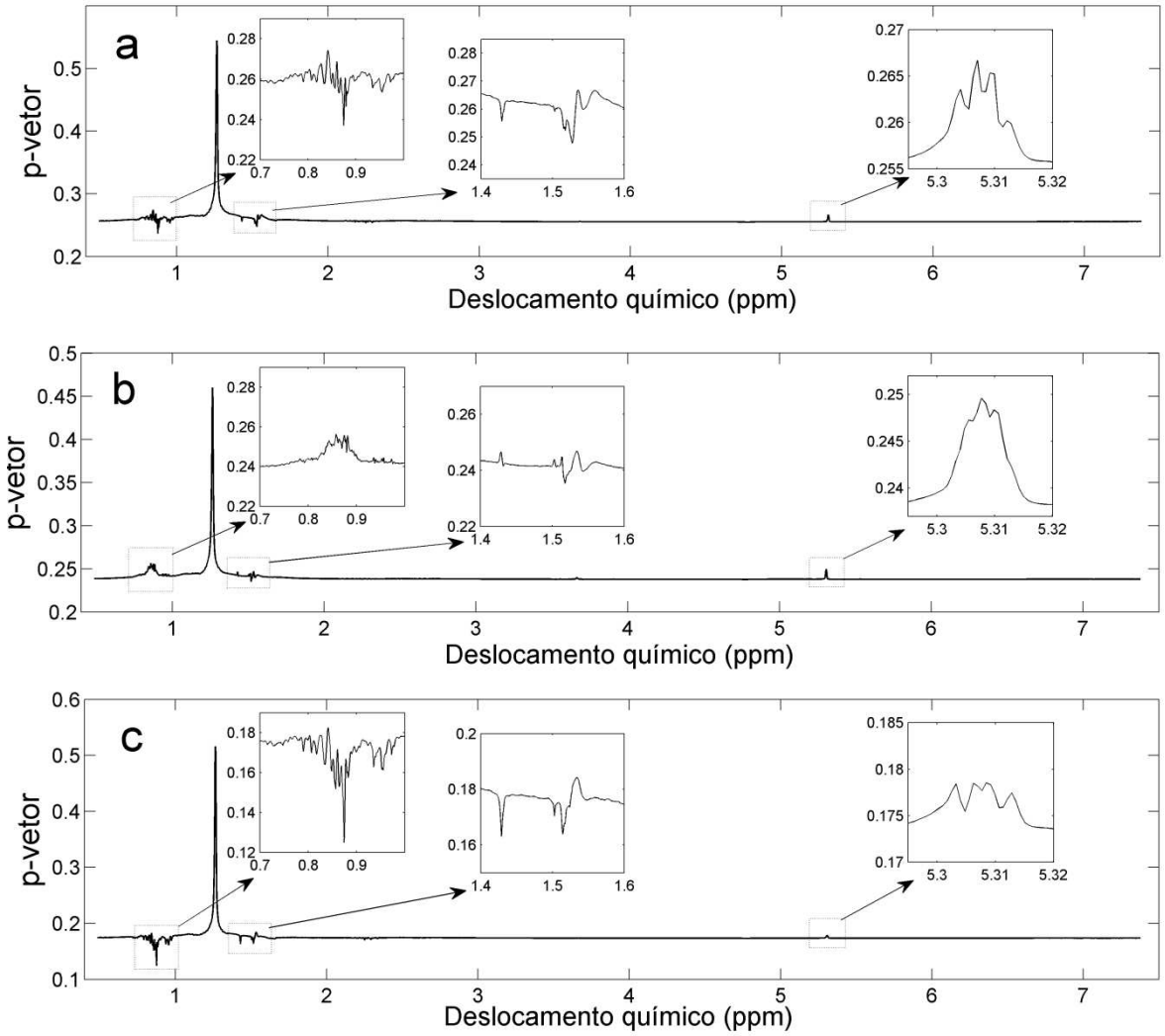

Figura 4.9. Gráfico do p-vetor médio do modelo ensemble SVR em função do deslocamento químico para as propriedades: T10% (a), T50% (b) e T90% (c).

4.5 Conclusões

Este trabalho demonstra a viabilidade da determinação das temperaturas de 10%, 50% e 90% de volume destilado em petróleos por espectroscopia de RMN de ^{1}H aliada à regressão por vetores de suporte. A utilização do método *boosting* ensemble permite estimar os intervalos de confiança para estes modelos, assim como possibilita a identificação de amostras anômalas no conjunto de calibração.

Na modelagem da propriedade T10%, uma amostra anômala é identificada, contudo, a mesma amostra exibe comportamento normal na modelagem das propriedades T50% e T90%. Usando a informação dos vetores de suporte é possível identificar regiões espectrais relacionadas com a propriedade de interesse, tornando os modelos SVR mais interpretáveis.

5 Aplicação 3: Determinação de saturados, aromáticos, resinas e asfaltenos em petróleo por RMN de ^{13}C e regressão por vetores de suporte com seleção de variáveis por algoritmo genético

5.1 Introdução

O petróleo é constituído predominantemente por hidrocarbonetos e compostos oxigenados, nitrogenados e sulfurados em menores quantidades.[1-3] Os hidrocarbonetos variam de moléculas simples, com poucos átomos de carbono, a moléculas complexas com alta massa molar.[37] Devido a esta mistura complexa, sua composição é medida por grupos de compostos com propriedades semelhantes, tais como: saturados, aromáticos, resinas e asfaltenos (SARA). Os saturados compreendem os alcanos de cadeia normal e ramificada (parafínicos) e cicloalcanos (naftênicos); os aromáticos são constituídos por moléculas aromáticas, cicloalcano-aromáticos (nafteno-aromáticos) e, usualmente, compostos cíclicos de enxofre; as resinas e asfaltenos são formadas por compostos policíclicos de alta massa molar, contendo uma maior quantidade de heteroátomos em sua composição. Os asfaltenos são insolúveis em alcanos leves como hexano e heptano. Assim, as resinas podem ser separadas dos asfaltenos.[79]

Devido a predominância de hidrocarbonetos, a técnica de ressonância magnética nuclear de próton (RMN de ^{1}H) vem sendo utilizada para estimar propriedades físico-químicas de petróleos e seus derivados.[24,80] Na determinação de SARA em petróleos a espectroscopia de RMN de ^{13}C pode ter uma maior potencialidade em relação ao RMN de ^{1}H. Os átomos de carbono no interior das estruturas de resinas e asfaltenos (Figura 5.1) não são identificados pela técnica de RMN de ^{1}H por não estarem ligados a prótons, entretanto, a técnica de RMN de ^{13}C obtém sinal destes átomos de carbono.

Um espectro de RMN de ^{13}C contém uma grande quantidade de variáveis, podendo passar de 65 mil, dependendo da faixa espectral analisada. Em aplicações quimiométricas, tratar esta grande quantidade de dados para várias amostras gera um alto custo computacional. Assim, é importante a utilização de

algum método de seleção de variáveis para gerar modelos mais rápidos e parcimoniosos. Métodos determinísticos de seleção de variáveis tais como: utilização de intervalos espectrais, combinação de intervalos são indicados quando o espectro amostral é continuo, como a espectroscopia no infravermelho. Espectros de RMN tem variáveis com caráter discreto, assim o método probabilístico de seleção de variáveis GA (algoritmo genético) é mais indicado.

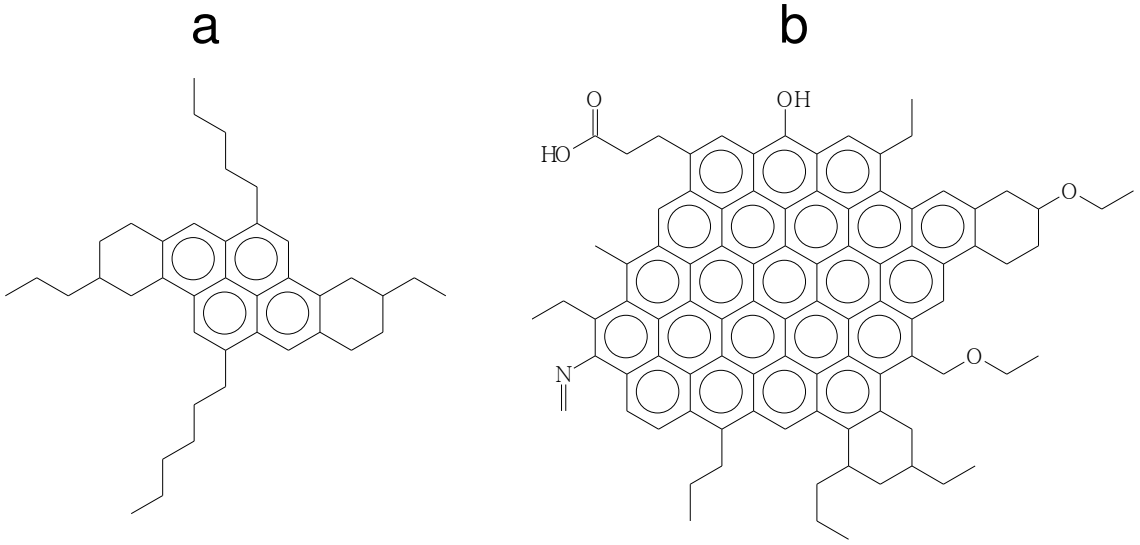

Figura 5.1. Exemplo da estrutura de uma molécula de resina (a) e asfalteno (b).

O algoritmo genético (GA do inglês, *Genetic Algorithm*) é uma técnica de seleção de variáveis amplamente aplicada em quimiometria.[81] Um diferencial do GA é a possibilidade de otimização de um subconjunto de variáveis espectrais simultaneamente aos parâmetros de otimização do modelo de calibração.

5.1.2 Objetivo

Determinar o teor de saturados, aromáticos, resinas e asfaltenos em petróleo bruto utilizando espectroscopia de RMN de ^{13}C associada a regressão por vetores de suporte utilizando algoritmo genético para selecionar variáveis espectrais e otimizar os parâmetros do modelo SVR simultaneamente.

5.2 Algoritmo genético

O algoritmo genético é uma técnica probabilística de busca e otimização matemática inspirada no princípio Darwiniano da evolução das espécies. Seu processo de otimização baseia-se no princípio de sobrevivência dos indivíduos mais aptos, reprodução e mutação dos mesmos.[82] De acordo com a teoria de Darwin, o princípio de seleção privilegia os indivíduos mais aptos com maior probabilidade de reprodução, passando seu código genético às próximas gerações.

Estes princípios biológicos foram motivadores para o desenvolvimento de algoritmos matemáticos de busca e otimização. Eles podem ser usados para encontrar solução numérica em problemas com grande número de variáveis, sendo esta uma de suas principais vantagens. Em aplicações químicas, normalmente o algoritmo genético é utilizado para selecionar variáveis em espectroscopia.[81,83-87] Neste trabalho o GA foi utilizado para selecionar variáveis e otimizar os parâmetros do modelo SVR simultaneamente.

O ponto de partida para a utilização do GA é a representação matemática do problema. Após a codificação, o algoritmo é inicializado e busca de forma iterativa pontos ótimos dentro do domínio amostral. Uma representação esquemática de algoritmo genético é mostrada na Figura 5.2. As etapas segue a ordem lógica:

1. Codificação das variáveis

Cada cromossomo artificial é representado por uma sequência de códigos binários (0's e 1's). Este tipo de codificação facilita o processo de otimização. Para otimizar parâmetros do modelo SVR e selecionar variáveis do espectro de RMN de ^{13}C simultaneamente, foram utilizados 15 dígitos binários para representar um gene (parâmetro a ser utilizado) do modelo SVR e apenas um dígito binário para cada variável espectral. Visto que para seleção de variáveis existem apenas duas possibilidades de decisão: ser ou não selecionada. Normalmente, codificações de GA utilizam 0 para variável não selecionada e 1 para variável selecionada.

A função resposta do problema consiste em obter um conjunto de parâmetros e variáveis do espectro de RMN de ^{13}C que melhor descrevam o

conjunto de dados de calibração. Este é alcançado pela minimização do RMSECV. Assim, o GA foi utilizado para obter mínimo valor de RMSECV calculado pelo procedimento "5-*fold*".

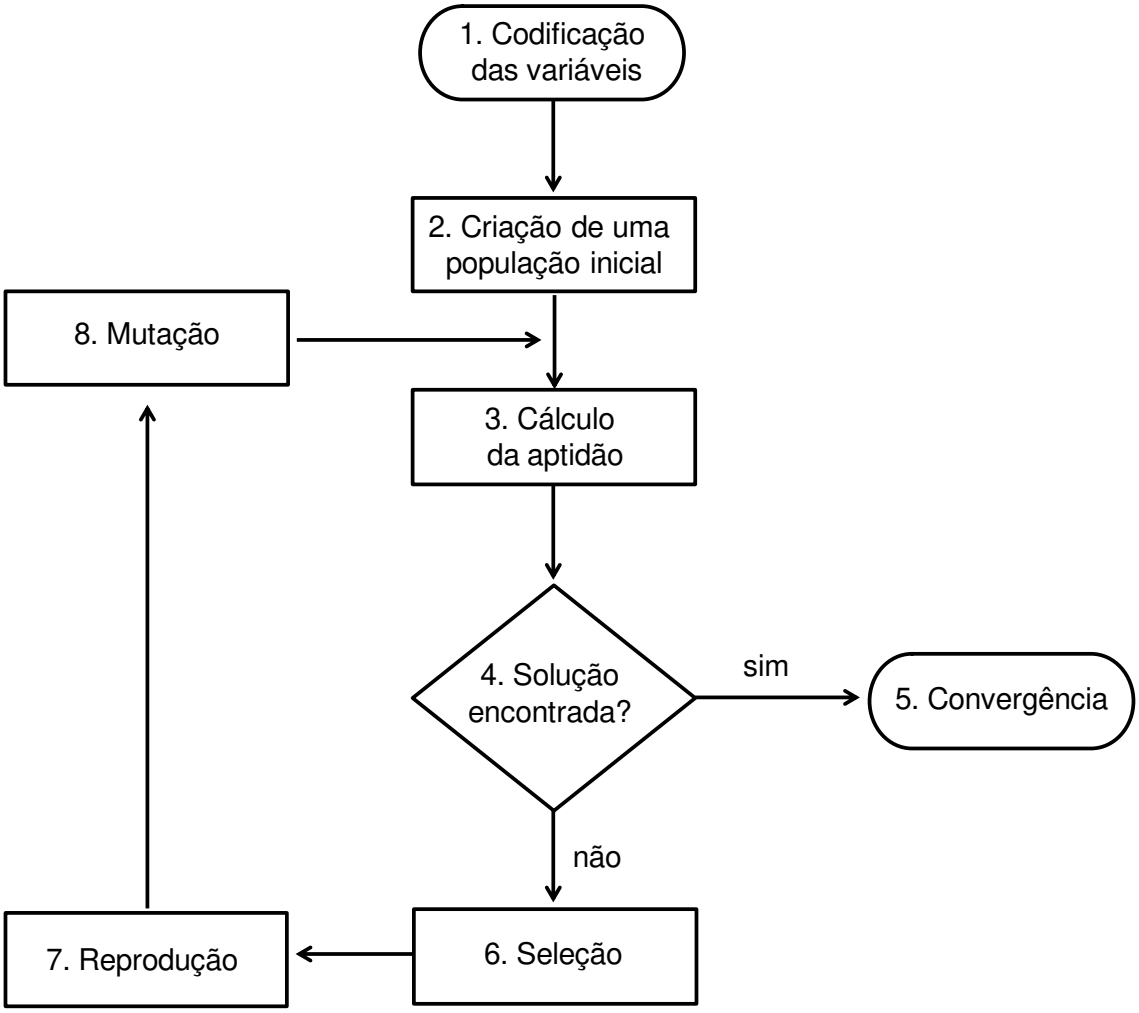

Figura 5.2. Fluxograma das principais etapas de um algoritmo genético básico.

2. Criação de uma população inicial

Para inicialização dos cálculos, uma população de indivíduos é criada de forma randômica para garantir que não há nenhum tipo de influência tendenciosa do meio externo. O número de indivíduos da população inicial depende do problema. Para seleção de variáveis em espectroscopia, aproximadamente 25% do total de variáveis espectrais representa um bom valor. Entretanto, em espectros de RMN de ^{13}C o número de variáveis é bem superior e adotar este

critério levaria a um alto custo computacional, que talvez fosse desnecessário. Nestes casos, duas medidas podem ser adotadas. Primeiro, reduzir o número de variáveis espectrais, excluindo regiões de menor interesse. Outra possibilidade é reduzir o tamanho da população inicial para algo em torno de 200 a 500 indivíduos, mas aumentando o número de gerações para uma boa convergência do algoritmo.

3. Cálculo da aptidão

Depois de criada a população inicial, o desempenho de cada indivíduo é avaliado, obtendo-se como resposta uma solução para o problema (aptidão). Neste caso, cada solução representa um valor de RMSECV. Indivíduos com menores valores de RMSECV são mais aptos, ou seja, podem estar mais próximos do mínimo global.

4. Avaliação da resposta

Todas as possíveis soluções obtidas agora devem ser avaliadas. Caso alguma delas satisfaz um critério de convergência, que pode ser certo valor mínimo de RMSECV ou porcentagem de convergência da população para certo valor, o algoritmo se encerra e o indivíduo mais apto (com melhor resposta) representa a resposta do problema. Em muitos casos estes critérios não são satisfeitos, assim adota-se um número máximo de gerações. Neste trabalho foi adotado número máximo de 300 gerações.

5. Convergência

Caso o critério de convergência adotado seja satisfeito o algoritmo é encerrado. Este critério é adotado para evitar cálculos desnecessários quando uma boa resposta para o problema foi alcançada. Caso o critério de convergência não seja satisfeito a etapa seguinte é a seleção dos indivíduos.

6. Seleção

Cada indivíduo da população apresenta uma solução para problema. Os indivíduos mais aptos (menores valores RMSECV) são selecionados, pois estão mais próximos do mínimo global procurado. Nesta etapa, uma porcentagem de indivíduos é descartada por não apresentarem resultados satisfatórios.

7. Reprodução

Apenas os indivíduos mais aptos são selecionados para reprodução e construção da próxima geração (população). Nesta etapa do algoritmo é aplicada uma função para aumentar a probabilidade de cruzamento dos indivíduos mais aptos. O objetivo da reprodução é transformar a população a cada sucessiva geração. O cruzamento dos indivíduos mais aptos é necessário para que a nova população se diversifique mantendo características adaptativas adquiridas pelas gerações anteriores. O cruzamento é o operador responsável pela recombinação de características dos pais durante a reprodução, permitindo que as próximas gerações herdem essas características. O operador genético para cruzamento pode ser de várias maneiras; as normalmente utilizadas são:

- Um ponto fixo: um ponto fixo de cruzamento é escolhido e a partir deste ponto as informações genéticas dos pais são trocadas. As informações anteriores a este ponto dos pais são trocados, gerando dois novos filhos;
- Dois pontos fixos: a ideia é semelhante à de um ponto fixo, mas neste caso são utilizados 2 pontos. Quanto maior o número de pontos de cruzamento, menos características dos pais serão herdadas pelos filhos;
- Máscara: uma máscara de cruzamento é aplicada aos pais para gerar os dois novos filhos. A máscara pode conter vários pontos de cruzamento.

Neste trabalho foi utilizado o operador de ponto fixo com taxa de cruzamento de 80%.

8. Mutação

A mutação assegura a probabilidade alcançar qualquer ponto do domínio de busca, além de contornar o problema de mínimos locais, não deixando que a população convirja muito cedo para mínimos locais. A mutação é caracterizada pela alteração no valor de um código binário no gene de um indivíduo. Ela é aplicada aleatoriamente com baixa probabilidade de ocorrência, normalmente menor que 1%. Devido a grande quantidade de genes de cada população, vários indivíduos da população subsequente podem ter um de seus genes alterado aleatoriamente. Assim neste trabalho foi adotada uma taxa de mutação de 1%.

Quando a solução encontrada não converge para o critério estabelecido (etapa 4 para 5), as etapas de 3 a 8 são repetidas. Cada repetição desta é chamada de geração. Quanto maio o número de gerações maiores a chances de o algoritmo convergir para o mínimo global. Entretanto, número muito altos de gerações (maiores que 500) pode representar um gasto computacional extra desnecessário.

O algoritmo genético foi executado 100 vezes para reduzir a probabilidade de seleção de variáveis pouco informativas. Com a execução consecutiva do GA, variáveis importantes tendem a ser repetidamente selecionadas, ao passo que variáveis menos importantes são selecionadas ao acaso.

5.3 Metodologia

Neste trabalho foram utilizadas 65 amostras de petróleo com características físico-químicas muito diferentes, variando de extraleve a asfálticos. O teor de saturados, aromáticos, resinas e asfaltenos foram determinados por uma coluna cromatográfica preparativa. O procedimento consiste em adicionar 200 mg de petróleo em uma coluna de fracionamento de sílica-gel (230-400 mesh). A fração de saturados é obtida pela adição de 200 mL de hexano no topo da coluna. Após remoção da fração de saturados, 200 mL de uma mistura de hexano/diclorometano (1:1) é adicionada no topo da coluna para obtenção da fração dos aromáticos. Em seguida, é adicionado 200 mL de metanol para remoção da fração polar do petróleo, composto por resinas e asfaltenos. Devido a uma inexatidão na quantificação do teor de resinas e asfaltenos separadamente, estes foram determinados como uma única propriedade. Desta forma, foram construídos modelos para determinação de três conjuntos de dados: saturados, aromáticos e resinas mais asfaltenos.

Os espectros de RMN de ^{13}C foram medidos a 27 °C com um equipamento Varian MR-400, com campo magnético de 399,8 MHz. As amostras foram separadas em dois conjuntos: 45 para calibração e 20 para previsão.

Para construção dos modelos de calibração foi desenvolvido um método para seleção de variáveis espectrais e otimização dos parâmetros do modelo SVR simultaneamente por algoritmo genético. Um dos problemas na utilização do GA é sua convergência para mínimos locais. Para minimizar este efeito o GA foi executado 100 vezes (gerando 100 modelos GA-SVR) e as variáveis com maior frequência de seleção foram utilizadas para construção do modelo final. Os parâmetros dos modelos SVR foram selecionados pelo valor mediano.

Inicialmente um vetor de códigos binários contendo apenas 0 e 1 é criado. Os parâmetros do modelo SVR pertencem ao conjunto Real. Assim, para representação de um número Real, um conjunto de 16 códigos binários foi utilizado para representar cada parâmetro do modelo SVR a ser otimizado. As variáveis espectrais por outro lado, podem ser representadas apenas por um código binário, recebendo valor 0 caso a variável não seja escolhida para construção do modelo, ou 1 caso esta variável seja escolhida (Figura 5.3). Desta forma, durante a otimização os parâmetros do modelo SVR são simultaneamente otimizados ao passo que variáveis espectrais vão sendo selecionadas.

O GA foi configurado com: população inicial de 1024 indivíduos, máximo número de gerações de 200, taxa de mutação de 1%, inicialização do algoritmo com 15% do total de variáveis espectrais, otimização com procedimento de validação cruzada “5-*fold*”.

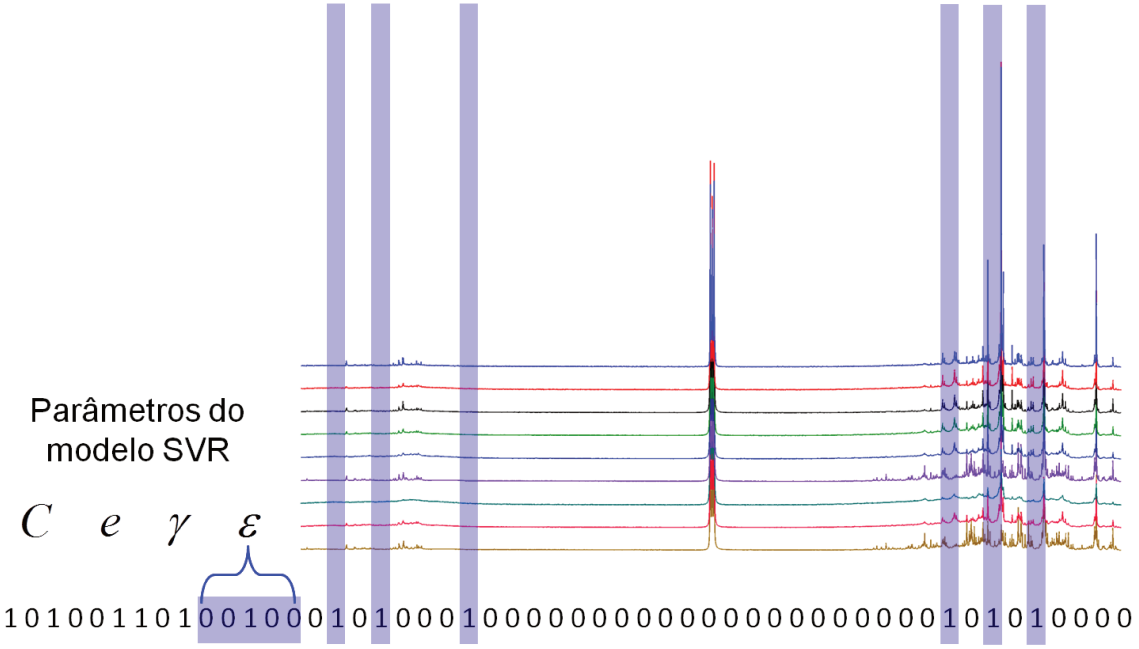

Figura 5.3. Exemplo ilustrativo da otimização de parâmetros do modelo SVR simultaneamente a seleção de variáveis de espectros de RMN de ^{13}C por algoritmo genético.

5.3 Resultados e discussões

A gravidade API é um importante parâmetro a ser determinado no petróleo. Normalmente, amostras mais pesadas (com menor gravidade API) têm maiores quantidades de resinas e asfaltenos. As 65 amostras de petróleo utilizadas variam a gravidade API de 10 a 55 (Figura 5.4). Estes óleos compreendem a faixa de extraleves a asfálticos.

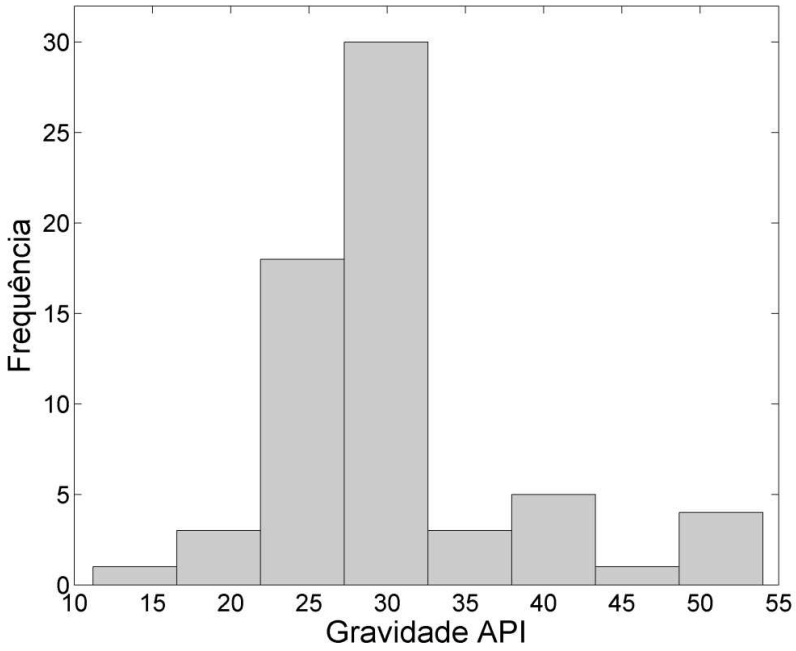

Figura 5.4. Histograma da gravidade API das 65 amostras de petróleos utilizadas.

Antes da construção dos modelos de calibração, os espectros de RMN de ^{13}C precisaram ser alinhados. Na Figura 5.5a é destacada o deslocamento dos sinais em duas regiões, enquanto na Figura 5.5b é mostrada estas regiões após alinhamento. O deslocamento observado está relacionado à modificações no ambiente químico do carbono 13 observado. Entretanto, em métodos quantitativos estes pequenos deslocamentos prejudicam a otimização dos modelos de calibração.

Em calibração multivariada, dados de espectroscopia NIR e MIR geralmente apresentam grande correlação entre os espectros amostrais (colinearidade entre as variáveis), assim, para seleção de variáveis, métodos determinísticos como: seleção de intervalos espectrais equidistantes ou combinação destes são preferencialmente utilizados. Estes métodos selecionam regiões espectrais pré-determinadas e constroem modelos de regressão com essas variáveis selecionadas. Entretanto, em dados de espectrometria de massas ou espectroscopia de ressonância magnética nuclear (de ^{1}H ou ^{13}C) os espectros apresentam menos variáveis colineares. Nestes casos, métodos probabilísticos de

seleção de variáveis espectrais pode gerar melhores resultados. Assim, o GA foi adotado para seleção de variáveis.

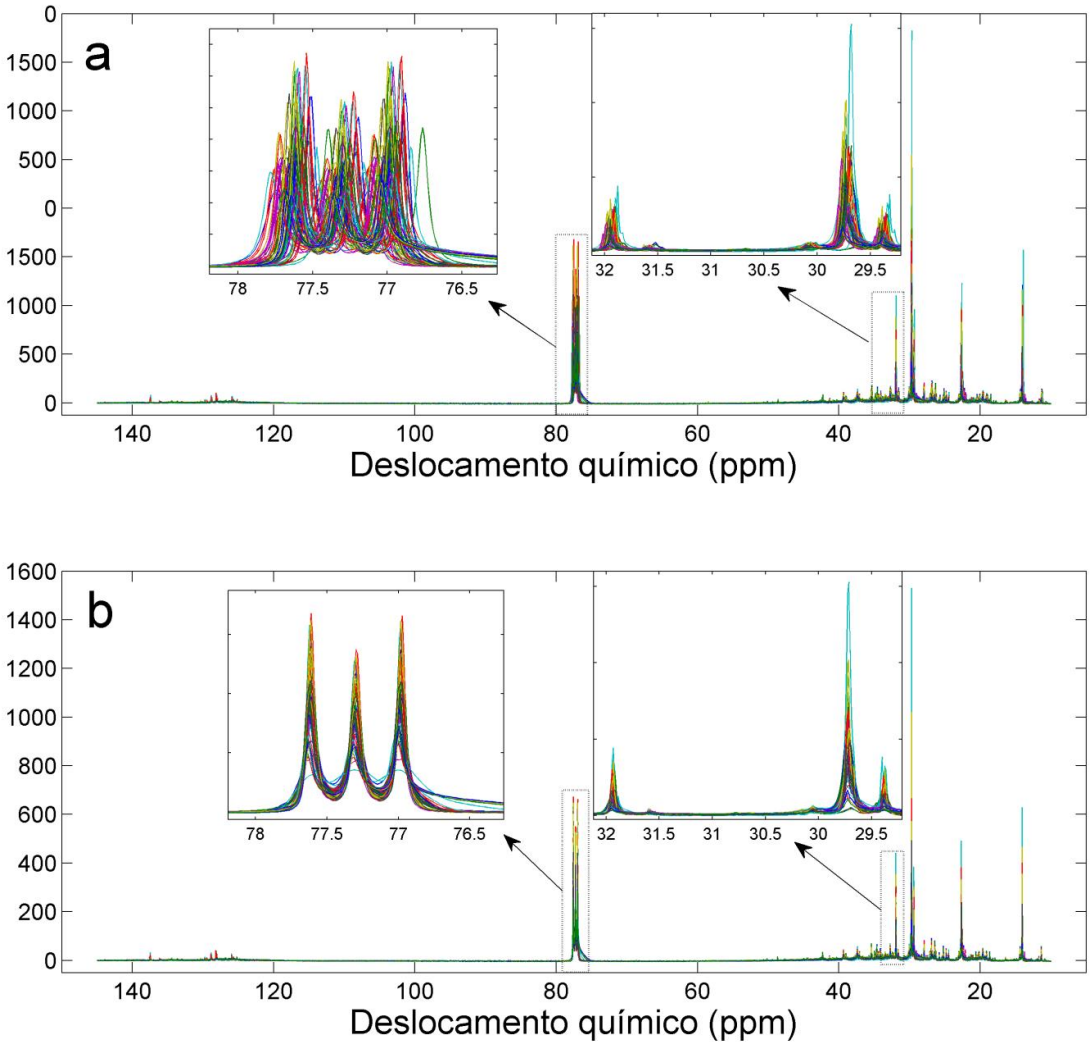

Figura 5.5. Espetros de RMN de ^{13}C de petróleos brutos originais (a) e após alinhamento utilizando o programa icoshift (b).

Um dos problemas na utilização do GA é sua convergência para mínimos locais. A cada modelo construído, diferentes variáveis espectrais e valores dos parâmetros SVR podem ser obtidos. Para minimizar este efeito, vários modelos GA são construídos e as variáveis selecionadas com maior frequência foram consideradas as mais importantes para o modelo de calibração. Os 100 modelos GA-SVR construídos selecionam variáveis em aproximadamente toda faixa de deslocamento químico estudado. Desta forma o modelo GA-SVR final deve ser

obtido com as variáveis mais frequentemente selecionadas. A Figura 5.6 mostra para cada parâmetro do petróleo calibrado o número de variáveis espectrais que foram selecionadas com maiores frequência: seleção de 30% a 80%, ou seja, variáveis que foram selecionadas em pelo menos 30 a 80 modelos GA. Assim, para cada parâmetro um diferente número de variáveis espectrais foram selecionados.

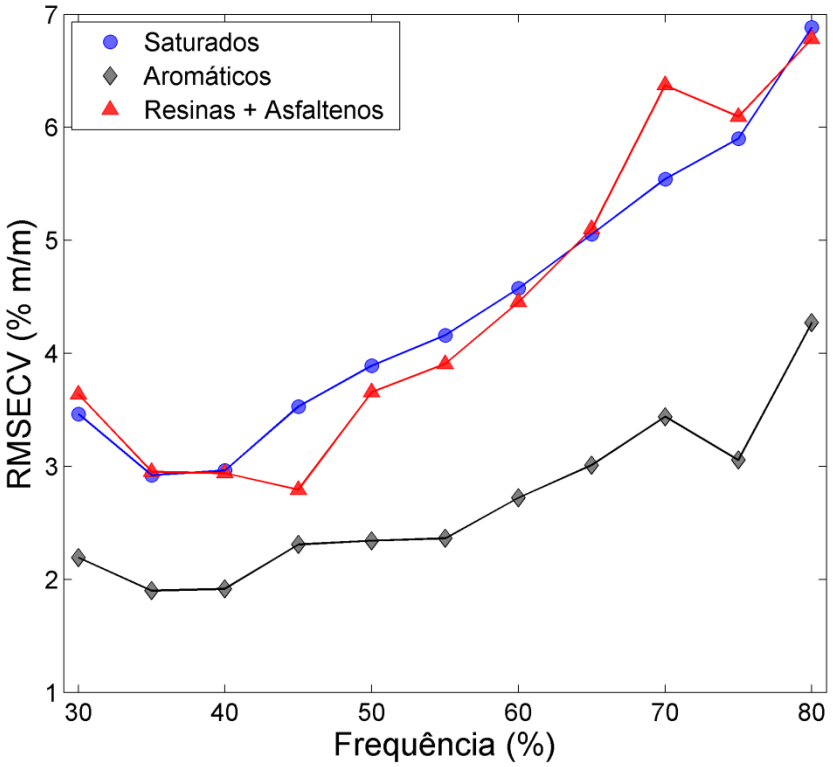

Figura 5.6. Gráfico do RMSECV em função da frequência de seleção do número de variáveis espectrais obtidas com execução de 100 modelos GA-SVR.

Em cada um dos 3 modelo de calibração construídos, foi verificado o número ótimo de variáveis espectrais (deslocamento químico) para minimização do erro de validação cruzada (Figura 5.6). No modelo de calibração para saturados e aromáticos, um RMSECV mínimo foi obtido quando utilizamos variáveis espectrais que foram selecionadas em pelo menos 35 dos 100 modelos GA construídos. Esta mesma frequência de seleção da variável não implica que os dois modelos terão o mesmo número de variáveis. O modelo de calibração para

resinas mais asfaltenos obteve um mínimo com frequência de seleção de pelo menos em 40 ou 45 modelos.

Em cada uma das 100 otimizações realizadas pelo GA, um diferente conjunto de parâmetros do modelo SVR é encontrado. O conjunto de valores: C, e, γ e ε utilizados no modelo GA-SVR final foi escolhido com base na mediana da distribuição (Figura 5.7).

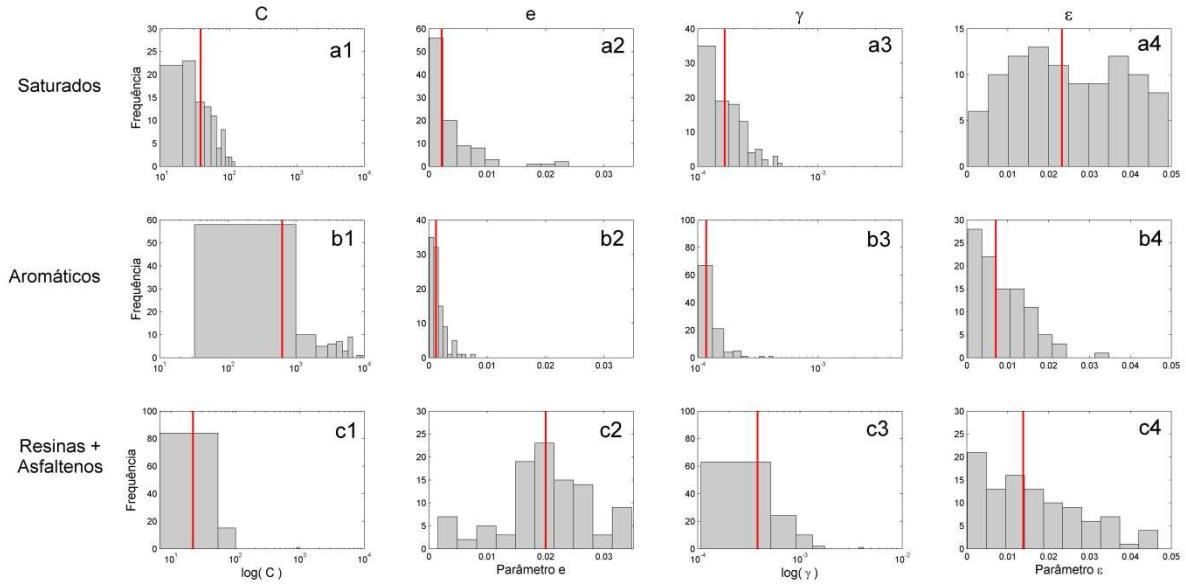

Figura 5.7. Histograma dos parâmetros do modelo SVR otimizados pelo algoritmo genético na determinação do teor de saturados, aromáticos e resinas mais asfaltenos. A linha vertical vermelha é o valor mediano da distribuição.

A região final do espectro de RMN de ^{13}C, apesar de apresentar baixa intensidade, tem grande importância para modelagem do teor de saturados e aromáticos, enquanto que o modelo para o teor de resinas e asfaltenos, as regiões iniciais e finais do espectro apresentam a mesma importância (Figura 5.8). Das 5530 variáveis iniciais o modelo GA-SVR final para saturados foi construído com 331 variáveis, sendo 132 referentes a região inicial do espectro e 199 referente a parte final (Figura 5.8b). O modelo GA-SVR para aromáticos foi construído com 363 variáveis sendo: 214 da região inicial e 149 da região final do espectro (Figura 5.8c). O modelo GA-SVR para resinas mais asfaltenos foi construído com 153

variáveis sendo: 84 da parte inicial e 69 da parte final do espectro de RMN de ^{13}C (Figura 5.8d).

Na Figura 5.8a é destacada a região onde há predominância de carbonos em grupos aromáticos e carbonos ligados a heteroátomos, como nitrogênio e oxigênio. Os deslocamentos químicas em 137 ppm, 129 ppm e 128 ppm podem ser atribuídos a carbonos aromáticos substituídos por metila, carbono aromático em junção de 3 anéis aromáticos e junção de 2 anéis aromáticos, respectivamente.[88] A região pouco abaixo de 80 ppm deve-se ao solvente $CDCl_3$. A região até 40 ppm é característica de grupos de carbonos alifáticos que estão relacionados aos saturados.[49] Esperava-se que a região até 40 ppm tivesse maior contribuição na modelagem dos saturados (Figura 5.8b) por estar relacionado a carbonos de grupos alifáticos, que constituem a fração de saturados. Entretanto, é observada uma maior contribuição da região com predominância de compostos aromáticos. Este fato pode estar relacionado a redução de compostos constituintes no petróleo. Apesar da menor intensidade relativa da região de aromáticos, durante a modelagem SVR as variáveis são padronizadas no intervalo entre 0 e 1, isso faz com que na entrada do algoritmo todas as variáveis espectrais tenham a mesma importância. Assim, as regiões do espectro selecionadas pelo GA estão relacionadas diretamente com a propriedade de interesse.

As variáveis selecionadas para a modelagem de aromáticos (Figura 5.8c) apresentam predominância da região relacionada aos aromáticos no espectro de RMN de ^{13}C como era esperado. Contudo, a modelagem de resinas mais asfaltenos também selecionou, em sua maioria, variáveis desta região (Figura 5.8d). Isto pode estar relacionado a predominância de heteroátomos associados a compostos de alta massa molar, que no caso do petróleo, corresponde a grupos aromáticos que constituem as resinas e asfaltenos.

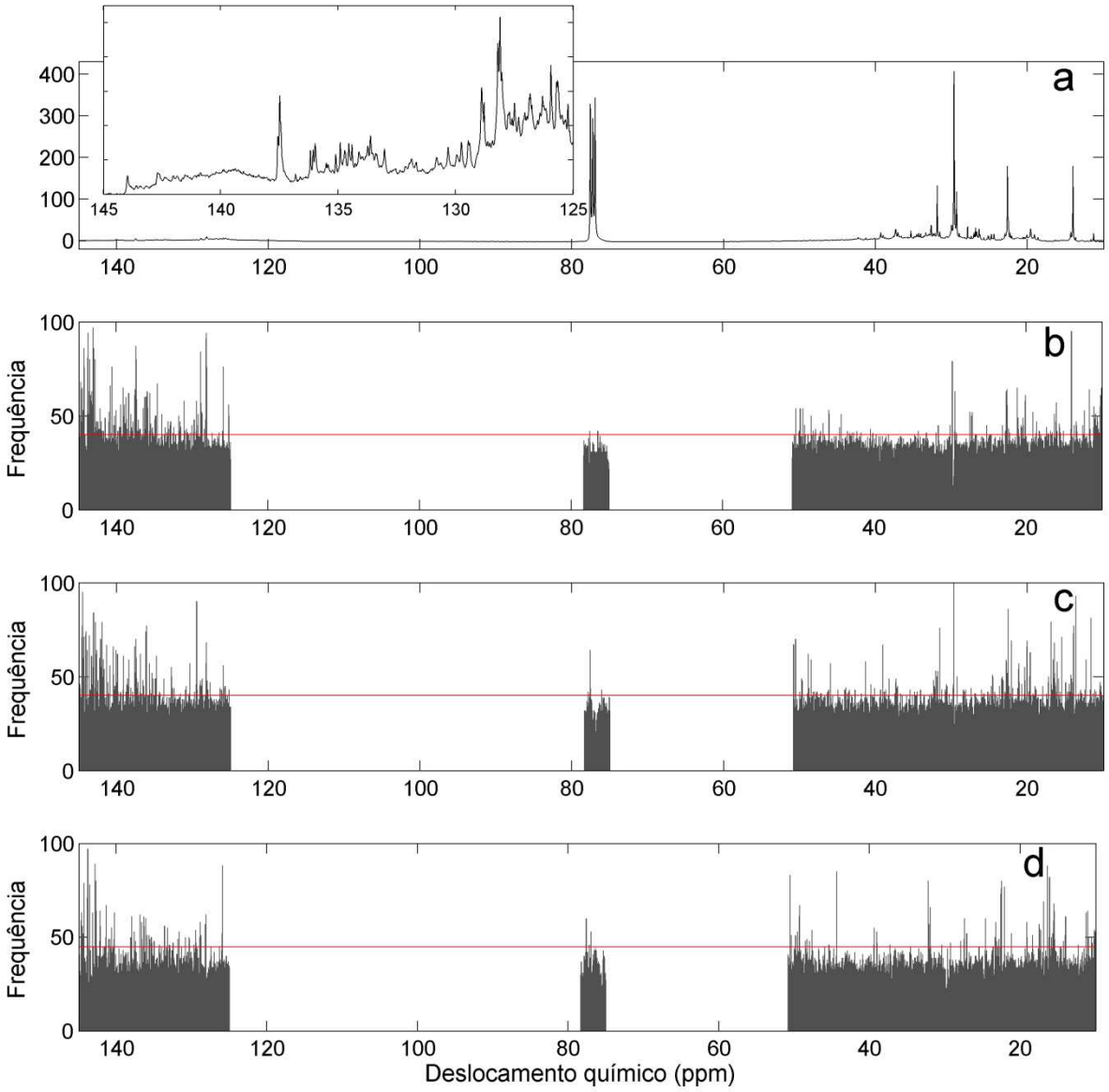

Figura 5.8. (a) espectro médio de RMN de C13 das amostras de calibração com região de predominância de carbonos aromáticos destacada à direita. Frequência das variáveis selecionadas para saturados (b), aromáticos (c) e resinas e asfaltenos (d). A linha horizontal em vermelho representa a frequência mínima de seleção para construção do modelo GA-SVR.

O modelo de calibração para teor de saturados apresenta erros de 3,0% m/m para validação cruzada e 4,4% m/m para previsão (Tabela 5.1). Apesar da diferença nos valores, estes não apresentam grandes diferenças quando se

observa o gráfico dos valores previstos pelo modelo em função dos valores de referência (Figura 5.9a). O menor RMSECV reflete um maior coeficiente de determinação para este conjunto, 0,9444 contra 0,9078 para o conjunto de previsão. Na modelagem do teor de aromáticos, uma diferença significativa é observada entre os erros de validação cruzada (1,9% m/m) e previsão (4,3% m/m). Devido ao menor teor de aromáticos nos petróleos trabalhados em comparação aos saturados, estes erros, apesar de apresentarem mesma magnitude que para a modelagem de saturados, refletem um menor coeficiente de determinação, R^2cv de 0,8974 e R^2p de 0,7298 (Tabela 5.1). Na Figura 5.9b, observa-se um ajuste inferior do modelo aos dados de referência, quando comparado à modelagem do teor de saturados (Figura 5.9a).

Tabela 5.1. Resultados dos parâmetros estatísticos dos modelos GA-SVR.

Parâmetro	Saturados	Aromáticos	Resinas + Asfaltenos
RMSECV (% m/m)	3,0	1,9	2,8
R^2cv	0,9444	0,8974	0,8990
RMSEP (% m/m)	4,4	4,3	3,7
R^2p	0,9078	0,7298	0,7737

Na modelagem do teor de resinas mais asfaltenos os resultados se assemelham à calibração do teor de aromáticos, com coeficientes de determinação de 0,8990 para calibração cruzada e 0,7737 para previsão (Tabela 5.1). Os erros para este modelo estão na mesma magnitude que para as duas calibrações anteriores, com 2,8% m/m e 3,7% m/m para validação cruzada e previsão, respectivamente.

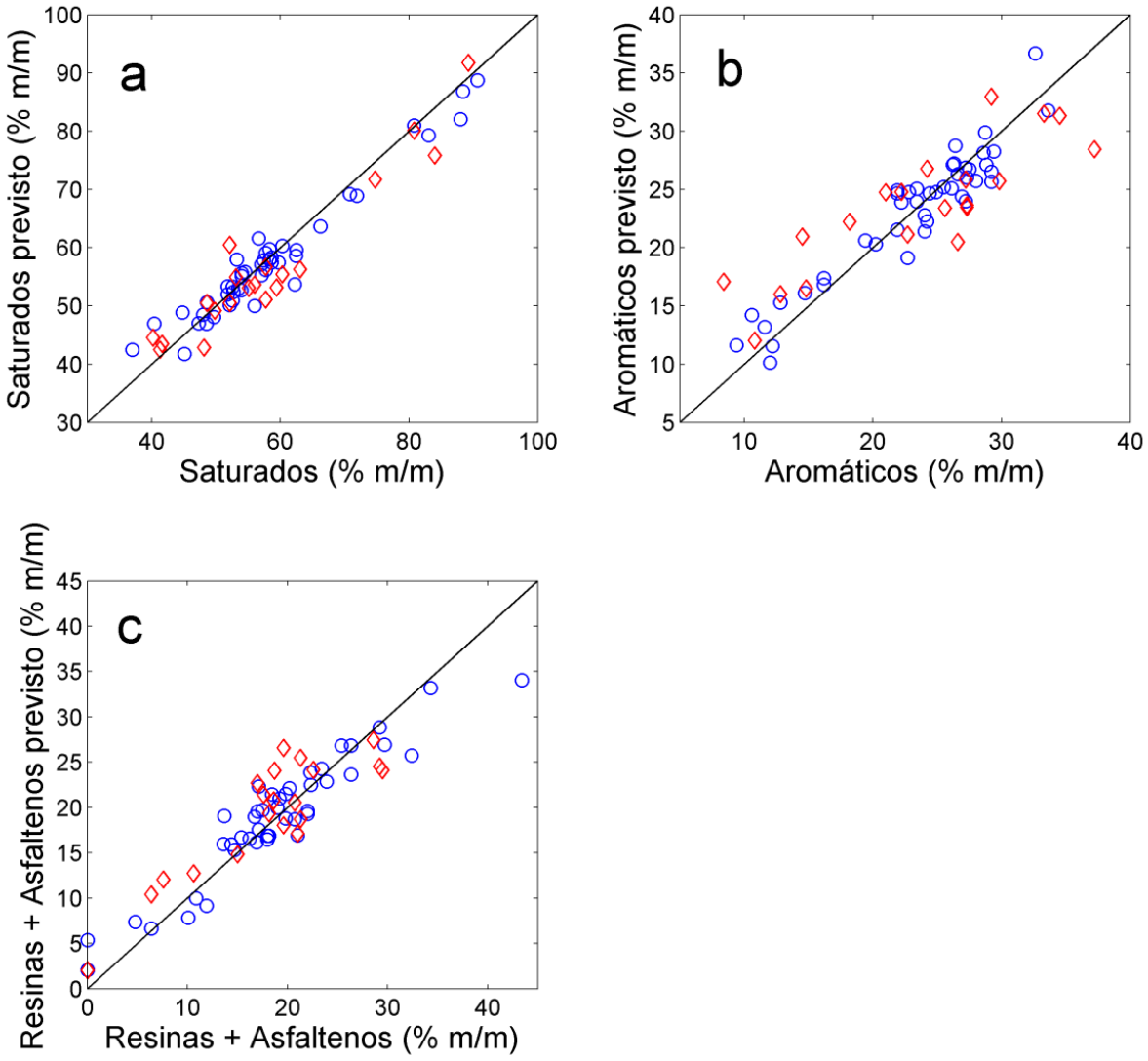

Figura 5.9. Gráfico dos teores de SARA medidos segundo a norma ASTM D 7169 pelos valores de estimados pelo modelo GA-SVR. (a) Saturados, (b) Aromáticos, (c) Resinas e Asfaltenos. Amostras de calibração (o) e previsão (◊).

Os resultados dos 3 modelos GA-SVR obtidos apresentam resultados aceitáveis para a indústria do petróleo, tendo em vista que os RMSEP's obtidos são menores de 5% m/m, que é o valor máximo de variação permitido na análise de SARA na indústria do petróleo. Outro ponto positivo é a boa relação linear obtida pelos modelos, com R^2 acima de 0,7 para o conjunto de teste.

5.5 Conclusões

Os erros de previsão para calibração do teor de saturados, aromáticos e resinas mais asfaltenos apresentam valores de aproximadamente 4% m/m. A propriedade SARA é muito importante para o monitoramento da qualidade de petróleo e poucos métodos são descritos na literatura capazes de realizar essa determinação usando poucas quantidades de amostras num tempo relativamente baixo (aproximadamente 2 horas). Portanto, a magnitude dos erros obtidos pelos modelos de calibração é satisfatória por estarem abaixo da margem aceita pela indústria do petróleo (5% m/m).

O algoritmo genético apresenta resultados satisfatórios para realizar simultaneamente a seleção de variáveis espectrais e otimização dos parâmetros do modelo de regressão por vetores de suporte. Porém, na aplicação do algoritmo genético foi necessário a otimização de vários modelos (100 nesse caso), para evitar a convergência do algoritmo para mínimos locais, e posterior combinação dos resultados para uma melhor seleção das variáveis.

6 Aplicação 4: Quantificação de biodiesel de gordura animal em biodiesel de soja e diesel B20 usando espectroscopia no infravermelho próximo e regressão por vetores de suporte com sinergismo de intervalos

6.1 Introdução

O biodiesel é atualmente o principal substituto do óleo diesel de petróleo, devido às suas características similares; ele tem sido utilizado em muitos países como combustível para motores diesel, geralmente em misturas de até 20% (v/v) em óleo diesel, o qual é também conhecido como B20.[5,89] O biodiesel é constituído por uma mistura de alquil ésteres de cadeia linear, obtida pela transesterificação de triglicerídeos de óleos vegetais e/ou gorduras animal com alcoóis de cadeia curta (metanol ou etanol).[90] A matéria prima utilizada para sua fabricação determina as estruturas moleculares de seus ésteres constituintes e impactam em sua qualidade. Algumas propriedades relacionadas à baixa temperatura (ponto de névoa, ponto de entupimento de filtro a frio e ponto de fluidez) de misturas biodiesel/diesel variam de acordo com a matéria-prima utilizada na produção do biodiesel. Variações nestas propriedades podem ser críticas, dependendo da região e temperatura climática de uso do combustível.

A reação de transesterificação de óleos vegetais ou gordura animal com metanol produz mistura de ésteres metílicos. Dependendo da matéria-prima utilizada, esta contém diferentes quantidades de ésteres metílicos de longa cadeia de carbonos saturados, que tendem a aumentar o ponto de névoa, ponto de entupimento de filtro a frio e ponto de fluidez.[91] Quanto maior o tamanho da cadeia e/ou o quantidade de saturados do biodiesel, maiores serão as temperaturas para estas propriedades. Como a gordura animal, em relação ao óleo vegetal, tem relativamente um maior teor de ácidos graxos saturados, esta matéria-prima do biodiesel possui maiores quantidades relativas de ésteres de cadeia longa de carbonos saturados e consequentemente maiores temperaturas de ponto de névoa, ponto de entupimento de filtro a frio e ponto de fluidez. Desta forma, a

gordura animal utilizada para produzir biodiesel determina suas propriedades de baixa qualidade, limitando o uso de gordura animal para biodiesel em climas relativamente frios.[91]

As baixas temperaturas para o ponto de névoa, ponto de entupimento de filtro a frio e ponto de fluidez são problemas importantes na utilização do diesel combustível com um elevado teor de biodiesel, tal como B20, a fim de evitar falhas no motor incluindo dificuldade de ignição e entupimento dos filtros de combustível. No Brasil, o biodiesel produzido usando óleo de soja e gordura animal como matérias-primas representaram aproximadamente 90% da produção total (ano de 2012). O uso de gordura animal como matéria-prima vem aumentado devido a vários aspectos atraentes, como menor custo e disponibilidade abundante.[92] Normalmente, misturas de biodiesel de soja com gordura animal são comercialmente disponíveis e utilizadas para produção do diesel B20 combustível, mas dependendo da proporção dos biodieseis mistura e as temperaturas operacionais, estas misturas podem causar falhas de funcionamento do motor. Desta forma, o desenvolvimento de métodos analíticos práticos e confiáveis para determinação do teor de biodiesel de gordura animal em biodiesel de soja ou diesel B20 pode facilitar o controle de qualidade do combustível e evitar falhas operacionais relacionados à qualidade do combustível em veículos com motores a diesel.

Devido à expansão da produção de biodiesel, vários estudos tem sido desenvolvidos para determinar seus parâmetros de qualidade,[14-18,91,92] seja em misturas de biodiesel ou misturas diesel-biodiesel.[8,65,93-96] Métodos envolvendo espectroscopia no infravermelho próximo (NIR), associada à regressão por mínimos quadrados parciais (PLS) vem mostrado grande potencial e são os preferíveis. No entanto, outras técnicas de regressão, como a regressão de vetores de suporte (SVR), podem gerar modelos mais robustos devido a sua capacidade de generalização e modelagem de problemas não lineares.[31,41,97]

Considerando o grande número de variáveis obtidas em espectros NIR, técnicas de seleção de variáveis baseadas em intervalos, podem melhorar o desempenho do modelo de calibração em relação às técnicas que empregam

espectro completo. Norgaard *et al*.[98] propuseram um método determinístico de seleção de regiões espectrais em modelos PLS, em que o espectro é dividido em *i* intervalos equidistantes. Uma combinação dos intervalos pode então ser realizada para conduzir a melhores previsões e gerar modelos mais parcimoniosos. Este método foi chamado de sinergismo de intervalos ou siPLS. A mesma idéia de seleção de variáveis por sinergismo de intervalos pode ser utilizada no desenvolvimento de modelos SVR numa abordagem chamada siSVR, em que a combinação de regiões espectrais pode melhorar a exatidão do modelo em relação ao método com o espectro completo.

6.1.2 Objetivo

Devido a grande importância do controle de qualidade de combustíveis contendo biodiesel de gordura animal, este trabalho tem como objetivo desenvolver um método de quantificação do biodiesel de gordura animal em mistura com biodiesel de soja e no diesel B20, utilizando calibração multivariada PLS e SVR baseada em espectro completo e abordagens com seleção de variáveis por sinergia de intervalos aplicados em espectros NIR.

6.2 Metodologia

Foram preparadas 99 amostras de misturas binárias de biodiesel de gordura animal em biodiesel de soja, com concentração de biodiesel de gordura animal variando de 0,00 a 69,00% m/m. Numa segunda etapa, estas mesmas amostras foram misturadas com diesel de petróleo S50 (diesel contendo no máximo 50 mgL^{-1} em massa, de enxofre) gerando outras 99 amostras de diesel B20. Para esta segunda etapa, o teor de biodiesel de gordura animal no B20 varia de 0,00 a 13,80% m/m.

Espectros na região do infravermelho próximo foram medidos em um espectrômetro Perkin Elmer 100N equipado com acessório de transflectância. Foi

utilizado o intervalo espectral de 9.000–4.000 cm^{-1}, com resolução de 4 cm^{-1} e 32 leituras por amostra.

Os modelos de calibração foram construídos separadamente para a mistura de biodieseis (biodiesel de gordura animal em biodiesel de soja) e no diesel B20. Para construção dos modelos de calibração foram empregadas 59 amostras com as 40 amostras restantes sendo utilizadas para previsão. Os espectros NIR foram pré-processados usando a correção multiplicativa de sinal (MSC).[58]

Na construção dos modelos com PLS a matriz de espectros foi centrada na média e o vetor de resposta contendo o teor de biodiesel de gordura animal, autoescalado. Na construção do modelo SVR a matriz de espectros e o vetor contendo a propriedade de interesse foram padronizados no intervalo entre 0 e 1. A função kernel RBF foi utilizada para mapeamento dos dados em alta dimensão.

Os modelos foram otimizados pelo procedimento de validação cruzada "7-*fold*". Para os modelos PLS e siPLS foi otimizado o número de variáveis latentes e para os modelos SVR e siSVR uma grade de pesquisa foi aplicada para otimizar os parâmetros C, ν e γ.

Modelos siPLS foram construídos nas mesmas condições que os modelos PLS, empregando o pacote iToolbox (disponível para download no site: http://www.models.life.ku.dk/ipls). Da mesma forma, os modelos siSVR foram construídos nas mesmas condições que o modelo SVR. Porém, suas rotinas foram desenvolvidas no próprio laboratório. Elaborou-se modelos com o espectro completo dividido em 10 intervalos e combinados dois a dois. Em todos os modelos, testes para erros sistemáticos e tendência foram aplicados e as exatidões entre modelos para um mesmo conjunto de dados foram comparadas por teste randômico para exatidão.

6.3 Resultados e discussões

A Figura 1 apresenta o espectro de NIR de biodiesel gorduras animais, com 31% m/m de biodiesel de soja, o biodiesel de soja puro e óleo diesel B20 com 20% m/m de biodiesel de soja, dividido em 10 intervalos equidistantes utilizados na construção de siPLS e modelos siSVR. Os números 1 a 10 referem-se ao número de intervalo.

No intervalo 10 é observada uma banda de absorção em torno de 4069 cm^{-1} com maior intensidade no diesel B20. Esta região é relativa à deformação C-H de cadeia alifática linear associada ao menor número de insaturações do biodiesel. Outra diferença pode ser observada no intervalo 9, em aproximadamente 4,662 cm^{-1}. Esta banda relacionada ao estiramento C-H, e combinação de estiramento C=O e C-H, com maior intensidade para o biodiesel de soja. No intervalo 7 é observada uma banda de absorção em aproximadamente 5865 cm^{-1} relacionada com o segundo harmônico de CH de grupo metil terminal, com menor intensidade para o diesel B20.

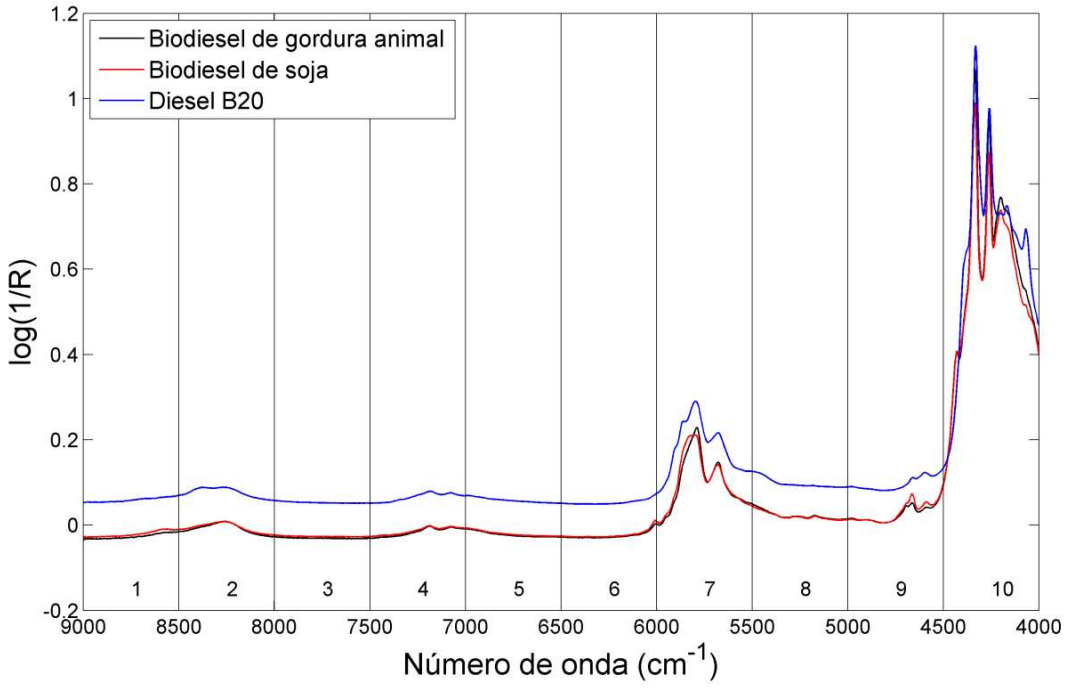

Figura 6.1. Espectro NIR de biodiesel de gordura animal com 31% m/m de biodiesel de soja (—), biodiesel de soja (—) e diesel B20 com 20% m/m de biodiesel de soja (—), dividido em 10 intervalos.

6.3.1 Quantificação de biodiesel de gordura animal em misturas com biodiesel de soja

A rápida determinação do teor de biodiesel de gordura animal biodiesel em mistura com biodiesel de soja é importante para controle de matérias-primas de biodiesel.

Os resultados dos parâmetros estatísticos dos modelos PLS, SVR, siPLS e siSVR são apresentados na Tabela 6.1. Nenhum dos modelos construídos apresentou evidências de erros sistemáticos ao nível de confiança de 95%. Os melhores resultados foram obtidos utilizando seleção de variáveis, com um RMSEP de 0,25% m/m e 0,18% m/m para os siPLS e modelos siSVR, respectivamente. Uma redução de mais de 50% no erro de previsão foi alcançada apenas selecionando regiões espectrais, demonstrando a importância em escolher adequadamente intervalos espectrais. O teste randômico para exatidão aplicado aos resíduos dos modelos siPLS e siSVR indicou haver diferença significativa, ao nível de confiança de 95%, nos resíduos dos dois modelos, com o modelo siSVR sendo mais exato.

Tabela 6.1. Resultados dos modelos de calibração para quantificação do teor de biodiesel de gordura animal em mistura com biodiesel de soja.

Modelo	Intervalo	Calibração			Previsão		
		RMSECV (%m/m)	Viés (%m/m)	R^2cv	RMSEP (%m/m)	Viés (%m/m)	R^2p
PLS (3)[a]	Todos	0,84	-0,022	0,9983	0,82	0,002	0,9983
siPLS (7)[a]	7 e 8	0,25	-0,013	0,9998	0,25	0,007	0,9999
SVR	Todos	0,46	-0,031	0,9995	0,45	-0,094	0,9995
siSVR	7 e 9	0,20	-0,023	0,9999	0,18	-0,056	0,9999

[a] Número de variáveis latentes utilizadas no modelo PLS.

O bom ajuste dos resultados previstos pelos modelos aos valores de referência é observado nos coeficientes de determinação acima de 0,99. Devido a

este ajuste, na Figura 6.2 são apresentados apenas os resíduos dos modelos de calibração em função dos valores de referência. Observa-se menores valores nos resíduos dos modelos com seleção de variáveis. Outro ponto importante está nos resíduos dos modelos PLS e siPLS, onde parece haver uma tendência quadrática, que subestima resultados do início e fim da curva, enquanto que na região em torno da média central, os resultados são superestimados. Um ajuste quadrático aos dados é mostrado em linha sólida preta. Em modelos corretamente ajustados espera-se resíduos aleatoriamente distribuídos em torno de zero. Para verificar esta tendência quadrática foi aplicado o teste de permutação.

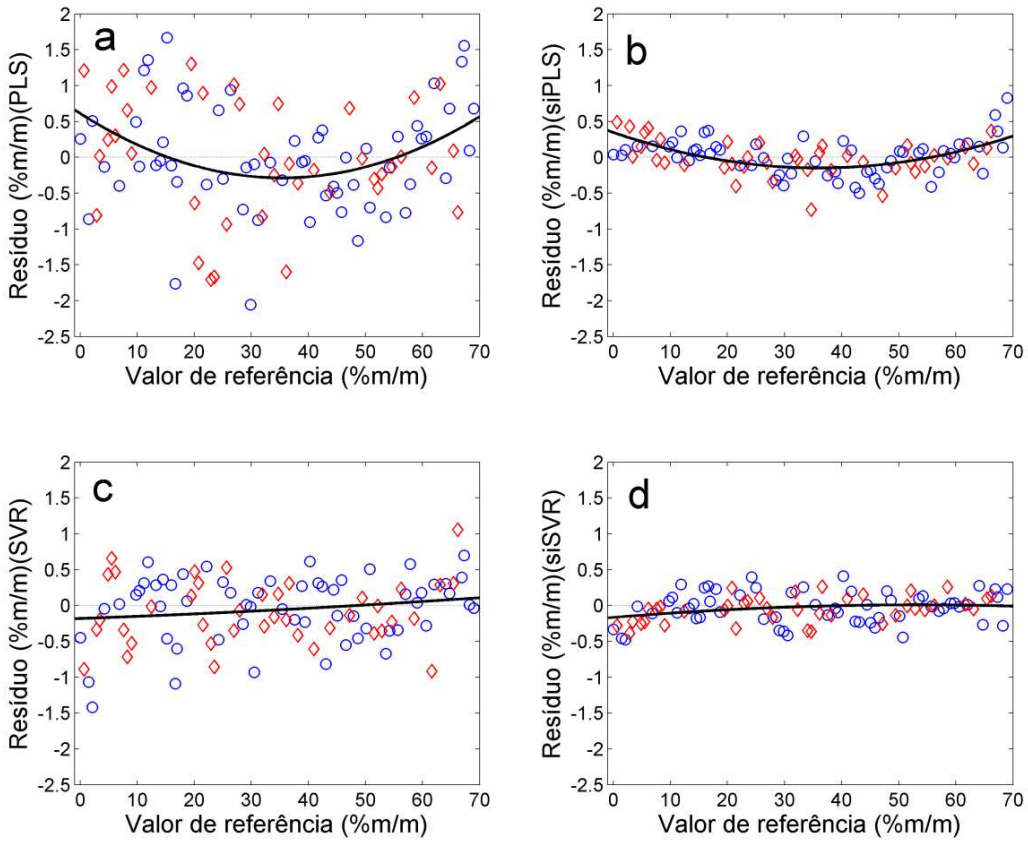

Figura 6.2. Gráfico dos resíduos dos modelos de calibração em função do teor de biodiesel de gordura animal em biodiesel de soja com ajuste quadrático aplicado aos dados (—). Modelo PLS (a), siPLS (b), SVR (c) e siSVR (d). Amostras de calibração (o) e previsão (◊).

Os histogramas com as distribuições dos coeficientes quadráticos permutados (100.000 permutações) são mostrados na Figura 6.3. A linha vermelha sólida na vertical representa o valor do coeficiente quadrático calculado para os resíduos originais. O p-valor do teste é determinado pela proporção da área à direita (porque o coeficiente quadrático obtido foi positivo) da linha vertical vermelha e a área total. Quanto maior o p-valor, menor a probabilidade da tendência observada ser devido ao acaso. Ao nível de significância de 5%, podemos afirmar que os modelos PLS (p-valor de 0,0389) e siPLS (p-valor de 0,0001) apresentam evidências de tendências quadrática em seus resíduos, enquanto os modelos SVR (p-valor de 0,2128) e siSVR (p-valor de 0,2444) não apresentam tendência nos resíduos. A presença de tendência nos resíduos de previsão é indício que informação importante não foi modelada e ainda estão presentes nos resíduos. O modelo siSVR não apresentou a mesma tendência que o modelo siPLS por modelar de forma mais eficiente informações não lineares relacionadas a propriedade de interesse.

O modelo siSVR selecionou os intervalos 7 e 9. O intervalo 7 tem duas bandas de absorção de interesse: 5.797 cm^{-1} referente ao primeiro sobretom de estiramento C-H de hidrocarboneto e 5.675 cm^{-1} referente ao primeiro sobretom de estiramento simétrico C-H de hidrocarboneto alifático.[99] O intervalo 9 contém uma banda de absorção em 4.662 cm^{-1} relacionada ao estiramento C-H, e combinação de estiramento C=O e C-H. O modelo siPLS selecionou o intervalo 7 assim como o modelo siSVR; entretanto, o intervalo 8 também foi selecionado. Nesta região aparentemente não apresenta bandas de interesse no espectro (Figura 6.1).

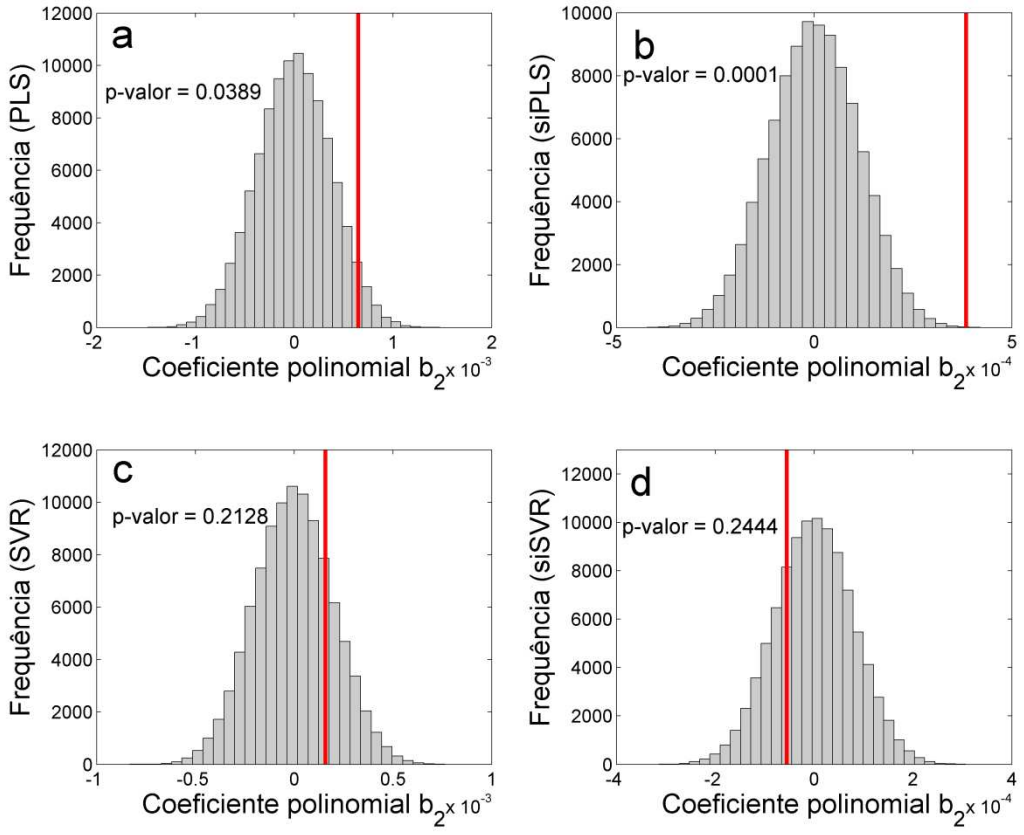

Figura 6.3. histograma dos coeficientes quadráticos do teste de permutação dos resíduos dos modelos: PLS (a), siPLS (b), SVR (c) e siSVR (d). A linha vermelha vertical sólida refere-se ao coeficiente quadrático ajustado aos dados.

6.3.2 Quantificação de biodiesel de gordura animal no diesel B20

Os modelos de calibração mostraram boa relação linear com os valores previstos (Tabela 6.2), apresentando coeficientes de determinação acima de 0,9, com exceção para o modelo PLS. Observa-se que, mesmo com a mudança da matriz, com adição de diesel de petróleo, o modelo siPLS e siSVR selecionaram os mesmo intervalos que na modelagem em mistura de biodieseis (ver Tabela 6.1). No entanto, o teste-t indicou haver evidências de erros sistemáticos, ao nível de confiança de 5%, nos resíduos dos modelos PLS e SVR

Tabela 6.2. Resultados dos modelos de calibração para quantificação do teor de biodiesel de gordura animal no diesel B20.

Modelo	Intervalo	Calibração			Previsão		
		RMSECV (%m/m)	Viés (%m/m)	R^2cv	RMSEP (%m/m)	Viés (%m/m)	R^2p
PLS (4)[a]	Todos	0,53	-0,0057	0,9829	0,55	0,0999	0,9817
siPLS (5)[a]	7 e 8	0,16	-0,0015	0,9984	0,15	0,0390	0,9986
SVR	Todos	0,36	-0,0179	0,9921	0,29	-0,1010	0,9953
siSVR	7 e 9	0,16	-0,0162	0,9984	0,10	-0,0213	0,9994

[a] Número de variáveis latentes utilizadas no modelo PLS.

Os resíduos de calibração e de previsão em função do teor de biodiesel de gordura animal no diesel B20 para os 4 modelos são mostrados na Figura 6.4. Novamente observa-se menores resíduos para os modelos com seleção de variáveis, com RMSEP de 0,15% m/m e 0,10% m/m para os modelos siPLS e siSVR, respectivamente. O teste randômico para exatidão aplicado aos resíduos dos modelos siPLS e siSVR indicou haver diferença significativa, ao nível de confiança de 95%, nos resíduos do dois modelos, com o modelo siSVR sendo mais exato.

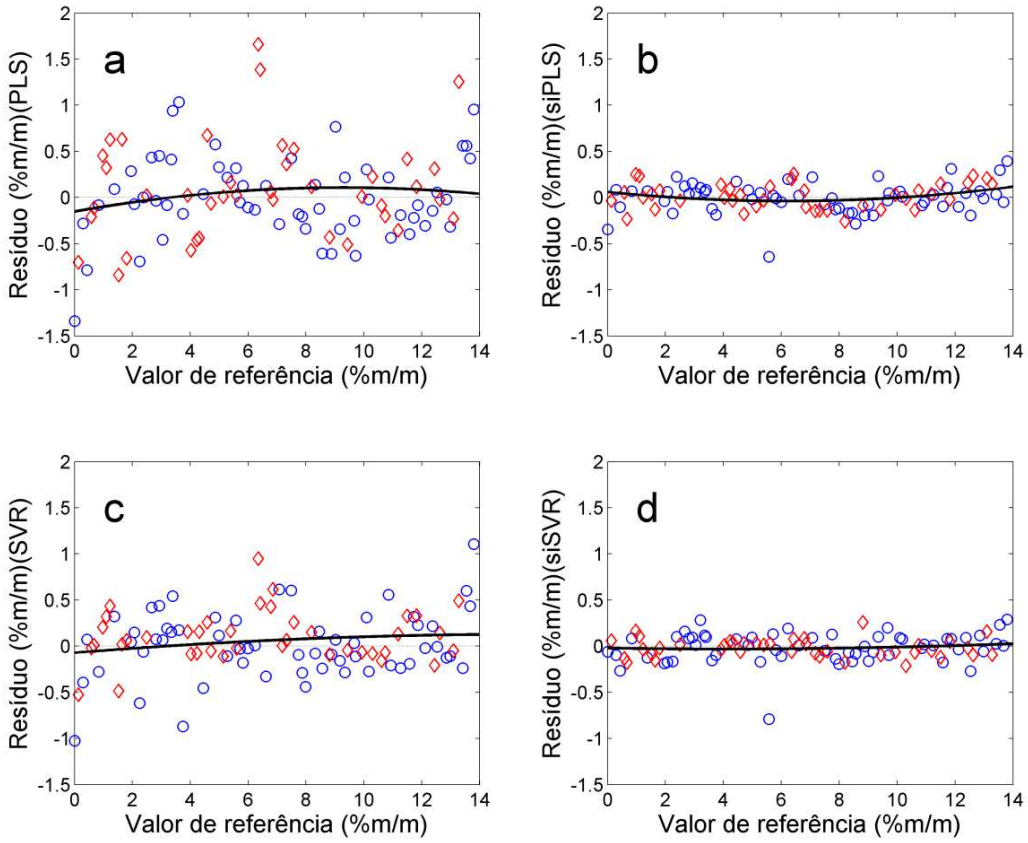

Figura 6.4. Gráfico dos resíduos dos modelos de calibração em função do teor de biodiesel de gordura animal no diesel B20 com ajuste quadrático aplicado aos dados (–). Modelo PLS (a), siPLS (b), SVR (c) e siSVR (d). Amostras de calibração (o) e previsão (◊).

Os resultados de previsão do modelo siPLS parecem apresentar as mesmas tendências observadas na modelagem em mistura de biodieseis. Aplicado o teste de permutação para avaliar tendências, esta tendência pode ser confirmada (Figura 6.5), porém, menos notória que na modelagem em mistura de biodieseis.

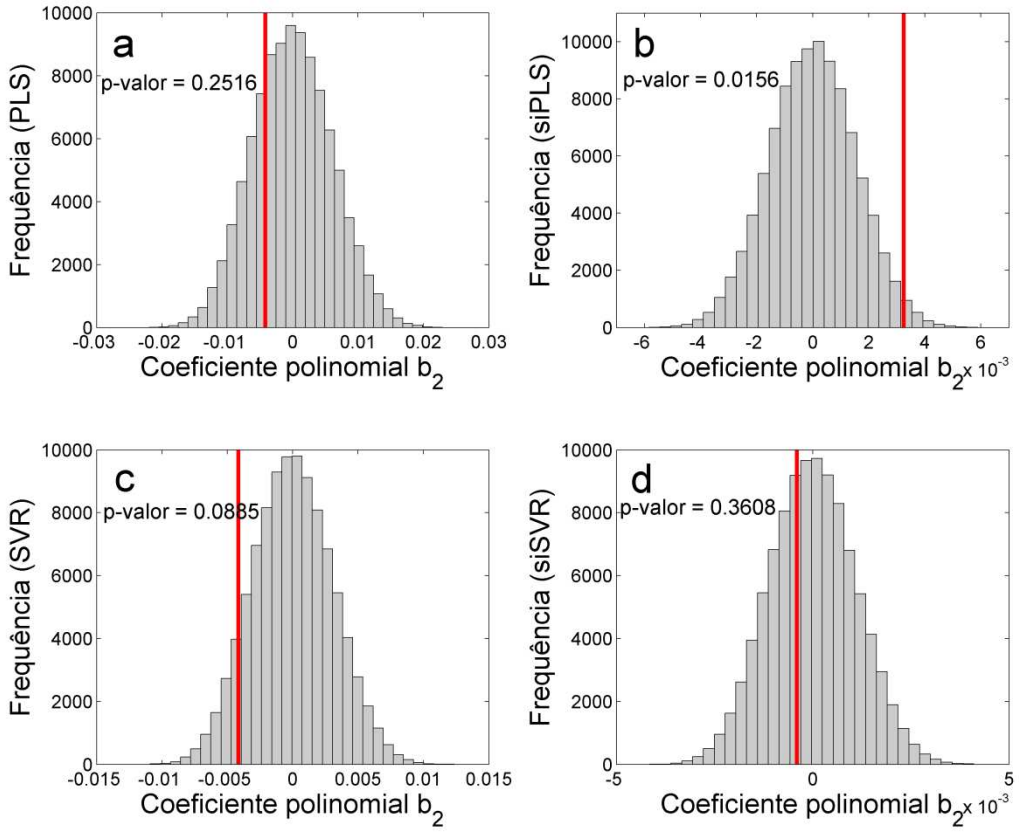

Figura 6.5. histograma dos coeficientes quadráticos do teste de permutação dos resíduos dos modelos: PLS (a), siPLS (b), SVR (c) e siSVR (d). A linha vermelha vertical sólida refere-se ao coeficiente quadrático ajustado aos dados.

6.4 Conclusões

A espectroscopia NIR e a calibração multivariada baseada em vetores de suporte combinado com seleção de variáveis por sinergismos de intervalos (siSVR) pode gerar uma metodologia adequada para quantificação do teor de biodiesel de gordura de animal em misturas com biodiesel de soja ou em diesel B20.

O método linear siPLS produziu modelos com tendência quadrática nos resíduos enquanto que o método siSVR produziu modelos com melhor exatidão e sem presença de tendência nos resíduos de previsão. O método siSVR seleciona melhor os intervalos espectrais que o modelo siPLS. Os resultados apresentados nessa aplicação foram publicados em um artigo no revista Talanta.[100]

7 Conclusões gerais

O método de regressão por vetores de suporte apresenta resultados semelhantes ou superiores ao método calibração multivariada PLS, que é amplamente aplicado em química analítica. O método SVR é indicado sempre quando os resultados do modelo PLS forem insatisfatórios ou apresentarem resíduos sistemáticos ou tendenciosos.

Em problemas envolvendo matrizes complexas, como petróleo, o método SVR é bastante indicado, pois estes sistemas normalmente não apresentam dependência linear dos espectros com a propriedade de interesse. Em meio relativamente mais simples como mistura de biodieseis ou diesel/biodiesel, o SVR também tem sua aplicação recomendada por gerar modelo sem erros de tendência e com boa exatidão. Um ponto crítico na modelagem SVR está na otimização dos parâmetros do modelo. O uso de grade de pesquisa e algoritmo genético são alternativas eficazes para construção de modelos mais exatos.

As figuras de mérito para o modelo SVR carecem de mais estudos, entretanto a metodologia *boosting* ensemble mostra-se eficaz para a determinação do intervalo de confiança na previsão do modelo. Este procedimento também auxilia na detecção de amostras anômalas e possibilita a modelagem nesta condição. O teste de permutação não paramétrico é uma excelente ferramenta para detecção da presença de erros de tendência em resíduos de modelos de regressão multivariada.

A seleção de variáveis pode melhorar a exatidão e gerar modelos mais parcimoniosos. Variáveis de espectros de infravermelho (NIR e MIR) podem ser consideradas contínuas, assim, métodos determinísticos de seleção de variáveis como o sinergismo de intervalos são recomendados. Em espectroscopia de RMN, os espectros tem caráter discreto, assim, métodos probabilísticos como o algoritmo genético podem gerar melhores resultados.

A técnica de ATR-FTIR associada à regressão por vetores de suporte foi eficiente para determinar a gravidade API, viscosidade cinemática e teor de água em petróleos. A espectroscopia NIR aliada à regressão por vetores de suporte

com sinergismos de intervalos mostrou-se uma metodologia adequada para quantificação do teor de biodiesel de gordura de animal em misturas com biodiesel de soja ou em diesel B20.

A espectroscopia de RMN de ^{1}H aliada à regressão por vetores de suporte como o método *boosting* ensemble geram boas estimativas das temperaturas referentes a 10%, 50% e 90% de volume destilado em petróleos, assim como o intervalos de confiança do valor estimado. A espectroscopia de RMN de ^{13}C aliada à regressão por vetores de suporte com seleção de variáveis por algoritmo genético apresentou-se uma metodológica eficiente para estimar os teores de saturados, aromáticos e resinas mais asfaltenos e petróleos.

A simplicidade na preparação da amostra, rapidez na análise, utilização de pequena quantidade de amostra e potencialidade em determinar, simultaneamente, algumas propriedades do petróleo com apenas uma única medida são as grandes vantagens para os químicos analíticos no uso de técnicas espectroscópicas aliadas a métodos quimiométricos.

8 Referências bibliográficas

[1] Speight, J. G. Handbook of petroleum product analysis. New Jersey: Wiley Interscience; 2002.

[2] Simanzhenkov, V.; Idem R. Crude oil chemistry. New York: Marcel Dekker, Inc; 2003.

[3] Lyons, W. C.; Plisga, G. J. Standard Handbook of Petroleum & Natural Gas Engeneering. 2nd ed. Amsterdam: Elsevier; 2005.

[4] Riazi, M. R. Characterization and properties of petroleum fractions, American Society for Testing and Materials (ASTM). 1st ed. Philadelphia, 2005.

[5] ANP 2012. Agência Nacional de Petróleo, Gás Natural e Biocombustíveis; http://nxt.anp.gov.br/NXT/gateway.dll/leg/resolucoes_anp/2011/janeiro/ranp%202%20-%202011.xml?f=templates$fn=document-frame.htm$3.0$q=$x=$nc=9362

[6] Balabin, R. M.; Lomakina, E.; Safieva, R. Z. Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. Fuel 2011, 90, 2007-2015.

[7] Balabin, R. M.; Safieva, R. Z. Near-Infrared (NIR) spectroscopy for biodiesel analysis: fractional composition, iodine value, and cold filter plugging point from one vibrational spectrum. Energy Fuels 2011, 25, 2373-2382.

[8] Gaydou, V.; Kister, J.; Dupuy, N. Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil. Chemometrics and Intelligent Laboratory Systems 2011, 106, 190-197.

[9] Breitkreitz, M. C.; Raimundo, I. M.; Rohwedder, J. J. R.; Pasquini, C.; Dantas, H. A.; José, G. E.; Araújo, M. C. U. Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration. Analyst 2003, 128, 1204-1207.

[10] Khanmohammadi, M.; Garmarudi, A. B.; de la Guardia, M. Characterization of petroleum-based products by infrared spectroscopy and chemometrics. Trac-Trends in Analytical Chemistry 2012, 35, 135-149.

[11] Li, J.Y.; Chu, X. L.; Tian, S. B.; Lu, W. Z. Research on determination of nitrogen content in petroleum using mid-infrared spectroscopy. China Petroleum Processing & Petrochemical Technology 2011, 13, 1-7.

[12] Li, J. Y.; Chu, X. L.; Tian, S. B. Research on determination of total acid number of petroleum sing mid-infrared attenuated total reflection spectroscopy. Energy Fuels 2012, 26, 5633-5637.

[13] Ruiz, M. D.; Bustamante, I. T.; Dago, A.; Hernandez, N.; Nunez, A. C.; Porro, D. A multivariate calibration approach for determination of petroleum hydrocarbons in water by means of IR spectroscopy. Journal of Chemometrics 2012, 24, 444-447.

[14] Felizardo, P.; Baptista, P.; Menezes, J. C.; Correia, M. J. N. Multivariate near infrared spectroscopy models for predicting methanol and water content in biodiesel, Analytica Chimica Acta 2007, 595, 107-113.

[15] Lira, L. F. B.; Vasconcelos, F. V. C.; Pereira, C. F.; Paim, A. P. S. Stragevitch, L.; Pimentel, M. F. Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration. Fuel 2010, 89, 405–409.

[16] Ferrão, M. F.; Viera, M. S.; Pazos, R. E. P.; Fachini, D.; Gerbase, A. E.; Marder, L. Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. Fuel 2011, 90, 701–706.

[17] Canha, N.; Felizardo, P.; Menezes, J. C.; Correia, M. J. N. Multivariate near infrared spectroscopy models for predicting the oxidative stability of biodiesel: effect of antioxidants addition. Fuel 2012, 97, 352-357.

[18] Alves, J. C. L.; Henriques, C. B.; Poppi, R. J. Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system. Fuel 2012, 97, 710-717.

[19] Monteiro, M. R.; Ambrozin, A. R. P.; Liao, L. M.; Boffo, E. F.; Pereira, E. R.; Ferreira, A. G. H-1 NMR and multivariate calibration for the prediction of biodiesel concentration in diesel blends. Journal of the American Oil Chemists Society 2009, 86, 581-585.

[20] Ramos, P. F. D.; de Toledo, I. B.; Nogueira, C. M.; Novotny, E. H.; Vieira, A. J. M.; Azeredo, R. B. D. Low field H-1 NMR relaxometry and multivariate data analysis in crude oil viscosity prediction. Chemometrics and Intelligent Laboratory Systems 2009, 99, 121-126.

[21] Zheng, X. Y.; Jin, Y. Q.; Chi, Y.; Ni, M. J. Simultaneous determination of water and oil in oil sludge by low-field H-1 NMR relaxometry and chemometrics. Energy Fuels 2010, 27, 5787-5792.

[22] Masili, A.; Puligheddu, S.; Sassu, L.; Scano, P.; Lai, A. Prediction of physical–chemical properties of crude oils by ^{1}H NMR analysis of neat samples and chemometrics. Magnetic Resonance in Chemistry 2012, 50, 729–738.

[23] Daniel Molina, V.; Uribe, U. N.; Murgich, J. Partial Least-Squares (PLS) correlation between refined product yields and physicochemical properties with the 1H Nuclear Magnetic Resonance (NMR) spectra of colombian crude oils. Energy Fuels 2007, 21, 1674-1680.

[24] Barbosa, L. L.; Kock, F. V. C.; Silva, R. C.; Freitas, J. C. C.; Lacerda Jr., V.; Castro, E. V. R. Application of low-field NMR for the determination of physical properties of petroleum fractions. Energy Fuels 2013, 27, 673-679.

[25] Höskuldsson, A. PLS regression methods. Journal of Chemometrics 1988, 2, 211-228.

[26] Kalivas, J. Interrelationships of multivariate regression methods using eigenvectors basis sets. Journal of Chemometrics 1999, 13, 111-132.

[27] Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 2001, 58, 109-130.

[28] Andersson, M. A comparison of nine PLS1 algorithms. Journal of Chemometrics 2009, 23, 518-522.

[29] Robertsson, G. Contributions to the problem of approximation of non-linear data with linear PLS in an absorption spectroscopic context. Chemometrics and Intelligent Laboratory Systems 1999, 47, 99-106.

[30] Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Page`s, J. NIR calibration in non-linear systems: different PLS approaches and Artificial Neural Networks. Chemometrics and Intelligent Laboratory Systems 2000, 50, 75-82.

[31] Fong, S. S.; Kiss, V. S. Brereton R. G. Self-organizing maps and support vector regression as aids to coupled chromatography: Illustrated by predicting spoilage in apples using volatile organic compounds. Talanta 2011, 83, 1269–1278.

[32] Sebald, D. J.; Bucklew, J. A. Support Vector Machine techniques for nonlinear equalization. IEEE Transactions on Signal Processing 2000, 48, 3217-3226.

[33] Cortes C.; Vapnik V. N. Support-Vector Networks. Machine Learning 1995, 20, 273–297.

[34] Chen, P. C.; Lee, K. Y.; Lee, T. J.; Huang, S. Y. Multiclass Support Vector Classification via Coding and Regression. Neurocomputing 2010, 73, 1501–1512.

[35] Wu, Y. C.; Lee, Y. S.; Yang, J. C. Robust and efficient multiclass SVM models for phrase pattern recognition. Pattern Recognition 2008, 41, 2874–2889.

[36] Smola, A. J.; Schölkopf, B. A tutorial on Support Vector Regression. Statistics and Computing 2004, 14, 199–222.

[37] Guimarães, R. C. L. Caracterização de petróleo e interpretação de resultados. Rio de Janeiro. Petrobras/CENPES/PDP/AP, 2004.

[38] Neto, B. B.; Scarmínio, I. S.; Bruns, R. E. 25 anos de Quimiometria no Brasil. Química Nova 2006, 29, 1401-1406.

[39] Vapnik, V. N. An overview of Statistical Learning Theory. IEEE Transactions on Neural Networks 1999, 5, 988-999.

[40] Schölkopf, B.; Sung, K. K.; Burges, C. J. C.; Girosi, F.; Niyogi, P.; Poggio, T.; Vapnik, V. N. Comparing Support Vector Machines with gaussian kernels to radial basis function classifiers. IEEE Transactions on Neural Networks 1997, 45, 2758-2765.

[41] Li, H.; Liang, Y.; Xu, Q. Support Vector Machines and its applications in chemistry. Chemometrics and Intelligent Laboratory Systems 2009, 95, 188–198.

[42] ASTM International. ASTM Standard E1655-12: Standard practices for infrared multivariate quantitative analysis; ASTM International: West Conshohocken, PA, 2012.

[43] Mevik, B. H.; Cederkvist, H. R. Mean squared error of prediction (MSEP) estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). Journal of Chemometrics 2004, 18, 422-429.

[44] Valderrama, P.; Braga, J. W. B.; Poppi, R. J. Estado da arte de figuras de mérito em calibração multivariada. Química Nova 2009, 32, 1278-1287.

[45] Valderrama, P.; Braga, J. W. B.; Poppi, R. J. variable selection, outlier detection, and figures of merit estimation in a Partial Least-Squares Regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by Near-Infrared spectroscopy. Journal of Agricultural and Food Chemistry 2007, 55, 8331-8338.

[46] Neto, B. B.; Scarminio, I. S.; Bruns, R. E. Como fazer experimentos: pesquisa e desenvolvimento na ciência e na indústria. 2 ed. Ed. Unicamp; 2001.

[47] Gaudio, A. C.; Zandonade, E. Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica. Química Nova 2001, 24, 658-671.

[48] Filgueiras, P. R.; Alves, J. C. L.; Sad, C. M. S.; Castro, E. V. R.; Dias, J. C. M.; Poppi, R. J. Evaluation of trends in residuals of multivariate calibration models by permutation test. Chemometrics and Intelligent Laboratory Systems 2014, 133, 33–41.

[49] Van der Voet, H. Comparing the predictive accuracy of models using a simple randomization test. Chemometrics and Intelligent Laboratory Systems 1994, 25, 313–323.

[50] Dantas, T. N. C.; Neto, A. A. D.; Moura, E. F. Microemulsion systems applied to breakdown petroleum emulsions. Journal of Petroleum Science and Engineering 2001, 32, 145–149.

[51] Fortuny, M.; Silva, E. B.; Filho, A. C.; Melo, R. L. F. V.; Nele, M.; Coutinho, R. C. C.; *et al.* Measuring salinity in crude oils: evaluation of methods and an improved procedure. Fuel 2008, 87, 1241–1248.

[52] Üstün, B.; Melssen, W. J.; Buydens, L. M. C. Visualisation and interpretation of Support Vector Regression models. Analytica Chimica Acta 2007, 595, 299–309.

[53] ISO 12185. Crude Petroleum and Petroleum Products - Determination of Density - Oscillating U-tube Method. Geneva: International Organization for Standardization. 1996.

[54] ASTM International. ASTM Standard D7042-04: Standard test method for kinematics viscosity in crude oil; ASTM International: West Conshohocken; 2004.

[55] Dias, J. C. M.; Aguiar, P. F.; Santos, M. F. P. Critério estatístico de aceitação da curva de viscosidade-temperatura de petróleos. VIII Seminário de Química, Rio de janeiro. 2004.

[56] ASTM International. ASTM Standard D4377-06: Standard test method for Karl Fischer in Crude Oil; ASTM International: West Conshohocken; 2006.

[57] Kennard, R. W.; Stone, L. A. Computer aided design of experiments. Technometrics 1969, 11, 137-148.

[58] Fearn, T.; Riccioli, C.; Garrido-Varo, A.; Guerrero-Ginel, J. E. On the geometry of SNV and MSC. Chemometrics and Intelligent Laboratory Systems 2009, 96, 22–26.

[59] Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R.S. PLS toolbox version 4.0 for use with Matlab. Wenatchee: Eigenvector Research, Inc.; 2006.

[60] Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machines; 2001. Software: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[61] Abbas, O.; Rebufa, C.; Dupuy, N.; Permanyer, A.; Kister, J. PLS regression on spectroscopic data for the prediction of crude oil quality: API gravity and aliphatic/aromatic ratio. Fuel 2012, 98, 5–14.

[62] Filgueiras, P. R.; Sad, M. C. S.; Loureiro, A. R.; Santos, M. F. P.; Castro, E. V. R.; Dias, J. C. M.; Poppi, R. J. Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. Fuel 2014, 116, 123-130.

[63] Laxalde, J.; Ruckebusch, C.; Devos, O.; Caillol, N.; Wahl, F.; Duponchel, L. Anal. Characterisation of heavy oils using near-infrared spectroscopy: optimisation of pre-processing methods and variable selection. Analytica Chimica Acta 2011, 705, 227–234.

[64] Olivieri, A. C.; Faber, N. K. M.; Ferré, J.; Boqué, R.; Kalivas, J. H.; Mark, H. Uncertainty estimation and figures of merit for multivariate calibration. IUPAC 2006, 78, 633–661.

[65] Rocha, W. F. C.; Nogueira, R.; Vaz, B. C. Validation of model of multivariate calibration: an application to the determination of biodiesel blend levels in diesel by near-infrared spectroscopy. Journal of Chemometrics 2012, 26, 456-461.

[66] Babamoradi, H.; Van den Berg, F.; Rinnan, A. Comparison of bootstrap and asymptotic confidence limits for control charts in batch MSPC strategies. Chemometrics and Intelligent Laboratory Systems 2013, 127, 102–111.

[67] Coskun, A.; Ceyhan, E.; Inal, T. C.; Serteser, M.; Unsal, I. The comparison of parametric and nonparametric bootstrap methods for reference interval computation in small sample size groups. Accreditation and Quality Assurance 2013, 18, 51-60.

[68] Imaizumi, Y.; Suzuki, N.; Shiraishi, H. Bootstrap methods for confidence intervals of percentiles from dataset containing nondetected observations using lognormal distribution. Journal of Chemometrics 2006, 20, 68-75.

[69] Serneels, S.; Van Espen, P. J. Bootstrap confidence intervals for trilinear partial least squares regression. Analytica Chimica Acta 2005, 544, 153-158.

[70] Cao, D. S.; Xu, Q. S.; Liang, Y. Z.; Li, H. D. The boodting: a new idea of building models. Chemometrics and Intelligent Laboratory Systems. 2010, 100, 1–11.

[71] Godoy, L. A. F.; Pedroso, M. P.; Ferreira, E. C.; Augusto, F.; Poppi, R. J. Prediction of the physicochemical properties of gasoline by comprehensive two-dimensional gas chromatography and multivariate data processing. Journal of Chromatography A 2011, 1218, 1663–1667.

[72] Galvão, R. K. H.; Araújo, M. C. U.; Martins, M. N.; José, G. E.; Pontes, M. J. C.; Silva, E. C.; Saldanha, T. B. An application of subagging for the improvement of prediction accuracy of multivariate calibration models. Chemometrics and Intelligent Laboratory Systems 2006, 81, 60-67.

[73] Breiman, L. Stacked regressions. Machine Learning 1996, 24, 49–64.

[74] Breiman, L. Bagging predictors. Machine Learning 1996, 24, 123–140.

[75] Schapire, R. E. The strength of weak learn ability. Machine Learning 1990, 5, 197-227.

[76] Freund, Y. Boosting a weak algorithm by majority. Information and Computation 1995, 121, 256-286.

[77] ASTM International. ASTM Standard D7169-11: Standard test method for boiling point distribution of samples with residues such as crude oils and atmospheric and vacuum residues by high temperature gas chromatography; ASTM International: West Conshohocken; 2011.

[78] Savorani F.; Tomasi G.; Engelsen S. B. icoshift: a versatile tool for the rapid alignment of 1D NMR spectra. Journal of Magnetic Resonance 2010, 202, 190–202.

[79] Thomas, J. E. Fundamentos de engenharia de petróleo. Rio de Janeiro: 2ª Ed. Interciência: Petrobra., 2004.

[80] Sanchez-Minero, F.; Ancheyta, J.; Silva-Oliver, G.; Flores-Vale. S. Predicting SARA composition of crude oil by means of NMR. Fuel 2013, 110, 318-321.

[81] Niazi, A.; Leardi, R. Genetic algorithms in chemometrics. Journal of Chemometrics 2012, 26, 345-351.

[82] Goldberg, D. E. Genetic Algorithms in search, optimization and machine learning, Reading, Addison-Wesley; 1989.

[83] Costa Filho, P. A.; Poppi, R. J. Algoritmo genético em química. Química Nova 1999, 22, 405-411.

[84] Leardi, R. Application of genetic algorithm–PLS for feature selection in spectral data sets. Journal of Chemometrics 2000, 14, 643-655.

[85] Xin, N.; Gu, X.; Wu, H.; Hu, Y.; Yang, Z. Application of Genetic Algorithm-Support Vector Regression (GA-SVR) for quantitative analysis of herbal medicines. Journal of Chemometrics 2012, 26, 353-360.

[86] Fei, Q.; Li, M.; Wang, B.; Huan, Y.; Feng, G.; Ren, Y. Analysis of cefalexin with NIR spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection. Chemometrics and Intelligent Laboratory Systems 2009, 97, 127–131.

[87] Lucasius, C. B.; Beckers, M. L. M.; Kateman, G. Genetic algorithms in wavelength selection: a comparative study. Analytica Chimica Acta 1994, 286, 135-153.

[88] Gillet, S.; Rubini P.; Delpuech J. J.; Escalier J. C.; Valentin P. Quantitative carbon-13 and proton nuclear magnetic resonance spectroscopy of crude oil and

petroleum products. I. Some rules for obtaining a set of reliable structural parameters. Fuel 1981, 60, 221-225.

[89] Randazzo, M. L.; Sodré, J. R. Exhaust emissions from a diesel powered vehicle fuelled by soybean biodiesel blends (B3–B20) with ethanol as an additive (B20E2–B20E5). Fuel 2011, 90, 98–103.

[90] Lôbo, I. P.; Ferreira, S. L. C.; Cruz, R. S. Biodiesel: parâmetros de qualidade e métodos analíticos. Química Nova 2009, 32, 1596-1608.

[91] Teixeira, L. S. G.; Couto, M. B.; Souza, G. S.; Andrade Filho, M.; Assis, J. C. R.; Guimarães, P. R. B.; Pontes, L. A. M.; Almeida, S. Q.; Teixeira, J. S. R. Characterization of beef tallow biodiesel and their mixtures with soybean biodiesel and mineral diesel fuel. Biomass and Bioenergy 2010, 34, 438–441.

[92] Araújo, B. Q.; Nunes, R. C. R.; Moura, C. V. R.; Moura, E. M.; Citó, A. M. G. L.; Santos Júnior, J. R. Synthesis and characterization of beef tallow biodiesel. Energy Fuels 2010, 24, 4476–4480.

[93] Fernandes, D. D. S.; Gomes, A. A.; Costa, G. B.; Silva, G. W. B.; Véras, G. Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection. Talanta 2011, 87, 30-34.

[94] Vasconcelos, F. V. C.; Souza Jr., P. F. B.; Pimentel, M. F.; Pontes, M. J. C.; Pereira, C. F. Using near-infrared overtone regions to determine biodiesel content and adulteration of diesel/biodiesel blends with vegetable oils. Analytica Chimica Acta 2012, 716, 101-107.

[95] Pontes, M. J. C.; Pereira, C. F.; Pimentel, M. F.; Vasconcelos, F. V. C.; Silva, A. G. B. Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification. Talanta 2011, 85, 2159-2165.

[96] Soares, I. P.; Rezende, T. F.; Silva, R. C.; Castro, E. V. R.; Fortes, I. C. P. Multivariate calibration by variable selection for blends of raw soybean oil/biodiesel from different sources using Fourier Transform Infrared spectroscopy (FTIR) spectra data. Energy Fuels 2008, 22, 2079-2083.

[97] Zhu, D.; Ji, B.; Meng, C.; Shi, B.; Tu, Z.; Qing, Z. The performance of v-Support Vector Regression on determination of soluble solids content of apple by acousto-optic tunable filter near-infrared spectroscopy. Analytica Chimica Acta 2007, 598, 227-234.

[98] Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Applied Spectroscopy 2000, 54, 413–419.

[99] Workman Jr., J.; Weyer, L. Practical guide to interpretive near infrared spectroscopy. Boca Raton: CRC Press; 2008.

[100] Filgueiras, P. R.; Alves, J. C. L.; Poppi, R. J. Quantification of animal fat biodiesel in soybean biodiesel and B20 diesel blends using near infrared spectroscopy and synergy interval Support Vector Regression. Talanta 2014, 119, 582-589.

Apêndice A – Formulação de um problema de classificação binária por um separador linear

A solução de um problema de classificação binária por um hiperplano linear, como descrito pelo SVM, pode ter como resposta várias soluções. Ou seja, vários hiperplanos podem separar as classes sem erros. O objetivo do SVM é encontrar o hiperplano com máxima margem entre os vetores (amostras) limitantes do hiperplano de separação (OSH, do inglês, *optimal separating hyperplane*) (Figura A.1).

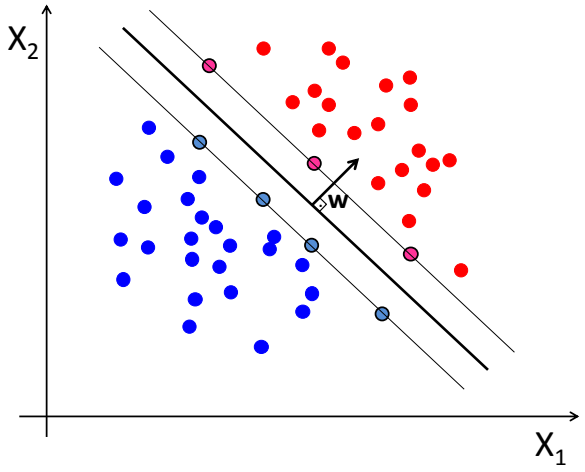

Figura A.1. Gráfico ilustrativo da utilização de um hiperplano como classificador primário no método SVM.

As amostras limitantes de classe são chamadas de vetores de suporte, e possuem uma distância fixa ao OSH, +1 ou -1 caso a amostra pertença à classe "+1" ou "-1" respectivamente. Atribuindo $\mathbf{x}^+$ e $\mathbf{x}^-$ as vetores pertencentes aos hiperplanos limitantes das classes "+1" ou "-1" respectivamente, e perpendiculares ao OSH, temos para estes vetores a formulação:

$$\mathbf{w}\cdot\mathbf{x}^+ + b = +1 \tag{A.1}$$

$$\mathbf{w}\cdot\mathbf{x}^- + b = -1 \tag{A.2}$$

O vetor $\mathbf{w}$ é ortogonal ao hiperplano de separação. Subtraindo a Equação A.1 de A.2, temos:

$$\mathbf{w}\left(\mathbf{x}^{+}-\mathbf{x}^{-}\right)=2 \tag{A.3}$$

A margem entre os vetores limitantes de classe (M) é por definição a distância mínima entre os dois hiperplanos paralelos ao OSH. Podemos

$$M=\left|\mathbf{x}^{+}-\mathbf{x}^{-}\right| \tag{A.4}$$

Substituindo a Equação A.4 em A.3, obtemos uma relação direta do vetor $\mathbf{w}$ com a margem.

$$M=\frac{2}{\|\mathbf{w}\|} \tag{A.5}$$

O problema de maximizar a margem (M) pode ser interpretado de duas formas; sempre sujeitas às restrições das Equações A.1 e A.2. Primeiro maximar a margem equivale a minimizar $\|\mathbf{w}\|$ ou, em segundo, maximizar uma equação pode ser convertido em minimizar sua equação inversa.

Podemos maximizar a margem de separação pela minimização de $\frac{1}{2}\|\mathbf{w}\|^{2}$.

Apêndice B – Otimização dos parâmetros do modelo SVR por grade de pesquisa

Uma grande dificuldade na utilização do método SVR consiste em otimizar seus parâmetros: C (constante de que pondera os erros de regressão), γ da função kernel (nesta Tese, foi utilizada somente a função kernel RBF) e ν para o método ν-SVR ou ε para o método e-SVR. Uma forma simples e objetiva para determinar estes parâmetros é utilizando uma grade de pesquisa associada ao procedimento de validação cruzada.

Na aplicação do método ν-SVR, três parâmetros podem ser otimizados $\{C, \gamma, \nu\}$, uma conjunto deste parâmetros é escolhido e com estes valores é realizado o procedimento de validação cruzada para determinação do erro associado ao conjunto de parâmetros escolhidos. Entretanto, o conjunto de parâmetros não é escolhido de forma aleatória. Em uma grade de pesquisa, o domínio de cada parâmetro é previamente definido. Nesta Tese os domínios utilizados foram:

- $C \in \left[10^{-4}, 10\right]$;
- $\gamma \in \left[10^{-4}, 10^{5}\right]$ e;
- $\nu \in \left[10^{-4}, 10^{0}\right]$ para (ν-SVR) ou $\varepsilon \in \left[10^{-4}, 10^{0}\right]$ para (e-SVR).

A partir dos domínios definidos para todos os parâmetros, deve-se determinar a dimensão da grade e o tipo de interpolação de dados. Para modelos baseado em Vetores de Suporte normalmente utiliza-se a interpolação logarítmica. Na otimização dos parâmetros $\{C, \gamma, \nu\}$ com 7 pontos (5 pontos interpolados mais as duas extremidades), teremos 7^3 conjuntos de parâmetros a serem verificados. Neste caso, cada parâmetros assumirá os valores:

- C {0,0001; 0,0007; 0,0046; 0,0316; 0,2154; 1,4678; 10};
- γ {0,0001; 0,0032; 0,1; 3,1623; 100; 3162,3; 100000};
- ν {0,0001; 0,0005; 0,0022; 0,01; 0,0464; 1}.

e a grade testará todas as combinações para estes valores.