UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

Explorando Métodos De Seleção De Variáveis E Fusão De Dados Em Regressão Por Vetores De Suporte: Uma Aplicação Em Petroleômica

Exploring Variable Selection Methods and Data Fusion in Support Vector Regression: An Application in Petroleomics

Pedro Henrique Pereira da Cunha

Tese de Doutorado em Química

Vitória
2024

Pedro Henrique Pereira da Cunha

Tese apresentada ao Programa de Pós-Graduação em Química do Centro de Ciências Exatas da Universidade Federal do Espírito Santo como requisito parcial para obtenção do Título de Doutor em Química

Área de Concentração: Química

Linha de Pesquisa: Química do Petróleo e Biocombustíveis.

Orientador: Prof. Dr. Paulo Roberto Filgueiras

VITÓRIA
2024

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Cunha, Pedro Henrique Pereira da, 1990-
C972e Explorando métodos de seleção de variáveis e fusão de dados em regressão por vetores de suporte : uma aplicação em petroleômica / Pedro Henrique Pereira da Cunha. - 2024.
156 p. : il.

Orientador: Paulo Roberto Filgueiras.
Tese (Doutorado em Química) - Universidade Federal do Espírito Santo, Centro de Ciências Exatas.

1. Máquinas de vetores de suporte. 2. Seleção de variáveis. 3. Fusão de dados. 4. Petróleo. 5. Aprendizagem de máquina. I. Filgueiras, Paulo Roberto. II. Universidade Federal do Espírito Santo. Centro de Ciências Exatas. III. Título.

CDU: 54

Explorando métodos de seleção de variáveis e fusão de dados em regressão por vetores de suporte: uma aplicação em Petroleômica

Pedro Henrique Pereira da Cunha

Tese submetida ao Programa de Pós-Graduação em Química do Centro de Ciências Exatas da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do Grau de Doutor(a) em Química.

Aprovada em 28/03/2024 por:

_____
Prof. Dr. Paulo Roberto Filgueiras
Orientador(a)
UFES

_____
Prof. Dr. Murilo de Oliveira Souza
IFES

_____
Prof. Dr. Lucas Mattos Duarte
UFF

_____
Prof.(a) Dr.(a) Mariana Ramos de Almeida
UFMG

_____
Prof. Dr. Wanderson Romão
UFES

Universidade Federal do Espírito Santo
Vitória, março de 2024

Datas e horários baseados em Brasília, Brasil
Sincronizado com o NTP.br e Observatório Nacional (ON) em 02/04/2024 às 12:23:26

Ata de defesa de doutorado da Pedro H. P. da Cunha

Data e Hora de Criação: 02/04/2024 às 10:25:47

Documentos que originaram esse envelope:
- Ata.pdf (Arquivo PDF) - 1 página(s)
- Folho de rosto.pdf (Arquivo PDF) - 1 página(s)
- Registro.pdf (Arquivo PDF) - 1 página(s)

Hashs únicas referente à esse envelope de documentos

[SHA256]: a8a8860c0c60fd9d7e2b75d3f14d89f6f94781dd2cf012e32e6e3dd0ddb56504
[SHA512]: f7cb35b7a8b1531403b1971d8fef8abf9e97297c1ada0edf08246596240c4418b8922eaed7962931941a8b4cf496abd4d42067e80d57aa2ba9b791a2d9bf4b9d

Lista de assinaturas solicitadas e associadas à esse envelope

ASSINADO - Lucas Mattos Duarte (duartelucas@id.uff.br)
Data/Hora: 02/04/2024 - 10:43:29, IP: 200.156.108.230, Geolocalização: [-22.897620, -43.126116]
[SHA256]: 2126f5ebaa9159fedda5c99981e212766fbd8f168bc67c57e167f6936e5d5d84

ASSINADO - Mariana Ramos de Almeida (mariramosalmeida@gmail.com)
Data/Hora: 02/04/2024 - 12:23:26, IP: 150.164.16.44
[SHA256]: 6d83e4d37d0e660617fdc9feb309a955247dfaee4bb07fd08b6ba4fb089a8332

ASSINADO - Murilo de Oliveira Souza (m.quimic@gmail.com)
Data/Hora: 02/04/2024 - 12:21:24, IP: 179.109.143.118, Geolocalização: [-20.781469, -41.688164]
[SHA256]: 060fb17ba7837d52291c501e943337873dacaa2fa7bf94e6e94b6ca1d7a3b1d7

ASSINADO - Paulo Roberto Filgueiras (paulo.filgueiras@ufes.br)
Data/Hora: 02/04/2024 - 10:31:52, IP: 179.176.216.206
[SHA256]: 9ee84977fdebb67399c0eefc034653d61024de8c1d377b2bdde1f77f7de64115

Paulo Roberto Filgueiras

ASSINADO - Wanderson Romão (wandersonromao@gmail.com)
Data/Hora: 02/04/2024 - 10:45:49, IP: 200.137.75.113
[SHA256]: d0aeac66d1a0833a689be4d8dce8b251738f00c5b9efd3c13ac76399d6ba4157

Wanderson Romão

Histórico de eventos registrados neste envelope

02/04/2024 12:23:26 - Envelope finalizado por mariramosalmeida@gmail.com, IP 150.164.16.44
02/04/2024 12:23:26 - Assinatura realizada por mariramosalmeida@gmail.com, IP 150.164.16.44
02/04/2024 12:22:59 - Envelope visualizado por mariramosalmeida@gmail.com, IP 150.164.16.44
02/04/2024 12:21:24 - Assinatura realizada por m.quimic@gmail.com, IP 179.109.143.118
02/04/2024 10:45:49 - Assinatura realizada por wandersonromao@gmail.com, IP 200.137.75.113
02/04/2024 10:43:29 - Assinatura realizada por duartelucas@id.uff.br, IP 200.156.108.230
02/04/2024 10:42:36 - Envelope visualizado por duartelucas@id.uff.br, IP 200.156.108.230
02/04/2024 10:31:52 - Assinatura realizada por paulo.filgueiras@ufes.br, IP 179.176.216.206
02/04/2024 10:31:49 - Envelope visualizado por paulo.filgueiras@ufes.br, IP 179.176.216.206
02/04/2024 10:27:29 - Envelope registrado na Blockchain por paulo.filgueiras@ufes.br, IP 179.176.216.206
02/04/2024 10:27:28 - Envelope encaminhado para assinaturas por paulo.filgueiras@ufes.br, IP 179.176.216.206
02/04/2024 10:25:48 - Envelope criado por paulo.filgueiras@ufes.br, IP 179.176.216.206

AGRADECIMENTOS

Primeiramente, gostaria de expressar minha gratidão pela persistência, paciência e esforço de uma pessoa fundamental para a construção desta tese e a conclusão deste trabalho: a mim mesmo. Em segundo lugar, gostaria de agradecer a Luiza Machado Fischer, que durante esses anos de pós-graduação me apoiou nos momentos difíceis e compartilhou os momentos felizes, dando-me força nos momentos de fraqueza e motivos para sorrir quando tudo parecia desabar.

Também gostaria de expressar minha gratidão às primeiras pessoas que acreditaram em mim: Leila Vaz Pereira, minha mãe, e Alvaro Roque Tosta da Cunha, meu pai. Ambos foram fundamentais na minha educação e desenvolvimento, não apenas fornecendo alimentação e educação acadêmica básica, mas também me permitindo acreditar em mim mesmo.

Isabela Pereira da Cunha, minha eterna companheira e parceira, aquela que me acordava às 8 horas da manhã de sábado para jogar bola e hoje traz tanto orgulho para a família. Gabriel Pereira Lopes, o irmão que nunca tive, que esteve sempre ao meu lado em todas as fases da vida, desde as brincadeiras de pega-pega, conhecendo The Sims, aprendendo a beber e hoje, juntos, tornando-nos dois adultos. Luana Tavares, minha prima e irmã mais velha, sempre me aconselhando e puxando minha orelha quando necessário.

Aos meus amigos de longa data, Ivan, Rafael Colombi e Kun Woo, que me proporcionaram alívio do estresse e mantiveram minha saúde mental, proporcionando momentos de felicidade e companheirismo. E aos vários amigos que me proporcionaram momentos de felicidade: João Souza, Gian Paulo, Letícia, Bárbara Miranda, Victor Rafael, Rafael Scalfoni e Yohann.

À Márcia Helena, Gabriely Folli, Madson Zanoni e a todos os meus amigos de laboratório, que compartilharam momentos de felicidade, tristeza e resiliência ao longo desses mais de seis anos de vida acadêmica. Vocês foram fundamentais para que eu permanecesse nesse caminho.

Agradeço também às empresas de fomento CNPq, CAPES e FAPES, pelo incentivo à pesquisa e apoio financeiro na forma de bolsa de doutorado. A Petrobras pelas amostras utilizadas nesse trabalho.

Ao pessoal da BAT, British American Tobacco, uma empresa sólida nas Américas que expandiu meus horizontes para a indústria.

Meus sinceros agradecimentos à minha banca examinadora, composta pelo Prof. Dr. Wanderson Romão, Prof. Dr. Lucas Mattos Duarte, Profa. Dra. Mariana Ramos de Almeida e Prof. Dr. Murilo de Oliveira Souza, que aceitaram participar da minha defesa.

Ao meu orientador, Paulo Roberto Filgueiras, que me conheceu na iniciação científica e me aceitou no mestrado e doutorado, sempre disponível para ensinar e com a paciência necessária para lidar com as melhores perguntas.

"*A vida não é fácil para nenhum de nós. Temos que ter persistência e, acima de tudo, confiança em nós mesmos.*"

Marie Curie

"*Se cheguei até aqui foi porque me apoiei no ombro dos gigantes.*"

Isac Newton

“*Se todas as partes do porco fossem perfeitas, não teríamos cachorro-quente.”*

Steven Universe

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

airPLS - método iterativamente adaptativo por mínimos quadrados ponderados e penalizados (*adaptive iteratively reweighted penalized least squares*)

AM - Aprendizagem de Máquina

ANN - rede neurais artificiais (*artificial neural network*)

API - *American Petroleum Institute*

Aparp - Gráfico de Resíduos Parciais Aumentados de Mallows (Mallows Augmented Partial Residual Plot)

ASA-VIF - algoritmo de busca angular e fator de inflação de variância (a*ngular search algorithm and variance inflation factor*)

ASTM - *ASTM International*

CARS - amostragem adaptativa competitiva reponderada (*competitive adaptive reweighted sampling*)

CENPES - centro de pesquisa Leopoldo Américo Miguel de Mello

CLS - clássico de quadrados mínimos (*Classical Least Squares)*

DHA - análise detalhada de hidrocarbonetos (*detailed hydrocarbon analysis*)

dw - Teste de Durbin–Watson

FTIR - espectroscopia no infravermelho por transformada de Fourier (*Fourier transform infrared*)

GA - algoritmo genético (*genetic algorithm*)

HTGC - cromatografia gasosa de alta temperatura (*high temperature gas chromatography*)

iPLS - mínimos quadrados parciais por intervalo (*interval partial least squares*)

iSVR – regressão por vetores de suporte por intervalo (*interval support vector regression*)

ISO - *International Organization for Standardization*

IUPAC - União Internacional de Química Pura e Aplicada (*International Union of Pure and Applied Chemistry*)

KPLS – kernel combinado com regressão por componentes principais (*Kernel-Partial Least Squares*)

LCPS-PLS - Calibração Local por Seleção de Percentil com PLS (Local Calibration by

Percentile Selection PLS)

LIBS - espectroscopia de emissão por plasma induzido por laser (laser-induced breakdown spectroscopy)

LS-SVM - máquinas de vetores de suporte por mínimos quadrados (l*east-squares support-vector machines*)

MB-PLS - mínimo quadrados parciais multiblocos (*multiblock partial least squares*)

MIR - espectroscopia no infravermelho médio (*mid infrared*)

MLR - regressão linear múltipla (*multiple linear regression*)

MPA - modelo de análise de população (*model population analysis*)

MSC - correção de dispersão multiplicativa (*multiplicative scatter correction*)

NIR - espectroscopia no infravermelho próximo (*near infrared spectroscopy*)

NISPA - Análise de subjanela permutada com ruído incorporado (*noise incorporated subwindow permutation analysis*)

NMR - ressonância magnética nuclear (*nuclear magnetic resonance*)

NT - teor de nitrogênio total

OPLS - projeções ortogonais a mínimos quadrados parciais (*orthogonal projections to latent structure*)

OSH - hiperplano de separação ótima (*optimal sepparting hyperplane*)

PC – Componente principal (*principal component)*

PCA - análise de componentes principais (*principal component analysis*)

PCR - regressão por componentes principais (*principal components regression*)

PLS - mínimos quadrados parciais (*partial least squares*)

PLS-DA - análise discriminante por mínimos quadrados parciais (*partial least squares-discriminant analysis*)

PSO - otimização por enxame de partículas (*particle swarm optimization*)

QSAR - relação quantitativa estrutura-atividade (*quantitative structure-activity relationship*)

QSRR - relação quantitativa estrutura-retenção (*quantitative structure-retention relationship*)

R^2 - coeficiente de determinação

RBF - função de base radial (RBF, do inglês *Radial-Basis Function)*

RF - Floresta Aleatória (Random Forest)

RMSEC - raiz quadrada do erro médio quadrático de calibração (*root mean square*

error of calibration)

RMSECV - raiz quadrada do erro médio quadrático de validação cruzada (*root mean square error of cross validation*)

RMSECV% - erro médio quadrático da raiz da porcentagem de validação cruzada (root mean square error *cross validation* percentage)

RMSEP - raiz quadrada do erro médio quadrático de previsão (*root mean square error of prevision*)

RMSEP% - erro médio quadrático da raiz da porcentagem de predição (root mean square error of prediction percentage)

S-PLS - mínimo quadrados parciais serial (*serial partial least squares*)

SARA - saturados, aromáticos, resinas e asfaltenos (*Saturation, Aromatic, Resins and Asphaltene*)

siPLS - mínimos quadrados parciais por intervalo sinérgico (*synergy interval partial least squares*)

siSVR - regressão por vetores de suporte por intervalo sinérgico (*synergy interval* support vector regression)

SPA - Análise de subjanela permutada (subwindow permutation analysis)

SPrA - algoritmo de projeções sucessivas (*Successive Projections Algorithm*)

SFC/TLC-FID - cromatografia em fluido supercrítico/cromatografia em camada delgada - detector por ionização de chama (Supercritical Fluid Chromatography / thin Layer Chromatography – Flame Ionization Detector)

SNV – variação normal padrão (*Standard Normal Variate*)

SVM - máquina de vetores de suporte (support vector machine)

SVC - classificação por vetores de suporte (support vector classification)

SVR - regressão por vetores de suporte (support vector regression)

UOP - *Universal Oil Products*

UVE - eliminação de variáveis não informativas (*uninformative variable elimination*)

VL - Variável Latente.

WAT - temperatura de aparecimento de cristal (*wax appearance temperature*)

LISTA DE SÍMBOLOS

b – escalar b

$\mathbf{C}$ – Constante que limita a variável de folga

cm^{-1} – comprimento de onda em centímetros

ErroN – Erro Normal, utilizado no cálculo do SPA e NISPA

ErroP – Erro Permutado, utilizado no cálculo do SPA e NISPA

g – Tamanho de grade de pesquisa

log – logaritmo

N – Número total de amostras

n – número de amostras

min – minuto

mL – mililitro

Mj/Kg – megajoule por kilograma

MHz – mega hertz

$mol \cdot L^{-1}$ – concentração, mol por litro

pH - potencial de hidrogênio

ppm – parte por milhão

p -número de parâmetros a otimizar.

Q – Quantidade de variáveis numa subjanela

q – Loading do PLS

ξi – Variável de Folga

T – Tesla

X – Variável independente, Fonte analítica, o que é utilizado para prever

y – Variável dependente, o que deseja prever

y_i – valor medido

$\hat{y}_i$ – valor estimado

$\boldsymbol{w}$ – Peso do SVR

µs – microssegundo

ºF – graus fahrenheits

ºC – graus celsius.

% m/m - percentual massa/massa

ε – Constante de tolerância do SVM

α_i – Vetores do SVM

γ – Otimizar no SVM.

ρ – Densidade a 60ºF

v_{TR} – Viscosidade transformada

v - Viscosidade

RESUMO

A regressão por vetores de suporte (SVR) é considerada um método de aprendizado de máquina caixa-preta e tem se destacado na quimiometria nas últimas décadas, alcançando resultados superiores ou iguais a métodos já consolidados na academia. Sendo um método caixa-preta, torna-se difícil compreender a relação causa/efeito. Para resolver isso, pode-se aplicar a seleção de variáveis, uma estratégia que visa identificar as variáveis mais influentes na construção do modelo. Este trabalho propõe o desenvolvimento de dois métodos de seleção de variáveis. - Análise de subjanela permutada (SPA) e análise de subjanela permutada incorporada por ruído (NISPA) - para aplicar no SVR aliado ao infravermelho. SPA e NISPA forneceram os modelos mais exatos para viscosidade cinemática, saturados e teor aromático. O erro médio quadrático percentual de previsão (RMSEP) do SPA e NISPA foram, respectivamente, de 14,3% e 14,6% para viscosidade cinemática, 4,7% e 4,4% para teor de saturados, e 3,4% e 3,1% para teor aromático. Portanto, SPA e NISPA, além de obterem, em geral, modelos mais rápidos, exatos e parcimoniosos, revelaram as variáveis mais importantes para a construção dos modelos SVR. Outra forma de aperfeiçoar um modelo é a fusão de dados, porém, essa estratégia foi pouco estudada no SVR. Assim, foi estudada a fusão de dados utilizando NIR, MIR, RMN de ^{1}H e ^{13}C combinados utilizando fusão de baixo, médio e alto nível. Os modelos gerados pela fusão de dados apresentaram-se superiores os modelos sem, para a maioria dos testes. Na densidade API, a aplicação de fusão de médio nível utilizando PCA combinando MIR e NIR, desenvolveu um modelo com parâmetros melhores que o modelo sem fusão de dados. Ao aplicar fusão de médio nível com GA para predizer ponto de fluidez, combinando NIR e NMR de ^{1}H, conseguiu-se superar os modelos sem fusão, além de modelos encontrados na literatura. No nitrogênio total, a fusão de alto nível com MIR e NMR de ^{1}H conseguiu ser estatisticamente melhor que os modelos sem fusão de dados. Isso demonstra que é possível extrair novas informações para modelagem em SVR, utilizando a fusão de dados e obter modelos estatisticamente melhores que aqueles advindos a partir de fontes analíticas isoladas.

Palavras-chave: Máquina de vetores de suporte, Seleção de Variáveis, Fusão de Dados, Petróleo, Aprendizagem de Máquina.

ABSTRACT

Support Vector Regression (SVR) is considered a black-box machine learning method and has stood out in chemometrics over the past decades, achieving results superior or equal to methods already established in academia. As a black-box method, it is challenging to understand the cause/effect relationship. To address this, variable selection can be applied, a strategy that aims to identify the most influential variables in building the model. This work proposes the development of two variable selection methods - Permutation Subwindow Analysis (SPA) and Noise-Incorporated Permutation Subwindow Analysis (NISPA) - to apply in SVR combined with infrared. SPA and NISPA provided the most accurate models for kinematic viscosity, saturates, and aromatic content. The root mean square error of prediction (RMSEP) for SPA and NISPA were, respectively, 14.3% and 14.6% for kinematic viscosity, 4.7% and 4.4% for saturates content, and 3.4% and 3.1% for aromatic content. Therefore, SPA and NISPA, in addition to generally obtaining faster, more accurate, and more parsimonious models, revealed the most important variables for building SVR models. Another way to improve a model is data fusion, but this strategy has been little studied in SVR. Thus, data fusion was studied using NIR, MIR, and NMR of ^{1}H and ^{13}C combined using low, medium, and high-level fusion. The models generated by data fusion were superior to the models without fusion for most tests. In API density, the application of medium-level fusion using PCA combining MIR and NIR developed a model with better parameters than the model without data fusion. By applying medium-level fusion with GA to predict pour point, combining NIR and NMR of ^{1}H, it was possible to surpass models without fusion, as well as models found in the literature. In total nitrogen, high-level fusion with MIR and NMR of ^{1}H proved to be statistically better than models without data fusion. This demonstrates that it is possible to extract new information for SVR modeling using data fusion and obtain statistically better models than those derived from isolated analytical sources.

Keywords: Support vector machine, Variable selection, data fusion, Petroleum, Machine Learning

SUMÁRIO

CAPÍTULO 1 INTRODUÇÃO

A regressão por vetores de suporte (SVR, do inglês *Support Vector Regression*)[1] foi desenvolvida a partir da adaptação do método biclasse máquina de vetores de suporte (SVM, do inglês *Support Vector Machine*)[2,3] com o diferencial de ser voltado para solução de problemas quantitativos, com a capacidade de obter resultados equivalentes, ou melhores, que outros métodos quimiométricos. O SVR se destaca pela sua capacidade generalista, não linearidade e habilidade de trabalhar com poucas amostras. Como consequência, o SVR ganhou destaque em diversas áreas como: meio ambiente,[4,5] alimentos,[6,7] petróleo,[8–11] forense[12,13] e outros.[14,15]

O gráfico apresentado na **Figura 1.1(a)** foi construído com informações de janeiro de 2000 à setembro de 2023, com base em uma pesquisa bibliométrica realizada nas bases de dados das plataformas *Web of Science* e Scopus, com a sintaxe que usou diferentes palavras-chave: [("*support vector machine**") OR ("*support vector regression**") OR ("*support vector classifier**")] AND [("*chemometric**") OR ("*Chemical**") OR ("*chemistry*")] NOT ("*ridge regression**") AND (“*Article*”). Pode-se afirmar que o uso do SVM tem crescido na área da quimiometria, chegando a acumular 6.812 artigos nesta faixa de tempo. Restringindo a pesquisa à área de regressão, utilizando as palavras-chave: [("*support vector regression**") OR ("*regression**" AND "*support vector machine*")] AND ("*chemometric**" OR "*Chemical**" OR "*chemistry*") NOT ("*ridge regression**") AND (“*Article*”), verifica-se um aumento do número de publicações, como mostra a **Figura 1.1 (a)**, demonstrando uma forte tendência de crescimento, principalmente após 2014.

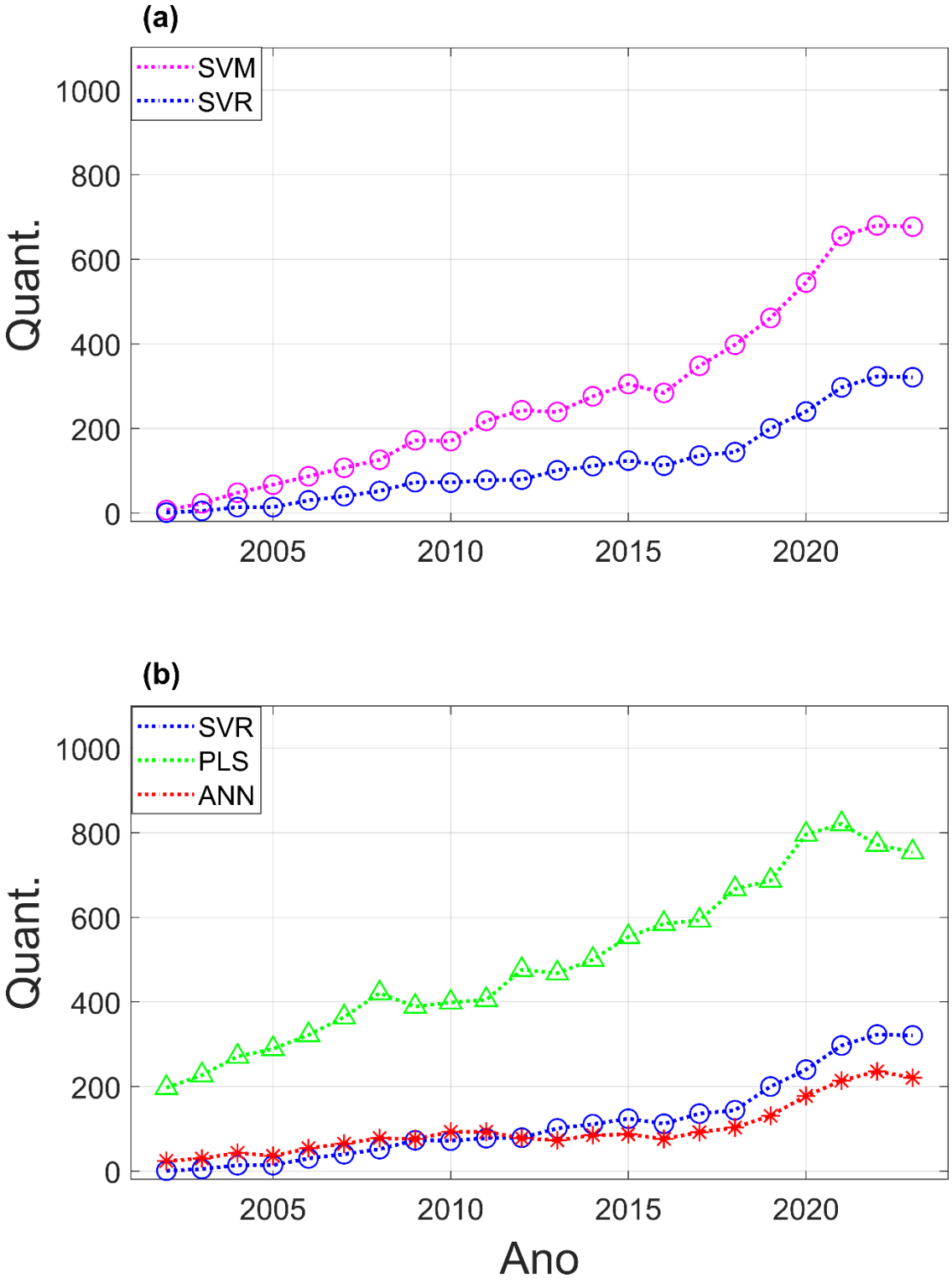


Figura 1. 1 Gráficos com a publicação anual de artigos entre 2000 a 2023 com base em SVM, SVR em Quimiometria (a) e SVR, PLS e ANN em Regressão e Quimiometria (b) através da plataforma Web of Science e Scopus.

O aumento no número de publicações sobre SVR pode ser decorrente tanto do avanço da tecnologia de hardware quanto da popularização da quimiometria e dos

algoritmos para construção dos modelos. Na **Figura 1.1(b)** é mostrada a evolução do uso do método SVR comparado à regressão por mínimos quadrados parciais (PLS do inglês *Partial Least Squares*),[16,17] método mais aplicado em Quimiometria, e às redes neurais artificiais (ANN, do inglês *Artificial Neural Network*),[14,18] que se trata de outro método não linear, amplamente utilizado em outras áreas da ciência e principalmente nas engenharias. Baseado em uma pesquisa de dados na Web Of Science e Scopus com as palavras-chaves: ("*partial least squares**") AND [("*chemometric**") OR ("*Chemical**") OR ("*chemistry*")] AND ("*Article*") NOT ("*Partial least squares discriminant analysis**") e [("*artificial neural network**") AND ("*regression**")] AND [("*chemometric**") OR ("*Chemical**") OR ("*chemistry*")] AND ("*Article*"), verifica-se que, em 2013, o SVR superou o ANN e se aproximou do PLS em número de publicações científicas. Entretanto, é provável que o SVR nunca alcance o mesmo patamar que o PLS, uma vez que este é um método mais simples e mais acessível.

Quando analisamos os cinco principais países que mais publicaram artigos sobre SVR, utilizando as mesmas palavras-chaves da pesquisa anterior, percebe-se que a maioria dos artigos é da Asia, com um destaque para China com 670 artigos e Irã com 112, como pode ser observado na **Figura 1.2.** O segundo é um dos principais países que trabalha com petróleo, Irã. Ainda entre os cinco países que mais publicam, temos EUA e Brasil, este último vem conquistando aos poucos seu espaço na quimiometria mundial.[19]

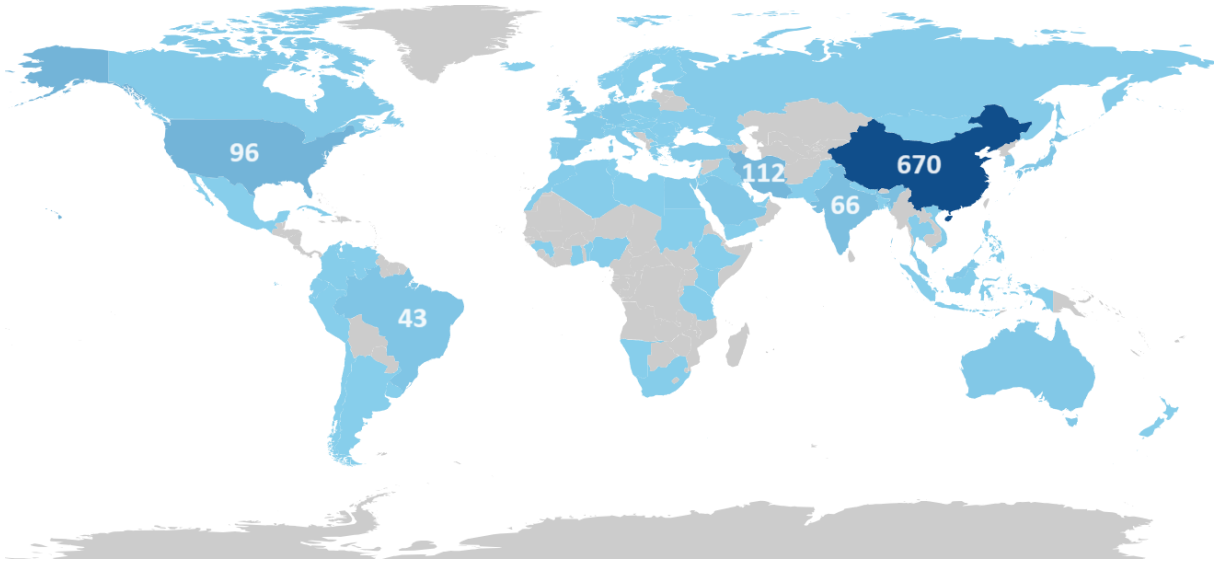


Figura 1. 2 Gráficos com a publicação total de artigos entre 2000 a 2023 por país com base em SVR em Quimiometria através da plataforma *Web of Science e Scopus*.

O SVR tem demonstrado ser uma ferramenta quimiométrica de grande valor e eficácia ao longo dos anos, tornando-se fundamental no arsenal dos químicos analíticos. Nesse contexto, os estudos realizados durante o doutorado foram direcionados para a divulgação, análise e aprimoramento dessa técnica, que atualmente possui mais de três décadas de história.

1.1 Objetivo

Objetivo Geral

Aprimorar o desempenho e a interpretabilidade da regressão por vetores de suporte por meio da utilização de espectros de infravermelho próximo e médio, bem como de ressonância magnética nuclear ^{1}H e ^{13}C, desenvolvendo métodos de seleção de variáveis e aplicando a fusão de dados.

Objetivos Específicos

- Desenvolver um método de seleção de variáveis baseada em permutação aplicável à regressão por vetores de suporte.
- Desenvolver um método de seleção de variáveis baseado na eliminação de variáveis não informativas, aplicável em máquinas de vetores de suporte.
- Aplicar seleção de variáveis em espectros de infravermelho médio e próximo para a predição de propriedades físico-químicas do petróleo utilizando modelagem de regressão por vetores de suporte.
- Desenvolver um método de fusão de dados de nível baixo, médio e alto, aplicável a regressão de vetores de suporte, fundindo espectros de infravermelho médio e próximo e ressonância magnética nuclear de hidrogênio e carbono 13.

1.2 Estrutura da Tese

Esta tese é fruto de um projeto de doutorado concebido com base em críticas recebidas durante a defesa do mestrado, buscando-se, ao mesmo tempo, aprimorar o método SVR, como introduzir inovações científicas ao tema específico e desmistificar a aplicação do SVR para solução de problemas químicos. No entanto, adaptações foram realizadas em decorrência de necessidades identificadas e da emergência de novas ideias e oportunidades.

Ao longo do progresso dos projetos de pesquisa, dedicou-se esforço à realização de uma bibliometria sobre a Regressão de Vetores de Suporte, explorando sua história de desenvolvimento, modificações matemáticas e estatísticas, aplicações ao longo dos anos e recentes avanços. O método foi abordado de maneira crítica, destacando aspectos positivos e negativos. Deseja-se no futuro transformar esse conhecimento em um artigo de revisão, utilizando uma linguagem concisa e acessível em português, com o propósito de ampliar a disseminação do SVR nos países de língua portuguesa. Os resultados desse estudo estão detalhados no **Capítulo 2 - Fundamentação Teórica**.

Alinhado ao objetivo central da pesquisa de doutorado, buscou-se aprimorar o desempenho do SVR por meio de duas abordagens distintas. A primeira envolveu a implementação do método de seleção de variáveis eliminação de variáveis não informativas (UVE, do inglês *uninformative variable elimination*), amplamente empregado na quimiometria para destacar informações relevantes, desconsiderando aquelas consideradas irrelevantes. Vale ressaltar que essa abordagem carece de estudos específicos no contexto do SVR. A aplicação da seleção de variáveis resultou na concepção de dois novos métodos, anteriormente restritos à Classificação por Vetores de Suporte (SVC, do inglês *support vector classification*). O estudo proporcionou resultados promissores, como na seleção das melhores variáveis do espectro de infravermelho médio na determinação da densidade API em petróleo. No entanto, há a necessidade de conduzir estudos adicionais empregando fontes analíticas discretas. Os resultados dessa investigação foram publicados na *Journal Of Chemometrics*[20] e encontram-se detalhados no **Capítulo 3**, intitulado **'Seleção de Variáveis por Permutação Aplicada em Modelos de Regressão por Vetores de Suporte'**.

A segunda estratégia adotada consistiu na aplicação da Fusão de Dados, uma

metodologia que tem recebido crescente atenção nos últimos anos. Destaca-se que, até o início da realização desta tese, não havia sido reportado na literatura científica o emprego de fusão de dados com modelos de SVR. O emprego de fusão de dados com informações oriundas de diferentes fontes analíticas para modelos de regressão por SVR revelou-se mais dispendiosa do que planejado, a princípio, demandando-se considerável esforço computacional e tempo. Isso resultou em ajustes nos planos iniciais, com uma redução no número de modelos para um método de fusão de dados que apenas combina os dados originais, a fusão de baixo nível, onde as fontes analíticas são concatenadas diretamente, e uma ampliação em outro tipo de fusão, que realiza processamentos e modelagem nos dados e combina resultados, conhecido como fusão de níveis médio, onde a informação de cada fonte analítica é extraída antes da fusão, e alto, onde os resultados dos modelos são fusionados (A Fusão será melhor explicada no tópico **2.4 Fusão de Dados**). Os resultados obtidos foram promissores; no entanto, carecem de uma análise mais aprofundada. Até o momento, os resultados não foram divulgados na forma de artigo científico e são apresentados no **Capítulo 4**, que trata da **Fusão de Dados em Regressão por Vetores de Suporte, aplicada na previsão de propriedades físico-químicas do petróleo**.

Paralelo aos objetivos da pesquisa de doutorado, foram direcionados esforços para adquirir experiência em diferentes áreas e contribuir para a divulgação científica. Neste último contexto, foram desenvolvidos materiais didáticos para o ensino de quimiometria em cursos de graduação, fornecendo rotinas, funções e dados para ensinar de maneira clara e concisa os principais métodos empregados na área. Atualmente, esse objetivo específico resultou na publicação de um artigo,[21] com um deles em avaliação e outros três em desenvolvimento. Em colaboração com o Laboratório de Café do IFES de Venda Nova do Imigrante, foram conduzidos diversos estudos relacionados ao café, abrangendo desde a análise das diferenças químicas decorrentes da aplicação de seis métodos distintos de fermentação,[22] até a determinação de blends de café,[23] entre outros temas abordados.[24–26] No âmbito da medicina, em colaboração com dois grupos acadêmicos, o Programa de Assistência Dermatológica e o Laboratório Computacional Inspirado na Natureza, buscou-se contribuir para a identificação de câncer de pele maligno por meio da utilização de um equipamento portátil de infravermelho próximo, empregando análises quimiométricas.[27] Adicionalmente, foram conduzidos estudos na área de alimentos em

geral, incluindo a determinação de alimentos adulterados,[28] entre outros.[29] Os detalhes dos estudos desenvolvidos simultaneamente estão disponíveis no **Capítulo 6**, intitulado '**Produção Científica'**.

1.3 Referência

1. Huang, Y., Zhang, J., Tze Ann, F. & Ma, G. Intelligent mixture design of steel fibre reinforced concrete using a support vector regression and firefly algorithm based multi-objective optimization model. *Constr. Build. Mater.* **260**, 120457 (2020).
2. Vapnik, V. Support-Vector Networks. *IEEE Expert. Syst. their Appl.* **7**, 63–72 (1992).
3. Vapnik, V. N. *The Nature of Statistical Learning Theory*. (Springer New York, 2000). doi:10.1007/978-1-4757-3264-1.
4. Lu, W.-Z. & Wang, W.-J. Potential assessment of the “support vector machine” method in forecasting ambient air pollutant trends. *Chemosphere* **59**, 693–701 (2005).
5. Leong, W. C., Kelani, R. O. & Ahmad, Z. Prediction of air pollution index (API) using support vector machine (SVM). *J. Environ. Chem. Eng.* **8**, 103208 (2020).
6. Xu, S., Zhao, Y., Wang, M. & Shi, X. Determination of rice root density from Vis–NIR spectroscopy by support vector machine regression and spectral variable selection techniques. *Catena* **157**, 12–23 (2017).
7. Zhou, X. *et al.* Development of deep learning method for lead content prediction of lettuce leaf using hyperspectral images. *Int. J. Remote Sens.* **41**, 2263–2276 (2020).
8. Bemani, A. *et al.* Modeling of cetane number of biodiesel from fatty acid methyl ester (FAME) information using GA-, PSO-, and HGAPSO- LSSVM models. *Renew. Energy* **150**, 924–934 (2020).
9. Alves, J. C. L. & Poppi, R. J. Quantification of conventional and advanced biofuels contents in diesel fuel blends using near-infrared spectroscopy and multivariate calibration. *Fuel* **165**, 379–388 (2016).
10. Voigt, M. *et al.* Using fieldable spectrometers and chemometric methods to determine RON of gasoline from petrol stations: A comparison of low-field 1H NMR@80 MHz, handheld RAMAN and benchtop NIR. *Fuel* **236**, 829–835

(2019).

11. Tuba, E., Ribic, I., Capor-Hrosik, R. & Tuba, M. Support Vector Machine Optimized by Elephant Herding Algorithm for Erythemato-Squamous Diseases Detection. *Procedia Comput. Sci.* **122**, 916–923 (2017).
12. Xu, C. *et al.* A novel strategy for forensic age prediction by DNA methylation and support vector regression model. *Sci. Rep.* **5**, 17788 (2015).
13. Rodrigues, C. & Bruni, A. VIDROS AUTOMOBILÍSTICOS COMO VESTÍGIOS DE CENA DE CRIME: UMA ABORDAGEM MULTIVARIADA. *Quim. Nova* **44**, 553–560 (2021).
14. Attia, K. A. M., Nassar, M. W. I., El-Zeiny, M. B. & Serag, A. Firefly algorithm versus genetic algorithm as powerful variable selection tools and their effect on different multivariate calibration models in spectroscopy: A comparative study. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **170**, 117–123 (2017).
15. Yuan, J. *et al.* Predicting the biological activities of triazole derivatives as SGLT2 inhibitors using multilayer perceptron neural network, support vector machine, and projection pursuit regression models. *Chemom. Intell. Lab. Syst.* **156**, 166–173 (2016).
16. Amsaraj, R., Ambade, N. D. & Mutturi, S. Variable selection coupled to PLS2, ANN and SVM for simultaneous detection of multiple adulterants in milk using spectral data. *Int. Dairy J.* **123**, 105172 (2021).
17. Mohammadi, M., Khanmohammadi Khorrami, M., Vatanparast, H., Karimi, A. & Sadrara, M. Classification and determination of sulfur content in crude oil samples by infrared spectrometry. *Infrared Phys. Technol.* **127**, 104382 (2022).
18. Najwa Mohd Rizal, N. *et al.* Comparison between Regression Models, Support Vector Machine (SVM), and Artificial Neural Network (ANN) in River Water Quality Prediction. *Processes* **10**, 1652 (2022).
19. Barros Neto, B. de, Scarminio, I. S. & Bruns, R. E. 25 anos de quimiometria no Brasil. *Quim. Nova* **29**, 1401–1406 (2006).
20. da Cunha, P. H. P. *et al.* Variable selection by permutation applied in support vector regression models. *J. Chemom.* **36**, 1–14 (2022).
21. Silveira Folli, G., Pereira da Cunha, P. H., Kuster Moro, M. & Roberto Filgueiras, P. Tutorial para aplicação didática de quimiometria em software gratuito – Parte I: Análise de Componentes Principais em dados de infravermelho médio e propriedades físico-químicas de amostras de petróleo. *Rev. Ifes Ciência* **9**, 01–

14 (2023).

22. Zani Agnoletti, B. *et al.* Effect of fermentation on the quality of conilon coffee (Coffea canephora): Chemical and sensory aspects. *Microchem. J.* **182**, (2022).
23. Vieira Lyrio, M. V. *et al.* SHS-GC-MS applied in Coffea arabica and Coffea canephora blend assessment. *Anal. Methods* **15**, 3499–3509 (2023).
24. Correia, R. M. *et al.* PORTABLE NEAR INFRARED SPECTROSCOPY APPLIED TO THE QUALITY CONTROL OF COFFEE ADULTERED BY GROUNDS. *Quim. Nova* **45**, 392–402 (2022).
25. de Paulo, E. H. *et al.* Study of coffee sensory attributes by ordered predictors selection applied to 1H NMR spectroscopy. *Microchem. J.* **190**, 108739 (2023).
26. Lyrio, M. *et al.* PERFIL VOLÁTIL DO Coffea arabica E Coffea canephora var. conilon POR SHS-GC-MS E QUIMIOMETRIA. *Quim. Nova* **X**, 1–10 (2024).
27. Loss, F. P. *et al.* Skin cancer diagnosis using NIR spectroscopy data of skin lesions in vivo using machine learning algorithms. (2024) doi:10.48550/arXiv.2401.01200.
28. Folli, G. S. *et al.* Food analysis by portable NIR spectrometer. *Food Chem. Adv.* **1**, 100074 (2022).
29. Zanoni, M. P., Cunha, P. H. P. da, Silveira Folli, G. & Roberto Filgueiras, P. O USO DO REDGIM PARA CARACTERIZAR E DISTINGUIR AZEITES EXTRAVIRGEM DE OLIVA ADULTERADOS COM DIFERENTES ÓLEOS VEGETAIS. *Rev. Ifes Ciência* **9**, 01–11 (2023).

CAPÍTULO 2 FUNDAMENTAÇÃO TEÓRICA

2.1 Quimiometria

Nas últimas décadas, os laboratórios de química, especialmente os de analítica, têm enfrentado uma sobrecarga devido ao grande volume de informações e resultados gerados por instrumentos analíticos modernos. Esse aumento na complexidade e multidimensionalidade dos dados é um reflexo do avanço exponencial na capacidade de processamento dos computadores, que não só transformou os laboratórios, mas também impactou profundamente nossas vidas pessoais.[1] Assim, criou-se a necessidade de desenvolver ferramentas que consigam extrair informações relevantes a partir destes dados e excluir informações repetidas ou sem importância. Neste cenário, surgiu a Quimiometria que, segundo a União Internacional de Química Pura e Aplicada (IUPAC), pode ser definida como “a aplicação de estatística para a análise de dados químicos, projeto de experimentos químicos e simulações”.[2]

Nesse sentido, pode-se dizer que a quimiometria está presente desde a Lei de Lambert-Beer, de 1792, em que é feita uma relação empírica entre absorção da luz e a propriedade do material. Todavia, não é isso que a comunidade acadêmica de quimiometria defende, considerando que quimiometria tem relação com dados multivariados. O primeiro artigo publicado a utilizar ferramentas estatísticas para o tratamento de dados químicos multivariados foi feito por Jurs, P. C., em 1969,[3] e em 1972 foi publicado o primeiro artigo contendo a palavra “quimiometria”, na revista sueca Kemisk Tidskrift.[4] Assim, afirma-se que a quimiometria iniciou-se na década de 70 junto com o avanço dos computadores nos laboratórios de química analítica.[1]

A Quimiometria, pode ser definida como um amálgama de matemática, estatística e programação alinhadas sob uma perspectiva química, nessa perspectiva, tem conquistado espaço no ambiente acadêmico, obtendo a confiança mesmo dos pesquisadores mais conservadores. Outra interpretação da Quimiometria a considera como um braço químico da Aprendizagem de Máquina (AM). Essa visão, entretanto, está parcialmente correta, uma vez que nem todo método quimiométrico pertence à AM, como o método análise de componentes principais (PCA, do inglês *principal components analysis*) e agrupamento hierárquico (HCA, do inglês *Hierarquical Cluster*

Analysis). No entanto, assim como em toda a ciência de dados, a Quimiometria se beneficiou do crescimento exponencial da AM desde a década de 90, destacando métodos mais complexos como Redes Neurais (ANN do inglês artificial neural network) e SVM.[5]

A Quimiometria pode ser dividida em três grandes áreas: métodos Supervisionados, Não-Supervisionados e Planejamento de Experimentos.[1] O Planejamento de Experimentos, também conhecido como Design de Experimentos, tem como objetivo realizar um número mínimo de experimentos para extrair o máximo de informações sobre um produto ou processo, fazendo uso de estatística multivariada. Nesse processo, diversos fatores relevantes são analisados simultaneamente, como quantidade de reagentes, produtos e pressão. Os dados resultantes podem ser empregados na avaliação e/ou otimização do processo.[6]

Os métodos não supervisionados são caracterizados pelo uso exclusivo da variável independente (X) para construir o modelo, sendo esta uma informação química, como por exemplo espectro de infravermelho e ressonância magnética nuclear (NMR, do inglês *nuclear magnetic resonance*). Essa metodologia pode ser empregada para análise de agrupamento, redução de dimensões e detecção de anomalias. A PCA é o método mais comumente utilizado nesse contexto.[7,8]

A PCA foi desenvolvida por Karl Pearson em 1901,[9] todavia a primeira publicação ocorreu somente em 1923, quando Fisher discriminou diferentes espécies de batatas com base no uso de fertilizantes[10] e a primeira aplicação em quimiometria foi em 1987,[11] onde foi utilizada para analisar a correlação entre petróleo e sua fonte. O método é frequentemente utilizado para melhorar a interpretação das variáveis num conjunto de dados amostral.[12] Em vez de analisar as variáveis separadamente, a PCA identifica variáveis correlacionadas e as combina em novas variáveis chamadas de Componentes Principais (PC do inglês *principal componente*). A primeira PC é construída para capturar a maior variação dos dados, e as PCs subsequentes são construídas com base no resíduo da análise anterior, visando capturar o máximo de restante.[13] As PCs são ortogonais entre si, o que significa serem perpendiculares em um espaço multidimensional, garantindo que cada PC capture uma informação única e não redundante da variação dos dados. A PCA é utilizada para projetar dados multivariados em um espaço de dimensão menor, frequentemente duas ou três PCs são utilizadas, permitindo a visualização da estrutura do espaço amostral original sem alterar as relações entre as amostras.[14]

Utilizando a PCA é possível diminuir o número de variáveis de um conjunto amostral, facilitando o processamento computacional, identificando variáveis importantes e analisando a relação química entre as amostras, o que possibilita a identificação de outliers e agrupamentos de amostras semelhantes.[1] Por ser um método de processamento simples e já consolidada, a PCA é um dos métodos mais utilizados na quimiometria, como demonstrado por Moro *et al.*, em 2020,[15] que utilizaram a PCA para redimensionar as variáveis e aplicá-las na fusão de médio nível para prever propriedades físico-químicas do petróleo e Correia *et al.*, em 2022,[16] que utilizaram microNIR e PCA para distinguir amostras de café com maior e menor teor de adulterante.

Dentre os métodos supervisionados, podemos categorizá-los em dois tipos: regressão multivariada (conforme será discutido no 2.1.1 *Regressão Multivariada*) e classificação multivariada. No segundo tipo o objetivo é treinar um modelo utilizando variáveis independentes para determinar uma variável dependente (y) com valores qualitativos, também chamado de classes. Um exemplo dessa aplicação é o estudo conduzido por Agnoletti B. Z., em 2022,[17] que utilizou análise discriminante por mínimos quadrados parciais (PLS-DA, do inglês *partial least squares-discriminant analysis*) e classificação por vetores de suporte (SVC, do inglês *support vector classification*) para avaliar o potencial impacto de diferentes processos de fermentação na qualidade do café.

2.1.1 *Regressão Multivariada*

A regressão multivariada, como o próprio nome sugere, utiliza métodos multivariados para prever variáveis dependentes quantitativas, tais como a concentração de uma substância,[18] uma propriedade físico-quimica[19] e a qualidade de um café.[20] Destacando-se da abordagem univariada pela capacidade de lidar com interferentes. Ao aplicarmos a Lei de Lambert-Beer, é necessário assumir que a variável utilizada é afetada apenas pelo que se deseja determinar. Entretanto, em contextos envolvendo matrizes complexas como o petróleo e o café, cuja composição é formada por uma diversidade de moléculas distintas, raramente é possível satisfazer tal pressuposto.

A regressão multivariada começa na obtenção da variável dependente de interesse. Os dados são organizados algebricamente em vetores e matrizes.

Considere o conjunto de dados $\{(x_1, y_1), \ldots, (x_n, y_m)\}$, com $x_n \in R^m$ sendo o vetor contendo o espectro para cada amostras e $y_n \in R$ o valor de referência para a amostra *i*. Em problemas qualitativos o y_n representa a classe à qual a amostra pertence, e em problemas quantitativos representa o valor da propriedade de interesse.[21]

As fontes analíticas são organizadas em matrizes de dados ($\mathbf{X}$) com cada amostra representada por uma linha. Cada variável é emparelhada na mesma coluna, dessa forma, todas amostras devem ter o mesmo número de variáveis. Para n amostras com m variáveis, a matriz de dados terá a dimensão $X_{(n,m)}$ (n linhas e m colunas). Após a organização dos dados, antes da aplicação de quaisquer métodos quimiométricos, pode ser necessário realizar pré-processamento nas fontes analíticas, como a aplicação da primeira derivada para dados de infravermelho,[22] alinhamento de espectros de RMN[23] e autoescalamento para concentrações de substâncias no café.[15] Esses pré-tratamentos visam facilitar a extração de informações relevantes e sua interpretação.

Os dados pré-processados são separados em dois grupos, conjunto de calibração e de teste, no caso da classificação multivariada é denominado treinamento e teste. A seleção das amostras para cada conjunto pode ser realizada manualmente, aleatoriamente ou com o auxílio de técnicas de seleção tal como o método Kennard-Stone, que utiliza a fonte analítica como referência. Esse algoritmo de seleção[24] começa selecionando as duas amostras com a maior distância Euclidiana entre si no $\mathbf{X}$ para o conjunto de treinamento. Nas amostras restantes, calcula-se a distância mínima com base às já selecionadas, a amostra com a maior distância mínima é retida, e o procedimento é repetido até que um determinado número de amostras seja selecionado.[25]

As amostras do conjunto calibração são utilizadas para construir os modelos multivariados, que tem a capacidade preditiva determinadas por parâmetros de avaliação. Os parâmetros de avaliação podem ser obtidos de duas formas distintas, interno, utilizando o próprio conjunto de calibração, como no caso da calibração cruzada, e podem ser externos, quando utilizamos o conjunto teste.[21]

Entre os primeiros métodos aplicados na calibração multivariada estão o método clássico de quadrados mínimos (CLS, do inglês Classical Least Squares) e a regressão linear múltipla (MLR, do inglês Multiple Linear Regression). Todavia, esses métodos perderam popularidade.[1] No caso do CLS, a necessidade de conhecer cada

espécie espectroscopicamente ativa na solução dificulta sua aplicação em amostras reais. A MLR, por sua vez, apresenta problemas de colinearidade; portanto, o número de variáveis deve ser menor do que o de amostras, o que raramente é alcançado em métodos espectroscópicos como infravermelho na região do médio (MIR, do inglês Mid Infrared),[26] infravermelho na região do próximo (NIR, do inglês Near Infrared)[27] e espectroscopia de ressonância magnética nuclear (NMR) de ^{1}H e ^{13}C.[28]

A regressão por mínimos quadrados parciais (PLS, do *inglês partial least squares*)[30] é o método de regressão multivariada mais utilizado na química, devido à sua capacidade de resolver rapidamente problemas lineares, com pouco esforço computacional e resultados eficientes. Mesmo em dados não lineares, o método pode ser adaptado para produzir resultados satisfatórios, como é o caso da utilização do Kernel PLS (KPLS, do inglês *kernel partial least squares*), que consiste em uma aplicação do mapeamento kernel, um conceito que será abordado posteriormente, antes da aplicação do PLS, visando uma linearização dos dados.[31] Além disso, podem ser realizados pré-processamentos,[32] na fonte analítica e seleção de variáveis.[33]

No PLS, o X é a variável independente e o y a variável dependente, para a predição é realizada considerando o(s) ponto(s) de melhor representatividade da fonte analítica como referência. As previsões são estatísticas e a presença de interferentes não influenciam nos resultados.[34]

O PLS decompõe a matriz $\mathbf{X}$ e o vetor $\mathbf{y}$, em matrizes de scores ($\mathbf{T}$ e $\mathbf{U}$) e de loadings ($\mathbf{P}$ e $\mathbf{q}$) e como resultado sobrando os resíduos ($\mathbf{E}$ e $\mathbf{f}$) de acordo com as equações:[35]

$$X = TP^T + E \qquad \textbf{Equação 2.1}$$

$$y = Uq^T + f \qquad \textbf{Equação 2.2}$$

As componentes principais, $\mathbf{T}$ e $\mathbf{U}$, têm a direção de variância dos dados originais maximizadas, essas direções não estão necessariamente correlacionadas e são ligeiramente rotacionadas de forma a aumentar a covariância entre $\mathbf{T}$ e $\mathbf{U}$. Depois da rotação, uma relação linear é determinada entre scores e cada variável latente (VL)[36]

$$U = BT \qquad \textbf{Equação 2.3}$$

Os coeficientes de $\mathbf{B}$ são calculados pela equação:

$$B = U^T T (T^T T)^{-1} \qquad \textbf{Equação 2.4}$$

Com os valores calculados podemos estimar as propriedades de novas amostras $\widehat{\boldsymbol{Y}}$ com a **Equação 2.5**:[35]

$$\hat{y} = BTq^T + f \qquad \textbf{Equação 2.4}$$

O objetivo do PLS é obter uma relação linear ótima entre os scores $\mathbf{U}$ e $\mathbf{T}$. Para isso, as componentes principais das variáveis dependentes e independentes devem ter um ângulo entre eles igual a zero, resultando em uma correlação máxima entre as variáveis. Posteriormente, esses componentes principais são denominados variáveis latentes.[37]

É essencial ter cautela ao determinar o número ótimo de VL, pois um número elevado pode levar a sobreajustes, resultando em modelos que se ajustam bem ao conjunto de calibração, mas que apresentam desempenho insatisfatório em conjuntos de testes e em futuras amostras. Por outro lado, um número excessivamente baixo de VL pode resultar em baixo poder preditivo. O princípio da parcimônia, ou navalha de Ockham, deve ser adotado pelos quimiometristas e usuários dos métodos quimiométricos. Este conceito, aplicado à análise multivariada, sugere a escolha de modelos mais simples, como o PLS com poucas VL, que sejam igualmente eficientes e possuam alto poder preditivo.[38]

O método PLS, apesar de ser um método antigo, mantém sua relevância na quimiometria, como demonstrado por Nascimento et al. em 2018,[39] ao utilizaram análise detalhada de hidrocarbonetos (DHA, do inglês *detailed hydrocarbon analysis*) e cromatografia a gás de alta temperatura (HTGC, do inglês *high temperature gas chromatography*) para determinar a pressão de vapor de reid e o ponto de fulgor utilizando PLS. Os resultados para a pressão de vapor de Reid foram promissores ao utilizar o HTGC, alcançando um coeficiente de determinação de predição (R^2p) de 0,99, o que é estatisticamente significativo em comparação com outros modelos. Os autores também concluíram que a relação entre HTGC e pressão de vapor de Reid é

alta, todavia depende de uma quantidade significativa de amostras, pois a redução do número de amostras e mesmo aliada a fusão de dados com DHA não produziu um resultado estatisticamente equivalente.

2.1.2 *Não linearidade*

Na química analítica um dado é considerado não linear quando sua variável dependente, as fontes analíticas (NIR, RMN, Raman, etc.), é linear em relação aos parâmetros e não em relação às variáveis independentes. São exemplos, os modelos baseados em funções geométricas, logarítmicas e polinomiais maiores que um.[40] A não linearidade pode ser confirmada com base em alguns testes, como; Teste de Durbin–Watson (dw),[41] o Wald–Wolfowitz,[22] e método Gráfico de Resíduos Parciais Aumentados de Mallows (Aparp, do inglês *Mallows Augmented Partial Residual Plot*).[40,42]

Existem algumas maneiras de linearizar dados, como a aplicação de pré-processamento na fonte analítica,[32] a seleção de variáveis para encontrar uma relação linear[33] e a aplicação de modelos locais. Esta última estratégia é baseada no princípio "dividir para conquistar", partindo do pressuposto de que um conjunto de dados não linear é um subconjunto de dados lineares que foram mesclados. Nesse contexto, Allegrini e Olivieri, em 2022,[43] utilizaram Calibração Local por Seleção de Percentil com PLS (LCPS-PLS, do inglês *Local Calibration by Percentile Selection PLS*) na determinação de gordura e umidade em carne, obtendo um ótimo resultado, reduzindo o erro médio quadrático da raiz da porcentagem de predição (RMSEP%, do inglês *root mean square error of prediction percentage*) de 0,30 para 0,17. No entanto, a abordagem mais convencional é a aplicação de metodologias de regressão multivariada não linear.

Entre os métodos não lineares destacam-se as Redes Neurais Artificiais (ANN, do inglês *Artificial Neural Networks*), a Floresta Aleatória (RF, do inglês *Random Forest*) e SVM. O método SVM será abordado com mais detalhes no tópico **2.2, Regressão por Vetores de Suporte**.

As ANNs, como o nome sugere, são baseadas nas redes neurais dos seres vivos, onde uma rede de neurônios recebe inputs, processa-os e produz outputs ponderados.[44] As ANN têm a capacidade de modelar tanto dados lineares quanto não lineares e lidam bem com dados faltantes. Em contraponto, exigem uma grande

quantidade de dados e apresentam risco de sofrer sobreajuste.[42] González-Viveros N. *et al*, em 2021,[18] demonstraram o potencial das ANNs ao utilizá-la para predizer sacarose, glicose e frutose em alimentos industrializados conseguindo bons parâmetros de avaliação, R^2p acima de 0,9 em todos casos.

A RF é um algoritmo não paramétrico que pode ser aplicado tanto para regressão quanto para classificação.[45] Este método treina e avalia várias árvores de decisão separadamente e, em seguida, pondera os resultados, aplicando votação no caso da classificação e média no caso da regressão. A RF tem a vantagem de produzir resultados exatos, modelar tanto linear quanto não linear e evitar o sobreajuste, mas apresenta desvantagens como dificuldade de interpretação, alta demanda computacional e tempo de otimização elevado.[46] Chen *et al*, em 2021,[47] utilizaram o RF aliado com espectroscopia de emissão por plasma induzido por laser (LIBS, do inglês *laser-induced breakdown spectroscopy*) e MIR para determinar de forma exata o pH do solo, com um R^2p de 0,98.

2.2 Regressão por vetores de suporte

Para compreender a regressão por vetores de suporte, é necessário entender o SVM. Este método foi proposto por Vladimir Vapnik, em 1992,[48] como um classificador para problemas binários. Em 1997, Scholkopf e Vapnik expandiram o SVM ao introduzir o mapeamento kernel, utilizando uma função de base radial (RBF, do inglês *Radial-Basis Function*),[49] que se tornou padrão no método. O mapeamento kernel capacita o SVM a lidar com problemas não lineares. Nesta técnica, o espaço amostral é transformado adicionando uma nova dimensão, permitindo que o SVM resolva problemas utilizando uma função linear.[50] Entretanto, ao aplicar o kernel, perde-se a relação causa/efeito entre as variáveis químicas (variáveis) e a propriedade predita. Üstün et al., em 2007,[51] propuseram superar essa limitação utilizando um truque programacional, obtendo uma variável denominada p-valor, que permite identificar as variáveis com maior influência na construção do modelo.

Atualmente, o SVM possui adaptações para lidar com problemas multiclasses (SVC),[52] que tem demonstrado potencial em diversas áreas, incluindo sensoriamento remoto,[53] diagnósticos médicos[54] e detecção de adulteração de alimentos.[55] Além disso, há uma adaptação para regressão. O primeiro artigo sobre SVM voltado para regressão foi escrito por Van Gestel *et al.* em 2001.[56] Nesse estudo, o método foi

aplicado para prever uma série temporal financeira e a volatilidade dos dados. No mesmo ano, Irena Nancovska[57] comparou o SVR com outros métodos não lineares na previsão do comportamento de elementos de referência de tensão elétrica, conseguindo demonstrar a aplicabilidade do método e sua capacidade de superar outros métodos não lineares. Embora o SVR e o SVM sejam métodos baseados em matemática e estatística, não serão abordados detalhes sobre esses aspectos neste capítulo. Essa parte relevante será discutida no **ANEXO A - REGRESSÃO POR VETORES DE SUPORTE - CÁLCULO**.

O primeiro artigo que aplicou SVR em dados químicos foi de Song *et al.*, em 2002.[58] Os autores tentaram utilizar modelos de relação quantitativa estrutura-retenção (QSRR, do inglês *Quantitative Structure-Retention Relationship*) para prever o tempo de retenção de proteínas em sistemas de cromatografia de troca aniônica. Os modelos QSRR foram previstos utilizando dois tipos de SVR e os resultados demonstraram a alta capacidade generalizadora do método.

Em 2004, Thissen U. *et al.*,[59] fizeram a primeira comparação entre PLS,[19,60,61] e SVR utilizando dados químicos. O objetivo era comparar os resultados dos dois métodos de regressão em espectros Raman e infravermelho na região do próximo para determinação da fração molar de etanol, água e isopropanol e misturas ternárias. Foi observado uma variação não linear do NIR com a mudança de temperatura, o que dificulta a modelagem com métodos lineares, como o PLS, e destaca a relevância de métodos não lineares como o SVR.[59] Para contornar o problema, fizeram a aquisição do espectro em diferentes temperaturas, com correção de efeito não linear e seleção de comprimento de onda para adaptar o método. Entretanto, ainda assim, o SVR forneceu o melhor resultado para previsão de todas as três frações molares. Os autores argumentam que o SVR tem vantagem sobre o PLS nessas situações devido a sua habilidade de tratar não linearidades. O PLS, mesmo usando metodologias para contornar a não linearidade, obteve raiz quadrada do erro médio da previsão (RMSEP, do inglês *Root Mean Square Error of Prediction*) duas vezes maior que o SVR em todas as situações estudadas.

Mei Hu *et al.*, em 2005,[62] aplicaram SVR, PLS e ANN na modelagem da relação quantitativa estrutura-atividade (QSAR, do inglês *Quantitative Structure-Activity Relationship*) de dois conjuntos peptídicos, dipeptídeos de sabores amargos e inibidores de enzima de conversão de angiotensina. Os autores utilizaram 8 escores da PCA extraídos de 70 vetores de propriedades hidrofóbicas, estéricas e eletrônicas.

Na determinação do QSAR de ambos os conjuntos, todos os métodos de regressão obtiveram modelos com boa capacidade preditiva, porém o SVR obteve R^2 maiores que os demais métodos. Segundo os autores, isso ocorreu devido à baixa quantidade de amostras (24 amostras) que tornam os modelos PLS e ANN sobreajustados, causando uma capacidade preditiva pobre nas amostras externas. Novamente o SVR se destaca, demonstrando maior capacidade de generalização que o PLS e o ANN em situações com número pequeno de amostras.

O primeiro artigo publicado de SVR na área de meio ambiente foi o de Lu W. e Wang W., em 2005,[63] que utilizaram o SVR e uma forma alternativa de ANN para prever a tendência de poluentes na atmosfera de Tokio. O estudo utilizou dados de diferentes áreas de estudo, desde químicos, como a concentração de poluentes como monóxido de carbono (CO) e óxidos de nitrogênio (NOx), até dados meteorológicos, como direção e velocidade do vento e radiação solar. As análises foram feitas a partir de duas coletas diferentes realizadas em julho e dezembro, com variações de hora, local e data. Ficou demonstrada a capacidade do SVR de lidar com dados complexos devido a sua tolerância a ruído, alta estabilidade e capacidade de generalização.

Em alimentos, a primeira publicação foi de Igor Kovalenko *et al.*, em 2006,[64] cujo principal objetivo foi a determinação de ácidos graxos de seis tipos em soja, utilizando NIR aliado aos métodos de SVR, ANN e PLS. O PLS apresentou os parâmetros de avaliação menos favoráveis, ao passo que o SVR, em grande parte dos casos, demonstrou desempenho superior. Esse resultado pode ser atribuído, provavelmente, à natureza não-linear do espectro de NIR nesse trabalho. No entanto, é importante ressaltar que o autor não conduziu um teste de comparação entre os modelos, o que dificulta a avaliação precisa das diferenças de desempenho.

Na área de alimentos, a maioria dos artigos focam principalmente na adulteração de alimentos.[65] Nesse contexto, destaca-se o trabalho inicial de Ferrão et al., em 2007,[66] que empregou a regressão de mínimos quadrados por vetores de suporte (LS-SVM, do inglês *Least-Squares Support Vector Regression*), uma adaptação do SVR, em conjunto com PLS, para prever três tipos de adulterantes em leite - amido, soro e sacarose - utilizando espectroscopia NIR. Na construção dos modelos, foram geradas amostras com um único adulterante e amostras binárias, contendo dois adulterantes. Como se tratava de adulterantes distintos, esperava-se um comportamento não-linear no espectro. Os resultados obtidos demonstraram melhor desempenho do SVM, especialmente na determinação da ausência de adulterantes nas amostras, aspecto

em que o PLS apresentou dificuldades.

Um dos destaques do SVR é a sua capacidade de modelagem em amostras complexas, assim como petróleo, que é uma substância formada pela mistura de diversas moléculas orgânicas nos três estados físicos, apresenta composição distinta de acordo com a origem geográfica e rochas geradoras.[67] Com essa variedade, é difícil o desenvolvimento de modelos bons para todos os tipos de amostra. A capacidade de generalização e a não linearidade do SVR, torna-o um método promissor para este tipo de matriz.

Um dos primeiros artigos de SVR aplicado a petróleo foi o de Balabin e colaboradores, em 2011,[68] que utilizou SVR, ANN e PLS, em espectros de NIR, para prever propriedades de gasolina, biocombustíveis e petróleo. Na determinação do teor de resinas, os modelos alcançaram valores de RMSEP iguais a 0,26, 0,30 e 0,71 wt% com SVR, ANN e PLS, respectivamente e, para teor de parafinas alcançaram 0,12, 0,13 e 0,35 wt%, respectivamente. Para os autores, os resultados mostram a vantagem de métodos não lineares para predição de propriedades de petróleo, além de evidenciar o SVR como uma metodologia promissora para matrizes complexas.

SVR e quimiometria também tem se expandiram na área de Ciências Forenses, nos últimos anos.[69–71] O primeiro trabalho ao usar SVR foi de Cheng Xu *et al.*, em 2015,[72] o artigo utilizou onze sítios de DNA que sofreram metilação para determinar a idade do indivíduo da amostra. Essa determinação é bastante importante na área forense, todavia é dificultada quando não se tem acesso ao corpo do indivíduo. Dessa forma, utilizando amostras de DNA, com auxílio de quimiometria, pode-se chegar a uma idade aproximada. Ao aplicar ANN e SVR em onze sítios de DNA, obteve-se um erro de 3,9 e 2,0 anos, respectivamente. Ao reduzir esses sítios, o SVR foi novamente aplicado, obtendo-se um erro de 2,8 e 4,23 anos para 6 e 3 sítios. Assim, foi demonstrado que o SVR, mesmo com a redução de variáveis, consegue obter bons resultados, o que demonstra um potencial ainda pouco explorado na área da forense.

Em aplicações recentes, é interessante destacar a área de análise de combustíveis. No estudo conduzido onde Leal A. L. *et al*, em 2020,[73] aplicaram SVR em dados de NMR de ^{1}H e NIR para predizer dez propriedades de gasolina. O NMR de ^{1}H demonstrou um desempenho superior na determinação do Éter Metil Terc-Butilílico em comparação ao infravermelho, apresentando um R^2p de 0,936 contra 0,933 obtido pelo PLS. No entanto, é importante notar que a diferença está abaixo do limite de reprodutibilidade, não tendo diferença estatística. Por outro lado, a NMR de

^{1}H e o NIR obtiveram resultados equivalentes ao analisar benzeno, oxigênio e olefinas. Nesse contexto, os dados de infravermelho se mostraram superiores aos de NMR de ^{1}H, entretanto, ambas as fontes analíticas trouxeram bons resultados na aplicação do SVR.

Recentemente, o artigo de Limin Li e Wee Shong Chin, em 2021,[74] utilizou uma modificação na espectroscopia Raman, aplicação de nanotubos de prata para permitir análises rápidas, para prever adulterante em leite, tanto leite líquido quanto leite em pó para adultos, comparando PLS e SVR, mesmo utilizando uma faixa de concentração considerada baixa (0,5 a 100 ppm), os resultados foram promissores. Foram obtidos valores de RMSEP de 2,0 ppm e 6,1 ppm para SVR e PLS, respectivamente, sendo melhores que resultados encontrados na literatura, na época da publicação, para leite líquido e no leite em pó o RMSEP foi de 7,3 ppm e 3,9 ppm para SVR e PLS, respectivamente. Apesar de o método precisar de melhorias, já demonstrou um potencial obtendo bons resultados, tendo baixo custo e sendo rápido.

É relevante ressaltar também a crescente tendência do emprego de equipamentos portáteis na química analítica.[75] Santos *et al.*, em 2022,[76] utilizaram um espectrômetro de NIR portátil aliado a quimiometria para quantificar a porcentagem de óleo leve, pesado e de motor em misturas ternarias com o objetivo de identificar óleos adulterados. Para quantificação de óleo leve e pesado, os modelos SVR apresentaram erros de previsão menores que os PLS, com RMSEP iguais a 4,0 e 6,2 wt%, respectivamente, para óleo leve e iguais a 0,9 e 2,9 wt%, para óleo pesado. Já para quantificação de óleo de motor, SVR apresentou maior erro, com RMSEP igual a 6,2, contra 4,8 wt% para PLS. Apesar da menor resolução do NIR portátil em comparação ao instrumento de bancada, o que pode causar o baixo desempenho dos modelos, evidenciados resultados mostram que, mesmo com um número reduzido de variáveis (absorbâncias medidas em 125 números de onda), o SVR conseguiu obter bons resultados possibilitando determinar adulteração em amostras de óleo no próprio campo.

Outra tendência recente é a aplicação em imagens, cujos dados têm alta dimensão na quimiometria. Levi N., Karnieli A. e Paz-Kagan T., em 2022,[77] que utilizaram imagens de satélite para avaliar os efeitos das atividades humanas na qualidade do solo em regiões áridas de Israel. Os autores utilizaram os dados das imagens em modelos SVR para prever doze propriedades físicas, químicas e biológicas que juntas compõem o chamado Índice de Qualidade de Solo, destacando

os seguintes resultados para pH, P(log10) e condutividade elétrica, que obtiveram R^2p de 0,733, 0,733 e 0,843 e RMSEP 0,136, 0,101 e 0,319, valores considerados bons para os autores. A combinação de fontes analíticas de alta dimensão e o método SVR se mostrou uma forma prática e rápida para avaliar a qualidade do solo.

Os diversos casos apresentados mostram que o SVR é versátil, podendo ser utilizado em diversas áreas, matrizes e\ou fontes analíticas, incluindo NIR portátil, Raman, dados meteorológicos, imagens de satélite, NMR, entre outros. No **Anexo B**, intitulado **Artigos científicos que envolvem SVR**, é possível identificar outras aplicações interessantes e variadas do método.

2.3 Seleção de variáveis

Fontes analíticas de elevado número de variáveis, muitas variáveis são poucos informativas, redundantes e ruídos, assim criando esforço computacional desnecessário, além de prejudicar a exatidão dos modelos. Para lidar com essa problemática é essencial realizar a seleção de variáveis, extraindo apenas os termos mais relevantes. Essa seleção pode ser realizada através de métodos que variam desde a simples observação da fonte de dados até abordagens mais avançadas, como a implementação do algoritmo genético (GA, do inglês *genetic algorithm*).[78]

No contexto do SVR, a seleção de variáveis apresenta uma vantagem significativa. Devido à aplicação do mapeamento kernel durante a modelagem, a informação sobre qual variável possui maior importância no modelo é perdida, resultando em uma desvantagem conhecida como "caixa preta". No entanto, os métodos de seleção de variáveis são uma forma de abordar esse problema. Eles têm a capacidade de analisar a relação causa-efeito entre as variáveis da fonte analítica e as variáveis dependentes, efetivamente quebrando a "caixa preta" do SVM. Essa abordagem não apenas aprimora a interpretação do modelo, mas também proporciona insights valiosos sobre quais variáveis contribuem significativamente para as previsões, tornando o processo de modelagem mais transparente e compreensível.[51]

O método GA trata-se de um método de seleção de variáveis, um dos mais corroborados na literatura.[78–80] Com grande potencial em aplicar em SVR, como foi demonstrado por Filgueiras *et al.*, em 2016,[81] compararam o método de seleção GA aplicado a PLS e SVR. O modelo GA-SVR conseguiu obter resultados satisfatórios

frente ao PLS, além de conseguir determinar quais variáveis do espectro são importantes para a previsão das propriedades.Com a aplicação do GA-SVR ficou demonstrando a causa-efeito entre as propriedades fisico-quimicas citadas do petróleo e partes do espectro de NMR de ^{13}C. No entanto, o GA pode enfrentar o problema dos mínimos locais, que ocorre quando um método de otimização, ou seleção, encontra um ponto com valor inferior aos vizinhos, mas que não é necessariamente o menor valor possível. Para contornar esse problema, os autores repetiram a seleção de variáveis utilizando o GA 100 vezes, acumulando as variáveis escolhidas em cada modelo. Essa abordagem resultou em um gráfico de frequência de seleção, identificando as variáveis mais consistentes na seleção.[81]

O GA não é a única abordagem para a seleção de variáveis, como demonstrado por Xu *et al.*, em 2017.[82] Eles aplicaram quatro métodos distintos: o GA, a UVE, o algoritmo de projeções sucessivas (SPrA, do inglês Successive Projections Algorithm) e a amostragem ponderada adaptativa competitiva (CARS, do inglês Competitive Adaptive Reweighted Sampling). O objetivo central do estudo foi destacar o potencial do CARS, um método que também se baseia nos princípios darwinianos.

O CARS inicia criando um subconjunto com algumas variáveis selecionadas aleatoriamente. Em seguida, aplica uma função exponencial decrescente para remover as variáveis com baixo coeficiente de regressão do PLS. Posteriormente, realiza uma amostragem adaptativa ponderada para reduzir o número de variáveis. Este procedimento é repetido várias vezes, e o melhor modelo é escolhido ao final.[82] O UVE, por sua vez, adiciona uma matriz de ruído à matriz principal, construindo um modelo no PLS. As variáveis originais com coeficientes de regressão menores, ou seja, menos importantes, são eliminadas com base na média dos coeficientes de regressão dos ruídos. No método SPrA, são selecionadas as variáveis com o mínimo de redundância, buscando minimizar a colinearidade por meio de operações de projeção em um espaço vetorial.[82]

Além desses métodos, novas abordagens vêm ganhando destaque, como o uso do algoritmo de busca angular combinado com fator de inflação de variância (ASA-VIF, do inglês *Angular Search Algorithm And Variance Inflation Factor*). Este método foca em remover variáveis altamente correlacionadas, preservando apenas aquelas que não compartilham informações iguais. O ASA-VIF combina a aplicação do ASA, obtendo o cosseno entre as variáveis e preservando aquelas com os menores cossenos, com a utilização do VIF para remover variáveis com alta colinearidade.

Subconjuntos são então criados e avaliados para determinar as melhores variáveis. Estudos recentes, como o de Folli *et al*.,em 2020,[22] aplicaram o ASA-VIF com sucesso em espectros de MIR, NIR e NMR de ^{1}H para prever propriedades físico-químicas do petróleo, destacando sua eficácia ao reduzir efetivamente o número de variáveis.

Outro estudo relevante foi conduzido por Cunha *et al.*, em 2022.[67] Nesse trabalho, os autores testaram seleções de variáveis baseadas em permutação, comparando-as a outras técnicas (GA, ASA e siSVR) usando SVR em conjunto com espectroscopia MIR. A análise de subjanela permutada (SPA, do inglês *subwindow permutation analysis*) e análise de subjanela permutada com ruído incorporado (NISPA, do inglês *noise incorporated subwindow permutation analysis*), ambos adaptados para SVR, demonstraram potencial para igualar e superar outros métodos de seleção testados no estudo. Notavelmente, o SPA, utilizando apenas 3,4% das variáveis, alcançou resultados equivalentes aos modelos que selecionaram mais de um terço ao prever a densidade API. Além disso, o NISPA conseguiu selecionar a menor quantidade de variáveis, contribuindo para a obtenção do melhor modelo.

A seleção de variáveis é crucial para aprimorar modelos não-lineares e, no caso do SVR, atua como uma ferramenta para desvendar a "caixa preta", revelando as variáveis de maior relevância na criação do modelo. No entanto, dada a complexidade do SVR, com diversos parâmetros a serem testados e otimizados, a seleção de variáveis ainda demanda mais estudos para uma compreensão mais aprofundada de sua aplicação e eficácia.

2.4 Fusão de dados

Uma abordagem que vem ganhado destaque em quimiometria é a fusão de dados.[83] Esta técnica teve origem na década de 70, inicialmente desenvolvida para uso militar, onde diversas fontes de informação eram combinadas para identificar alvos como aeronaves, mísseis e formações militares.[84] A fusão de dados vem sendo aplicadas em diversas áreas, como, robótica, química analítica, entre outras.[85]

Essa técnica parte do pressuposto de que ao combinar fontes diferentes, o número de informações relevantes aumenta, o que melhora a performance do modelo. Isso ocorre devido a informações complementares e sinérgicas das diferentes fontes, quando comparado a um modelo de uma única fonte analítica. Na área da química, onde existem diversas fontes analíticas que oferecem informações distintas sobre a

mesma matriz amostral, surge a possibilidade de aprimorar modelos com baixo desempenho.[86] Um exemplo disso é combinar MIR, NIR e NMR de ^{1}H para predizer uma propriedade físico-química do petróleo.[15]

Semelhante a outras abordagens, é importante compreender as fontes analíticas utilizadas durante o processo para combinar informações relevantes e evitar informações redundantes e ruído. Caso contrário, o tempo de processamento pode ser drasticamente aumentado e os critérios de avaliação podem ser prejudicados. Outro ponto importante, é o cuidado com a sobreposição de uma fonte analítica sobre a outra e o antagonismo entre fontes.[87]

Devido a sua complexidade a fusão de dados é desmembrada em três tipos, baixo, médio e alto nível. Estes três tipos se diferenciam no momento que a fusão (concatenação) dos dados ocorrem, sendo o de baixo nível ocorre inicialmente, médio nível ocorre após uma extração prévia das amostras e o de alto nível que ocorre após o desenvolvimento do modelo.[86]

2.4.1 Fusão de baixo nível

A fusão de baixo nível, também chamada de *sensor fusion*, é a mais simples das três. Esta abordagem combina duas, ou mais, fontes analíticas após serem devidamente pré-processados e as concatena, resultando em uma nova matriz que possui tantas linhas (amostras) quanto as matrizes originais e tantas colunas quanto as variáveis das fontes combinadas. A fusão ocorre antes de qualquer seleção de variáveis, redução de dimensão (variável) e modelagem,[88] conforme demonstrado na **Figura 2.1.**

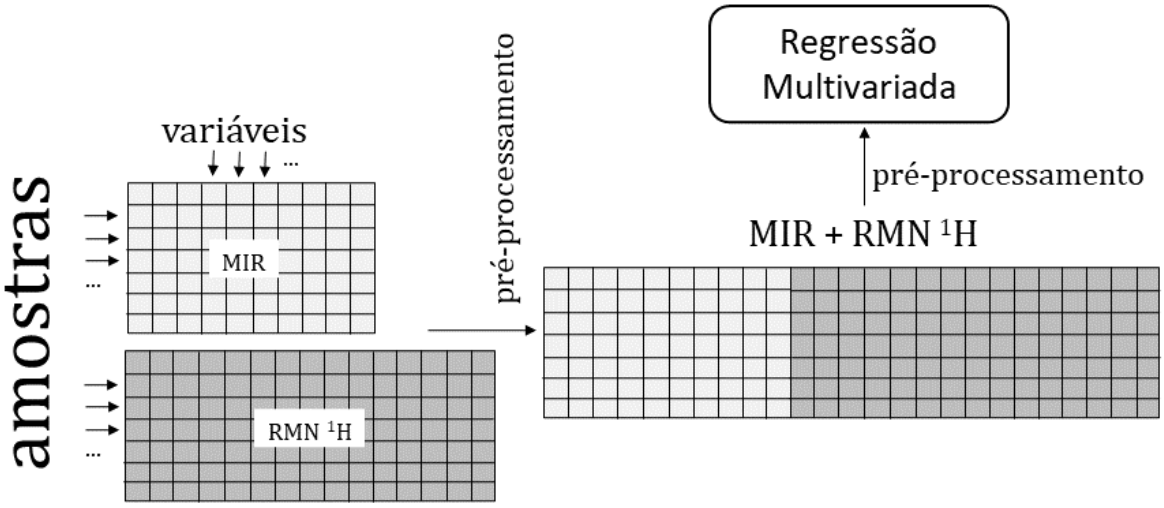


Figura 2. 1 Esquema da fusão de nível baixo para um modelo de regressão usando MIR e NMR de ^{1}H.

Neste tipo de fusão, pode ocorrer o problema da predominância de uma única fonte analítica, semelhante ao que ocorre quando combinamos NMR de ^{1}H e MIR, devido à variância do NMR ser maior, a AM tende a favorecer as informações desta fonte e desconsiderar, em partes, o MIR. Dessa forma, perde-se o sinergismo e a complementariedade das variáveis, perdendo o sentido da aplicação da fusão. Uma solução para este problema é a aplicação de autoescalamento após a concatenação, o que transforma todas as variáveis da matriz para terem variância igual a um, garantindo assim que todas variáveis contribuem igualmente na construção do modelo.[86]

Apesar de sua simplicidade, a fusão de nível baixo pode resultar em matrizes com um número considerável de variáveis, aumentando consideravelmente a complexidade computacional e o tempo de processamento para o desenvolvimento do modelo, além disso, o número de variáveis irrelevante e redundantes também aumenta. Diante disso, é possível aplicar uma seleção de variáveis, como GA, SPA e NISPA, ou reduzir a dimensionalidade da matriz aplicando PCA, para eliminar variáveis desnecessárias e simplificar a modelagem.[89]

2.4.2 Fusão de nível médio

Também chamada de *feature fusion*, a fusão de nível médio extrai as informações relevantes de cada fonte analítica antes de combinar as matrizes.[86] Essa

extração pode ser feita utilizando PCA,[86] seleção de variáveis[87] e até uma AM, como PLS.[15] Após a extração das variáveis as fontes analíticas são concatenadas e o modelo é desenvolvida, como exemplificado na **Figura 2.2**.

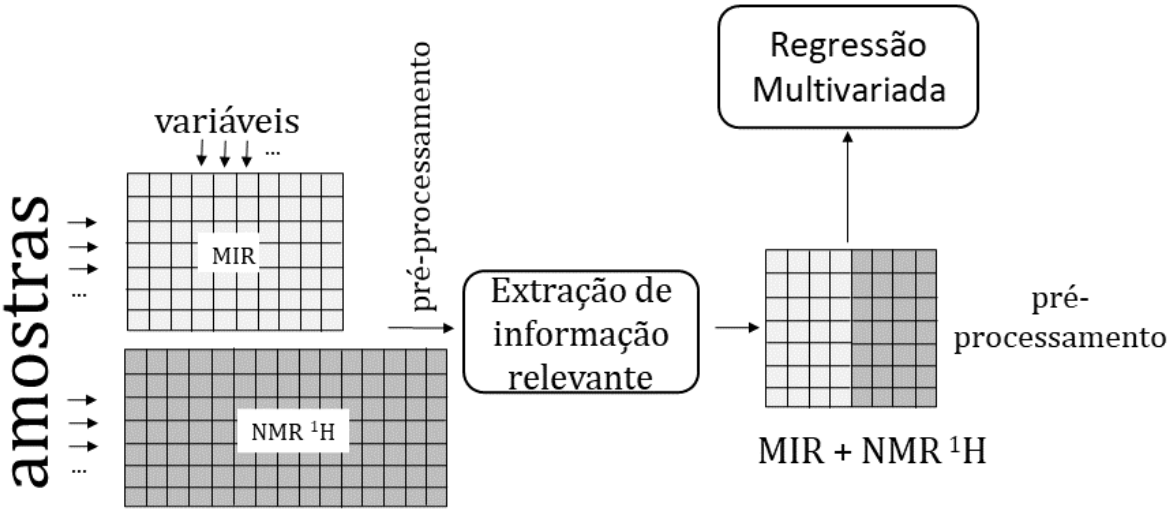


Figura 2. 2 Esquema da fusão de médio nível para um modelo de regressão usando MIR e NMR de ^{1}H.

Em comparação a fusão de baixo nível a de médio tem poucos problemas em relação a grande quantidade de variáveis, devido a extração de informação feita na etapa anterior que elimina informações redundantes e irrelevantes,[90] contudo, tem o risco de remover variáveis que contribuiriam com o sinergismo do modelo.[86] Para contornar isso, Geurts *et al.*, em 2016,[91] sugeriu uma estratégia de “reciclagem” para reavaliar as variáveis não selecionadas e detectar possíveis sinergismo.

Semelhante a estratégia anterior tem-se que ter cuidado com a predominância de uma fonte analítica, podendo, novamente, resolver o problema utilizando o autoescalamento após a concatenação dos dados, entretanto, ainda pode ocorrer a predominância devido a quantidade de variáveis de uma mesma fonte analítica.[92]

2.4.3 Fusão de nível alto

Na fusão de alto nível, também conhecida como "*decision fusion*", um modelo é construído para cada fonte analítica. Nesse método, a fusão é realizada com base nas

decisões de cada modelo para cada amostra, conforme exemplificado na **figura 2-3**. Dessa forma, cada modelo pode ser tratado individualmente, permitindo a aplicação de diferentes AM ao mesmo tempo, bem como pré-processamento distintos e seleções de variáveis para cada modelo.[15,89]

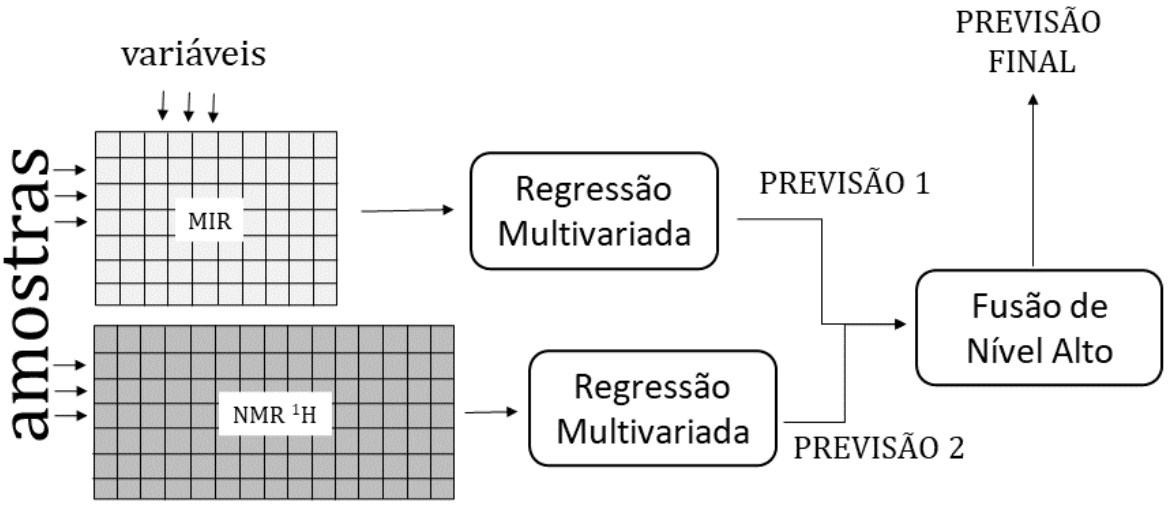


Figura 2. 3 Esquema da fusão de alto nível para um modelo de regressão usando MIR e NMR de ^{1}H.

Para determinar a resposta da fusão, é necessário aplicar um método para definir a resposta da fusão. Por exemplo, no caso da regressão, em que lidamos com resultados quantitativos, é possível utilizar uma média aritmética simples ou uma média ponderada com base em algum parâmetro de avaliação. Já na classificação, como utilizamos resultados qualitativos, o sistema de votação pode ser empregado, onde a classe mais votada é escolhida, e em caso de empate, um parâmetro de avaliação pode ser utilizado para desempatar.[90]

Uma desvantagem da fusão de alto nível está na perda do sinergismo, uma vez que cada fonte analítica é modelada de forma individual as informações novas que surgiriam pelo sinérgismo são perdidas. Por outro lado, problemas relacionados à predominância de uma única fonte não ocorrem nesse tipo de fusão, uma vez que todos os dados são transformados no mesmo número de resposta.[93]

Devido à determinação por decisão a fusão de alto nível pode ser confundida com a aplicação do modelo de análise populacional (MPA, do inglês *Model Population Analysis*).[94] No início do MPA, é criado um grupo de subconjuntos, onde cada subconjunto possui uma quantidade aleatória de amostras e variáveis da matriz original, a partir dos quais são desenvolvidos submodelos. Os submodelos são

avaliados com base em parâmetros de avaliação e o resultado é uma média de todos os submodelos. A diferença entre o MPA e a fusão de dados de alto nível está na utilização de uma única fonte analítica.

2.5 Referência

1. Ferreira, M. M. C. *Quimiometria: conceitos, métodos e aplicações*. (Editora da Unicamp, 2015). doi:10.7476/9788526814714.
2. IUPAC. Chemometrics. in *IUPAC Compendium of Chemical Terminology* vol. 69 1140 (IUPAC, 1997).
3. Jurs, P. C., Kowalski, B. R. & Isenhour, T. L. Computerized Learning Machines Applied to Chemical Problems Molecular Formula Determination from Low Resolution Mass Spectrometry. *Anal. Chem.* **41**, 21–27 (1969).
4. Kowalski, B. R. & Bender, C. F. Pattern recognition. Powerful approach to interpreting chemical data. *J. Am. Chem. Soc.* **94**, 5632–5639 (1972).
5. Géron, A. *Mãos à obra aprendizado de máquina com Scikit-Learn, Keras & TensorFlow: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes*. (Alta Books, 2021).
6. De Souza, A. M. & Poppi, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: UM tutorial, parte I. *Quim. Nova* **35**, 223–229 (2012).
7. Correia, R. M. *et al.* Portable near infrared spectroscopy applied to quality control of Brazilian coffee. *Talanta* **176**, 59–68 (2018).
8. Lemos, M. F. *et al.* Chemical and sensory profile of new genotypes of Brazilian Coffea canephora. *Food Chem.* **310**, 125850 (2020).
9. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
10. Fisher, R. A. & Mackenzie, W. A. Studies in crop variation. II. The manurial response of different potato varieties. *J. Agric. Sci.* **13**, 311–320 (1923).
11. Kvalheim, O. M. Oil—source correlation by the combined use of principal component modelling, analysis of variance and a coefficient of congruence. *Chemom. Intell. Lab. Syst.* **2**, 127–136 (1987).
12. Hotelling, H. Analysis of a complex of statistical variables into principal

components. *J. Educ. Psychol.* **24**, 417–441 (1933).

13. Clarke, B., Fokoue, E. & Zhang, H. H. *Principles and Theory for Data Mining and Machine Learning*. (Springer New York, 2009). doi:10.1007/978-0-387-98135-2.
14. Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Comput. Stat.* **2**, 433–459 (2010).
15. Moro, M. K. *et al.* FTIR, 1H and 13C NMR data fusion to predict crude oils properties. *Fuel* **263**, 116721 (2020).
16. Correia, R. M. *et al.* PORTABLE NEAR INFRARED SPECTROSCOPY APPLIED TO THE QUALITY CONTROL OF COFFEE ADULTERED BY GROUNDS. *Quim. Nova* **45**, 392–402 (2022).
17. Zani Agnoletti, B. *et al.* Effect of fermentation on the quality of conilon coffee (Coffea canephora): Chemical and sensory aspects. *Microchem. J.* **182**, (2022).
18. González-Viveros, N., Gómez-Gil, P., Castro-Ramos, J. & Cerecedo-Núñez, H. H. On the estimation of sugars concentrations using Raman spectroscopy and artificial neural networks. *Food Chem.* **352**, 129375 (2021).
19. Rainha, K. P. *et al.* Determination of API Gravity and Total and Basic Nitrogen Content by Mid- and Near-Infrared Spectroscopy in Crude Oil with Multivariate Regression and Variable Selection Tools. *Anal. Lett.* **52**, 2914–2930 (2019).
20. de Paulo, E. H. *et al.* Study of coffee sensory attributes by ordered predictors selection applied to 1H NMR spectroscopy. *Microchem. J.* **190**, 108739 (2023).
21. Filgueiras PR. Regressão Por Vetores De Suporte Aplicado Na Determinação De Propriedades Físico-Química De Petróleo E Biocombustíveis. (UNICAMP, 2014).
22. Folli, G. S. *et al.* Variable selection in support vector regression using angular search algorithm and variance inflation factor. *J. Chemom.* **34**, 1–16 (2020).
23. de Paulo, E. H. *et al.* Determination of gross calorific value in crude oil by variable selection methods applied to 13C NMR spectroscopy. *Fuel* **311**, 122527 (2022).
24. Kennard, R. W. & Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **11**, 137–148 (1969).
25. Honorato, F. A., Neto, B. D. B., Martins, M. N., Galvão, R. K. H. & Pimentel, M. F. Transferência de calibração em métodos multivariados. *Quim. Nova* **30**, 1301–1312 (2007).
26. Mohammadi, M., Khanmohammadi Khorrami, M., Vatanparast, H., Karimi, A. & Sadrara, M. Classification and determination of sulfur content in crude oil

samples by infrared spectrometry. *Infrared Phys. Technol.* **127**, 104382 (2022).

27. Loss, F. P. *et al.* Skin cancer diagnosis using NIR spectroscopy data of skin lesions in vivo using machine learning algorithms. (2024) doi:10.48550/arXiv.2401.01200.
28. Lovatti, B. P. O., Nascimento, M. H. C., Neto, Á. C., Castro, E. V. R. & Filgueiras, P. R. Use of Random forest in the identification of important variables. *Microchem. J.* **145**, 1129–1134 (2019).
29. Voigt, M. *et al.* Using fieldable spectrometers and chemometric methods to determine RON of gasoline from petrol stations: A comparison of low-field 1H NMR@80 MHz, handheld RAMAN and benchtop NIR. *Fuel* **236**, 829–835 (2019).
30. Daniel Molina, V., Angulo, R., Dueñez, F. Z. & Guzmán, A. Partial least squares (PLS) and multiple linear correlations between heithaus stability parameters (P o) and the colloidal instability indices (CII) with the 1H nuclear magnetic resonance (NMR) spectra of colombian crude oils. *Energy and Fuels* **28**, 1802–1810 (2014).
31. de Almeida, V. E. *et al.* Vis-NIR spectrometric determination of Brix and sucrose in sugar production samples using kernel partial least squares with interval selection based on the successive projections algorithm. *Talanta* **181**, 38–43 (2018).
32. Blanco, M., Coello, J., Iturriaga, H., Maspoch, S. & Pagès, J. NIR calibration in non-linear systems: different PLS approaches and artificial neural networks. *Chemom. Intell. Lab. Syst.* **50**, 75–82 (2000).
33. Durand, A., Devos, O., Ruckebusch, C. & Huvenne, J. P. Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton-viscose textiles. *Anal. Chim. Acta* **595**, 72–79 (2007).
34. 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* **98**, 5–14 (2012).
35. Brereton, R. G. Introduction to multivariate calibration in analytical chemistry. *Analyst* **125**, 2125–2154 (2000).
36. Geladi, P. Notes on the history and nature of partial least squares (PLS) modelling. *J. Chemom.* **2**, 231–246 (1988).

37. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).
38. Bramer, M. *Principles of Data Mining*. (Springer London, 2016). doi:10.1007/978-1-4471-7307-6.
39. Nascimento, M. H. C. *et al.* Determination of flash point and Reid vapor pressure in petroleum from HTGC and DHA associated with chemometrics. *Fuel* **234**, 643–649 (2018).
40. Centner, V., de Noord, O. E. & Massart, D. L. Detection of nonlinearity in multivariate calibration. *Anal. Chim. Acta* **376**, 153–168 (1998).
41. Galvan, D. *et al.* Compact low-field NMR spectroscopy and chemometrics applied to the analysis of edible oils. *Food Chem.* **365**, 130476 (2021).
42. Zareef, M. *et al.* An Overview on the Applications of Typical Non-linear Algorithms Coupled With NIR Spectroscopy in Food Analysis. *Food Eng. Rev.* **12**, 173–190 (2020).
43. Allegrini, F. & Olivieri, A. C. Linear or non-linear multivariate calibration models? That is the question. *Anal. Chim. Acta* **1226**, 340248 (2022).
44. Salehi, M., Zare, A. & Taheri, A. Artificial Neural Networks (ANNs) and Partial Least Squares (PLS) Regression in the Quantitative Analysis of Respirable Crystalline Silica by Fourier-Transform Infrared Spectroscopy (FTIR). *Ann. Work Expo. Heal.* **65**, 346–357 (2021).
45. Zhang, S., Tan, Z., Liu, J., Xu, Z. & Du, Z. Determination of the food dye indigotine in cream by near-infrared spectroscopy technology combined with random forest model. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **227**, 117551 (2020).
46. Xu, M. *et al.* Fermentation process monitoring of broad bean paste quality by NIR combined with chemometrics. *J. Food Meas. Charact.* **16**, 2929–2938 (2022).
47. Chen, T., Men, J., Zhao, M., Zhang, T. & Li, H. The spectral fusion of laser-induced breakdown spectroscopy (LIBS) and mid-infrared spectroscopy (MIR) coupled with random forest (RF) for the quantitative analysis of soil pH. *J. Anal. At. Spectrom.* **36**, 1084–1092 (2021).
48. Vapnik, V. Support-Vector Networks. *IEEE Expert. Syst. their Appl.* **7**, 63–72 (1992).
49. Scholkopf, B. *et al.* Comparing support vector machines with Gaussian kernels

to radial basis function classifiers. *IEEE Trans. Signal Process.* **45**, 2758–2765 (1997).

50. Li, H., Liang, Y. & Xu, Q. Support vector machines and its applications in chemistry. *Chemom. Intell. Lab. Syst.* **95**, 188–198 (2009).
51. Üstün, B., Melssen, W. J. & Buydens, L. M. C. Visualisation and interpretation of Support Vector Regression models. *Anal. Chim. Acta* **595**, 299–309 (2007).
52. Li, C. *et al.* Multimodal deep support vector classification with homologous features and its application to gearbox fault diagnosis. *Neurocomputing* **168**, 119–127 (2015).
53. Li, P., Dong, L., Xiao, H. & Xu, M. A cloud image detection method based on SVM vector machine. *Neurocomputing* **169**, 34–42 (2015).
54. Savio, A. & Graña, M. Local activity features for computer aided diagnosis of schizophrenia on resting-state fMRI. *Neurocomputing* **164**, 154–161 (2015).
55. Peng, D. *et al.* Detection and quantification of adulteration of sesame oils with vegetable oils using gas chromatography and multivariate data analysis. *Food Chem.* **188**, 415–421 (2015).
56. Van Gestel, T. *et al.* Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Trans. Neural Networks* **12**, 809–821 (2001).
57. Nancovska, I. Support vector regression for voltage reference elements monitoring. in *VIMS 2001. 2001 IEEE International Workshop on Virtual and Intelligent Measurement Systems (IEEE Cat. No.01EX447)* 45–50 (IEEE, 2001). doi:10.1109/VIMS.2001.924899.
58. Song, M. *et al.* Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *J. Chem. Inf. Comput. Sci.* **42**, 1347–1357 (2002).
59. Thissen, U., Pepers, M., Üstün, B., Melssen, W. J. & Buydens, L. M. C. Comparing support vector machines to PLS for spectral regression applications. *Chemom. Intell. Lab. Syst.* **73**, 169–179 (2004).
60. Souza, A. M. De, Breitkreitz, M. C., Filgueiras, P. R., Rohwedder, J. J. R. & Poppi, R. J. Experimento didático de quimiometria para calibração multivariada na determinação de paracetamol em comprimidos comerciais utilizando espectroscopia no infravermelho próximo: um tutorial, parte II. *Quim. Nova* **36**, 1057–1065 (2013).

61. de Paulo, E. H. *et al.* Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties. *Fuel* **279**, 118462 (2020).
62. MEI, H., Zhou, Y., Liang, G. & Li, Z. Support vector machine applied in QSAR modelling. *Chinese Sci. Bull.* **50**, 2291 (2005).
63. Lu, W.-Z. & Wang, W.-J. Potential assessment of the “support vector machine” method in forecasting ambient air pollutant trends. *Chemosphere* **59**, 693–701 (2005).
64. Kovalenko, I. V., Rippke, G. R. & Hurburgh, C. R. Measurement of soybean fatty acids by near-infrared spectroscopy: Linear and nonlinear calibration methods. *J. Am. Oil Chem. Soc.* **83**, 421–427 (2006).
65. Balabin, R. M. & Smirnov, S. V. Melamine detection by mid- and near-infrared (MIR/NIR) spectroscopy: A quick and sensitive method for dairy products analysis including liquid milk, infant formula, and milk powder. *Talanta* **85**, 562–568 (2011).
66. Ferrão, M. F., Mello, C., Borin, A., Maretto, D. A. & Poppi, R. J. LS-SVM: A new chemometric tool for multivariate regression. Comparison of LS-SVM and PLS regression for determination of common adulterants in powdered milk by NIR spectroscopy. *Quim. Nova* **30**, 852–859 (2007).
67. da Cunha, P. H. P. *et al.* Variable selection by permutation applied in support vector regression models. *J. Chemom.* **36**, 1–14 (2022).
68. Balabin, R. M. & Lomakina, E. I. Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. *Analyst* **136**, 1703 (2011).
69. Rodrigues, C. & Bruni, A. VIDROS AUTOMOBILÍSTICOS COMO VESTÍGIOS DE CENA DE CRIME: UMA ABORDAGEM MULTIVARIADA. *Quim. Nova* **44**, 553–560 (2021).
70. Bovens, M. *et al.* Chemometrics in forensic chemistry — Part I: Implications to the forensic workflow. *Forensic Sci. Int.* **301**, 82–90 (2019).
71. Kumar, R. & Sharma, V. Chemometrics in forensic science. *TrAC Trends Anal. Chem.* **105**, 191–201 (2018).
72. Xu, C. *et al.* A novel strategy for forensic age prediction by DNA methylation and support vector regression model. *Sci. Rep.* **5**, 17788 (2015).

73. Leal, A. L., Silva, A. M. S., Ribeiro, J. C. & Martins, F. G. Using Spectroscopy and Support Vector Regression to Predict Gasoline Characteristics: A Comparison of 1H NMR and NIR. *Energy and Fuels* **34**, 12173–12181 (2020).
74. Li, P., Ma, J. & Zhong, N. Raman spectroscopy combined with support vector regression and variable selection method for accurately predicting salmon fillets storage time. *Optik (Stuttg).* **247**, 167879 (2021).
75. Brasil, Y. L., Cruz-Tirado, J. P. P. & Barbin, D. F. Fast online estimation of quail eggs freshness using portable NIR spectrometer and machine learning. *Food Control* **131**, 108418 (2022).
76. Santos, F. D. *et al.* Characterization of crude oils with a portable NIR spectrometer. *Microchem. J.* **181**, 107696 (2022).
77. Levi, N., Karnieli, A. & Paz-Kagan, T. Airborne imaging spectroscopy for assessing land-use effect on soil quality in drylands. *ISPRS J. Photogramm. Remote Sens.* **186**, 34–54 (2022).
78. Asghari, A., Khorrami, M. K. & Garmarudi, A. B. Comparison between partial least square and support vector regression with a genetic algorithm wavelength selection method for the simultaneous determination of some oxygenate compounds in gasoline by FTIR spectroscopy. *Infrared Phys. Technol.* **105**, 103177 (2020).
79. Mohammadi, M. *et al.* Genetic algorithm based support vector machine regression for prediction of SARA analysis in crude oil samples using ATR-FTIR spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **245**, 118945 (2021).
80. Bemani, A. *et al.* Modeling of cetane number of biodiesel from fatty acid methyl ester (FAME) information using GA-, PSO-, and HGAPSO- LSSVM models. *Renew. Energy* **150**, 924–934 (2020).
81. Filgueiras, P. R. *et al.* Determination of Saturates, Aromatics, and Polars in Crude Oil by 13C NMR and Support Vector Regression with Variable Selection by Genetic Algorithm. *Energy and Fuels* **30**, 1972–1978 (2016).
82. Xu, S., Zhao, Y., Wang, M. & Shi, X. Determination of rice root density from Vis–NIR spectroscopy by support vector machine regression and spectral variable selection techniques. *Catena* **157**, 12–23 (2017).
83. Ferrer-Cid, P., Barcelo-Ordinas, J. M., Garcia-Vidal, J., Ripoll, A. & Viana, M. Multisensor Data Fusion Calibration in IoT Air Pollution Platforms. *IEEE Internet*

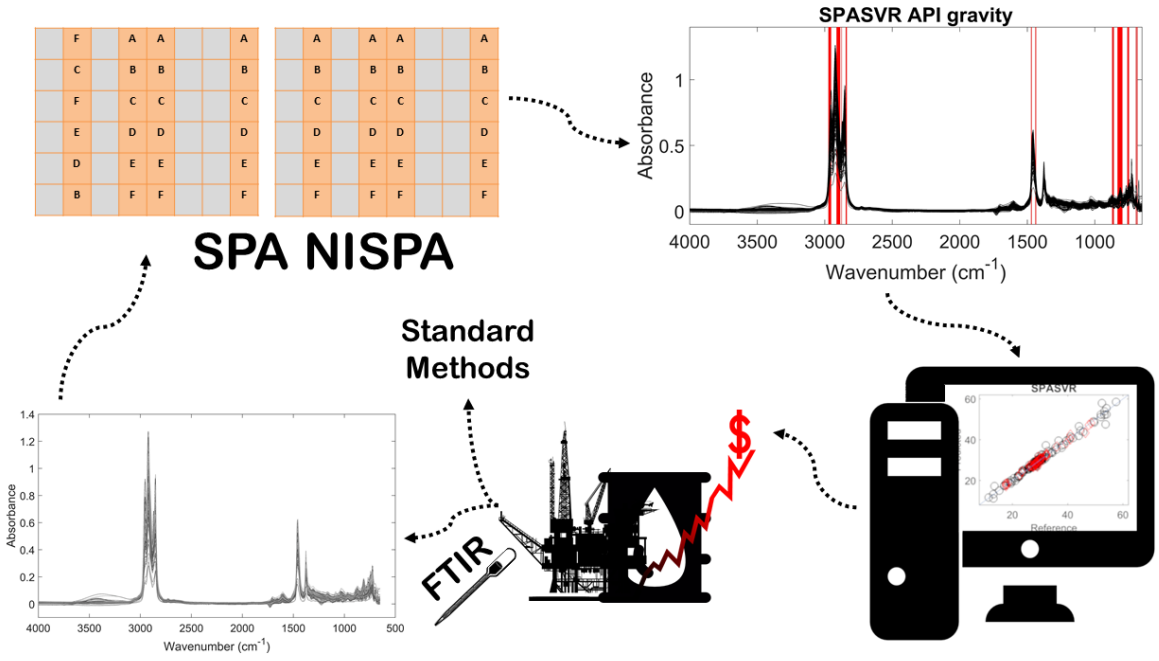
Things J. **7**, 3124–3132 (2020).

84. Khaleghi, B., Khamis, A., Karray, F. O. & Razavi, S. N. Multisensor data fusion: A review of the state-of-the-art. *Inf. Fusion* **14**, 28–44 (2013).
85. Lahat, D., Adali, T. & Jutten, C. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* **103**, 1449–1477 (2015).
86. Smolinska, A., Engel, J., Szymanska, E., Buydens, L. & Blanchet, L. General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences. in *Data Handling in Science and Technology* vol. 31 51–79 (2019).
87. Yang, B. *et al.* Rapid prediction of yellow tea free amino acids with hyperspectral images. *PLoS One* **14**, e0210084 (2019).
88. Li, Y., Zhang, J.-Y. & Wang, Y.-Z. FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng. *Anal. Bioanal. Chem.* **410**, 91–103 (2018).
89. Assis, C. *et al.* A data fusion model merging information from near infrared spectroscopy and X-ray fluorescence. Searching for atomic-molecular correlations to predict and characterize the composition of coffee blends. *Food Chem.* **325**, 126953 (2020).
90. Borràs, E. *et al.* Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Anal. Chim. Acta* **891**, 1–14 (2015).
91. Geurts, B. P. *et al.* Improving high-dimensional data fusion by exploiting the multivariate advantage. *Chemom. Intell. Lab. Syst.* **156**, 231–240 (2015).
92. Yu, S. *et al.* Qualitative and quantitative assessment of flavor quality of Chinese soybean paste using multiple sensor technologies combined with chemometrics and a data fusion strategy. *Food Chem.* **405**, 134859 (2022).
93. Floudas, N., Polychronopoulos, A., Aycard, O., Burlet, J. & Ahrholdt, M. High Level Sensor Data Fusion Approaches For Object Recognition In Road Environment. in *2007 IEEE Intelligent Vehicles Symposium* 136–141 (IEEE, 2007). doi:10.1109/IVS.2007.4290104.
94. Li, H. D., Liang, Y. Z., Xu, Q. S. & Cao, D. S. Model population analysis for variable selection. *J. Chemom.* **24**, 418–423 (2010).

CAPÍTULO 3 SELEÇÃO DE VARIÁVEIS POR PERMUTAÇÃO APLICADA EM MODELOS DE REGRESSÃO POR VETORES DE SUPORTE

Variable selection by permutation applied in support vector regression models

10.1002/cem.3444



• As técnicas de seleção de variáveis SPA e NISPA apresentaram um grande potencial quando combinadas com MIR, resultando em desempenho superior ou equivalente aos modelos sem seleção de variáveis, ao mesmo tempo em que selecionam apenas as variáveis relevantes.

• O SPA demonstrou a capacidade de determinar a propriedade da densidade API utilizando apenas 3,5% do espectro MIR, o que contraria a literatura até então.

• O NISPA surgiu da combinação do SPA e UVE, devido a necessidade de encontrar um ponto de otimização de forma fácil e rápida, obtendo resultados semelhantes em menos tempo.

Resumo

Support Vector Regression (SVR) pode ser considerada como um método de aprendizado de máquina caixa-preta. Assim, identificar a relação de causa/efeito e o sinergismo entre as variáveis mais importantes é uma tarefa difícil. Este estudo demonstra o potencial de dois métodos de seleção de variáveis por permutação - análise de sub-janela permutada (SPA) e análise de sub-janela permutada incorporada por ruído (NISPA) - para superar essa limitação. A aplicação desses dois métodos de seleção de variáveis no SVR é pouco explorada na literatura, principalmente para problemas de regressão. Os algoritmos foram aplicados em dados de FTIR (espectroscopia de infravermelho médio de transformada de Fourier) de amostras de óleo bruto para estimar a gravidade API, a viscosidade cinemática a 50°C, teor de saturados, aromáticos, resinas e teor de asfaltenos. Os resultados foram comparados com outros métodos de seleção de variáveis. SPA e NISPA forneceram os modelos mais exatos para viscosidade cinemática, saturados e teor aromático. O erro percentual médio quadrático de previsão (RMSPEP) do SPA e NISPA foram, respectivamente, 14,26% e 14,62% para viscosidade cinemática, 4,7% e 4,4% para teor de saturados, e 3,4% e 3,1% para teor aromático. Em relação à previsão da API, apesar de obter uma exatidão similar aos outros métodos de seleção, o SPA produziu um modelo mais simplificado, usando apenas 3,5% das 3351 variáveis totais, com RMSEP igual a 1,0 e R^2p a 0,981. Portanto, SPA e NISPA, além de levarem a obtenção de modelos, em geral, modelos mais rápidos, exatos e parcimoniosos, revelaram as variáveis mais importantes para a construção dos modelos SVR.

Keywords: quimiometria, análise de subjanela permutada com ruído, análise de subjanela permutada, máquina de vetores de suporte, seleção de variáveis

3.1 Introdução

A Máquina de vetores de suporte (SVM) foi desenvolvida na década de 90,[1,2] inicialmente para resolver problemas de classificação binária. Atualmente, existem algumas adaptações para este método, incluindo o método SVM aplicado a problemas de regressão, ou seja, a Regressão de vetores de suporte (SVR). Estes métodos têm sido utilizados para caracterização de óleo bruto devido ao seu bom desempenho na resolução de matrizes complexas, que frequentemente representam problemas não-lineares.[3–8] A principal característica do SVM, que lhe permite lidar com dados de matrizes complexas, é a capacidade de resolver problemas não-lineares. Isso se deve à aplicação de mapeamento de kernel combinado com uma função objetivo. O kernel transforma o conjunto de dados de entrada (X) de dimensão $(\mathrm{N}\times\mathrm{M})$, onde N é o número de amostras e M é o número de variáveis, para uma nova dimensão $(\mathrm{N}\times\mathrm{N})$, tornando a matriz de dados de entrada uma matriz quadrada, alterando a dimensão dos dados.[9] Dessa forma, a correlação entre o modelo SVM e os dados de entrada é perdida, impedindo a identificação da relação de causa/efeito entre a resposta química (variáveis) e a propriedade prevista; assim, o SVR é conhecido como uma caixa preta.[10]

Üstün *et al.*[9]propuseram dois métodos diferentes para resolver este problema usando conjuntos de dados de uma simulação e de ressonância magnética nuclear (NMR, do inglês *nuclear magnetic resonance*) de tecido cerebral. O primeiro método utilizou o produto dos dados de entrada $\mathrm{X}(\mathrm{N}\times\mathrm{M})$ e a matriz de kernel (NxN), obtendo uma matriz de correlação $\mathrm{R}(\mathrm{N}\times\mathrm{M})$, que demonstra a contribuição de cada variável de entrada para a construção do kernel. Os valores contidos em R variam de -1 a 1, considerando que quanto mais próximo de zero menos contribui; e quanto mais próximo de 1 mais contribui positivamente, ou -1, negativamente. O segundo método utilizou os vetores de suporte do modelo SVR, $\alpha(1\mathrm{xN})$, com a matriz de entrada $\mathrm{X}(\mathrm{N}\times\mathrm{M})$, para obter um vetor de peso, vetor p. Este vetor tem a dimensão $(1\times\mathrm{M})$, que revela as variáveis mais importantes do modelo SVR. Ambos os métodos mostraram resultados promissores ao selecionar variáveis importantes e quebrar a caixa preta do SVR. No entanto, também seleciona variáveis não importantes, incluindo ruído adicionado nas amostras artificiais.

Por outro lado, Postma *et al.*[11] sugerem uma maneira diferente de abrir a caixa preta do kernel, fazendo uma transformação mais transparente. Amostras artificias foram aplicadas a cinco conjuntos de dados, onde dois eram dados sintéticos e os outros três eram reais. A aplicação de kernel combinado com mínimo quadrados parciais (KPLS, do inglês *Kernel-Partial Least Squares*) e SVR tinha o objetivo de traçar a trajetória e descobrir a importância de cada variável. O método mostrou grande potencial para quebrar a caixa preta, mas não completamente.

Outra maneira de analisar a relação causa/efeito é determinar a influência das variáveis na exatidão do modelo.[12] Métodos de seleção de variáveis podem ser usados para gerar modelos mais rápidos e parcimoniosos devido à redução do número de variáveis. O sinergismo em intervalo em SVR (siSVR)[13,14] é um procedimento de seleção de variáveis que separa o conjunto de dados espectrais em intervalos e constrói modelos SVR com base na combinação de dois ou mais intervalos. Assim, os intervalos mais sinérgicos, que fornecem um resultado mais exato, podem ser determinados.

No entanto, este método é limitado a testes em intervalos, o que pode resultar na seleção de variáveis sem sinal analítico ou com informações redundantes. Para resolver esse problema, as variáveis podem ser selecionadas individualmente. Como exemplo, o método do algoritmo genético (GA) utiliza o conceito de seleção natural para selecionar individualmente as variáveis mais importantes.[15,16] O GA começa com uma população inicial, onde cada "indivíduo" possui uma cadeia cromossômica, codificada em 0 e 1, indicando quais variáveis serão consideradas em seu modelo. Assim, os indivíduos mais adaptados são aqueles que obtêm os parâmetros mais exatos na validação cruzada, os menos adaptados são eliminados e substituídos por versões modificadas dos mais ‘adaptadas’ - uma ação chamada mutação. Este procedimento de GA é repetido várias vezes para identificar as variáveis com a maior frequência de seleção, ou seja, as variáveis mais significativas. Essas variáveis devem conter informações relevantes sobre a propriedade modelada de interesse.

Filgueiras *et al.*[17] aplicaram GA-SVR, SVR combinado com GA, em espectros de ressonância magnética nuclear (RMN de ^{13}C) para estimar o teor de saturados, aromáticos e polares no petróleo bruto. Os autores selecionaram regiões de carbono aromático e parafínico, construindo os modelos GA-SVR. Assim, os autores obtiveram os melhores resultados para GA-SVR em comparação com modelos de mínimos quadrados parciais (GA-PLS). No entanto, o GA tem a desvantagem de selecionar

variáveis que podem levar a um mínimo local; para evitar isso, é possível executar o GA diversas vezes com o objetivo de conseguir obter modelos diferentes, que fujam no mínimo local, e analisar a frequência que uma variável é selecionada.

Outra seleção de variáveis é o algoritmo de busca angular e fator de inflação de variância (ASA-VIF). Este método de seleção de variáveis tem como principal objetivo eliminar variáveis correlacionadas, permanecendo apenas aquelas que não possuem informações compartilhadas.[18] Os espectros químicos possuem um vasto número de variáveis que possuem informações correlacionadas.[19] Assim, o número de variáveis pode diminuir drasticamente. O SVR tem uma limitação de ranqueamento devido ao número de amostras, geralmente muito menor do que o número de variáveis.[18] Dessa forma, a redução drástica de variáveis proporciona uma otimização da modelagem do SVR, associada ao ASA. O algoritmo ASA é aplicado inicialmente para obter o cosseno do ângulo entre as variáveis. As variáveis menos correlacionadas (menores cossenos) são preservadas no conjunto de amostras. Depois, o VIF é aplicado para eliminar variáveis que ainda apresentam alta multicolinearidade. Em seguida, subconjuntos de amostras aleatórias são criados para calcular o menor erro quadrático médio de validação cruzada (RMSECV).[18] No entanto, o ASA-VIF é demorado e de processamento difícil, principalmente para matrizes grandes, devido ao cálculo de correlação e multicolinearidade realizado.

A análise de subjanela permutada (SPA)[20] determina a importância de cada variável por meio de uma simples permutação.[21] O método destrói a informação de uma variável trocando seus conteúdos e posteriormente constrói o modelo SVR. Os modelos das variáveis originais e permutadas são comparados usando o teste U, que você pode ver com melhores detalhes no **ANEXO C**, o que permite determinar se a variável permutada pode influenciar significativamente a previsão da propriedade, o que possibilita analisar a importância de cada variável separadamente.

SPA já foi aplicada em SVM para fins de classificação.[22] Uma de suas desvantagens é a dificuldade em determinar o menor número de variáveis que gerarão o melhor modelo. Este problema pode ser resolvido pelo método de análise de subjanela permutada com ruído incorporado (NISPA), adicionando ruído. Wang *et al.*[23] criaram o NISPA usando uma modificação baseada na eliminação de variáveis não informativas (UVE). A principal diferença entre os métodos SPA e NISPA é o procedimento baseado em UVE para excluir variáveis não informativas. Os autores foram capazes de selecionar as variáveis mais importantes, obtendo os melhores

resultados do que por outros métodos de seleção de variáveis.

Existem vários métodos de seleção de variáveis aplicados em SVR, como CARS-SVR e UVE-SVR em dados de espectroscopia NIR,[15] algoritmo vagalume e GA-SVR em dados espectrais de ultravioleta (UV),[16] ASA-VIF-SVR em espectroscopia no infravermelho médio (MIR), no infravermelho próximo (NIR) e RMN de ^1H,[18] e GA-SVR em espectroscopia de RMN.[17] No entanto, SPA e NISPA em SVR ainda são pouco aplicados. Propomos utilizar os métodos SPA e NISPA em modelos SVR para prever propriedades físico-químicas do petróleo bruto. O potencial dos procedimentos de permutação foi comparado com métodos modernos e estabelecidos, como GA, intervalos e ASA.

3.2 Experimental

Neste artigo, o modelo SVR foi comparado a seis métodos de seleção de variáveis aplicados no SVR: SPA-SVR, NISPA-SVR, GA-SVR, iSVR, siSVR e ASA-SVR. Esses modelos foram aplicados para estimar seis propriedades físico-químicas do petróleo bruto: densidade API, viscosidade cinemática a 50 ºC e teor de saturado, aromático, resina e asfaltenos (SARA do inglês *Saturation, Aromatic, Resins and Asphaltene*) no petróleo bruto. Os modelos foram avaliados quanto à sua capacidade de estimar as propriedades físico-químicas.

3.2.1 Amostras

Neste estudo, um total de 200 amostras de petróleo bruto da bacia sedimentar da costa brasileira foram utilizadas. Todas as propriedades físico-químicas foram determinadas seguindo métodos padrão listados na **Tabela 3.1**, a distribuição amostral pode ser visualizada na **Figura 3.1**. As amostras foram divididas em conjuntos de calibração (154) e previsão (46) pelo algoritmo de Kennard-Stone, respeitando os intervalos das propriedades físico-químicas.

Tabela 3. 1 Número de amostras nos conjuntos de calibração e de previsão para cada modelo de propriedade físico-química

Propriedade	*Quant. Conjunto Calibração*	*Quant. Conjunto Teste*	*Faixa de medida*
Densidade API	154	40	11,4-57,5
Viscosidade 50ºC (ν)	154	40	0,6 – 97,71 mm^2s^{-1}
Saturado	145	40	29,2 - 90,6 wt%
Aromático	145	40	7,0 - 37,20 wt%
Resina	141	37	0 - 42,8 wt%
Asfalteno	144	37	0 - 10,3 wt%

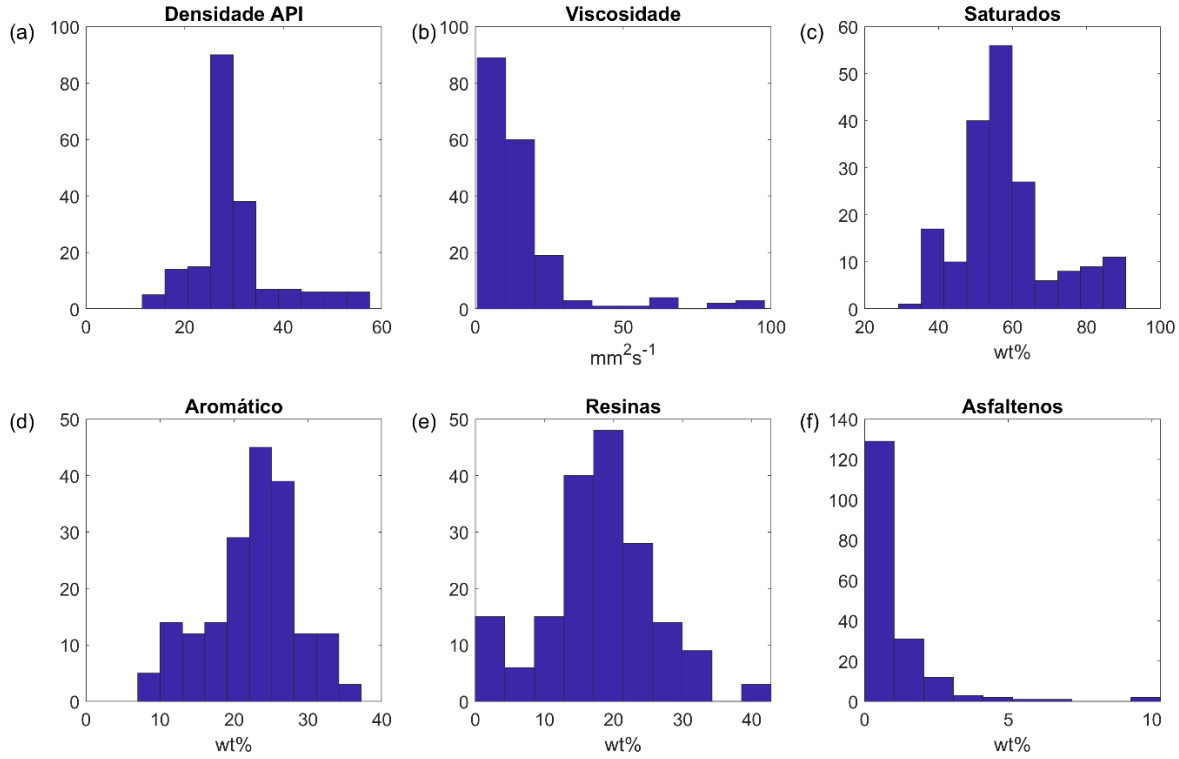


Figura 3. 1 Histograma da distribuição amostral. (Fonte: Elaboração própria)

3.2.2 Propriedade Fisico-Quimicas

Um viscosímetro digital Anton Paar (modelo Stabinger SVM 3000) com um limite de detecção de 0,0002 g cm^{-3} a 20 ºC foi usado para determinar a gravidade API e a viscosidade cinemática a 50 ºC, de acordo com a ISO 12185,[24] e ASTM D7042,[25] respectivamente. Cinco mililitros de amostra foram adicionados à célula de medição do viscosímetro e as análises foram realizadas em duplicata a 40 ºC e 50 ºC.

Para determinar a densidade API, utilizou-se triplicatas, os resultados de densidade foram convertidos para uma densidade equivalente a 20 ºC e aplicados na Equação 3.1, onde ρ é a densidade a 60 ºF.[24]

$$API = \frac{141,5}{\rho} - 131,5 \quad \text{.....Equação 3.1}$$

O teor de asfaltenos foi determinado de acordo com a ASTM D6560[26] e o teor de saturados, aromáticos e resinas foi analisado por Cromatografia com Fluido Supercrítico/Cromatografia em Camada Delgada com Detector de Ionização por Chama (SFC/TLC-FID).[27,28]

3.2.3 FTIR

Espectroscopia de infravermelho médio com transformada de Fourier (FTIR do inglês *Fourier transform mid infrared spectroscopy*) foram realizados em um espectrômetro, modelo Spectrum 400 da PerkinElmer, no Centro de Competência em Química do Petróleo da Universidade Federal do Espírito Santo. Os espectros foram adquiridos em triplicata, na região de infravermelho médio (4000 a 650 cm^{-1}), com 32 varreduras e uma resolução de 1 cm^{-1}. A análise durou aproximadamente 10 minutos e foi necessária 1 mL de amostra, após obteve a média das triplicatas e utilizou no trabalho.

3.2.4 Quimiometria

Uma série de pré-tratamentos foi testada e aplicada na calibração do SVR. O conjunto de previsão foi usado para determinar a exatidão e os parâmetros de desempenho dos modelos. A modelagem foi feita usando o software MATLAB versão R2013a e utilizando o LibSVM[29] adaptado para grade de pesquisa com códigos próprios.

3.2.4.1 Tratamento dos Dados

As amostras foram separadas em calibração e teste utilizando o método k-fold,

antes da calibração, os espectros de FTIR foram pré-processados para reduzir variações indesejáveis. Os seguintes métodos de pré-tratamento foram testados para cada propriedade: primeira derivada, airPLS, Correção de Espalhamento Multiplicativo (MSC do inglês *Multiplicative Scatter Correction*), Variável Normalizada Padrão (SNV do inglês *Standard Normal Variate*), normalização euclidiana e centralização média. A primeira derivada e o airPLS também foram combinados com MSC, SNV, normalização e centralização. Isso resultou em um total de 18 pré-tratamento testados, além da ausência. A partir dessas combinações, os modelos foram comparados.

Antes da modelagem da viscosidade cinemática, os valores medidos foram aplicados à **Equação 3.2**.[30]

$$\boldsymbol{v_{TR} = log(log(v + 0.7))} \text{ .....} \textbf{Equação 3.2}$$

onde $\boldsymbol{v_{TR}}$ e v são os valores de viscosidade cinemática transformada e medida, respectivamente.

3.2.4.2 SVR

SVM é um método de aprendizado de máquina, desenvolvido por Vapnik.[1] O SVM trabalha mapeando vetores de entrada (amostras) com o objetivo de encontrar um hiperplano que separe as classes sem erros, usando o princípio da minimização do risco, obtendo o Hiperplano de Separação Ótima (OSH, do inglês Optimal Separating Hyperplane). O OSH em uma aplicação para regressão, SVR, é obtido pela **Equação 3.3**.

$$\boldsymbol{y_i = w\phi(x_i) + b,} \text{ .....} \textbf{Equação 3.3}$$

onde $\boldsymbol{y_i}$ é o valor quantitativo, x_i é o vetor de amostra e $\boldsymbol{w}$ é o vetor de peso.[17] Nesta equação, uma constante de tolerância é aplicada, "ε", admitindo possíveis erros. A constante C atribui um peso aos erros da função $(\xi i, \xi * \geq 0)$,[5] de acordo com a equação:

$$Min\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad \textbf{Equação 3.4}$$

$$\text{Sujeito a} \begin{cases} y_i - f(\phi(\mathbf{x}_i), \mathbf{w}) \leq \varepsilon + \xi_i^* \\ f(\phi(\mathbf{x}_i), \mathbf{w}) - y_i \leq \varepsilon + \xi_i \\ \xi_i, \xi* \geq 0, i = 1, \ldots, n \end{cases} \quad \textbf{Equação 3.5}$$

As constantes C e ϵ precisam ser otimizadas durante a construção do modelo. Neste estudo, foi utilizado grade de pesquisa (*Gried Search*) de tamanhos 8 e 20. O problema da **Equaçaõ 3.5** pode ser resolvido aplicando a multiplicação de Lagrange, como demonstrado por Smola A. J.,[31] cujo resultado é a **Equação 3.6**.

$$f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b \quad \textbf{Equação 3.6}$$

onde $K(\mathbf{x}_i, \mathbf{x})$ é a função de kernel, responsável pelo espaço de alta dimensão no modelo SVM. Existem alguns tipos de função de kernel.[32] Neste estudo, utilizamos a função de base radial (RBF, do inglês *Radial-Basis Function*) e a sigmodal. Na grade de pesquisa foram otimizados os seguintes parâmetros, C, ϵ e γ, da função de kernel.

3.2.4.3 Validação do modelo

Os modelos foram avaliados pelo coeficiente de determinação (R^2), que representa uma medida estatística da concordância entre os valores medidos pelos testes experimentais padrão e os valores previstos pelos modelos de regressão. Ele varia de 0 a 1, e quanto mais próximo o valor estiver de 1, melhor o modelo se ajusta aos dados experimentais. A determinação desse parâmetro pode ser calculada de acordo com a **Equação 3.7**.

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \textbf{Equação 3.7}$$

Em que n é o número de amostras, y_i são os valores medidos, $\hat{y}_i$ é o valor estimado por um modelo de regressão e $\bar{y}$ é a média dos valores medidos. Os modelos também foram avaliados pelos parâmetros RMSEC, RMSECV e RMSEP, que

consistem, respectivamente, no erro quadrático médio da calibração, da validação cruzada e da predição. Esses parâmetros podem ser estimados de acordo com a **Equação 3.8**.

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \textbf{Equação 3.8}$$

Além disso, o erro médio quadrático da raiz da porcentagem de predição (RMSPE do inglês *root mean square error of prediction percentage*) foi usado para comparar os modelos de viscosidade cinemática a 50°C. Utilizamos o RMSPE, em vez do RMSE, para mitigar o erro devido à alta variabilidade da variável. Os erros de validação cruzada (RMSPECV) e de predição (RMSPEP) também foram calculados de acordo com a **Equação 3.9**.

$$\boldsymbol{RMSPE} = \mathbf{100}.\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad \textbf{Equação 3.9}$$

3.2.4.4 Seleção de variáveis

A seleção de variáveis foi usada para determinar quais variáveis são mais representativas na modelagem e desenvolver modelos mais parcimoniosos.[15,31] Além disso, os métodos de seleção de variáveis permitiram descobrir as variáveis mais importantes em uma calibração quebrando a caixa-preta do SVM. Em nosso estudo, quatro métodos de seleção foram aplicados e comparados com os métodos SPA e NISPA. A regressão por vetores de suporte por intervalo sinérgico (siSVR do inglês *synergy interval support vector regression*) foi realizado usando 10 intervalos utilizando combinando de 2 a 9 e foi utilizando a regressão por vetores de suporte em intervalos (iSVR do inglês *interval support vector regression*) utilizando somente um intervalo.

O GA-SVR, combinação de GA com SVR, foi usado em duas etapas diferentes: A primeira, utilizando uma população inicial de 200 e 100 gerações, sendo executados 100 vezes. Em seguida, com base nesses 100 modelos, uma análise de frequência das variáveis foi usada como critério para construir os novos modelos. O modelo com o menor valor de RMSECV na segunda etapa foi selecionado.

O ASA visa remover variáveis altamente correlacionadas. Inicialmente, a matriz X foi centralizada na média e o coeficiente de correlação foi estimado. Em seguida, o VIF das variáveis foi calculado. Estabelecemos que, se o VIF fosse menor que 50, as variáveis não estavam relacionadas e permaneciam no conjunto de dados. Caso contrário, se fosse maior que 50, as variáveis estavam totalmente correlacionadas e, portanto, eram eliminadas.[18]

3.2.4.5 Análise de permutação de subjanela (SPA)

O SPA é um método de seleção de variáveis que determina a importância de cada através de pequenos conjuntos de variáveis, conhecidos como subjanelas.[20] No SPA, a primeira etapa consiste em criar uma subjanela com Q variaveis, através de uma amostragem Monte Carlo. Esta subjanela representada na **Figura 3.2** pelo primeiro bloco, no qual cada linha se refere a uma amostra e cada coluna uma variável. Além disso, as colunas mais escuras são as colunas selecionadas pelo Monte Carlo. A segunda etapa é desenvolver o modelo para cada subjanela, obtendo o seu RMSECV, ou ErroN (Erro Normal), em seguida, é repetido o processo anterior Q vezes, com a diferença que cada vez uma variavel é permutada, destruindo sua informação e obtendo um novo RMSECV, ou ErroP (Erro Permutado), na **Figura 3.2**, a permutação da variável é representando pelo embaralhamento das letras de A a F.

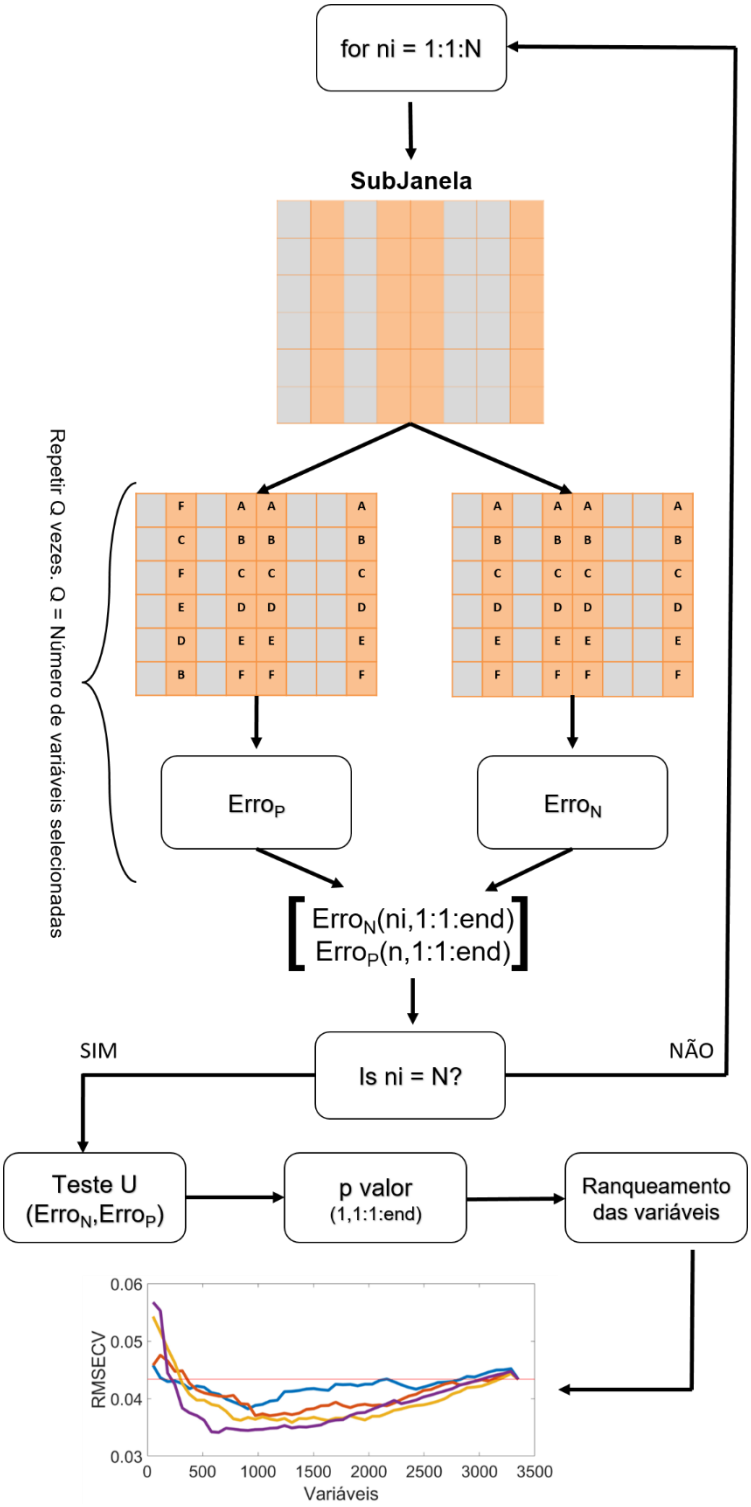


Figura 3. 2 Funcionamento da Análise de Subjanela Permutada (SPA). (Fonte: Elaboração própria)

As etapas subsequentes foram repetidas múltiplas vezes, permitindo a reutilização de uma variável em duas ou mais iterações, enquanto os valores de ErroN e ErroP foram acumulados e analisados por meio de um teste U,[23] para cada variável. Esse procedimento teve como objetivo obter um valor p, revelando o impacto da

destruição da informação de cada variável na exatidão do modelo. Neste estudo, o número de permutações foi definido em 10.000, determinado após diversos testes de modelo que variaram de 500 a 100.000 permutações. Devido à aleatoriedade intrínseca do processo, essa otimização mostrou-se necessária. O teste U, ou teste de Mann-Whitney, determina se a diferença entre dois resultados é estatisticamente significativa.[9]

Utilizando o p valor do teste U é construído um ranqueamento das variáveis separando da mais importante para a menos importante, através desse ranking, 50 novos modelos foram construídos. O número total de modelos foi testado, variando entre – 10, 20, 50, 100 e 200 – e 50 se mostrou mais vantajoso. No SPA, o primeiro modelo contém todas variáveis, o segundo remove 66 variáveis menos importante e o terceiro modelo remove mais 66 e assim por diante. Um gráfico de RMSECV versus número de variáveis é utilizado para determinar o número ótimo de variáveis, último gráfico da **Figura 3.2**. Além disso, otimizou, também, o número de variáveis iniciais (Q), utilizando falores de 12, 24, 48, 96 e 192, para obter o melhor resultado de SPA.

3.2.4.6 Ruído incorporado a análise de permutação de subjanela (NISPA)

Para obter o número ótimo de variáveis, foram gerados 50 novos modelos no final do SPA, o que exigiu um longo tempo de processamento. O NISPA foi uma alternativa ao SPA,[15] adicionando ruído aos dados, seguindo os princípios do UVE, cuja metodologia já está bem estabelecida. O NISPA adiciona ruído na matriz espectral, no início da seleção de variáveis, assim como o UVE faz.[33] O mesmo número de variáveis é adicionado, dobrando o número de colunas, ou seja, teremos uma parte informação relevante (espectro original) e outra ruído irrelevante. O próximo passo foi semelhante ao mostrado na **Figura 3.2**, modificando apenas o último passo. O NISPA utiliza o p valor, obtido pelo ruído, com melhor desempenho como faixa de corte para remover as variáveis menos importantes.[33]

3.3 Resultados e discussão

Neste trabalho, analisamos a eficácia dos métodos de seleção por permutação, SPA-SVR e NISPA-SVR, juntamente com outros métodos de seleção de variáveis,

para prever seis propriedades do petróleo: densidade API, viscosidade cinemática a 50 °C, e teores de saturados, aromáticos, resinas e asfaltenos. Os modelos de seleção de variáveis por permutação apresentaram resultados semelhantes ou superiores aos modelos dos outros métodos de seleção.

3.3.1 Infravermelho

A **Figura 3.3** mostra os espectros infravermelhos de amostras de petróleo. As amostras apresentaram um perfil químico semelhante, com bandas de absorção em regiões características de compostos hidrocarbonetos. As bandas em 3450 e 1650 cm^{-1} correspondem, respectivamente, ao estiramento e à vibração das ligações químicas O–H. Na faixa de 3100 a 3000 cm^{-1}, podemos ver as bandas de baixa absorção atribuídas aos estiramentos das ligações =C–H. Os picos de absorbância máxima, na faixa de 3000 a 2800 cm^{-1}, são devidos ao estiramento das ligações H–C dos grupos CH_2 em 2922 e 2852 cm^{-1} e dos grupos CH_3 em 2953 cm^{-1}. As bandas de baixa absorção em 1603 cm^{-1} são de vibrações aromáticas de ligações C=C. As bandas entre 1454 e 1375 cm^{-1} reforçam a presença de alcanos, CH_3 e grupos alifáticos. As bandas nas regiões de 1300 a 1100 cm^{-1} e de 900 a 720 cm^{-1} estão relacionadas às vibrações de compostos aromáticos e com deformações angulares fora do plano das ligações C–H.[34,35]

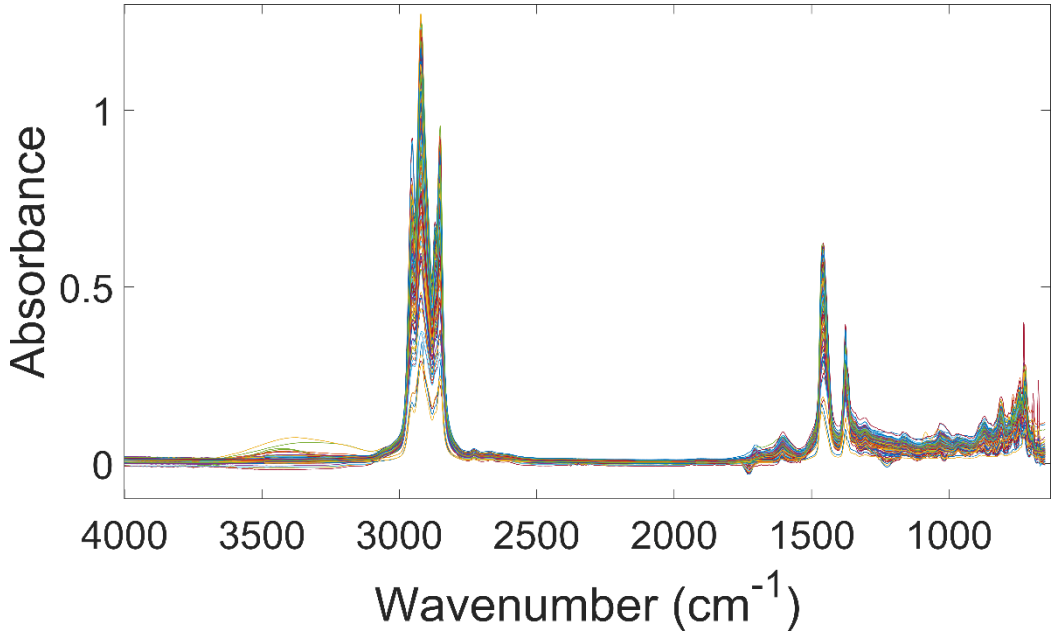


Figura 3. 3 Espectro MIR das amostras sem pré-tratamento.

3.3.2 Densidade API

A densidade API é a propriedade mais difundida para a caracterização do petróleo bruto na indústria.[36] Para esta propriedade, o melhor pré-tratamento foi primeira derivada e SNV e nenhum método de seleção de variáveis antes da aplicação do SVR. Para os modelos com seleção de variáveis, utilizamos o airPLS, seguido de normalização como pré-tratamento. A **Figura 3.4** mostra os valores medidos e previstos da densidade API para todos os modelos desenvolvidos. Nota-se que com exceção do ASA-VIF todas as seleções de variáveis obtiveram um bom ajuste na linha de referência. Todos apresentaram um ajuste linear elevado, com valores de R^2p acima de 0,95, como mostrado na **Tabela 3.2**.

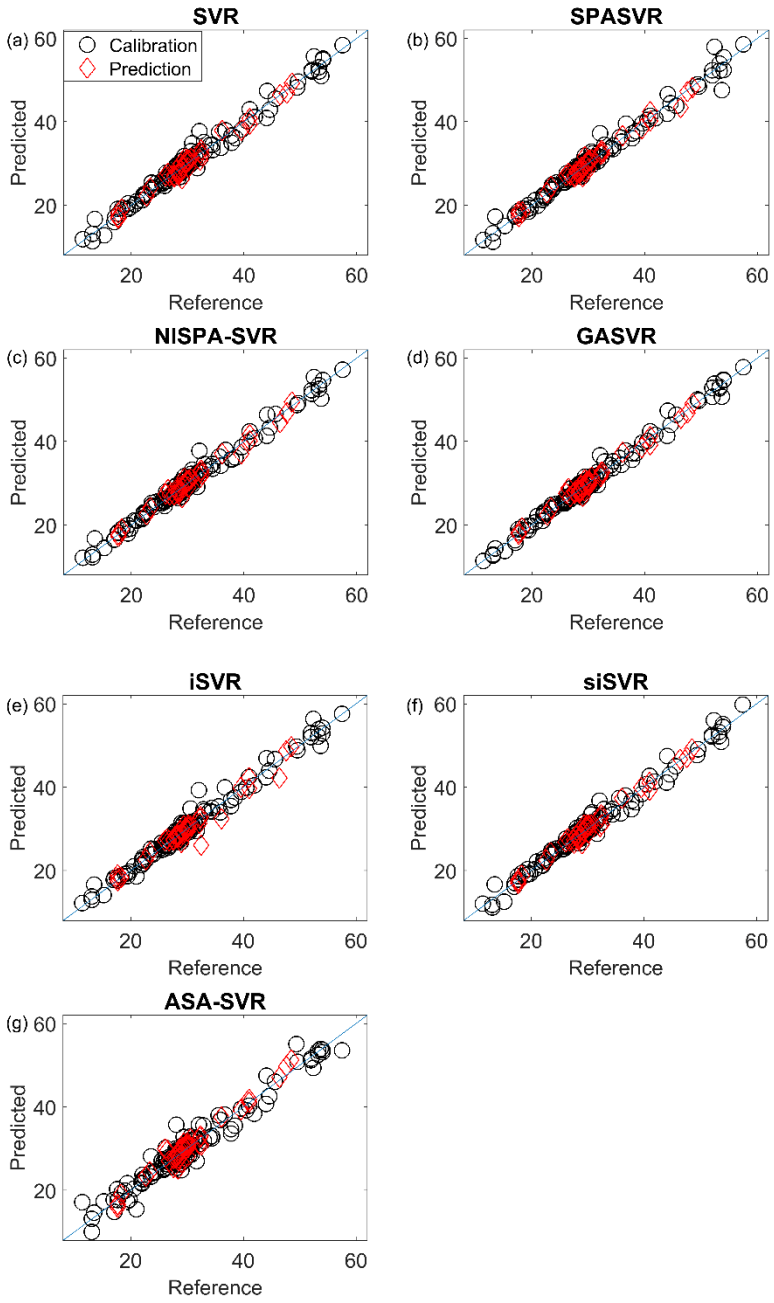


Figura 3. 4 Gráfico de valores Medido e Predito para densidade API. Sem seleção de variáveis (a), SPA (b), NISPA (c), GA (d), iSVR (e), siSVR (f) e ASA-VIF (g).

Entre os modelos destacados, o SPA-SVR alcançou um RMSEP igual a 0,9 API e um R^2p igual a 0,983, resultados semelhantes aos dos outros modelos, mas utilizando cerca de 117 variáveis, o que corresponde a 3,5% do total. NISPA-SVR e GA-SVR apresentaram RMSEP igual a 0,9 e R^2p igual a 0,982 e 0,986, respectivamente, e ambos selecionaram mais de 1000 variáveis, sem selecionar uma região específica. Os modelos ASA e iSVR apresentaram valores de RMSEP mais altos, próximos de 1,50 API.

Tabela 3. 2 Parâmetros de avaliação para densidade API.

Model	Number of Variables	RMSECV	RMSEP	R^2cv	R^2p
SVR	3351	1.2	1.0	0.979	0.983
SPA-SVR	117	1.2	0.9	0.981	0.983
NISPA-SVR	1870	1.1	1.0	0.982	0.982
GA-SVR	1001	1.0	0.9	0.987	0.986
iSVR	335	1.3	1.6	0.979	0.955
siSVR	3016	1.2	0.9	0.980	0.984
ASA-SVR	14	1.9	1.5	0.952	0.967

Rainha *et al.*[36] determinaram densidade API por FTIR, com seleção de variáveis por intervalo de mínimos quadrados parciais (iPLS), mínimos quadrados parciais de intervalo de sinergia (siPLS, do inglês *synergy interval partial least squares*), amostragem adaptativa competitiva reponderada (CARS-PLS) e projeções ortogonais para estruturas latentes (OPLS). Os autores obtiveram um modelo OPLS com RMSEP de 1,0 e R^2p de 0,982, que é semelhante ao nosso resultado. Ainda de acordo com os mesmos autores, a seleção de variáveis contribuiu pouco para melhorar a exatidão dos modelos, selecionando um número grande de variáveis de diferentes regiões, indicando que a informação química importante para prever a densidade API está espalhada por todo o espectro FTIR.[36] Em nosso estudo, o SPA não apenas selecionou um baixo número de variáveis, mas também revelou a importância de três regiões, como mostrado na **Figura 3.5A**.

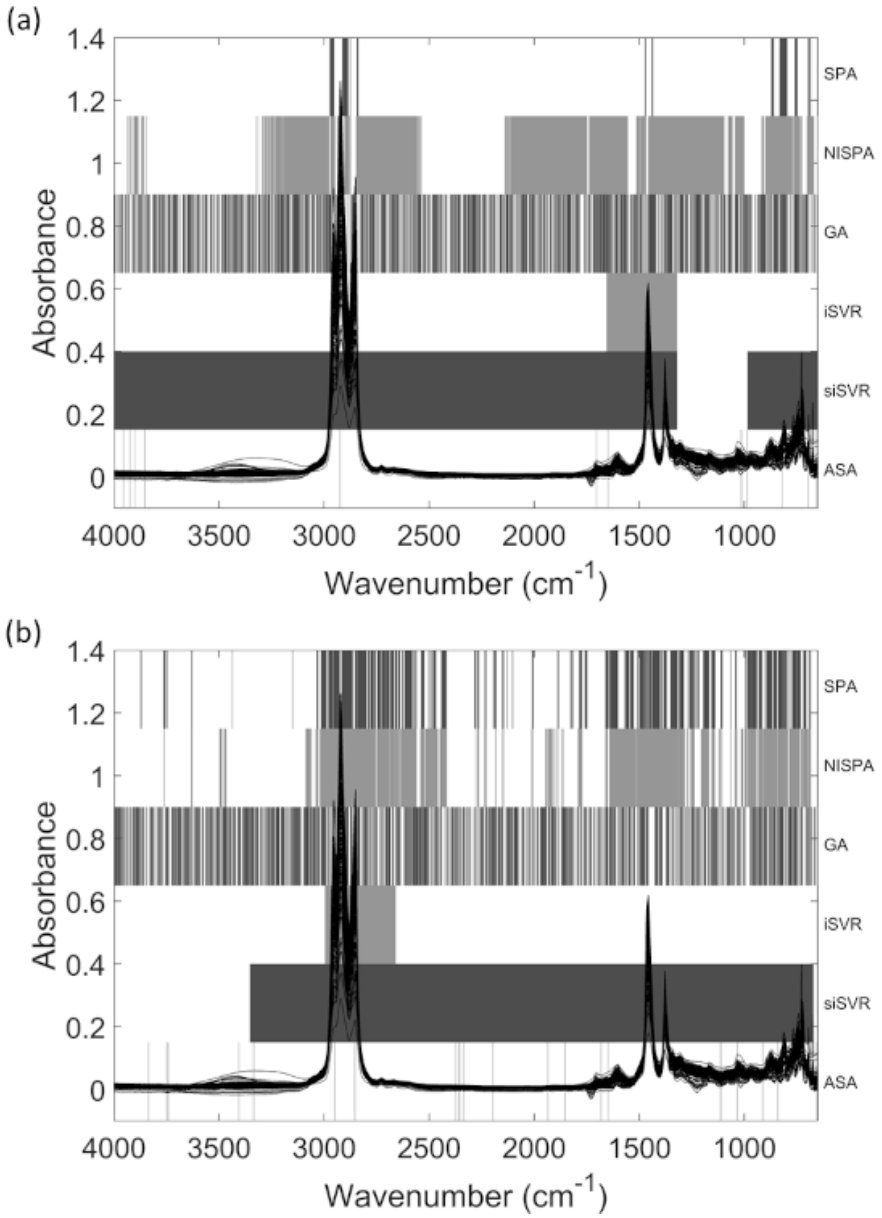


Figura 3. 5 Variáveis selecionadas para densidade API (a) e viscosidade cinemática (b).

As regiões selecionadas pelo SPA foram a região de parafina (3000–2800 cm^{-1}); as bandas em 1454 e 1375 cm^{-1}, resultantes de alcano CH_3 e alifático CH_3 e a região de 900 a 650 cm^{-1}, resultante de substituição no benzeno (**Figura 3.5A**),[37] o que sugere que essas regiões são mais relevantes para prever densidade API no petróleo. Corroborando nossos resultados, Paulo *et al*.[38] também selecionaram a região de parafina usando espectros de ressonância magnética nuclear (NMR de ^{1}H e ^{13}C) para prever a gravidade API.

3.3.3 Viscosidade cinemática a 50ºC

Outra propriedade importante no óleo investigado em nosso estudo foi a viscosidade cinemática, sendo o melhor pré-tratamento a primeira derivada, seguida pelo MSC. A **Figura 3.6** apresenta os dados de viscosidade cinemática medidos e previstos, sem transformação para modelagem. As amostras dentro da faixa de 0 a 50 mm²/s estão próximas da linha de referência, enquanto as outras amostras, dentro da faixa de 50 a 60 mm²/s, estão consideravelmente distantes. No entanto, os modelos que utilizaram permutação melhor previram a propriedade nesta faixa, como pode ser visto nos parâmetros de avaliação na **Tabela 3.3**.

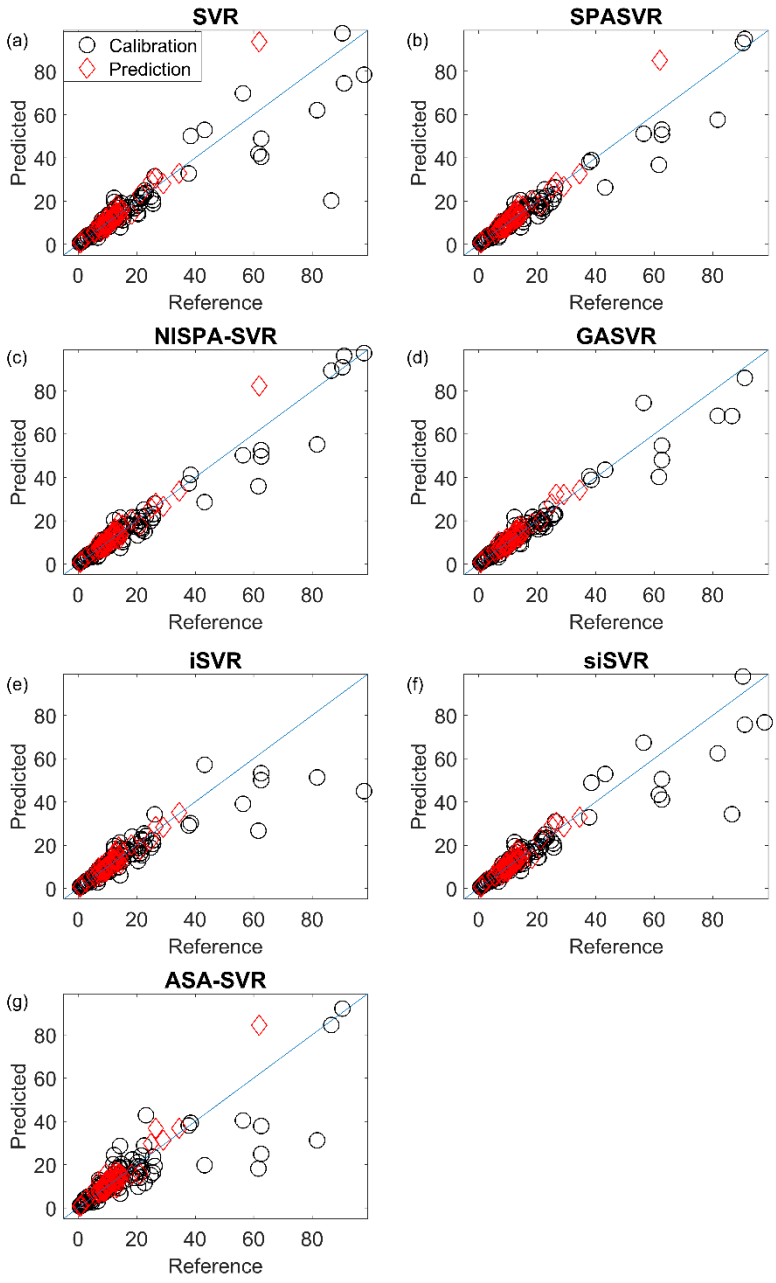


Figura 3. 6 Gráfico de valores Medido e Predito para viscosidade cinemática. Sem seleção de variáveis (a), SPA (b), NISPA (c), GA (d), iSVR (e), siSVR (f) e ASA-VIF (g).

A seleção de variáveis com permutação foi o único método que proporcionou uma melhoria em comparação com o modelo SVR puro (sem seleção de variáveis), alcançando um RMSPEP de aproximadamente 14,3% e 14,6% para SPA-SVR e NISPA-SVR, respectivamente. Vale ressaltar que o RMSPEP foi utilizado apenas para a viscosidade cinemática, em virtude da ampla faixa de variação desta propriedade. No entanto, com base no teste t,[39] não há diferença significativa, indicando que é

possível identificar regiões importantes para o modelo; embora isso não implique em uma melhoria na exatidão.

Tabela 3. 3 Parâmetros de avaliação para viscosidade cinemática.

Model	Number of Variables	RMSPECV	RMSPEP	R²cv	R²p
SVR	3351	19.9	15.8	0.975	0.988
SPA-SVR	975	19.4	14.3	0.964	0.989
NISPA-SVR	1421	10.0	14.6	0.964	0.988
GA-SVR	989	17.4	16.1	0.978	0.988
iSVR	335	27.1	16.3	0.967	0.988
siSVR	2346	17.6	16.2	0.981	0.988
ASA-SVR	21	33.6	23.2	0.930	0.975

Semelhante aos modelos de gravidade API, três regiões foram mais frequentemente selecionadas para os modelos de viscosidade cinemática SPA e NISPA, como pode ser visto na **Figura 3.5B**. Rocha *et al.*[4] utilizaram SVR para prever viscosidade cinemática com dados de cromatografia gasosa/espectrometria de massa (GC-MS) e obtiveram um R²p de 0,58. Filgueiras *et al.*[6] utilizaram SVR em FTIR e obtiveram R²p de 0,8584. Em nosso estudo, obtivemos um R²p maior que 0,98 com os modelos SVR e SPA-SVR.

3.3.4 Teor de saturados

Para a previsão do teor de saturados, o pré-tratamento SNV foi o melhor para todos os modelos. A **Figura 3.7** mostra os valores medidos e previstos para o teor de saturados. Em todos os modelos, algumas amostras estão longe da linha de referência, principalmente entre 40-47%, o que pode indicar valores discrepantes. O teste de Grubbs foi usado para identificar valores discrepantes, mas nenhuma amostra foi selecionada.[17]

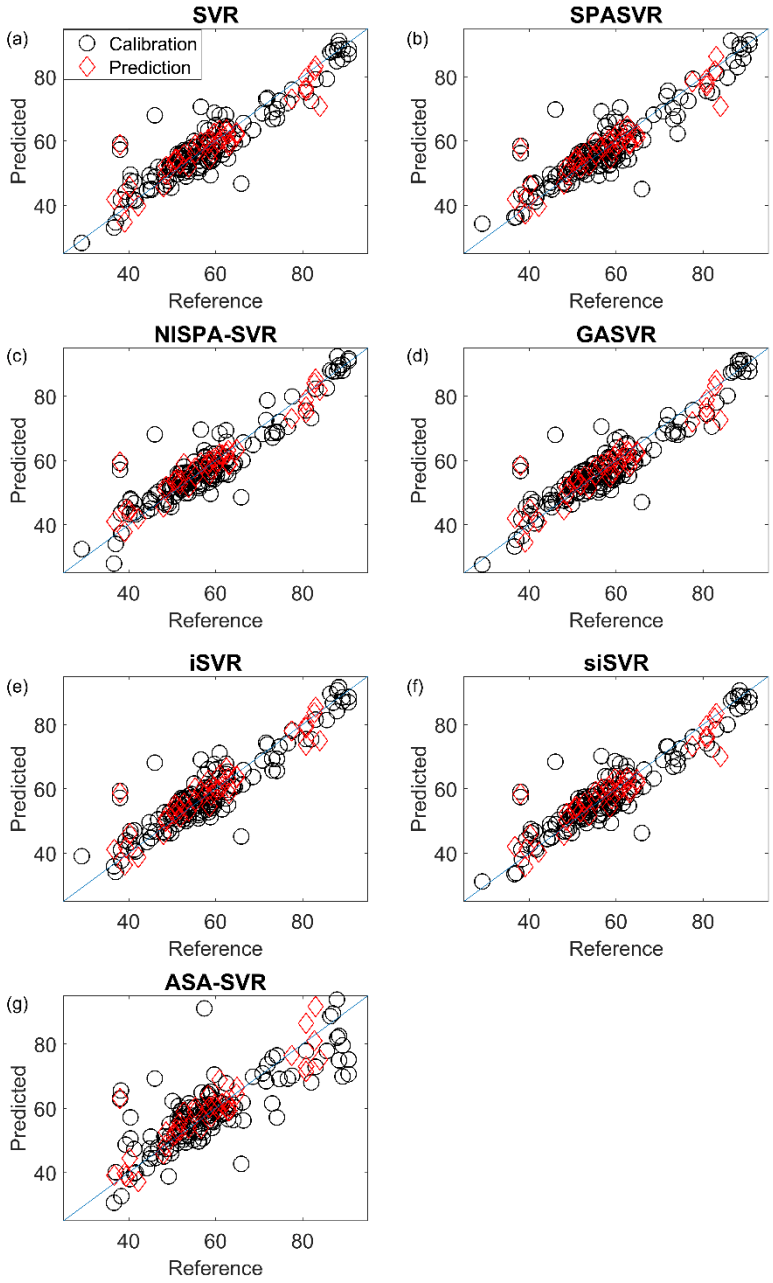


Figura 3. 7 Gráfico de valores Medido e Predito para teor de saturados. Sem seleção de variáveis (a), SPA (b), NISPA (c), GA (d), iSVR (e), siSVR (f) e ASA-VIF (g).

Os métodos de seleção de variáveis não proporcionaram uma melhoria significativa em relação ao modelo puro de SVR. No entanto, o modelo NISPA-SVR utilizou apenas 214 variáveis (cerca de 6,4% do total) e obteve o menor valor de RMSEP (4,4 wt%) e um R^2p igual a 0,893, como pode ser visto na **Tabela 3.4**. No entanto, os métodos de seleção de variáveis —NISPA e iSVR— revelaram as variáveis mais importantes para a construção do modelo, apresentadas na **Figura 3.8A**.

Tabela 3. 4 Parâmetros de avaliação para saturados.

Model	Number of Variables	RMSECV	RMSEP	R^2cv	R^2p
SVR	3351	5.1	4.8	0.838	0.869
SPA-SVR	2427	5.2	4.7	0.831	0.878
NISPA-SVR	214	5.0	4.4	0.847	0.893
GA-SVR	758	5.0	4.7	0.844	0.876
iSVR	335	5.2	4.6	0.836	0.882
siSVR	3016	5.1	4.9	0.840	0.869
ASA-SVR	5	8.0	5.7	0.618	0.817

A **Figura 3.8A** mostra as regiões selecionadas no NISPA-SVR: região de parafina (3000–2800 cm^{-1}), grupos alcano e alifáticos (1500–1200 cm^{-1}), e região de substituição no benzeno (900–680 cm^{-1}).[37] As duas primeiras regiões estão diretamente relacionadas à propriedade de saturados, enquanto a última está inversamente relacionada. Moro *et al.*[40] utilizaram PLS e fusão de dados para prever o teor de saturados em petróleo, usando três métodos de espectroscopia, FTIR, RMN ^1H e RMN ^{13}C. Quando utilizado apenas o MIR, obtiveram R^2p igual a 0,76 e RMSEP igual a 5,82 wt%, o que é menos exato do que nosso modelo. Os autores obtiveram o melhor modelo para o teor de saturados através da fusão de nível médio com PCA de espectros de FTIR e RMN ^1H, com R^2p igual a 0,85 e RMSEP igual a 5,12 wt%, no entanto, também menos exato do que o nosso.

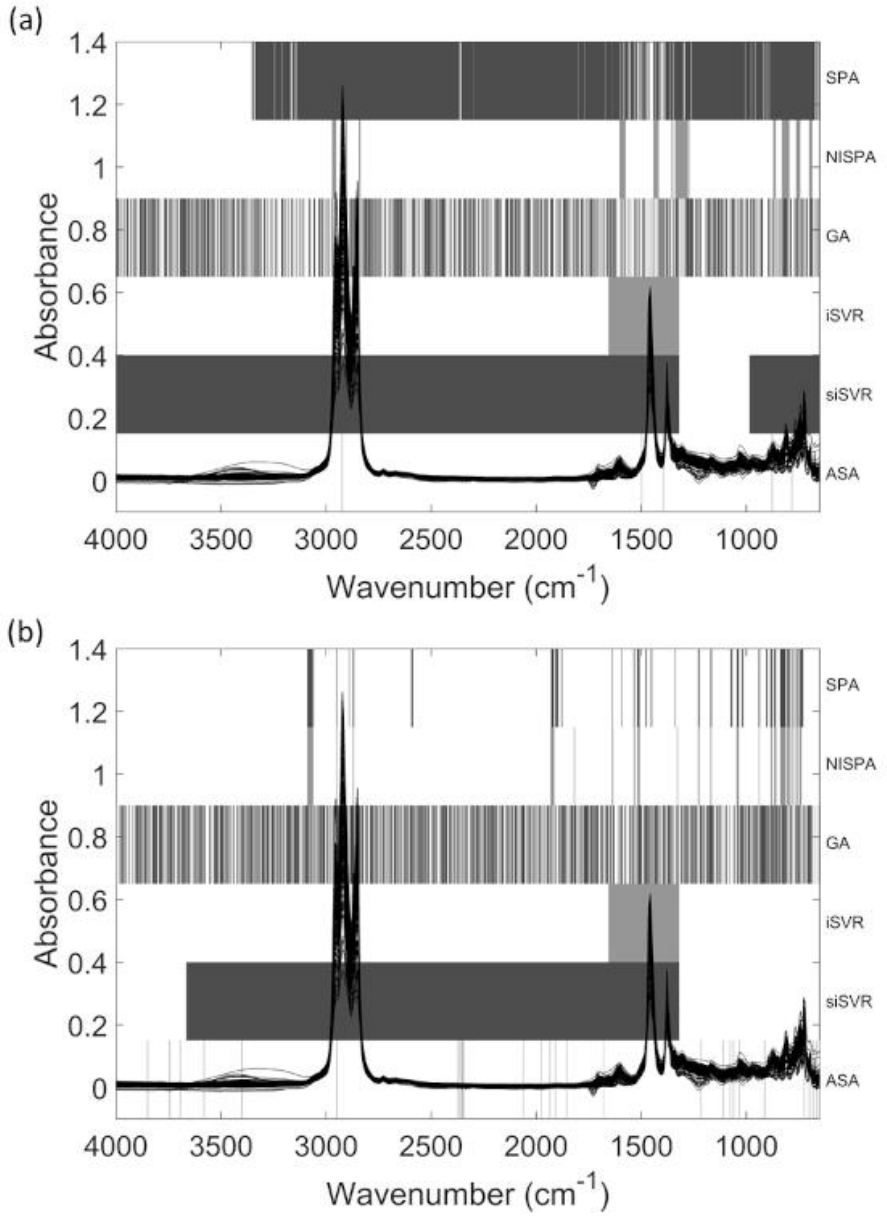


Figura 3. 8 Variáveis selecionadas para teor de saturados (a) e aromáticos (b).

3.3.5 Teor de aromáticos

Para o teor de aromáticos, os pré-tratamento airPLS e SNV foram combinados para melhorar os parâmetros de avaliação dos modelos de calibração. A **Figura 3.9** mostra os valores medidos e previstos para essa propriedade, na qual identificamos uma má aproximação das amostras com a linha de referência; apesar das amostras manterem uma distância constante, o que é refletido nos parâmetros de avaliação apresentados na **Tabela 3.5**. O modelo mais exato foi o NISPA-SVR, com um RMSEP igual a 3,1 wt% e um R^2p igual a 0,742, usando apenas 159 variáveis (4,7% do total).

Esses resultados são melhores do que os do modelo puro de SVR, cujos valores de RMSEP e R^2p foram 3,4 wt% e 0,703, respectivamente. Assim, fica evidenciado que mesmo reduzindo o número de variáveis a performance é mantida.

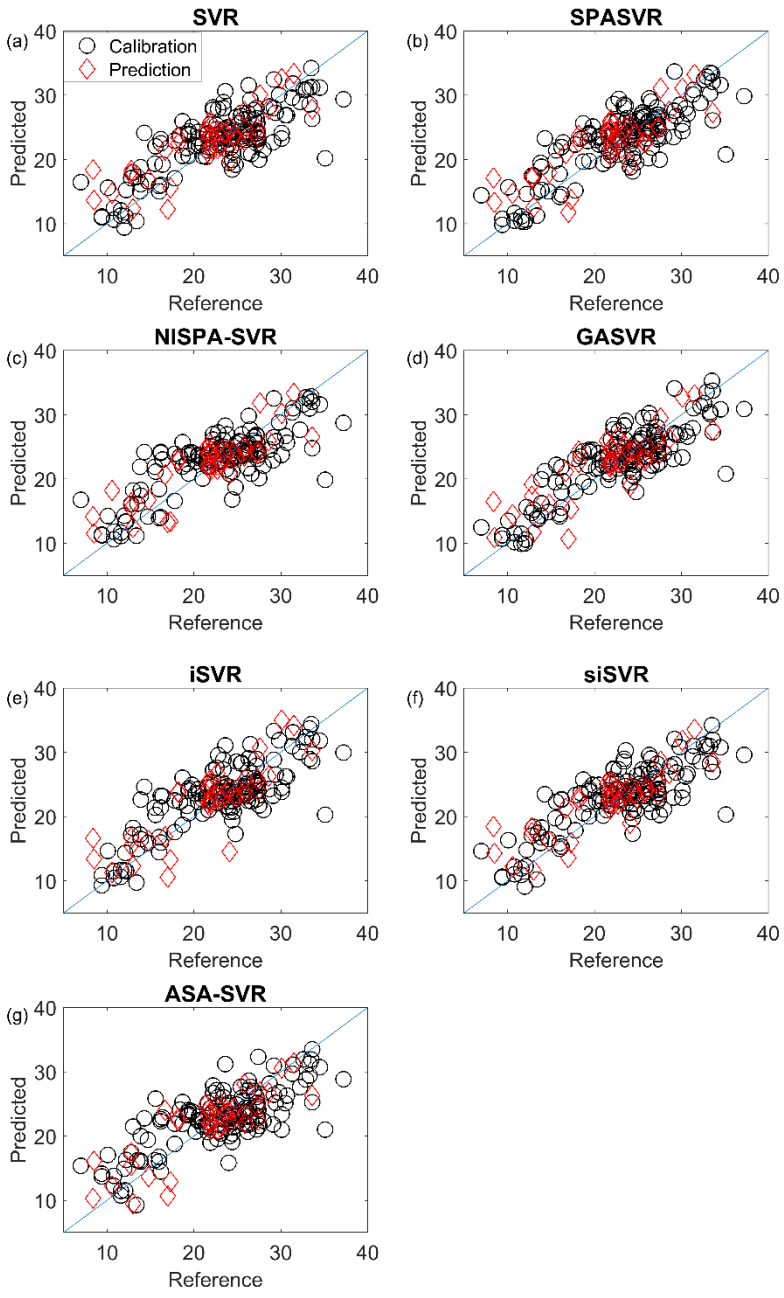


Figura 3. 9 Gráfico de valores Medido e Predito para teor de aromáticos. Sem seleção de variáveis (a), SPA (b), NISPA (c), GA (d), iSVR (e), siSVR (f) e ASA-VIF (g).

A **Figura 3.8B** mostra as variáveis selecionadas pelo modelo NISPA-SVR. Destacamos a região de 900 a 650 cm^{-1}, atribuída a compostos aromáticos, especialmente anéis de benzeno. Também destacamos a região próxima a 1450 cm^{-}

[1], atribuída a grupos alqueno, e 2800–3000 cm^{-1}, atribuída a parafinas.

Tabela 3. 5 Parâmetros de avaliação para aromáticos.

Model	Number of Variables	RMSECV	RMSEP	R²cv	R²p
SVR	3351	3.8	3.4	0.629	0.703
SPA-SVR	183	3.6	3.4	0.664	0.692
NISPA-SVR	159	3.8	3.1	0.625	0.742
GA-SVR	919	3.4	3.4	0.703	0.698
iSVR	335	3.7	3.5	0.650	0.688
siSVR	2345	3.6	3.7	0.664	0.623
ASA-SVR	28	4.0	3.2	0.572	0.703

Moro *et al.*[40] também utilizaram PLS e fusão de dados para determinar o teor de aromáticos. Usando dados de MIR e RMN ^{13}C, obtiveram um modelo com R²p e RMSEP iguais a 0,67 e 3,66 wt%, respectivamente. Obtivemos modelos mais exatos, utilizando apenas dados de MIR. Mohammadi *et al.*[41] utilizaram MIR combinado com GA-PLS e GA-SVR para prever aromáticos e resinas em 12 amostras de óleo do Iraque. O GA foi usado tanto para otimizar os parâmetros do SVR quanto para selecionar variáveis, reduzindo-as de 1738 para 422 (25%). Eles alcançaram um R²p igual a 0,840 e 0,860 para GA-PLS e GA-SVR, respectivamente. Embora esse modelo seja mais exato do que o nosso, o GA-SVR não pode determinar as variáveis de MIR mais importantes, enquanto o NISPA e o ASA-VIF conseguem.

3.3.6 Teor de resinas

Para modelar o teor de resina, o melhor pré-tratamento foi o MSC. O gráfico de dados medidos e preditos (**Figura 3.10**) mostra um ajuste ruim para amostras com teor de resina próximo de 0 wt% em todos os modelos. A seleção de variáveis não proporcionou uma melhoria considerável. O melhor modelo foi obtido com o GA-SVR, com RMSEP igual a 4,1 wt% e R²p igual a 0,753, usando 949 variáveis (**Tabela 3.6**). Esses resultados são ligeiramente melhores do que os obtidos com o SVR puro, que apresentou RMSEP e R²p iguais a 4,2 wt% e 0,739, respectivamente.

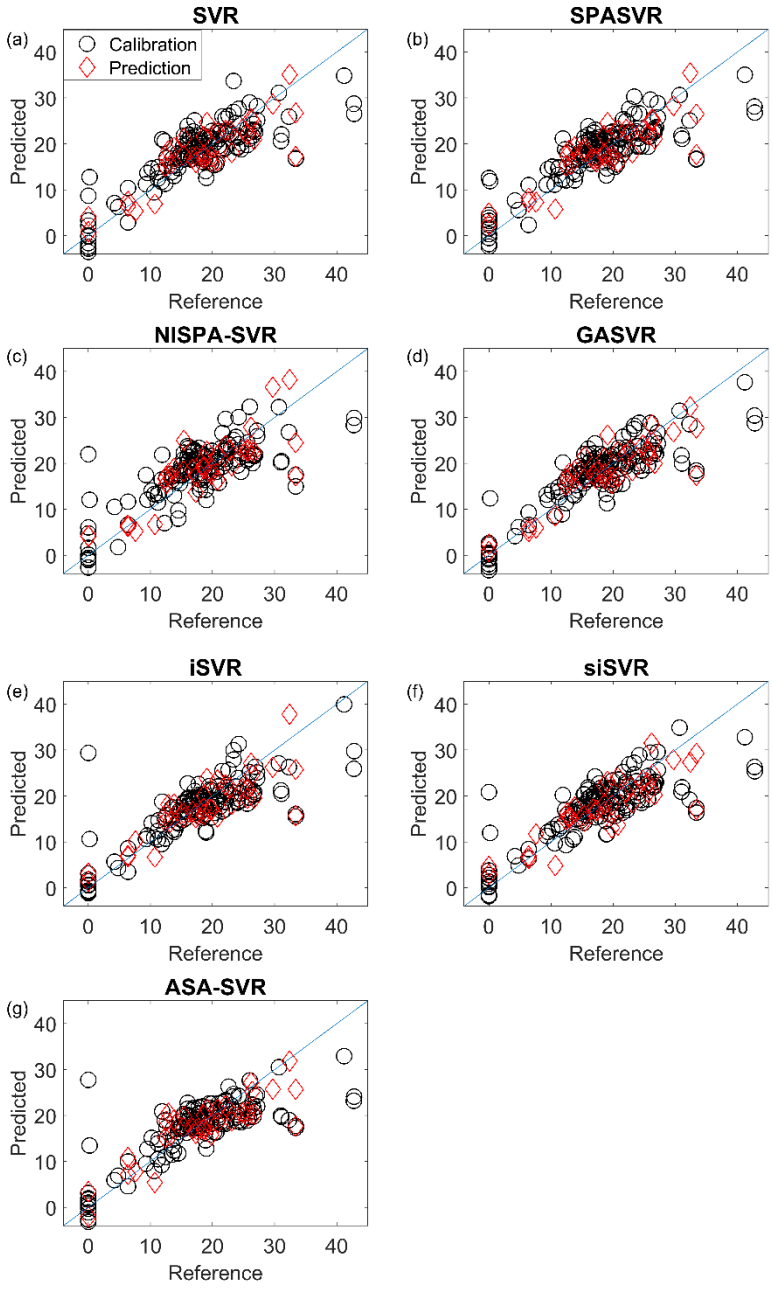


Figura 3. 10 Gráfico de valores Medido e Predito para teor de resinas. Sem seleção de variáveis (a), SPA (b), NISPA (c), GA (d), iSVR (e), siSVR (f) e ASA-VIF (g).

O SPA apresentou um resultado melhor do que o NISPA, com RMSEP igual a 4,2 wt% e R^2p igual a 0,737, usando 1889 variáveis. Embora o GA-SVR tenha alcançado um resultado melhor, o NISPA-SVR e o ASA-SVR também obtiveram resultados satisfatórios e mostraram a importância das bandas de alcanos nas cadeias laterais (3000–2800 cm^{-1}), em cadeias alifáticas e alcano (1450–1350 cm^{-1}),[34,35] e na região de substituição do anel de benzeno (900–650 cm^{-1}) (**Figura 3.11**) para a previsão de resina em petróleo, usando FTIR.

Tabela 3. 6 Parâmetros de avaliação para resinas.

Model	Number of Variables	RMSECV	RMSEP	R^2cv	R^2p
SVR	3351	4.6	4.2	0.690	0.739
SPA-SVR	1889	4.6	4.2	0.695	0.737
NISPA-SVR	36	5.4	4.6	0.604	0.689
GA-SVR	949	4.2	4.1	0.740	0.753
iSVR	335	5.0	4.5	0.643	0.715
siSVR	670	5.0	4.4	0.647	0.727
ASA-SVR	13	5.3	4.4	0.595	0.722

Mohammadi *et al.*[41] obtiveram um R^2p de 0,800 e 0,911 e RMSEP de 1,013 e 0,780 para GA-PLS e SVR, respectivamente, para prever o teor de resina. Esses resultados são melhores do que os nossos, mas os autores utilizaram um conjunto de amostras com uma pequena faixa de teor de resina (6,5 a 15,69 wt%), diferente dos nossos dados (0 a 42,8 wt%).

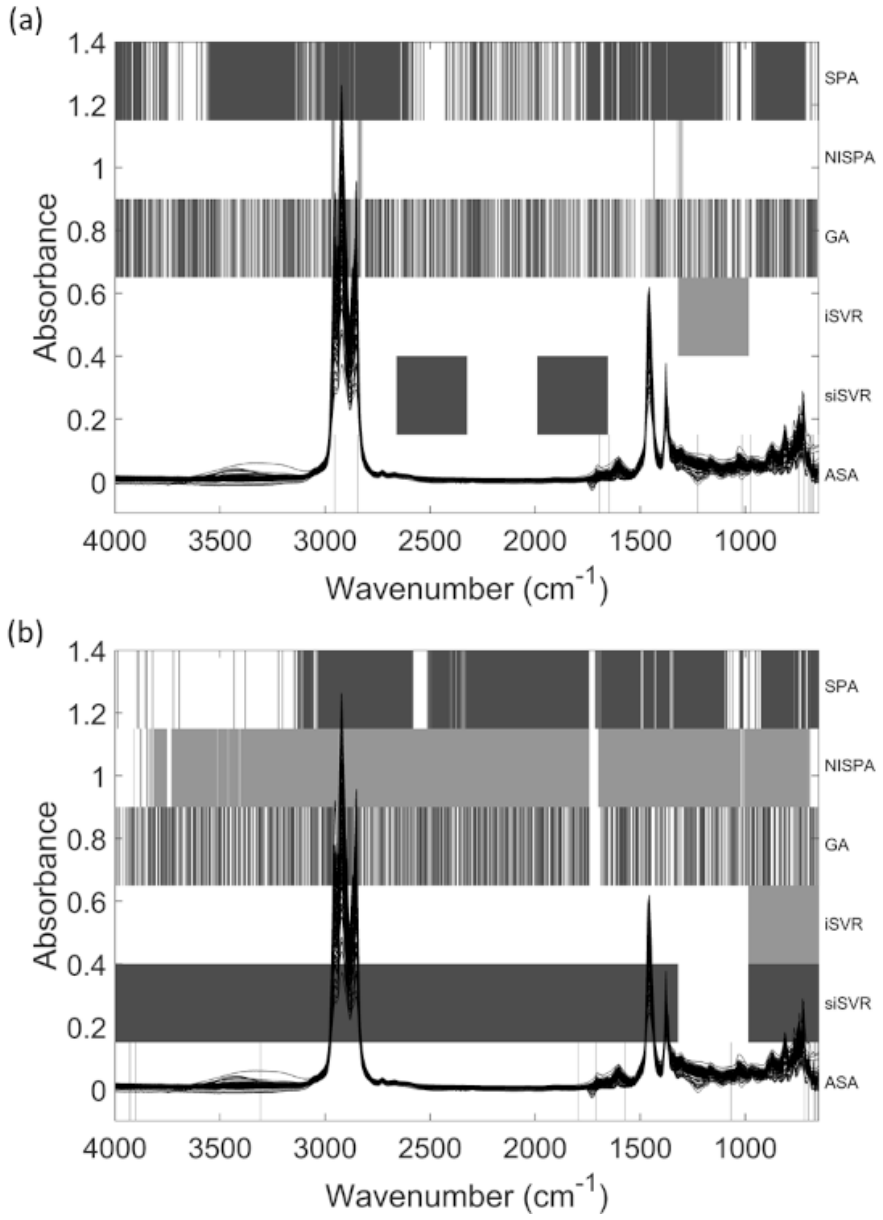


Figura 3. 11 Variáveis selecionadas para teor de resinas (a) e asfaltenos (b).

3.3.7 Teor de asfaltenos

Para a previsão do teor de asfaltenos, o melhor modelo foi obtido com o pré-tratamento de centralização na média. A **Tabela 3.6** mostra os parâmetros de avaliação dos modelos de previsão de asfaltenos. O gráfico de dados medidos e previstos (Figura 3.12) mostra um ajuste ruim para amostras com teor de asfaltenos acima de 5%, o que não ocorre no modelo GA-SVR. O modelo mais exato foi obtido com o ASA-SVR, usando apenas 12 variáveis (0,43% do total), alcançando um RMSEP de 0,43 wt% e R^2p de 0,828, enquanto o modelo SVR puro alcançou um

RMSEP igual a 0,51 wt% e R²p igual a 0,731. O melhor modelo usando permutação apresentou um RMSEP de 0,44 wt% e R²p de 0,806.

Tabela 3. 7 Parâmetros de avaliação para asfaltenos.

Model	Number of Variables	RMSECV	RMSEP	R²cv	R²p
SVR	3351	0.64	0.51	0.827	0.731
SPA-SVR	2000	0.72	0.44	0.771	0.806
NISPA-SVR	3006	0.66	0.46	0.817	0.786
GA-SVR	1002	0.44	0.48	0.915	0.757
iSVR	335	0.66	0.49	0.807	0.834
siSVR	3016	0.59	0.51	0.855	0.724
ASA-SVR	12	0.74	0.43	0.760	0.828

A maioria dos modelos selecionou um número grande de variáveis em todo o espectro e não em uma região específica, exceto o ASA-SVR, que selecionou algumas regiões como, a região de 900 a 650 cm^{-1}, relacionada à substituição do anel de benzeno, que está diretamente relacionada aos asfaltenos (Figura 3.11B). Além disso, selecionou região entre 1500-1700 cm-1 que tem sinais relacionados moléculas asfaltênicas, como estiramento C=C em anéis aromáticos.

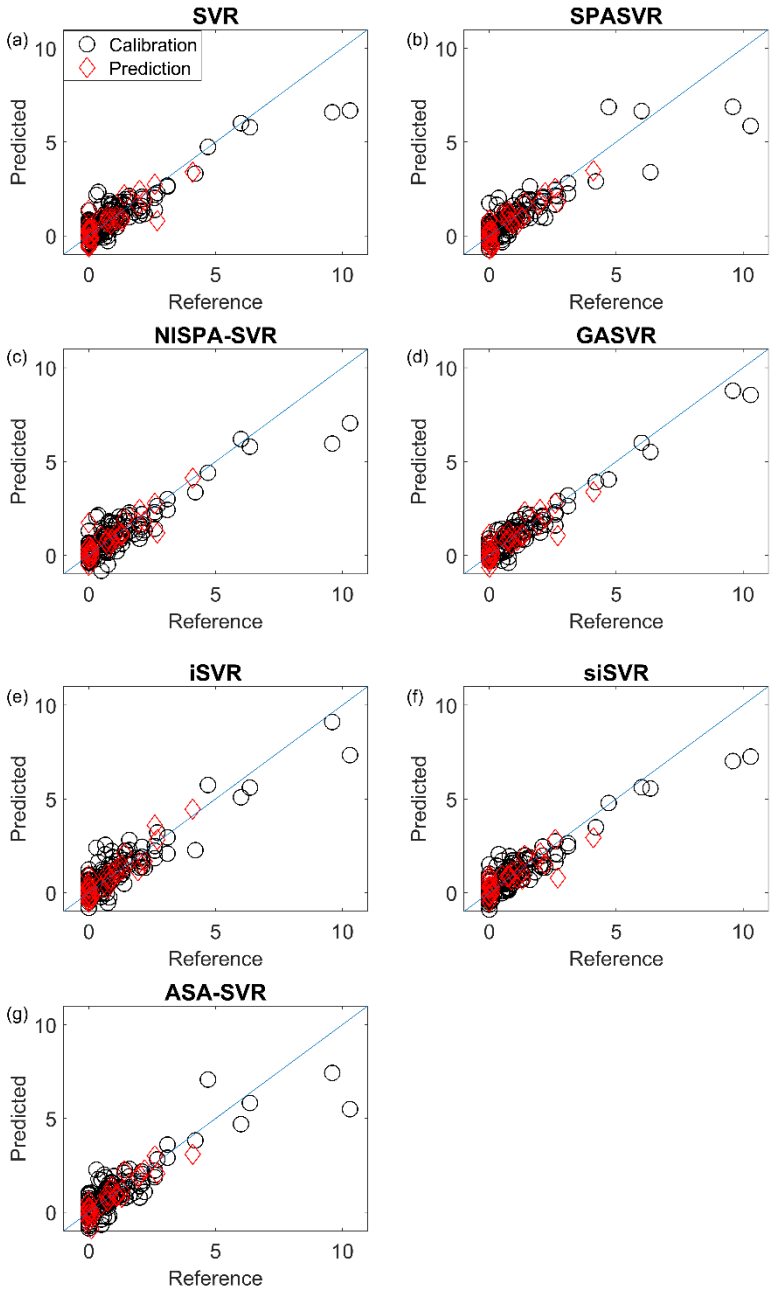


Figura 3. 12 Gráfico de valores Medido e Predito para teor de asfaltenos. Sem seleção de variáveis (a), SPA (b), NISPA (c), GA (d), iSVR (e), siSVR (f) e ASA-VIF (g).

Laxalde *et al.*[34] combinaram a faixa espectral do infravermelho próximo (NIR) e médio (MIR) para prever a composição SARA em óleos pesados. Usando métodos multibloco, eles obtiveram modelos multibloco-PLS (MB-PLS) e serial-PLS (S-PLS) para prever asfaltenos, obtendo valores baixos de RMSEP (0,97 e 0,94 wt%, respectivamente), contudo, maiores do que os que obtivemos (**Tabela 3.7**). Os autores também afirmaram que as contribuições de C–H das vibrações de flexão planar e vibrações de estiramento aromático de C=C nos espectros de MIR são importantes

para a determinação de compostos altamente insaturados.

3.4 Conclusão

Os métodos de seleção de variáveis baseados em permutação, SPA e NISPA, foram aplicados em modelos SVR e comparados com outros métodos de seleção de variáveis. SPA e NISPA apresentam potencial para selecionar as principais variáveis no espectro FTIR para prever propriedades físico-químicas de óleos crus. Para densidade API, os outros métodos de seleção de variáveis não conseguiram alcançar uma redução tão grande no número de variáveis e obter, ao mesmo tempo, um modelo de alta exatidão, como SPA e NISPA fizeram. SPA reduziu o número de variáveis para apenas 3,5% do total original, mantendo apenas as variáveis relacionadas aos carbonos de cadeias alifáticas de compostos de parafina e relacionadas a moléculas de benzeno substituídas, sugerindo que essas regiões estão intimamente relacionadas com a densidade API.

Para a previsão de saturados, além de fornecer o modelo mais exato, NISPA reduziu consideravelmente o número de variáveis e demonstrou a importância das regiões de parafina, alcanos e grupos alifáticos nos espectros de MIR para a modelagem dessa propriedade. As mesmas variáveis foram selecionadas para densidade API e teor de saturados, devido à relação entre essas duas propriedades. Para aromáticos, SPA e NISPA, além de reduzir drasticamente o número de variáveis, selecionaram as variáveis diretamente atribuídas a compostos aromáticos. Para os teores de resinas e asfaltenos, os métodos baseados em permutação não obtiveram os modelos mais exatos, mas selecionaram regiões espectrais importantes com um resultado satisfatório.

Os resultados comprovam que as seleções de variáveis SPA e NISPA podem identificar variáveis relevantes em modelos SVR, ao contrário dos outros métodos de seleção. Isso é uma vantagem relevante, pois quando a informação química não é perdida durante a modelagem das propriedades físico-químicas o espectro e as propriedades podem ser relacionados, tornando o modelo mais confiável e compreensível.

3.5 Referência

1. Vapnik, V. Support-Vector Networks. *IEEE Expert. Syst. their Appl.* **7**, 63–72 (1992).
2. Scholkopf, B. *et al.* Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* **45**, 2758–2765 (1997).
3. Hemmati-Sarapardeh, A., Aminshahidy, B., Pajouhandeh, A., Yousefi, S. H. & Hosseini-Kaldozakh, S. A. A soft computing approach for the determination of crude oil viscosity: Light and intermediate crude oil systems. *J. Taiwan Inst. Chem. Eng.* **59**, 1–10 (2016).
4. Rocha, W. F. de C. & Sheen, D. A. Determination of physicochemical properties of petroleum derivatives and biodiesel using GC/MS and chemometric methods with uncertainty estimation. *Fuel* **243**, 413–422 (2019).
5. Ghorbani, M., Zargar, G. & Jazayeri-Rad, H. Prediction of asphaltene precipitation using support vector regression tuned with genetic algorithms. *Petroleum* **2**, 301–306 (2016).
6. Filgueiras, P. R. *et al.* Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. *Fuel* **116**, 123–130 (2014).
7. Alves, J. C. L. & Poppi, R. J. Quantification of conventional and advanced biofuels contents in diesel fuel blends using near-infrared spectroscopy and multivariate calibration. *Fuel* **165**, 379–388 (2016).
8. Bemani, A. *et al.* Modeling of cetane number of biodiesel from fatty acid methyl ester (FAME) information using GA-, PSO-, and HGAPSO- LSSVM models. *Renew. Energy* **150**, 924–934 (2020).
9. Üstün, B., Melssen, W. J. & Buydens, L. M. C. Visualisation and interpretation of Support Vector Regression models. *Anal. Chim. Acta* **595**, 299–309 (2007).
10. Huang, Y., Zhang, J., Tze Ann, F. & Ma, G. Intelligent mixture design of steel fibre reinforced concrete using a support vector regression and firefly algorithm based multi-objective optimization model. *Constr. Build. Mater.* **260**, 120457 (2020).
11. Postma, G. J., Krooshof, P. W. T. & Buydens, L. M. C. Opening the kernel of kernel partial least squares and support vector machines. *Anal. Chim. Acta* **705**,

123–134 (2011).

12. Chen, S.-W. W., Li, Z.-R. R. & Li, X.-Y. Y. Prediction of antifungal activity by support vector machine approach. *J. Mol. Struct. THEOCHEM* **731**, 73–81 (2005).
13. Teye, E. & Huang, X. Novel Prediction of Total Fat Content in Cocoa Beans by FT-NIR Spectroscopy Based on Effective Spectral Selection Multivariate Regression. *Food Anal. Methods* **8**, 945–953 (2015).
14. 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* **119**, 582–589 (2014).
15. Xu, S., Zhao, Y., Wang, M. & Shi, X. Determination of rice root density from Vis–NIR spectroscopy by support vector machine regression and spectral variable selection techniques. *Catena* **157**, 12–23 (2017).
16. Attia, K. A. M., Nassar, M. W. I., El-Zeiny, M. B. & Serag, A. Firefly algorithm versus genetic algorithm as powerful variable selection tools and their effect on different multivariate calibration models in spectroscopy: A comparative study. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **170**, 117–123 (2017).
17. Filgueiras, P. R. *et al.* Determination of Saturates, Aromatics, and Polars in Crude Oil by 13C NMR and Support Vector Regression with Variable Selection by Genetic Algorithm. *Energy and Fuels* **30**, 1972–1978 (2016).
18. Folli, G. S. *et al.* Variable selection in support vector regression using angular search algorithm and variance inflation factor. *J. Chemom.* **34**, 1–16 (2020).
19. ASTM E1655. Standard Practives for Infrared Multivariate Quantitative Analysis. at https://doi.org/10.1520/E1655-05 (2012).
20. Li, H.-D. *et al.* Recipe for revealing informative metabolites based on model population analysis. *Metabolomics* **6**, 353–361 (2010).
21. Bin, J. *et al.* An efficient variable selection method based on variable permutation and model population analysis for multivariate calibration of NIR spectra. *Chemom. Intell. Lab. Syst.* **158**, 1–13 (2016).
22. Yang, B. *et al.* Rapid prediction of yellow tea free amino acids with hyperspectral images. *PLoS One* **14**, e0210084 (2019).
23. Wang, Q., Li, H. D., Xu, Q. S. & Liang, Y. Z. Noise incorporated subwindow permutation analysis for informative gene selection using support vector

machines. *Analyst* **136**, 1456–1463 (2011).

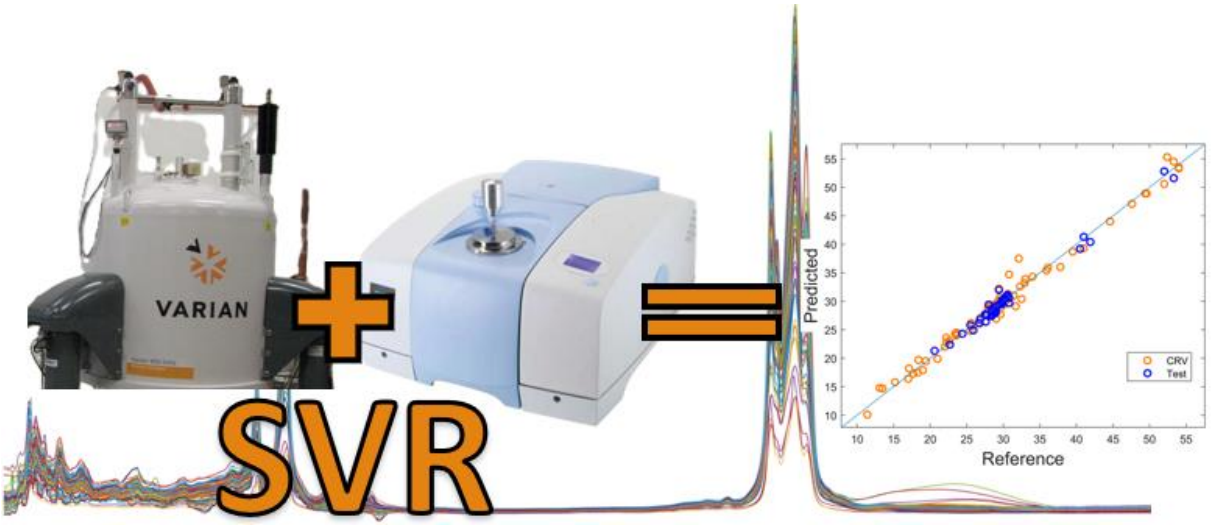
24. ISO 12185. Crude petroleum and petroleum products – determination ofdensity – oscillating U-tube method. at (1996).
25. ASTM. Standard D7042: Test Method for Dynamic Viscosity and Density of Liquids by Stabinger Viscometer (and the Calculation of Kinematic Viscosity). *Am. Natl. Stand. Inst.* **12a**, 1–11 (2013).
26. ASTM D6560. Standard Test Method for Determination of Asphaltenes (Heptane Insolubles) in Crude and Petroleum ProductsASTM International. STM International: West Conshohocken, PA. *ASTM Int.* 1–6 (2013) doi:10.1520/D6560-12.2.
27. Ferreira, P. S. *et al.* SAP fractions from light, medium and heavy oils: Correlation between chemical profile and stationary phases. *Fuel* **274**, 117866 (2020).
28. ASTM D2549. ASTM D2549. Standard Method for Separation of Representative Aromatics and Nonaromatics Fractions of High-Boiling Oils by Elution Chromatography. *Man. Hydrocarb. Anal. 6th Ed.* **02**, 379-379–6 (2008).
29. Chang, C.-C. & Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2001).
30. Dias, J. C. M. & Aguiar, P. F. a Statistical Method for Acceptance of Crude Oil Viscosity-Temperature Curves. *Brazilian J. Pet. Gas* **5**, 019–024 (2011).
31. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
32. Leong, W. C., Kelani, R. O. & Ahmad, Z. Prediction of air pollution index (API) using support vector machine (SVM). *J. Environ. Chem. Eng.* **8**, 103208 (2020).
33. Andries, J. P. M., Heyden, Y. Vander & Buydens, L. M. C. Improved variable reduction in partial least squares modelling by Global-Minimum Error Uninformative-Variable Elimination. *Anal. Chim. Acta* **982**, 37–47 (2017).
34. Laxalde, J., Caillol, N., Wahl, F., Ruckebusch, C. & Duponchel, L. Combining near and mid infrared spectroscopy for heavy oil characterisation. *Fuel* **133**, 310–316 (2014).
35. Hongfu, Y., Xiaoli, C., Haoran, L. & Yupeng, X. Determination of multi-properties of residual oils using mid-infrared attenuated total reflection spectroscopy. *Fuel* **85**, 1720–1728 (2006).
36. Rainha, K. P. *et al.* Determination of API Gravity and Total and Basic Nitrogen Content by Mid- and Near-Infrared Spectroscopy in Crude Oil with Multivariate

Regression and Variable Selection Tools. *Anal. Lett.* **52**, 2914–2930 (2019).

37. Silverstein, R. M.; Webster F. X.; Kiemle, D. J. *Spectrometric Identification of Organic Compounds*. vol. 7 ed (2005).
38. de Paulo, E. H. *et al.* Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties. *Fuel* **279**, 118462 (2020).
39. van der Voet, H. Comparing the predictive accuracy of models using a simple randomization test. *Chemom. Intell. Lab. Syst.* **25**, 313–323 (1994).
40. Moro, M. K. *et al.* FTIR, 1H and 13C NMR data fusion to predict crude oils properties. *Fuel* **263**, 116721 (2020).
41. Mohammadi, M. *et al.* Genetic algorithm based support vector machine regression for prediction of SARA analysis in crude oil samples using ATR-FTIR spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **245**, 118945 (2021).

CAPÍTULO 4 FUSÃO DE DADOS EM REGRESSÃO POR VETORES DE SUPORTE APLICADO NA PREVISÃO DE PROPRIEDADES FÍSICO-QUÍMICAS DO PETRÓLEO

DATA FUSION IN SUPPORT VECTOR REGRESSION APPLIED TO THE PREDICTION OF PHYSICO-CHEMICAL PROPERTIES OF PETROLEUM



• A fusão de dados tem mostrado potencial promissor na aplicação da regressão por vetores de suporte, em conjunto com a espectroscopia de infravermelho e ressonância magnética nuclear.

• Apesar da necessidade de alto processamento computacional, a fusão de dados tem o potencial de aprimorar modelos de regressão ao combinar diferentes fontes analíticas, permitindo a previsão de várias propriedades da matriz de petróleo, tanto físicas quanto químicas.

• O infravermelho próximo e médio, apesar de serem fontes analíticas semelhantes, mostraram potencial de sinergia, assim como a ressonância magnética nuclear de carbono e hidrogênio.

• A fusão de dados apresenta um grande potencial; contudo, devido à alta demanda computacional da regressão por vetores de suporte, sua aplicação deve ser

feita com cuidado e preparação, para não demandar tempo de máquina desnecessariamente.

Resumo

A Fusão de Dados trata-se de uma nova estratégia para aperfeiçoar modelos, demonstrando potencial na calibração multivariada, todavia, a aplicação em máquina de vetores de suporte (SVM, support vector machine) são escassas, o que aponta a necessidade de mais estudos. Essa estratégia combina fontes analíticas diferentes com o objetivo de utilizar o sinergismo entre informações para desenvolver modelos mais exatos. O vigente estudo desenvolveu modelos de regressão de vetores de suporte (SVR, support vector regression) com fusão de dados de baixo, médio e alto nível, combinando espectroscopia de infravermelho médio (MIR do inglês *middle infrared*) e próximo (NIR do inglês *near infrared*) com ressonância magnética nuclear de hidrogênio (NMR de H^1 do inglês *nuclear magnetic ressonance*) e carbono (NMR de ^{13}C) para predizer nove propriedade físico-químicas do petróleo (densidade API, ponto de fluidez, WAT, teor de saturados, aromáticos, polares, enxofre, nitrogênio total e poder calorífico). Quando necessário seleção de variáveis foi utilizado análise de componentes principais (PCA, do inglês *Principal Component Analysis*) e o algoritmo genético (GA, do inglês Genetic algorithm). Os modelos gerados por fusão conseguiram se igualar, ou superar, os modelos individuais (sem fusão). Na predição da densidade API, a fusão de médio nível com PCA combinando MIR e NIR desenvolveu um modelo com parâmetros melhores, RMSEP de 0,9 e R^2p de 0,985, superando o melhor modelo sem fusão de dados, RMSEP 1,2 e R^2p 0,969. Isso evidencia que mesmo fontes analíticas semelhantes, como infravermelhos, podem se beneficiar do sinergismo da fusão de dados. Ao aplicar fusão de médio nível com GA para predizer ponto de fluidez, combinando NIR e NMR de H^1, conseguiu superar os modelos sem fusão, além de modelos encontrados na literatura, com parâmetros de avaliação de R^2p 0,899 e RMSEP de 7 para o modelo com fusão e R^2p 0,839 e RMSEP 8 para sem fusão. No nitrogênio total a fusão de alto nível com MIR e NMR de ^1H, R^2p 0,945 e RMSEP 0,0245, conseguiu ser estatisticamente melhor que os modelos sem fusão de dados, R^2p 0,0304 e RMSEP 0,931. Neste estudo ficou demonstrado o potencial de utilizar fusão de dados em SVR para conseguir modelos estatisticamente melhores que aqueles sem fusão, tanto para propriedades físicas quanto químicas do petróleo.

Palavras-chave: data fusion, support vector machine, crude oil, chemometrics, variable selection

4.1 Introdução

A Regressão Multivariada utiliza fontes analíticas, como espectros de infravermelho, para extrair informações e construir um modelo de regressão.[1] Entretanto, em algumas situações, uma única fonte pode não conter todas as informações necessária para prever satisfatoriamente a propriedade desejada,[2] neste caso, é interessante aplicar diversas fontes analíticas combinadas, o que é conhecido como fusão de dados.[3–7]

A fusão de dados teve seu começo na década de 70, originalmente concebida para aplicação militar, visando identificar alvos como aeronaves, mísseis ou formações militares.[8] Deste então, o método encontrou uma ampla gama de aplicações em diversas áreas, como, robótica, química, reconhecimento de padrões, entre outras.[9] Na química, onde um grande número de fontes analíticas oferece informações distintas sobre a matriz amostral, a fusão de dados ganhou destaque..

Assim, modelos construídos a partir de uma única fonte analítica podem apresentar limitada capacidade preditiva, entretanto, ao combinar diferentes fontes numa fusão de dados apropriada, é possível obter modelos estatisticamente mais exatos.[10] Portanto, é importante compreender a natureza dos dados, visando fundir apenas informações complementares, evitando redundâncias, que podem atrapalhar a aplicação do método. A fusão de dados é classificada em três níveis a depender da forma que é aplicada: baixo, médio e alto nível.[11]

A fusão de baixo nível consiste na combinação dos dados brutos provenientes de diferentes fontes analíticas em uma única matriz antes da criação do modelo, ou alguma extração de dados. É importante, no entanto, levar em conta a escala e a variância dos dados fusionados para evitar o problema da predominância de uma fonte sobre as outras. Isso pode ocorrer quando uma fonte analítica parece ser mais relevante para o modelo do que as demais para o algoritmo utilizado.[10,12] Além disso, o problema da predominância pode ocorrer quando os conjuntos de dados têm muita discrepância no número de variáveis. Como a fusão de infravermelho médio (MIR do inglês *middle infrared*), com 3351 variáveis e a ressonância magnética nuclear de carbono (NMR de ^{13}C do inglês *nuclear magnetic ressonance*), com 61606 variáveis. Este tipo de fusão apresenta desafios como utilizar informações redundantes, ruído e interferência, limitando o efeito sinérgico desejado.[13] Esta forma de fusão resulta frequentemente em uma matriz de grande dimensão, exigindo um poder

computacional substancial.

A fusão de médio nível pode ser utilizada para contornar os desafios mencionados anteriormente. Neste método, a informação relevante de cada fonte analítica é extraída antes da concatenação e construção do modelo, podendo se utilizar de técnicas como análise de componentes principais (PCA), regressão por mínimos quadrados parciais (PLS)[12] e a aplicação de seleção de variáveis.[13] Isso resulta numa grande redução do número de variáveis, eliminando informações redundantes e ruído.[14,15] Sua desvantagem está no risco de excluir variáveis que poderiam contribuir sinergicamente com o modelo[10] e ainda persiste o problema da predominância de uma fonte analítica com base na quantidade de variáveis.

A fusão de alto nível inicia-se no desenvolvimento de modelos individuais para cada fonte analítica. Nesse processo, as previsões de cada amostra em cada modelo são combinadas para formar uma única previsão. Essa combinação é então realizada por meio de uma média aritmética simples, ponderada e, em alguns casos, por uma estimativa condicionada bayesiana, quando falamos de modelos quantitativos.[16] Como a fusão ocorre após a criação dos modelos, é possível aplicar diferentes seleções de variáveis e pré-tratamentos nos conjuntos de dados de modo a otimizar os resultados. Essa estratégia resolve o problema de predominância de dados, pois cada modelo gera uma única predição. No entanto, tem-se a desvantagem da perda de sinergismo entre as informações, uma vez que apenas as predições são fundidas e não os dados em si.[10]

Recentemente, a fusão de dados vem sendo amplamente estudada, como Yang, B. *et al*, em 2019, [6] demonstra ao utilizar fusão de dados em imagens hiperespectrais para predizer amino ácidos livres em chá amarelo. O autor utilizou o espectro característico e as variáveis de textura aliado a SVR de desenvolveu modelos com R^2p de 0,74 e 0,81, ao fundir ambas as fontes analíticas conseguiu R^2p de 0,87, uma melhora significativa.

Na indústria petrolífera, Moro *et al*., em 2019,[12] aplicou fusão de dados em MIR, NMR de ^{13}C e ^{1}H em conjunto com PLS para predizer propriedades físico-químicas do petróleo. Foram testadas técnicas de fusão de baixo e médio nível, com a aplicação de PCA e dos loadings do PLS. Os resultados indicaram que a fusão de médio nível foi a mais eficaz, provavelmente devido a capacidade de selecionar as variáveis mais relevantes de cada espectro para então fundir e predizer a propriedade desejada. Além disso, vale destacar os resultados para predizer acidez total, com a fusão de

médio nível com loadings obtendo um R²p 0,91, superando o melhor modelo sem fusão com R²p de 0,80.

Apesar do aumento de publicações nos últimos anos, a maioria dos trabalhos sobre fusão se concentram na área de alimentos e saúde.[17] Verifica-se uma carência de estudos envolvendo fusão de dados na área de quimiometria para petróleo e derivados.[18,19] Nesta área, a Regressão por Vetores de Suporte (SVR) vem se destacando devido ao bom desempenho que apresenta em matrizes complexas e não lineares.[20–23]

Diante disso, a fusão de dados e o SVR se apresentam como duas estratégias quimiométricas com alto potencial, cada uma com funções e vantagens distintas.[24] Todavia, a aplicação de ambas em conjunto carece de estudos. Em um estudo recente, Xu S. *et al*., em 2022,[25] empregaram língua eletrônica, nariz eletrônico, colorímetro e análise de textura para predizer qualidade da pasta de soja chinesa utilizando PLS e SVR. A qualidade foi avaliada em cinco categorias: Aroma, Sabor, Cor, Textura e Nota Global. A fusão de dados foi aplicada neste último conseguindo R²p de 0,96 para SVR e 0,93 para PLS. O SVR conseguiu os melhores resultados, sugerindo que os dados podem ter uma relação não-linear. A fusão de dados combinada com SVR já demonstrou bons resultados para estimar parâmetros bioquímicos da cultura da soja,[26] calibrar sensores que medem a poluição do ar,[7] para prever a concentração de aminoácidos livres em chá amarelo,[6] entre outros estudos. Nesses trabalhos foram utilizadas metodologias para reduzir a dimensão das fontes analíticas, mas não foram utilizadas técnicas tradicionais de análise química, como infravermelho, cromatografia e ressonância magnética nuclear.

Tendo em vista o potencial da fusão de dados e do SVR, o objetivo deste estudo é aplicar ambos os métodos em conjunto, fazendo as devidas adaptações, utilizando fontes analíticas tradicionais como dados de infravermelho médio e próximo e ressonância magnética nuclear de carbono e hidrogênio com intuito de prever propriedades físico-químicas do petróleo.

4.2 Metodologia

4.2.1 Amostras

Para estudar a viabilidade da fusão de dados aplicada ao SVR, foi utilizado um

total de 119 amostras de petróleo bruto, provenientes da bacia sedimentar da costa brasileira. As amostras cujo resultados da análise de alguma propriedade estava ausente, não foram incluídas para modelagem de tais propriedades, de modo que a faixa de medição e a quantidade de amostras utilizadas em cada caso ficaram conforme apresentado na **Tabela 4.1**.

Tabela 4. 1 Número de amostras de calibração e predição e faixa por propriedade.

Prop	Técnica	Unidade	Cal	Teste	Faixa	Tipo
densidade API	ISO 12185		83	36	11,4 --- 54	Física
Ponto de Fluidez	ASTM D97	ºC	38	15	-35 --- 49	Físico
WAT	DSC (@1°C/min)				7,96 --- 38,39	Física
Saturado	SFC/TLC-FID	wt%	78	33	36,6 --- 90	Química
Aromático	SFC/TLC-FID	wt%	78	33	7 --- 37,2	Química
Polares	SFC/TLC-FID	wt%	78	33	0 --- 43,4	Química
Enxofre	D 4294 (Horiba)	wt%			7 --- 37,2	Química
Nitrogênio Total	ASTM D4629-17	wt%			0.0005 --- 0.4700	Química
Poder Calorifico	D 4809	MJ/kg	81	34	41,2 --- 46,5	Físico

4.2.2 Propriedades Físico-Químicas

Os ensaios analíticos para determinação das propriedades físico-químicas foram feitos no centro de pesquisa Leopoldo Américo Miguel de Mello (CENPES). A determinação da densidade API foi feita com base na norma ISO 12185.[27]

Foi utilizado um viscosímetro digital Anton Paar (modelo Stabinger SVM 3000) com limite de detecção de 0,0002 g/cm^{-3} a 20 ºC. Um volume de 5 mL de amostra foi adicionado na célula de medição do viscosímetro e as análises foram feitas em duplicata nas temperaturas de 40 e 50ºC. Para determinar a densidade API, os resultados de viscosidade foram convertidos para a densidade equivalente a 20 ºC e

aplicados na **Equação 4.1**, em que ρ é a densidade a 60 ºF.[27]

$$API = \frac{141{,}5}{\rho 60^{\circ}F} - 131{,}5 \qquad \textbf{Equação 4.1}$$

O ponto de fluidez foi determinado utilizando o método padrão ASTM D97, este método permite expressar o ponto de fluidez com uma incerteza de 3 ºC. A temperatura de aparecimento de cristal (WAT do inglês *wax appearance temperature*) foi determinada utilizando o método por Calorimetria Diferencial de Varredura (DSC, do inglês *Differential Scanning Calorimetry*) seguindo a metodologia ASTM D4419.

O teor de asfaltenos foi determinado com base na norma ASTM 6560. Uma amostra de óleo foi misturada com n-heptano e aquecida por 60 minutos sob refluxo. Após o repouso por 150 minutos, o precipitado, contendo asfaltenos e ceras, foi coletado com papel de filtro e as ceras removidas por lavagem com n-heptano em um extrator. Os asfaltenos foram diluídos em tolueno aquecido e, em seguida, o solvente foi evaporado em um retrovaporador. Em seguida, os asfaltenos foram aquecidos na estufa a 120 ºC até que sua massa ficasse constante e, por fim, pesados. Os maltenos (saturados, aromáticos e resinas), que ficaram na parte solúvel do n-heptano, foram analisados por cromatografia em fluido supercrítico/cromatografia em camada delgada - detector por ionização de chama (SFC/TLC-FID do inglês, *Supercritical Fluid Chromatography / thin Layer Chromatography – Flame Ionization Detector*).[28] O teor de polares foi determinado por meio do somatório de teor de asfalteno e resinas.

O Teor de Enxofre foi determinado com base na ASTM D4294, utilizando espectrometria de fluorescência de raios X por dispersão de energia com analisador automático HORIBA, SFLA-2800.[29] O Nitrogênio total foi determinado pelo método padrão ASTM D4629-17, consistindo na combustão das amostras, seguida pela determinação do nitrogênio por quimioluminescência. O poder calorífico foi determinado utilizando a ASTM D4809, uma amostra ponderada do petróleo é colocada em combustão dentro de um calorímetro, a principal desvantagem desse método está no fato de ser destrutivo.[30,31]

4.2.3 Aquisição dos espectros

Foram obtidos espectros de MIR (4000 a 650 cm^{-1}), NIR (10000 a 4000 cm^{-1}),

NMR de ^{1}H (0 a 10 ppm) e NMR de ^{13}C (0 a 240 ppm) de todas as amostras, em triplicata. Os ensaios de espectroscopia na região do infravermelho foram realizados em um equipamento Spectrum 400 da PerkinElmer. No caso do MIR, as análises foram feitas com 32 scans, resolução de 4 cm^{-1} e intervalo de leitura de 4000-650 cm^{-1}. Para NIR, foram utilizados 128 scans, com resolução de 8 cm^{-1} e intervalo de leitura de 8500 a 4000 cm^{-1}, totalizando 3001 variáveis. Cada análise utilizou 1 mL de amostra e durou, aproximadamente, 1,5 e 10 min, para NIR e MIR, respectivamente.

Os espectros de NMR de hidrogênio foram medidos utilizando um equipamento Varian, com o campo magnético de 9,4 T, nas seguintes condições: largura de pulso de 90° e 11,7 μs, tempo de espera 7 s, atraso de relaxamento de 2,556 s, tempo de espera 1 s, largura espectral fixa de 6410,3 Hz e 64 transientes, a 25 °C. As amostras foram dissolvidas em 0,6 mL de diclorometano-d2 com 20 mg do óleo. No espectro de carbono, utilizou-se frequência de 100,51 MHz, largura espectral de 25510,2 Hz, tempo de espera de 7 s e ângulo de 90º e 14,2μs. As amostras foram dissolvidas em clorofórmio deuterado, CDCl3, com Cr (acac) 3 0,05 $mol \cdot L^{-1}$.

4.2.4 Quimiometria

Os modelos computacionais foram conduzidos no software MATLAB versão R2013a. Foram utilizados os espectros médios de cada amostra após um pré-tratamento determinado que variou conforme a propriedade a ser estimada. As amostras foram separadas em conjuntos de calibração, utilizado para o desenvolvimento dos modelos, e de teste, para avaliação dos modelos, com base no método k-fold. Em seguida, foram aplicados aos espectros do conjunto de amostras de calibração as diferentes estratégias de fusão e a calibração com SVR. Após a regressão, os espectros do conjunto de teste foram utilizados para determinar a exatidão e os parâmetros de desempenho dos modelos.

4.2.4.1 Processamento de dados

Antes da construção dos modelos, os espectros de infravermelho foram pré-processados com o objetivo de reduzir variações indesejadas. Para a correção de linha de base, foi utilizada a primeira derivada e o AirPLS. Para fins de teste, foram utilizados os pré-tratamentos centralização, normalização, autoescalamento, variação

normal padrão (SNV, do inglês Standard Normal Variate) e correção de espalhamento multiplicativo (MSC, do inglês Multiplicative Scatter Correction). Para obtenção de um melhor resultado, foram feitas combinações das técnicas de linha de base e pré-tratamentos, resultando em um total de 14 combinações diferentes.

Nos dados de NMR utilizou-se SNV, autoescalamento e centralização, além de, combinar SNV com centralização, dando um total de sete combinações. A partir dessas combinações, os modelos foram desenvolvidos, comparados e os melhores discutidos.

4.2.4.2 SVR

A máquina de vetores de suporte tem como objetivo fazer uma separação binaria, começando mapeando os vetores de entrada (amostras) com o objetivo de encontrar um hiperplano que separe as classes sem erro, usando o princípio da minimização dos riscos, obtendo um hiperplano de separação ótima (OSH). Entretanto, precisamos aplicar uma modificação para adaptar a classificação binária para regressão, para cada amostra $\boldsymbol{x}_i$ é número constante positivo, d, é adicionar e subtraindo da propriedade química de interesse y_i, formando assim dois grupos. O OSH aplicado na regressão, SVR, está destacado na **Equação 4.2**.

$$y_i = \boldsymbol{w} \cdot \phi(\boldsymbol{x}_i) + b \qquad \textbf{Equação 4.2}$$

Onde $\boldsymbol{w}$ é o vetor peso.[32] Nessa equação aplica-se uma constante de folga, ε, admitindo possíveis erros. A constante C controla os erros da função (ξi,ξ* ≥ 0),[22] ocorrendo a seguinte equação:

$$Min\frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \qquad \textbf{Equação 4.3}$$

$$\text{Subject to} \begin{cases} y_i - f(\boldsymbol{x}_i, \boldsymbol{\omega}) \leq \varepsilon + \xi_{\mathrm{i}}^* \\ f(\boldsymbol{x}_i, \boldsymbol{\omega}) - y_i \leq \varepsilon + \xi_{\mathrm{i}} \\ \xi\mathrm{i}, \xi* \geq 0\ , \mathrm{i} = 1, \ldots, \mathrm{n} \end{cases} \qquad \textbf{Equação 4.4}$$

As constantes C e ε precisam ser otimizadas durante a construção do modelo, neste trabalho utilizou-se a pesquisa de grade, utilizando entre 8, 16 e 20. O Problema

da **Equação 4.4** pode ser resolvido aplicando multiplicadores de Lagrange, como foi explicado por Smola A. J.,[33] e o resultado pode ser visto na **Equação 4.5:**

$$f(\boldsymbol{x}) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)K(\boldsymbol{x}_i, \boldsymbol{x}) + b \qquad \textbf{Equação 4.5}$$

Onde $K(\boldsymbol{x}_i, \boldsymbol{x})$ é a função kernel aplicada nos dados de entrada.

4.2.4.3 Fusão de baixo nível

A fusão dos espectros a nível baixo levou a onze combinações diferentes, a saber: MIR + NIR, MIR + NMR ^{1}H, MIR + NMR ^{13}C, NIR + NMR ^{1}H, NIR + NMR ^{13}C, NMR ^{1}H + NMR ^{13}C, MIR + NIR + NMR ^{1}H, MIR + NIR + NMR ^{13}C, MIR + NMR ^{1}H + NMR ^{13}C, NIR + NMR ^{1}H + NMR ^{13}C e MIR + NIR + NMR ^{1}H + NMR ^{13}C. Antes da fusão, os espectros foram pré-tratados. O pré-tratamento escolhido para cada espectro foi aquele que forneceu o melhor resultado para modelo sem fusão.

A correção de variância foi feita utilizando autoescalamento, isso é feito com o objetivo de deixar ambos os espectros com a mesma variância, deixando que nenhuma se sobreponha na outra na construção do modelo. Após pré-tratamento e fusão, os modelos foram desenvolvidos e sua qualidade avaliada.

4.2.4.4 Fusão de médio nível com PCA

Da mesma forma que os procedimentos anteriores, foram avaliados onze tipos de combinações de fusão de médio nível, entretanto, em vez de usar um único pré-tratamento, foram estudados os três mais eficazes, totalizando 243 modelos. Após a aplicação desses pré-tratamentos em cada conjunto de dados, foi construído um modelo PCA e seus principais escores extraídos. Selecionado apenas os primeiros escores, garantindo que representassem ao menos 95% da variância total dos dados. Para eliminar o problema da predominância foi aplicado autoescalamento. Por fim, os escores foram concatenados e um modelo foi gerado utilizando o método SVR.

4.2.4.5 Fusão de médio nível com GA

Semelhante com a abordagem anterior, inicialmente aplicou o método GA para selecionar os dados isolados, somente no melhor pré-tratamento. Após identificar as melhores variáveis em cada conjunto, estas foram autoescaladas e concatenadas numa única matriz e um novo modelo SVR foi gerado.

4.2.4.6 Fusão de alto nível

Nesse tipo de fusão, as predições dos melhores modelos, sem fusão, para cada amostra são somadas utilizando as combinações possíveis apresentadas no subtópico 4.2.4.3. Em seguida, é calculada uma média, que pode ser aritmética simples ou ponderada. O valor resultante dessa média é considerado o novo valor predito pela fusão. Devido às duas formas distintas de calcular a média, são considerados dois tipos de fusão de alto nível.

4.2.4.8 Avaliação de modelos

Os modelos foram avaliados por meio do coeficiente de determinação (R^2), que consiste em uma medida estatística do grau de concordância entre os valores medidos pelos ensaios experimentais padrão e os valores previstos pelos modelos. O valor varia de 0 a 1 e quanto mais próximo de 1 melhor é o ajuste do modelo. O R^2 é calculado conforme a **Equação 4.7**.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \textbf{Equação 4.7}$$

Onde n é o número de amostras, y_i é o valor observado, $\hat{y}_i$ é o valor estimado e $\underline{y}$ é a média das observações. Os modelos também foram avaliados com base nos parâmetros raiz quadrada do erro quadrático médio de validação cruzada (RMSECV) e raiz quadrada do erro quadrático médio de previsão (RMSEP). Esses parâmetros podem ser calculados conforme a **Equação 4.8**.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{df}} \qquad \textbf{Equação 4.8}$$

Para avaliar a exatidão do modelo de calibração, o SVR costuma utilizar o RMSECV. Além disso, foi aplicado o teste randômico de exatidão para analisar os melhores modelos, utilizou o teste bicaudal, com 5% de nível de significância ($\alpha > 0.05$) e 500 mil permutações, com o objetivo de determinar estatisticamente qual o melhor modelo.[34]

4.3 Resultados e discussão

Neste estudo, empregamos o SVR em conjunto com a fusão de dados para estimar propriedades físico-químicas do petróleo, tais como densidade API, ponto de fluidez, WAT, teor de saturados, aromáticos, polares, enxofre, nitrogênio total e poder calorífico. Observou-se que, para todas as propriedades, o modelo mais exato foi obtido através da fusão de dados, conforme apresentado a seguir.

4.3.1 Espectros

A **Figura 4.1** apresenta os espectros de infravermelho utilizados neste estudo. Todos os espectros têm perfil químico semelhante. As bandas de alta absorbância na faixa 3000-2800 cm^{-1}, são devido ao estiramento da ligação C–H dos grupos CH_2, e dos grupos CH_3 em 2953 cm^{-1}. As bandas em 1454 e 1375 cm^{-1} indicam a presença de grupos alcanos, CH_3 de alifáticos. Nas faixas 1300-1100 cm^{-1} e 900-720 cm^{-1} as bandas estão relacionadas às vibrações de compostos aromáticos, com deformações angulares fora do plano.[35,36]

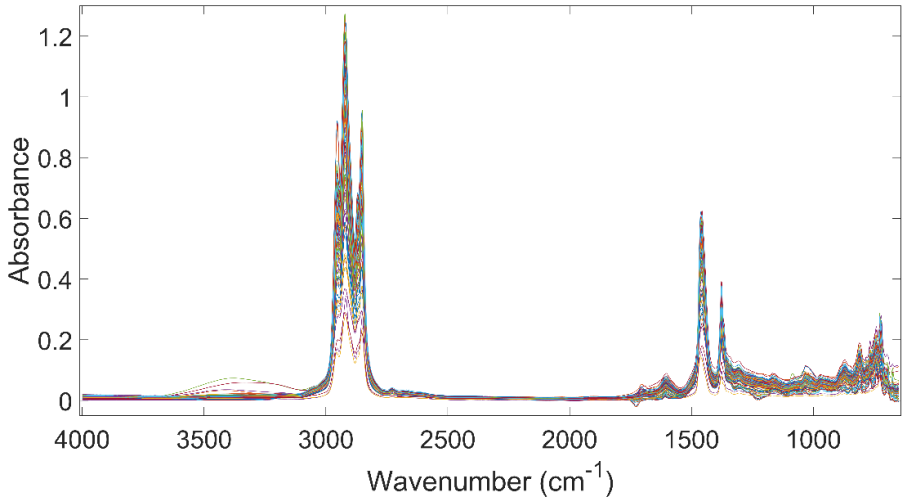


Figura 4. 1 Espectros MIR de petróleos brutos. (Fonte: Elaboração Própria)

A **Figura 4.2** apresenta os espectros de infravermelho próximo, nota-se que os espectros sofreram um processamento chamado espalhamento o que torna necessário um tratamento adequado. Podemos constatar quatro bandas em destaque, a 4000 a 4500 e 7700 cm^{-1} são referentes a bandas combinadas de C-H e as bandas 5500 a 6250 e 8000 a 9000 cm^{-1} são referentes a primeira e segunda bandas harmônicas de C-H.

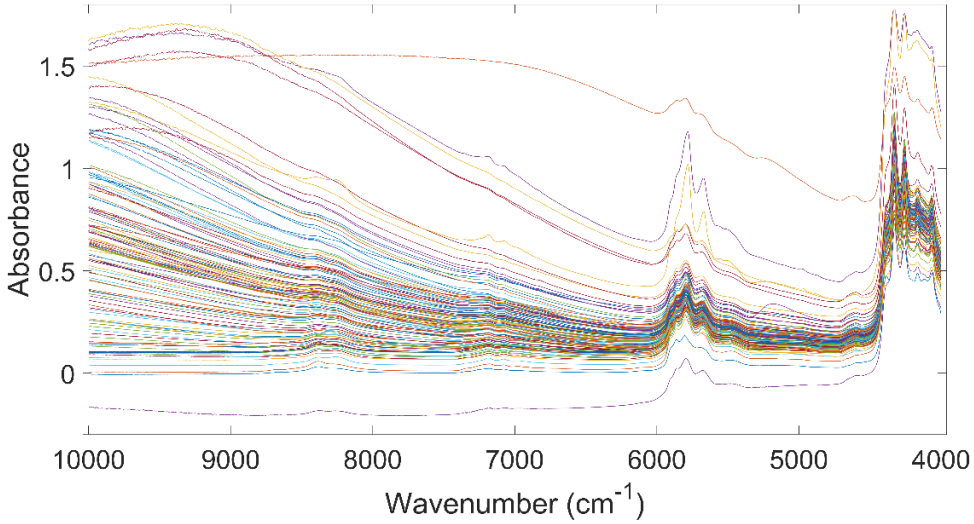


Figura 4. 2 Espectros NIR de petróleos brutos. (Fonte: Elaboração própria)

Na **Figura 4.3** podemos ver os espectros agrupados de NMR de ^{1}H. Nota-se sinal entre 0,50 e 1,00 ppm, indicando presença de CH_3 de cadeias alifáticas. O sinal

em destaque em 1,26 ppm demonstra CH_2 de cadeias alifáticas, além disso, os sinais entre 2.0 e 2.5 ppm se referem a CH_3 α-aromático e os pequenos sinais entre 7.2 e 7.4 ppm à cadeia di-aromática. O pequeno sinal em 1,42 ppm refere-se a β-aromático CH_2 e o em 5,31 ppm ao solvente utilizado.[12]

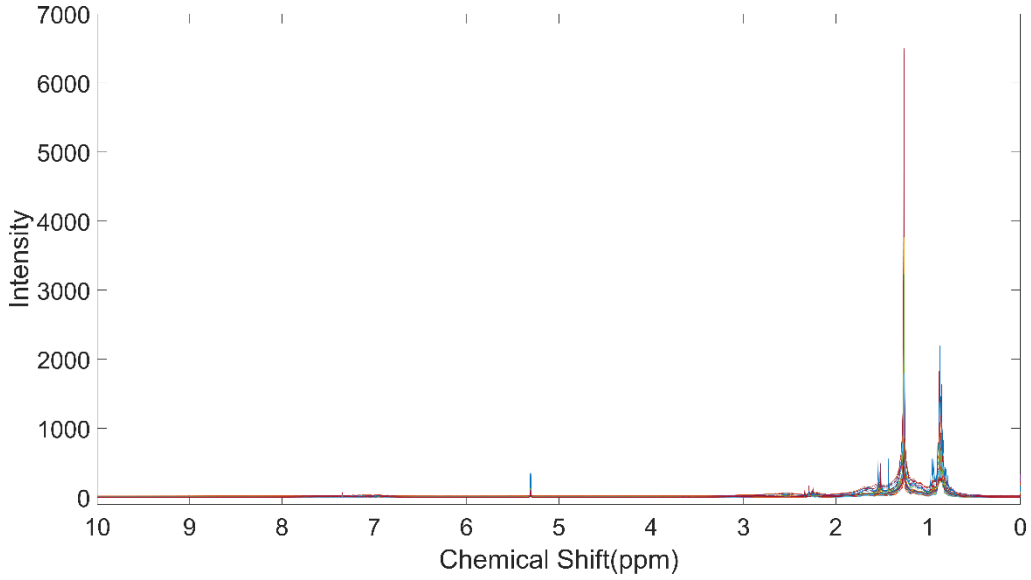


Figura 4. 3 Espectros NMR ^1H de petróleos brutos. (Fonte: Elaboração própria)

Na **Figura 4.4** estão dispostos os espectros de NMR ^{13}C das amostras utilizadas neste estudo. Na região de 40~60 ppm, os sinais são decorrentes principalmente, de carbono terciário e quaternário de parafinas. Os sinais abaixo de 40 ppm correspondem a carbonos de cadeia alifáticas. A região com sinais fracos entre 120 e 140 ppm corresponde à carbonos aromáticos. O sinal mais intenso entre 70~80 ppm pertence ao solvente utilizado na amostra.[12]

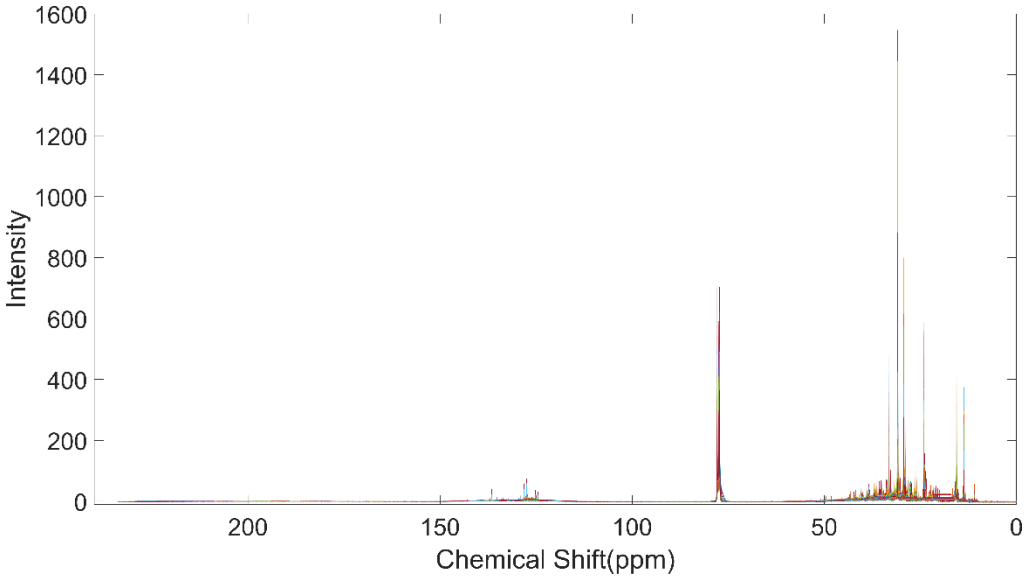


Figura 4. 4 Espectros NMR ^{13}C de petróleos brutos. (Fonte: Elaboração própria)

4.3.2 Densidade API

A densidade API é um indicador importante na classificação do petróleo, influenciando diretamente na qualidade e valor econômico, além de determinar o uso adequado do óleo na indústria. Os melhores modelos desenvolvidos estão dispostos na **Tabela 4.2**. O modelo mais exato foi obtido a partir da fusão de médio nível com PCA, utilizando os espectros de MIR e NIR, obtendo RMPEP e R^2p, respectivamente de 0,9 e 0,985. Entre os modelos sem fusão, o melhor resultado foi a partir dos espectros de MIR cujos valores de RMSEP e R^2p de 1,2 e 0,969. Aplicou-se o teste randômico de exatidão para comparar os modelos e se constatou que os valores de RMSEP apresentam diferenças estatísticas significativas. A aplicação de PCA anterior a concatenação demonstrou ser eficaz na extração das informações relevantes dos espectros, o que contribuí para obtenção de um modelo de maior exatidão. Além disso, evidenciou que, apesar de semelhantes, os espectros de infravermelho contêm informações complementares com potencial sinérgico.

Tabela 4. 2 Parâmetros de avaliação para densidade API.

Espectro	Fusão	RMSECV	RMSEP	R^2cv	R^2p
MIR	---	1,3	1,2	0,980	0,969
NIR	---	1,8	1,0	0,964	0,979
NMR ^1H	---	1,7	0,9	0,967	0,982
NMR ^{13}C	---	2,4	1,2	0,945	0,967
Mir ^1H	B - Bruto	1,5	1,0	0,976	0,979
Mir Nir ^1H	B - Bruto	1,7	1,0	0,969	0,979
Todos	B - Bruto	2,4	0,9	0,943	0,982
Mir NIR	M - PCA	1,2	0,9	0,982	0,985
Mir Nir ^1H	M - PCA	1,3	0,9	0,980	0,984
Tudo	M - PCA	1,4	0,9	0,977	0,985
Mir ^{13}C	M - GA	1,8	1,0	0,967	0,977
Mir ^1H ^{13}C	M - GA	1,7	0,9	0,977	0,982
Tudo	M - GA	1,7	0,9	0,970	0,982
Nir ^1H	A - Média	1,2	0,9	0,983	0,984
Nir ^1H ^{13}C	A - Média	1,4	0,9	0,982	0,983
Tudo	A - Média	1,3	0,9	0,984	0,983
Nir ^1H	A - Pond	1,2	0,9	0,983	0,984
Nir ^1H ^{13}C	A - Pond	1,3	0,9	0,982	0,983
Tudo	A - Pond	1,3	0,9	0,984	0,983

Ao analisar a Tabela 4.2, observa-se uma maior frequência da combinação de MIR com NMR, provavelmente devido à complementaridade, ou sinergia, entre essas técnicas analíticas. O MIR destaca-se por suas bandas relacionadas a parafinas, alcanos e compostos alifáticos,[37] enquanto o NMR fornece informações sobre parafinas e compostos naftênicos.[38]

4.3.3 Ponto de fluidez

Os modelos para a determinação do ponto de fluidez encontram-se na **Tabela 4.3**. Esta propriedade é importante na indústria, pois determina a temperatura em que o petróleo deixa de fluir pela ação da gravidade, ou seja, a temperatura que a viscosidade impede o escoamento do petróleo.[39] O modelo que aplica fusão de alto nível e média combinado aos dados de MIR, NIR e NMR H^1, demonstrou a maior

exatidão entre os modelos, obtendo um RMSEP igual a 8 ºC e R²p 0,899. Entre os modelos sem fusão, o modelo com MIR se destacou, apresentando um RMSEP 9 ºC e R²p de 0,839. Testes randômicos não revelaram diferenças estatística significativa na exatidão entre esses dois modelos.

Tabela 4. 3 Parâmetros de avaliação para ponto de fluidez.

Espectro	Fusão	RMSECV	RMSEP	R²cv	R²p
MIR	---	8	8	0,865	0,839
NIR	---	11	9	0,730	0,863
^{1}H	---	13	9	0,611	0,898
^{13}C	---	18	12	0,237	0,650
Mir NIR	B - Bruto	9	10	0,808	0,810
Mir Nir ^{1}H	B - Bruto	12	9	0,671	0,819
Todos	B - Bruto	16	10	0,357	0,764
Nir ^{13}C	M - PCA	14	12	0,492	0,787
Nir ^{1}H ^{13}C	M - PCA	15	12	0,437	0,807
Tudo	M - PCA	13	13	0,600	0,637
Nir ^{1}H	M - GA	11	7	0,682	0,899
Mir Nir ^{1}H	M - GA	10	10	0,762	0,815
Tudo	M - GA	15	10	0,482	0,760
Nir ^{1}H	A - Média	11	9	0,722	0,911
Mir Nir ^{1}H	A - Média	9	8	0,809	0,899
Tudo	A - Média	10	8	0,782	0,868
Nir ^{1}H	A - Pond	11	9	0,728	0,911
Mir Nir ^{1}H	A - Pond	9	8	0,822	0,900
Tudo	A - Pond	9	8	0,818	0,877

No estudo de Rodrigues *et al.*, em 2018,[39] a cromatografia gasosa de alta temperatura foi empregada para estimar o ponto de fluidez, resultando em RMSEP de 12 ºC e R²p de 0,711. Esses parâmetros foram inferiores aos obtidos no presente estudo. Os autores sugerem que essa disparidade ocorre devido à grande variação na propriedade e nas moléculas dentro de uma mesma classe de amostras.

Em 2009, Peinder P. *et al.*,[3] empregaram a fusão de dados combinando a espectroscopia de infravermelho com NMR e PLS para predizer propriedades do petróleo. No caso do ponto de fluidez, o modelo utilizando apenas NMR de H¹

apresentou um RMSEP de 9,2 ºC, sendo o melhor dos modelos sem fusão, enquanto, com fusão destacou-se o com MIR, NMR H^1 e ^{13}C resultando em um RMSEP de 10,3 ºC. Neste caso, observou-se uma desvantagem em relação ao parâmetro de avaliação, em contraste com os resultados deste estudo, que apresentou uma melhora. Todavia, é importante notar que os estudos citados aplicaram PLS, enquanto utilizamos SVR, indicando a possível presença de relações não-lineares, o que justifica os resultados mais promissores obtidos através deste método.

4.3.4 WAT

A temperatura de aparecimento de cristal (WAT do inglês *wax appearance temperature*) é importante na indústria de petróleo por indicar a temperatura que as parafinas cristalizam, dessa forma, a indústria consegue se precaver quanto a entupimentos em suas refinarias e na extração do óleo. Como podemos observar na **Tabela 4.4**, o melhor modelo foi a fusão de médio nível com PCA combinado entre MIR e NMR ^{13}C que obteve um RMSEP de 2,66 e um R^2p de 0,4461, parâmetros de avaliação superiores aos obtidos no modelo sem fusão, que foi o MIR com RMSEP de 3,10 e R^2p de 0,317, contudo, não foi encontrado diferença estatística entre estes modelos, apesar de ser encontrado entre o modelo com fusão e os sem fusão com NMR. O que corrobora com a ideia de que NMR ^{13}C e MIR são técnicas com informações sinérgicas, ou complementares, ao menos quando o assunto é WAT. Todavia, esses resultados não obtiveram um resultado adequado para aplicação que exige o mínimo de 0,700 de R^2.

Tabela 4. 4 Parâmetros de avaliação para WAT.

Espectro	Fusão	RMSECV	RMSEP	R²cv	R²p
MIR	---	3,42	3,10	0,589	0,317
NIR	---	3,65	3,11	0,520	0,291
^{1}H	---	3,40	3,61	0,594	0,209
^{13}C	---	4,28	3,27	0,347	0,212
Nir ^{1}H	B - Bruto	3,43	3,03	0,577	0,296
Mir Nir ^{1}H	B - Bruto	3,46	3,16	0,577	0,267
Todos	B - Bruto	3,71	2,95	0,506	0,334
Mir ^{13}C	M - PCA	4,05	2,66	0,419	0,446
Mir ^{1}H ^{13}C	M - PCA	3,67	2,94	0,532	0,334
Tudo	M - PCA	3,43	2,99	0,592	0,315
^{1}H ^{13}C	M - GA	3,92	3,11	0,447	0,330
Mir ^{1}H ^{13}C	M - GA	3,82	3,01	0,474	0,354
Tudo	M - GA	3,59	2,95	0,543	0,337
Mir NIR	A - Média	3,33	3,01	0,611	0,327
Mir Nir ^{13}C	A - Média	3,46	2,95	0,584	0,321
Tudo	A - Média	3,29	3,00	0,633	0,308
Mir NIR	A - Pond	3,33	3,01	0,613	0,328
Mir Nir ^{13}C	A - Pond	3,40	2,95	0,601	0,325
Tudo	A - Pond	3,22	2,98	0,649	0,314

4.3.5 Teor de saturados

O teor de saturados é indicativo da proporção de moléculas saturadas na composição do petróleo, contendo compostos como n-parafinas, iso-parafinas e naftenos. Essas moléculas saturadas desempenham um papel crucial na fluidez do petróleo, sendo associadas a petróleos de baixa densidade.[28,32] Os modelos desenvolvidos para determinação de saturados estão dispostos na **Tabela 4.5.** O modelo com maior exatidão foi desenvolvido a partir da fusão de MIR, NIR e NMR de ^{13}C na fusão de nível alto atingindo um RMSEP igual a 5,1 wt% e R²p igual a 0,840. Entre os modelos sem fusão, o mais exato foi aquele a partir de dados de NMR ^{13}C com RMSEP igual a 6,0 wt% e R²p 0,784. O teste randômico de exatidão acusou diferença significativa entre estes dois modelos, então podemos dizer que a fusão desenvolveu um modelo melhor. Novamente, os espectros de infravermelho tiveram

maior destaque na modelagem, isso pode ocorrer pois no infravermelho temos diversas bandas relacionadas a parafinas, grupos alcanos e alifáticos.

Tabela 4. 5 Parâmetros de avaliação para teor de saturados.

Espectro	Fusão	RMSECV	RMSEP	R²cv	R²p
MIR	---	5,8	5,1	0,833	0,825
NIR	---	6,5	5,3	0,784	0,836
^{1}H	---	5,9	6,2	0,824	0,747
^{13}C	---	7,2	6,0	0,731	0,784
Mir NIR	B - Bruto	7,0	5,7	0,750	0,798
Mir Nir ^{1}H	B - Bruto	6,8	6,3	0,762	0,756
Todos	B - Bruto	7,1	6,4	0,746	0,758
Nir ^{13}C	M - PCA	7,6	5,8	0,711	0,775
Mir Nir ^{1}H	M - PCA	6,7	5,9	0,766	0,781
Tudo	M - PCA	7,1	6,0	0,737	0,783
Mir NIR	M - GA	6,2	5,6	0,801	0,798
Mir Nir ^{13}C	M - GA	7,1	5,7	0,737	0,811
Tudo	M - GA	7,0	6,0	0,751	0,786
Mir NIR	A - Média	5,8	4,9	0,831	0,851
Mir Nir ^{13}C	A - Média	6,0	5,1	0,818	0,840
Tudo	A - Média	5,7	5,2	0,834	0,828
Mir NIR	A - Pond	5,8	4,9	0,831	0,851
Mir Nir ^{13}C	A - Pond	5,9	5,1	0,820	0,841
Tudo	A - Pond	5,7	5,2	0,836	0,829

Comparando com os resultados apresentados no Capítulo 3, os obtidos com a fusão de dados não demonstraram maior exatidão, sugerindo que a seleção de variáveis pode ser a melhor abordagem. Além disso, Rodrigues et al. (2018),[39] utilizando dados de cromatografia gasosa, obtiveram um RMSEP de 6,8 wt% e um R²p de 0,692 para a previsão de saturados, indicando um modelo menos exato do que os obtidos neste estudo. Para a previsão de saturados, Moro et al. (2020)[12] aplicaram fusão de dados de nível baixo e médio em dados de MIR, NMR de ^{1}H e NMR de ^{13}C e utilizaram PLS. Os autores obtiveram resultados semelhantes aos deste estudo, com RMSEP de 5,12 wt% e R²p de 0,85, e uma melhoria significativa com a aplicação da fusão de dados.

4.3.6 Teor de aromáticos

O teor de aromáticos, como sugere, reflete a proporção de hidrocarbonetos aromáticos no petróleo, exercendo impacto direto na sua qualidade e valor econômico. Este índice serve como um indicador significativo da maturidade do óleo,[40] estabilidade de emulsões[41] e, em conjunto com o teor de saturados, no ponto de fluidez.[42] Os modelos desenvolvidos para determinação do teor de aromáticos estão dispostos na **Tabela 4.6.** O modelo mais exato foi desenvolvido a partir dos espectros de MIR, com RMSEP e R^2p, respectivamente, iguais a 3,9 wt% e 0,585. Entre os modelos que utilizaram fusão, o melhor resultado foi a partir da fusão de MIR e NIR com fusão de alto nível, obtendo RMSEP e R^2p, respectivamente, iguais a 3,9 wt% e 0,578. Os espectros que proporcionaram os melhores resultados foram MIR e NMR de ^{13}C. O primeiro devido às diversas bandas relacionadas aos compostos aromáticos,[23] já o segundo devido aos sinais de carbono quaternários e grupos funcionais sem átomos de hidrogênio, carregando assim bastante informação sobre moléculas aromáticas.[38]

Tabela 4. 6 Parâmetros de avaliação para teor de aromáticos.

Espectro	Fusão	RMSECV	RMSEP	R^2cv	R^2p
MIR	---	3,3	3,9	0,750	0,585
NIR	---	3,6	4,2	0,702	0,527
^1H	---	3,5	4,3	0,721	0,469
^{13}C	---	3,7	4,2	0,685	0,525
Nir ^{13}C	B - Bruto	3,6	4,3	0,699	0,491
Nir ^1H ^{13}C	B - Bruto	3,5	4,5	0,723	0,459
Todos	B - Bruto	3,5	4,4	0,709	0,474
Mir ^1H	M - PCA	3,7	4,0	0,698	0,539
Mir ^1H ^{13}C	M - PCA	3,5	4,0	0,733	0,556
Tudo	M - PCA	3,4	4,3	0,744	0,488
Mir ^{13}C	M - GA	3,5	4,1	0,723	0,527
Mir Nir ^{13}C	M - GA	3,5	4,2	0,723	0,518
Tudo	M - GA	3,4	4,3	0,737	0,493
Mir NIR	A - Média	3,2	3,9	0,768	0,577
Mir Nir ^{13}C	A - Média	3,2	4,0	0,768	0,573
Tudo	A - Média	3,1	4,0	0,790	0,561
Mir NIR	A - Pond	3,2	3,9	0,769	0,578
Mir Nir ^{13}C	A - Pond	3,2	4,0	0,769	0,575
Tudo	A - Pond	3,0	4,0	0,791	0,564

Paulo E. H. *et al*. (2020)[38] utilizaram dados de NMR de ^{13}C, seleção de variáveis e PLS para determinar o teor de aromáticos. Enquanto o modelo a partir do espectro integral apresentou RMSEP de 3,35 wt%, a seleção de variáveis trouxe uma melhora considerável, com RMSEP de 2,86 wt%, resultado semelhante ao obtido neste estudo. Moro *et al*. (2020)[12] também aplicaram fusão para predizer o teor de aromáticos. O melhor resultado foi a partir da fusão de MIR e NMR de ^{13}C a nível médio, utilizando PLS para extração das informações, obtendo RMSEP e R^2p, respectivamente, iguais a 3,66 wt% e 0,67. Os autores não conseguiram um bom modelo somente com MIR, se comparados com os modelos deste estudo, entretanto, seu método de regressão foi linear, o que pode ter limitado a qualidade de seus resultados.

4.3.7 Teor de polares

O teor de polares, consiste na combinação de resinas e asfaltemos do petróleo, ou seja, compostos aromáticos policíclicos de alto peso molecular,[32] e tem relação com parafinas e heteroátomos do petróleo.[38] Os modelos desenvolvidos para determinação do teor de polares estão dispostos na **Tabela 4.7**. A fusão de MIR e NIR a nível baixo proporcionou o melhor resultado, com RMSEP e R^2p, respectivamente, iguais a 5,0 wt% e 0,647. Entre os modelos sem fusão, o mais exato foi aquele desenvolvido com os espectros de NIR, obtendo RMSEP e R^2p, respectivamente, iguais a 5,2 wt% e 0,623. O teste randômico não indicou diferença significativa entre as exatidões, contudo o modelo com fusão apresentou diferença estatística com o modelo sem fusão de NMR de ^{13}C. Neste estudo, o infravermelho se destacou como a fonte analítica mais eficaz, tanto de forma isolada quanto na fusão de dados, possivelmente devido à sua correlação com as parafinas.

Tabela 4. 7 Parâmetros de avaliação para teor de polares.

Espectro	Fusão	RMSECV	RMSEP	R²cv	R²p
MIR	---	5,1	5,2	0,718	0,605
NIR	---	5,6	5,2	0,661	0,623
^{1}H	---	6,0	5,5	0,619	0,545
^{13}C	---	6,7	5,5	0,517	0,571
Mir NIR	B - Bruto	5,3	5,0	0,697	0,647
Mir Nir ^{1}H	B - Bruto	5,7	5,1	0,653	0,606
Todos	B - Bruto	6,2	5,4	0,589	0,589
Nir ^{13}C	M - PCA	5,6	4,9	0,665	0,631
Mir Nir ^{1}H	M - PCA	5,1	5,4	0,717	0,596
Tudo	M - PCA	5,9	5,3	0,621	0,581
Mir NIR	M - GA	5,2	5,1	0,710	0,640
Mir Nir ^{13}C	M - GA	6,1	5,3	0,598	0,621
Tudo	M - GA	6,0	5,2	0,616	0,616
Mir NIR	A - Média	5,1	5,1	0,719	0,640
Mir Nir ^{1}H	A - Média	5,2	5,0	0,710	0,630
Tudo	A - Média	5,5	5,1	0,685	0,622
Mir NIR	A - Pond	5,1	5,1	0,720	0,640
Mir Nir ^{1}H	A - Pond	5,2	5,0	0,712	0,631
Tudo	A - Pond	5,4	5,1	0,693	0,623

Moro *et al*. (2020)[12] aplicou fusão para predizer o teor de polares. Seu melhor resultado foi a partir da fusão de NMR de ^{1}H e ^{13}C, com PCA a nível médio obtendo RMSEP e R²p, respectivamente, iguais a 6,60 wt% e 0,65, resultados semelhantes aos obtidos neste trabalho.

4.3.8 Enxofre

A concentração de enxofre tem um importante papel na determinação de qualidade do produto e no processo de refino. Elevados teores de enxofre comprometem catalisadores químicos, prejudicando o refino e os motores de veículos. Portanto, o conhecimento da concentração de enxofre assegura a produção de um produto final de qualidade.[43] Os modelos desenvolvido para predizer enxofre estão na

Tabela 4.8.

Tabela 4. 8 Parâmetros de avaliação para teor de enxofre.

Espectro	Fusão	RMSECV	RMSEP	R²cv	R²p
MIR	---	0,076	0,090	0,876	0,692
NIR	---	0,107	0,100	0,754	0,649
^{1}H	---	0,067	0,110	0,899	0,587
^{13}C	---	0,113	0,109	0,718	0,574
Mir NIR	B - Bruto	0,103	0,090	0,791	0,699
Mir Nir ^{1}H	B - Bruto	0,083	0,110	0,849	0,587
Todos	B - Bruto	0,114	0,106	0,711	0,598
Mir ^{1}H	M - PCA	0,105	0,091	0,763	0,689
Mir Nir ^{1}H	M - PCA	0,105	0,093	0,782	0,672
Tudo	M - PCA	0,108	0,097	0,750	0,644
Mir NIR	M - GA	0,091	0,095	0,842	0,676
Mir Nir ^{13}C	M - GA	0,102	0,105	0,774	0,598
Tudo	M - GA	0,103	0,106	0,763	0,601
Mir NIR	A - Média	0,079	0,086	0,864	0,721
Mir Nir ^{13}C	A - Média	0,081	0,089	0,860	0,708
Tudo	A - Média	0,072	0,091	0,890	0,694
Mir NIR	A - Pond	0,086	0,088	0,839	0,712
Mir Nir ^{13}C	A - Pond	0,089	0,091	0,829	0,693
Tudo	A - Pond	0,077	0,092	0,873	0,680

O modelo mais eficaz foi com fusão de alto nível combinando dados de infravermelho. Este modelo apresentou um RMSEP de 0,086 e um R²p de 0,721. Sem a fusão, o melhor modelo foi obtido apenas com o espectro MIR, resultando em um RMSEP de 0,090 e um R²p de 0,624. O destaque do MIR foi provavelmente devido as bandas relacionadas a óxidos de enxofre, como estiramento simétricos em 970-1030 cm^{-1}, e tioéteres, estiramento C-S 600-750 cm^{-1}.[44] Ao optar pela fusão de alto nível como melhor modelo, conclui-se que a combinação de resultados proporcionou um desempenho em comparação a abordagem que produz sinergismo de resultados.

Rocha *et al*. (2016)[43] analisou a concentração de enxofre utilizando infravermelho (MIR e NIR) e diversas técnicas de seleção de variáveis, obtendo o melhor resultado com iPLS, PLS por intervalos, e MIR com um RMSEP de 0,042 e

R^2p de 0,973, resultados superiores aos conseguidos neste estudo, contudo, justifica a eficiência do MIR na determinação de enxofre.

4.3.9 Teor de nitrogênio total

Entre os heteroátomos presentes no petróleo, o nitrogênio se destaca devido à sua capacidade de contaminar catalisadores durante o processo de refino, o que pode reduzir a qualidade dos derivados e desencadear reações indesejadas, como a contaminação de catalisadores. Assim, a determinação do teor de nitrogênio total (NT) antes do processo de refino é crucial.[45]

Os melhores modelos obtidos neste estudo estão na **Tabela 4.9**. O modelo mais eficaz, com fusão de dados, utilizou espectros de MIR e NMR de ^1H com fusão de alto nível com média aritmética, obtendo um RMSEP de 0,0244 wt% e um R^2p de 0,945. Por outro lado, o modelo sem fusão, que utilizou apenas MIR, apresentou um RMSEP de 0,0304 wt% e um R^2p de 0,931. Houve diferença estatisticamente significativa entre o modelo com fusão e os modelos sem fusão, com exceção do modelo MIR. Desse modo, o modelo com fusão foi escolhido como o melhor para NT, devido aos seus parâmetros de avaliação superiores. A fonte analítica que se destacou nos modelos foi o MIR, possivelmente devido às bandas N-H presentes.[46]

Tabela 4. 9 Parâmetros de avaliação para teor de nitrogênio total.

Espectro	Fusão	RMSECV	RMSEP	R²cv	R²p
MIR	---	0,0385	0,0304	0,914	0,931
NIR	---	0,0416	0,0342	0,899	0,889
^{1}H	---	0,0484	0,0283	0,863	0,923
^{13}C	---	0,0616	0,0376	0,778	0,864
Mir NIR	B - Bruto	0,0432	0,0273	0,891	0,928
Mir Nir ^{1}H	B - Bruto	0,0511	0,0292	0,847	0,919
Todos	B - Bruto	0,0550	0,0338	0,826	0,888
Mir NIR	M - PCA	0,0462	0,0277	0,878	0,926
Mir Nir ^{13}C	M - PCA	0,0549	0,0330	0,825	0,903
Tudo	M - PCA	0,0516	0,0337	0,845	0,901
Mir NIR	M - GA	0,0378	0,0276	0,916	0,926
Mir ^{1}H ^{13}C	M - GA	0,0514	0,0307	0,845	0,913
Tudo	M - GA	0,0525	0,0325	0,839	0,899
Mir ^{1}H	A - Média	0,0392	0,0244	0,910	0,945
Mir Nir ^{1}H	A - Média	0,0361	0,0256	0,924	0,940
Tudo	A - Média	0,0395	0,0268	0,910	0,933
Mir ^{1}H	A - Pond	0,0388	0,0245	0,912	0,945
Mir Nir ^{1}H	A - Pond	0,0367	0,0258	0,921	0,938
Tudo	A - Pond	0,0397	0,0269	0,909	0,933

Rainha *et al.*, em 2019,[45] utilizaram MIR e NIR em combinação com PLS e seleção de variáveis para determinar o NT. O melhor resultado foi obtido com o uso do NIR em conjunto com amostragem competitiva adaptativa (CARS, do inglês *competitive adaptive reweighted sampling*), resultando em um RMSEP de 0,0275 e um R²p de 0,951, resultados semelhantes aos obtidos neste estudo. Além disso, os autores, conseguiram demonstrar uma relação entre o índice de API e o nitrogênio total.

4.3.10 Poder calorifico

O poder calorifico indica a capacidade total de uma amostra de liberar energia, propriedade importante para analisar o potencial do petróleo em virar combustível. Os

modelos desenvolvidos para determinação do poder calorífico superior estão dispostos na **Tabela 4.10**. A fusão das quatro fontes analíticas utilizando alto nível proporcionou o melhor resultado, obtendo RMSEP e R²p, respectivamente, de 0,3 MJ/kg e 0,716. Entre os modelos sem fusão, os espectros MIR proporcionaram o melhor resultado, com RMSEP e R2p, respectivamente de 0,4 MJ/kg e 0,652. O teste randômico indicou que a diferença na exatidão dos modelos não é significativa.

Tabela 4. 10 Parâmetros de avaliação para poder calorifico**.**

Espectro	Fusão	RMSECV	RMSEP	R²cv	R²p
MIR	---	0,7	0,4	0,457	0,652
NIR	---	0,8	0,3	0,320	0,698
^{1}H	---	0,7	0,4	0,491	0,601
^{13}C	---	0,8	0,4	0,370	0,578
Mir ^{1}H	B - Bruto	0,7	0,4	0,467	0,617
Mir Nir ^{1}H	B - Bruto	0,8	0,4	0,383	0,587
Todos	B - Bruto	0,7	0,5	0,420	0,516
Mir ^{13}C	M - PCA	0,8	0,4	0,414	0,591
Mir ^{1}H ^{13}C	M - PCA	0,8	0,4	0,369	0,588
Tudo	M - PCA	0,8	0,4	0,370	0,561
Mir NIR	M - GA	0,8	0,4	0,413	0,638
Mir Nir ^{1}H	M - GA	0,8	0,3	0,314	0,703
Tudo	M - GA	0,7	0,5	0,472	0,545
Mir NIR	A - Média	0,7	0,3	0,420	0,716
Nir ^{1}H ^{13}C	A - Média	0,7	0,3	0,449	0,711
Tudo	A - Média	0,7	0,3	0,469	0,715
Mir NIR	A - Pond	0,7	0,3	0,440	0,709
Nir ^{1}H ^{13}C	A - Pond	0,7	0,3	0,460	0,714
Tudo	A - Pond	0,7	0,3	0,478	0,716

Os infravermelhos foram as fontes analíticas que conseguiram os melhores resultados nos modelos, provavelmente isso ocorreu devido a ligação desse espectro com ligações orgânicas ligadas diretamente a valor calorífico, como C-C, S-H, C-H,C-N, N-H e outros.[46] Paulo E. H. *et al*. (2022)[30] aplicaram a seleção de variáveis chamada de otimizador por exame de partículas (PSO, do inglês particle swarm optimization) e PLS em dados de NMR de ^{13}C para modelar poder calorífico, conseguindo um

resultado superior ao obtido neste estudo, com RMSEP e R2p, respectivamente, iguais 0,152 MJ/kg e 0,970 selecionando-se principalmente regiões dos espectros relacionadas às parafinas.

4.4 Conclusão

Espectros de quatro fontes analíticas distintas - MIR, NIR, NMR ^{1}H e NMR ^{13}C - foram empregadas para predizer nove propriedades físico-químicas do petróleo; densidade API, ponto de fluidez, WAT, teor de saturados, aromáticos, polares, enxofre, NT e poder calorifico. Para isso, foram aplicandas a estratégia de fusão de dados de diferentes níveis e o método não-linear de regressão, SVR. A extração de informação da fusão de nível médio foi executada aplicando-se PCA e GA.

A fusão de dados, associada ao SVR, proporcionou modelos com maior exatidão, ou equivalentes, aos modelos sem fusão de dados em todos os casos, obtendo-se diferença estatística com base no teste de exatidão randômico. Como a modelagem de densidade API demonstrou, o melhor modelo foi a fusão de médio nível com PCA, obtendo RMSEP de 0,9 e R^2p de 0,985, superando os modelos sem fusão no teste estatístico. Nessa abordagem, a PCA demonstrou conseguir extrair a informação relevante dos espectros.

A combinação de espectros de ressonância magnética nuclear e infravermelho produzem informações sinérgicas que favorecem os modelos de regressão não-linear. Como evidenciaram pelos modelos de NT, o melhor resultado foi obtido com a fusão de alto nível, que combinou MIR e NMR de ^{1}H, resultando num RMSEP de 0,0244 wt% e R^2p de 0,945, sendo resultados mais exatos e estatisticamente melhor que os modelos sem fusão. A fusão de alto nível não utiliza o sinergismo dos espectros para obter um bom resultado, mas sim a combinação dos resultados de modelos prontos, tratando-se, assim, de uma estratégia de rápida aplicação e avaliação se comparada as outras fusões.

A mescla de método não-linear com a fusão de dados demonstrou ser uma estratégia eficiente para o desenvolvimento de modelos mais exatos, sendo vantajosa para situações em que uma única fonte analítica não consegue alcançar a exatidão necessária para modelagem.

4.5 Referência

1. Ferreira, M. M. C., Antunes, A. M., Melgo, M. S. & Volpe, P. L. O. Quimiometria I: calibração multivariada, um tutorial. *Quim. Nova* **22**, 724–731 (1999).
2. Castanedo, F. A review of data fusion techniques. *Sci. World J.* **2013**, (2013).
3. de Peinder, P. *et al.* Partial least squares modeling of combined infrared, 1H NMR and 13C NMR spectra to predict long residue properties of crude oils. *Vib. Spectrosc.* **51**, 205–212 (2009).
4. Cheng, J. H. & Sun, D. W. Data fusion and hyperspectral imaging in tandem with least squares-support vector machine for prediction of sensory quality index scores of fish fillet. *LWT - Food Sci. Technol.* **63**, 892–898 (2015).
5. Chen, J., Yang, C., Zhu, H., Li, Y. & Gui, W. A novel variable selection method based on stability and variable permutation for multivariate calibration. *Chemom. Intell. Lab. Syst.* **182**, 188–201 (2018).
6. Yang, B. *et al.* Rapid prediction of yellow tea free amino acids with hyperspectral images. *PLoS One* **14**, e0210084 (2019).
7. Ferrer-Cid, P., Barcelo-Ordinas, J. M., Garcia-Vidal, J., Ripoll, A. & Viana, M. Multisensor Data Fusion Calibration in IoT Air Pollution Platforms. *IEEE Internet Things J.* **7**, 3124–3132 (2020).
8. Khaleghi, B., Khamis, A., Karray, F. O. & Razavi, S. N. Multisensor data fusion: A review of the state-of-the-art. *Inf. Fusion* **14**, 28–44 (2013).
9. Lahat, D., Adali, T. & Jutten, C. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* **103**, 1449–1477 (2015).
10. Smolinska, A., Engel, J., Szymanska, E., Buydens, L. & Blanchet, L. General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences. in *Data Handling in Science and Technology* vol. 31 51–79 (2019).
11. Steinmetz, V., Sévila, F. & Bellon-Maurel, V. A methodology for sensor fusion design: Application to fruit quality assessment. *J. Agric. Eng. Res.* **74**, 21–31 (1999).
12. Moro, M. K. *et al.* FTIR, 1H and 13C NMR data fusion to predict crude oils properties. *Fuel* **263**, 116721 (2020).
13. Li, Y., Zhang, J.-Y. & Wang, Y.-Z. FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng. *Anal. Bioanal. Chem.* **410**, 91–103 (2018).

14. Borràs, E. *et al.* Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Anal. Chim. Acta* **891**, 1–14 (2015).
15. Geurts, B. P. *et al.* Improving high-dimensional data fusion by exploiting the multivariate advantage. *Chemom. Intell. Lab. Syst.* **156**, 231–240 (2015).
16. Bevilacqua, M. *et al.* Recent chemometrics advances for foodomics. *TrAC - Trends Anal. Chem.* **96**, 42–51 (2017).
17. Smolinska, A. *et al.* Interpretation and visualization of non-linear data fusion in kernel space: Study on metabolomic characterization of progression of multiple sclerosis. *PLoS One* **7**, (2012).
18. Muhammad, A. & Azeredo, R. B. D. V. 1H NMR spectroscopy and low-field relaxometry for predicting viscosity and API gravity of Brazilian crude oils - A comparative study. *Fuel* **130**, 126–134 (2014).
19. Dearing, T. I., Thompson, W. J., Rechsteiner, C. E. & Marquardt, B. J. Characterization of Crude Oil Products Using Data Fusion of Process Raman, Infrared, and Nuclear Magnetic Resonance (NMR) Spectra. *Appl. Spectrosc.* **65**, 181–186 (2011).
20. Hemmati-Sarapardeh, A., Aminshahidy, B., Pajouhandeh, A., Yousefi, S. H. & Hosseini-Kaldozakh, S. A. A soft computing approach for the determination of crude oil viscosity: Light and intermediate crude oil systems. *J. Taiwan Inst. Chem. Eng.* **59**, 1–10 (2016).
21. Rocha, W. F. de C. & Sheen, D. A. Determination of physicochemical properties of petroleum derivatives and biodiesel using GC/MS and chemometric methods with uncertainty estimation. *Fuel* **243**, 413–422 (2019).
22. Ghorbani, M., Zargar, G. & Jazayeri-Rad, H. Prediction of asphaltene precipitation using support vector regression tuned with genetic algorithms. *Petroleum* **2**, 301–306 (2016).
23. Voigt, M. *et al.* Using fieldable spectrometers and chemometric methods to determine RON of gasoline from petrol stations: A comparison of low-field 1H NMR@80 MHz, handheld RAMAN and benchtop NIR. *Fuel* **236**, 829–835 (2019).
24. Folli, G. S. *et al.* Variable selection in support vector regression using angular search algorithm and variance inflation factor. *J. Chemom.* **34**, 1–16 (2020).
25. Yu, S. *et al.* Qualitative and quantitative assessment of flavor quality of Chinese soybean paste using multiple sensor technologies combined with chemometrics

and a data fusion strategy. *Food Chem.* **405**, 134859 (2022).

26. Maimaitijiang, M. *et al.* Unmanned Aerial System (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. *ISPRS J. Photogramm. Remote Sens.* **134**, 43–58 (2017).
27. ISO 12185. Crude petroleum and petroleum products – determination ofdensity – oscillating U-tube method. at (1996).
28. Ferreira, P. S. *et al.* SAP fractions from light, medium and heavy oils: Correlation between chemical profile and stationary phases. *Fuel* **274**, 117866 (2020).
29. Sad, C. M. S. *et al.* Limitations of the pour point measurement and the influence of the oil composition on its detection using principal component analysis. *Energy and Fuels* **28**, 1686–1691 (2014).
30. de Paulo, E. H. *et al.* Determination of gross calorific value in crude oil by variable selection methods applied to 13C NMR spectroscopy. *Fuel* **311**, 122527 (2022).
31. Astm. Standard Test Method for Heat of Combustion of Liquid Hydrocarbon Fuels by Bomb Calorimeter. *ASTM D 240-92 (reapproved 1997)* 144–151 (1997) doi:10.1520/D4809-18.mendations.
32. Filgueiras, P. R. *et al.* Determination of Saturates, Aromatics, and Polars in Crude Oil by 13C NMR and Support Vector Regression with Variable Selection by Genetic Algorithm. *Energy and Fuels* **30**, 1972–1978 (2016).
33. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
34. van der Voet, H. Comparing the predictive accuracy of models using a simple randomization test. *Chemom. Intell. Lab. Syst.* **25**, 313–323 (1994).
35. Hongfu, Y., Xiaoli, C., Haoran, L. & Yupeng, X. Determination of multi-properties of residual oils using mid-infrared attenuated total reflection spectroscopy. *Fuel* **85**, 1720–1728 (2006).
36. Laxalde, J., Caillol, N., Wahl, F., Ruckebusch, C. & Duponchel, L. Combining near and mid infrared spectroscopy for heavy oil characterisation. *Fuel* **133**, 310–316 (2014).
37. da Cunha, P. H. P. *et al.* Variable selection by permutation applied in support vector regression models. *J. Chemom.* **36**, 1–14 (2022).
38. de Paulo, E. H. *et al.* Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties. *Fuel* **279**, 118462 (2020).

39. Rodrigues, É. V. A., Silva, S. R. C., Romão, W., Castro, E. V. R. & Filgueiras, P. R. Determination of crude oil physicochemical properties by high-temperature gas chromatography associated with multivariate calibration. *Fuel* **220**, 389–395 (2018).
40. 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* **98**, 5–14 (2012).
41. Terra, L. A. *et al.* Laser desorption ionization FT-ICR mass spectrometry and CARSPLS for predicting basic nitrogen and aromatics contents in crude oils. *Fuel* **160**, 274–281 (2015).
42. Lovatti, B. P. O., Nascimento, M. H. C., Neto, Á. C., Castro, E. V. R. & Filgueiras, P. R. Use of Random forest in the identification of important variables. *Microchem. J.* **145**, 1129–1134 (2019).
43. Rocha, J. T. C. *et al.* Sulfur Determination in Brazilian Petroleum Fractions by Mid-infrared and Near-infrared Spectroscopy and Partial Least Squares Associated with Variable Selection Methods. *Energy and Fuels* **30**, 698–705 (2016).
44. Infrared and Raman Characteristic Group Frequencies: Tables and Charts. 3rd ed By George Socrates (The University of West London, Middlesex, U.K.). J. Wiley and Sons: Chichester. 2001. xviii + 348 pp. $185.00. ISBN: 0-471-85298-8. *J. Am. Chem. Soc.* **124**, 1830–1830 (2002).
45. Rainha, K. P. *et al.* Determination of API Gravity and Total and Basic Nitrogen Content by Mid- and Near-Infrared Spectroscopy in Crude Oil with Multivariate Regression and Variable Selection Tools. *Anal. Lett.* **52**, 2914–2930 (2019).
46. El-Bassoussi, A. A., Ahmed, M. H. M., Sayed, S. M. El, Basta, J. S. & Attia, E.-S. K. Characterization of Some Local Petroleum Residues by Spectroscopic Techniques. *Pet. Sci. Technol.* **28**, 430–444 (2010).

CAPÍTULO 5 CONCLUSÕES GERAIS

A presente tese explorou diversas abordagens inovadoras para a otimização de modelos de regressão por vetores de suporte (SVR), utilizando seleção de variáveis e a fusão de dados. Esses métodos não apenas permitem uma melhoria significativa na exatidão dos modelos, mas também proporcionam uma visão mais detalhada e específica sobre as contribuições individuais de cada variável e a sinergia entre diferentes fontes de dados. A aplicação dessas metodologias tem o potencial de transformar a análise quimiométrica, oferecendo ferramentas mais robustas e eficientes para a interpretação de dados complexos e a tomada de decisões informadas na indústria química e em outras áreas afins.

A seleção de variáveis aplicada ao SVR tem o potencial de aprimorar modelos existentes e revelar quais variáveis são mais importantes para uma determinada aplicação, melhorando não apenas os parâmetros de avaliação, mas também reduzindo a complexidade computacional. No **capítulo 03** foram aplicadas no SVR seleção de variáveis baseadas em permutação, SPA e NISPA, e os métodos foram comparados com outras seleções de variáveis consolidados na literatura. Ao aplicar SPA na densidade API, o método conseguiu reduzir as variáveis para 3,5% do total original, mantendo apenas as relacionadas a cadeias alifáticas de compostos de parafina e moléculas de benzeno substituídas, indicando uma forte relação com a densidade API. No caso dos saturados, o NISPA não apenas forneceu o modelo mais preciso, mas também destacou a importância das regiões de parafina, alcanos e grupos alifáticos nos espectros de MIR, com variáveis semelhantes às selecionadas para densidade API devido à relação entre essas propriedades.

A fusão de dados tem sido uma estratégia amplamente discutida na quimiometria, demonstrando um grande potencial na aplicação do SVR ao permitir a utilização de informações relevantes de diversas fontes analíticas, melhorando modelos e revelando sinergias entre diferentes fontes. A modelagem da densidade API mostrou que a fusão de médio nível com PCA foi a mais eficaz superando modelos sem fusão no teste estatístico. A combinação de espectros de ressonância magnética nuclear e infravermelho permitiu identificar informações de maior sinergismo que beneficiaram os modelos de regressão não-linear, destacando-se o Nitrogênio total, que aplicando fusão de alto nível, produziu os melhores modelos, superando

estatisticamente os modelos sem fusão. A integração de métodos não-lineares com a fusão de dados mostrou-se eficiente no desenvolvimento de modelos mais exatos.

Tanto a seleção de variáveis quanto a fusão de dados têm potencial para aprimorar modelos de SVR, cada estratégia com a sua vantagem e desvantagem, o melhor método para cada situação deve ser escolhido com cautela, utilizando como base um breve estudo e levando em consideração o potencial computacional e os dados.

CAPÍTULO 6 PRODUÇÃO CIENTÍFICA

Artigos Publicados

Objeto de identificação digital (DOI): 10.1016/j.foodchem.2021.131072
Periódico: Journal Of Chemometrics ISSN: 0308-8146 Vol. 1, p. e3282-16, 2020
Título: Variable selection in support vector regression using angular search algorithm and variance inflation fator.
Autores: FOLLI, GABRIELY S.; NASCIMENTO, MÁRCIA H.C.; DE PAULO, ELLISSON H.; **DA CUNHA, PEDRO H.P.**; ROMÃO, WANDERSON; FILGUEIRAS, PAULO R.

Objeto de identificação digital (DOI): 10.1016/j.fuel.2020.118462
Periódico: FUEL ISSN: 0016-2361 Vol. 279, p. 118462, 2020
Título: Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties.
Autores: DE PAULO, ELLISSON H.; FOLLI, GABRIELY S.; NASCIMENTO, MÁRCIA H.C.; MORO, MARIANA K.; **DA CUNHA, PEDRO H.P.**; CASTRO, EUSTÁQUIO V.R.; NETO, ALVARO CUNHA; FILGUEIRAS, PAULO R

Objeto de identificação digital (DOI): 10.21577/0100-4042.20170850
Periódico: Quimica Nova ISSN: 1678-7064 Vol. 45, 2021
Título: Infravermelho portátil na região do próximo (nir) aplicado no controle de qualidade de cafés adulterado por borra
Autores: Radigya M. Correia; **Pedro H. Cunha**; Bárbara Z. Agnoletti; Lucas L. Pereira; Fábio L. Partelli; Paulo R. Filgueiras; Valdemar Lacerda Jr.; e Wanderson Romão;

Objeto de identificação digital (DOI): 10.1016/j.fuel.2021.122527
Periódico: FUEL ISSN: 0016-2361 Vol. 311, p. 122527, 2021.
Título: Determination of gross calorific value in crude oil by variable selection methods applied to 13C NMR spectroscopy.
Autores: DE PAULO, ELLISSON H.; DOS SANTOS, FRANCINE D. ; FOLLI, GABRIELY S. ; SANTOS, LAYLA P. ; NASCIMENTO, MÁRCIA H.C. ; MORO,

MARIANA K. ; DA **CUNHA, PEDRO H.P.** ; CASTRO, EUSTÁQUIO V.R. ; CUNHA NETO, ALVARO ; FILGUEIRAS, PAULO R

Objeto de identificação digital (DOI): 10.1016/j.fuel.2020.118854
Periódico: FUEL ISSN: 0016-2361. Vol. 283, p. 118854, 2021
Título: Discrimination of oils and fuels using a portable NIR spectrometer
Autores: SANTOS, FRANCINE D.; SANTOS, LAYLA P.; **CUNHA, PEDRO H.P.**; BORGHI, FLÁVIA T.; ROMÃO, WANDERSON; CASTRO, EUSTÁQUIO V.R;. OLIVEIRA, ELCIO C.; FILGUEIRA, PAULO R.;

Objeto de identificação digital (DOI): 10.1002/CEM.3444
Periódico: Journal of Chemometrics ISSN: 0886-9383 Vol. 36 2022
Título: Variable selection by permutation applied in support vector regression models
Autores: Pedro H. P. da Cunha, Ellisson H. de Paulo, Marcia H. C. Nascimento, Mariana K. Moro, Gabriely Silveira Folli, Paulo R. Filgueiras.

Objeto de identificação digital (DOI): 10.1016/j.microc.2022.107966
Periódico: Microchemical Journal ISSN: 0026-265X Vol. 182, 2022
Título: Effect of fermentation on the quality of conilon coffee (Coffea canephora): chemical and sensory aspects
Autores: Barbara Z. Agnoletti ; Willian S. Gomes; Gustavo F. Oliveira; **Pedro Henrique Cunha**, Marcia H. C. Nascimento; Alvaro Cunha, Lucas L. P; Eustáquio V. R. Castro; Emanuele C. S. Oliveira; Paulo R. Filgueiras

Objeto de identificação digital (DOI): 10.1016/j.focha.2022.100074
Periódico: Food Chemistry Advances ISSN: 2772-753X Vol. 1. 2022
Título: Food analysis by portable NIR spectrometer
Autores: Gabriely S. Folli; Layla P. Santos; Francine D. Santos; **Pedro H.P. Cunha**; Izabela F. Schaffel; Flávia T. Borghi; Iago H.A.S. Barros; André A. Pires; Araceli V.F.N. Ribeiro; Wanderson Romão; Paulo R. Filgueiras;

Objeto de identificação digital (DOI): 10.1016/j.microc.2022.107696
Periódico: Microchemical Journal ISSN: 0026-265X Vol. 181, 2022
Título: Characterization of crude oils with a portable NIR spectrometer

Autores: Santos, Francine D., Vianna, Stéphany G.T., **Cunha, Pedro H.P.,** Folli, Gabriely S., de Paulo, Ellisson H., Moro, Mariana K., Romão, Wanderson, de Oliveira, Elcio C., Filgueiras, Paulo R.

Objeto de identificação digital (DOI): 10.1016/j.microc.2023.108739
Periódico: Microchemical Journal, ISSN: 0026-265X Vol. 190, 2022
Título: Study of coffee sensory attributes by ordered predictors selection applied to 1H NMR spectroscopy
Autores: de Paulo, Ellisson H., Nascimento, Márcia H.C., **da Cunha, Pedro H.P.**, Pereira, Lucas L., Oliveira, Emanuele C. da S., Filgueiras, Paulo R., Ferrão, Marco F.

Objeto de identificação digital (DOI): 10.1039/D3AY00510K
Periódico: Analytical Methods. ISSN: 1759-9660, Vol. 15, 2023
Título: SHS-GC-MS applied in Coffea arabica and Coffea canephora blend assessment
Autores: Lyrio, Marcos V. V., **da Cunha, Pedro H. P.,** Debona, Danieli G, Agnoletti, Bárbara Z., Araújo, Bruno Q., Frinhani, Roberta Q., Filgueiras, Paulo R., Pereira, Lucas L., de Castro, Eustáquio V. R.

Objeto de identificação digital (DOI): 10.36524/ric.v9i1.1868
Periódico: Revista Ifes Ciência, ISBN: 2359-4799, Vol 9
Título: Tutorial para aplicação didática de quimiometria em software gratuito – Parte I: Análise de Componentes Principais em dados de infravermelho médio e propriedades físico-químicas de amostras de petróleo
Autores: Folli, Gabriely S., **da Cunha, Pedro H. P.,** Moro, Mariana K;, Filgueiras, Paulo R.

Objeto de identificação digital (DOI): 10.1039/D3AY00711A
Periódico: Analytical Methods, ISSN: 1759-9660, Vol 15.
Título: Correlation analysis of modern analytical data – a chemometric dissection of spectral and chromatographic variables
Autores: Folli, Gabriely S., de Paulo, Ellisson H., Santos, Francine D., Nascimento, Márcia H. C., **da Cunha, Pedro H. P.**, Romão, Wanderson, Filgueiras, Paulo R.

Objeto de identificação digital (DOI): 10.21577/0100-4042.20230130
Periódico: Química Nova, ISBN: 01004042 , Vol. 47
Título: PERFIL VOLÁTIL DO Coffea arabica E Coffea canephora var. conilon POR SHS-GC-MS E QUIMIOMETRIA
Autores: Lyrio, Marcos, **da Cunha, Pedro H.**, Debona, Danieli, Agnoletti, Bárbara, Frinhani, Roberta, Oliveira, Emanuele, Filgueiras, Paulo R., Pereira, Lucas, de Castro, Eustáquio V.

Objeto de identificação digital (DOI): 10.48550/arXiv.2401.01200
Periódico: Arxiv
Título: Skin cancer diagnosis using NIR spectroscopy data of skin lesions in vivo using machine learning algorithms
Autores: Loss, Flavio P., **da Cunha, Pedro H.**, Rocha, Matheus B., Zanoni, Madson P.,de Lima, Leandro M., Nascimento, Isadora T., Rezende, Isabella, Canuto, Tania R. P., Vieira, Luciana de P., Rossoni, Renan, Santos, Maria C. S., Frasson, Patricia L., Romão, Wanderson, Filgueiras, Paulo R., Krohling, Renato A.

Objeto de identificação digital (DOI): 10.36524/ric.v9i3.2213
Periódico: Revista Ifes Ciência, ISBN: 2359-4799, Vol 9
Título: O USO DO REDGIM PARA CARACTERIZAR E DISTINGUIR AZEITES EXTRAVIRGEM DE OLIVA ADULTERADOS COM DIFERENTES ÓLEOS VEGETAIS
Autores: Zanoni, Madson P., **Cunha, Pedro H. P.**, Folli, Gabriely S., Filgueiras, Paulo R.

Artigos em submissão

Título: Tutorial para aplicação didática de quimiometria em software gratuito – Parte II: Regressão por Mínimos Quadrados Parciais (PLS) em dados de infravermelho médio e próximo para determinação de teor de adulterantes e propriedades físico-químicas.
Periódico: IFES-CIÊNCIA
Autores: Pedro H. P. da Cunha, Gabriely Silveira Folli, Sara Joaquina Inocencio Dionisio, Amanda Guedes Caldeira, Paulo R. Filgueiras.

Resumos publicados em anais 2020~2024

FOLLI, Gabriely S.; SANTOS, L. P.; PAULO, Ellisson H.; **CUNHA, Pedro H. P.;** SANTOS, F. D.; NASCIMENTO, MÁRCIA H.C.; ROMAO, Wanderson; FILGUEIRAS, P. R. Identificação de adulteração em azeite de oliva extra-virgem por micronir associada à SVM mult-class e SVM one-class com adição de outliers artificiais In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria. Campina Grande: XI Workshop de Quimiometria, 2020. p.184 – 184

FOLLI, G. S.; PAULO, E. H.; **CUNHA, P. H.**; SANTOS, L. P.; COELHO NETO, D. M.; LACERDA JR, V.; ROMAO, W.; FILGUEIRAS, P. R. Comparação da exatidão de modelos de calibração multivariada (PLS, SVR, PSO-PLS e ASA-VIF-SVR) em espectroscopia MIR, NIR e RMN de 1H. In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria, 2020. v. 1. p. 166.

CUNHA, P. H.; FOLLI, G. S.; PAULO, E. H.; SILVA, F. P.; OLIVEIRA, E. C. S.; FILGUEIRAS, P. R. Comparação de métodos de calibração multivariada aplicados em dados de ressonância magnética nuclear para a determinação de propriedades físico-químicas de mel. In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria, 2020. v. 1.

PAULO, E. H.; NASCIMENTO, M. H. C.; **CUNHA, P. H.**; PEREIRA, L. L.; OLIVEIRA, E. C. S.; FILGUEIRAS, P. R. Determinação de atributos sensoriais de café por método de seleção de variáveis em RMN de 1H. Resumo expandido em: XXI Encontro Latino Americano de Pós-Graduação (XXI EPG). INIC_2021 Anais Trabalhos Ciências Exatas e da Terra. ISSN 9786588226032.

PAULO, E. H.; FOLLI, G. S.; SANTOS, L. P.; NASCIMENTO, M. H. C.; **CUNHA, P. H.;** CUNHA NETO, A.; FILGUEIRAS, P. R. Seleção de variáveis aplicada a RMN de 13C para a determinação do poder calorífico superior em petróleo. INIC_2021 Anais Trabalhos Ciências Exatas e da Terra. Resumo expandido em: XXI Encontro Latino-

Americano de Pós-Graduação (XXI EPG). INIC_2021 Anais Trabalhos Ciências Exatas e da Terra. ISSN 9786588226032, p.114.

CUNHA, P. H. P.; PAULO, E. H.; FOLLI, G. S.; NASCIMENTO, M. H. C.; MORO. K. M.; FILGUEIRAS, P. R. Fusão de dados em regressão por vetores de suporte na determinação de propriedades físico-químicas de petróleo. In: VIII Encontro Capixaba de Química, 2021, Vitoria.

FOLLI, G. S.; BORGHI, F. T.; SANTOS, F. D.; SANTOS, L. P.; **CUNHA, P. H. P. DA**; RIBEIRO, A. V. F. N.; PIRES, A. A.; ROMAO, W.; FILGUEIRAS, P. R. Quantificação de adulterantes em alimentos por espectroscopia na região do infravermelho próximo com equipamento portátil. In: VIII Encontro Capixaba de Química, 2021, Vitória - ES. A Química e Sua Transversalidade na Ciência e na Vida, 2021.

CUNHA, P. H.; PAULO, E. H.; FOLLI, GABRIELY; FILGUEIRAS, P. R.. Aplicação De Fusão De Dados De Baixo E Médio Nível Em Regressão Por Vetores De Suporte Para Estimarpropriedades Físico-Químicas Do Petróleo. In: Vi Escola De Inverno De Quimiometria, 2023, Brasilia. Anais Da Escola De Inverno De Quimiometria, 2023. V. 1. P. 1.

FOLLI, GABRIELY; DE ALMEIDA, CAMILA M.; MOTTA, LARISSA C.; NASCIMENTO, M. H. C.; **CUNHA, P. H.**; SILVA, ANDRÉ; BATISTA JUNIOR, A. C.; OLIVEIRA, J. P.; BERNARDO, R. A.; CHAVES, ANDREA R.; ROMÃO, WANDERSON; FILGUEIRAS, P. R. Geração De Amostras Sintéticas Em Modelos De Classificação Por Máquinas De Vetores De Suporte Em Amostras Biológicas. In: Vi Escola De Inverno De Quimiometria, 2023, Brasilia. Anais Da Escola De Inverno De Quimiometria, 2023. V. 1. P. 1.

FOLLI, GABRIELY; **CUNHA, P. H.;** NASCIMENTO, M. H. C.; DE ALMEIDA, CAMILA M.; MOTTA, LARISSA C.; SILVA, ANDRÉ; BATISTA JUNIOR, A. C.; BERNARDO, R. A.; OLIVEIRA, J. P.; CHAVES, ANDREA R.; ROMÃO, WANDERSON; FILGUEIRAS, P. R. . Valiação Da Capacidade Preditiva Demodelos De Máquina De Vetores De

Suporte Em Classificação Bináriautilizando Permutação. In: Vi Escola De Inverno De Quimiometria, 2023, Brasilia. Anais Da Escola De Inverno De Quimiometria, 2023. V. 1. P. 1.

CALDEIRA, A. G.; DIONISIO, S. J. I.; FOLLI, GABRIELY; **CUNHA, P. H.;** FILGUEIRAS, P. R.. Tutorial De Quimiometria, Voltado Para O Ensino De Pls Na Graduação. In: Ix Encontro Capixaba De Química - Encaqui, 2023, Vitória. Anais Do Encontro Capixaba De Química. Vitória: Ufes, 2023. V. 1. P. 1.

PAULO, E. H.; MESSIAS, E.; FOLLI, GABRIELY; **CUNHA, P. H.**; FILGUEIRAS, P. R. Classificação De Méis Usando Rmn De 1h E Balanceamento De Classes Com Amostras Sintéticas. In: Ix Encontro Capixaba De Química - Encaqui, 2023, Vitória. Anais Do Encontro Capixaba De Química. Vitória: Ufes, 2023. V. 1. P. 1.

FOLLI, GABRIELY; PAULO, E. H.; MESSIAS, E.; NASCIMENTO, M. H. C.; **CUNHA, P. H.**; OLIVEIRA, E. C. S.; PEREIRA, LUCAS LOUZADA; FERRAO, M. F.; ROMÃO, WANDERSON; FILGUEIRAS, P. R.. Identificação De Cafés Especiais Por Rmn De 1h E Quimiometria. In: Ix Encontro Capixaba De Química - Encaqui, 2023, Vitória. Anais Do Encontro Capixaba De Química. Vitória: Ufes, 2023. V. 1. P. 1.

ZANONI, M. P.; **DA CUNHA, PEDRO H. P.**; FOLLI, GABRIELY; FILGUEIRAS, P. R. O Uso Do Redgim Para Caracterizar E Distinguir Azeites Extravirgem De Oliva Adulterados Com Diferentes Óleos Vegetais. In: Ix Encontro Capixaba De Química - Encaqui, 2023, Vitória. Anais Do Encontro Capixaba De Química. Vitória: Ufes, 2023. V. 1. P. 1.

Apresentação de trabalho em congresso 2020~2024

CUNHA, P. H.; FOLLI, G. S.; PAULO, E. H.; SILVA, F. P.; OLIVEIRA, E. C. S.; FILGUEIRAS, P. R. Comparação de métodos de calibração multivariada aplicados em dados de ressonância magnética nuclear para a determinação de propriedades físico-químicas de mel. In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria, 2020. v. 1.

CUNHA, P. H.; PAULO, E. H.; FOLLI, G. S.; NASCIMENTO, M. H. C.; MORO. K. M.; FILGUEIRAS, P. R. Fusão de dados em regressão por vetores de suporte na determinação de propriedades físico-químicas de petróleo. In: VIII Encontro Capixaba de Química, 2021, Vitoria.

CUNHA, P. H. P., DE PAULO, E. H., FOLLI, G. S., FILGUEIRAS, P. R. Aplicação de fusão de dados de baixo e médio nível em regressão por vetores de suporte para estimar propriedades físico-químicas do petróleo. In: VI Escola de Inverno de Quimiometria, 2023, Brasilia.

ANEXO A – REGRESSÃO POR VETORES DE SUPORTE - CÁLCULO

A FUNDAMENTAÇÃO MATEMÁTICA DE VETORES DE SUPORTE

Nos primeiros artigos sobre SVM, amostras são denominadas vetores, isso pois se caracterizam como um vetor linha onde cada coluna representa o valor de atributo (propriedade medida sobre a amostra ou valores espectrais). Assim, a concatenação dos vetores ($\mathbf{x_i}$) forma a matriz de dados $\mathbf{X}$. A Primeira descrição do SVM foi na solução de um problema de classificação binária. O SVM tem como princípio o mapeamento dos vetores de entradas (amostras utilizadas no conjunto de calibração), com o objetivo de encontrar um hiperplano que separe as amostras nas classes. A **Equação A1** mostra como a função é normalmente empregada;[1]

$$f(\boldsymbol{x}) = \mathbf{w} \cdot \mathbf{x} + b \quad \text{..... Equação A.1}$$

Nesta função, desejamos encontrar o escalar b e o vetor $\mathbf{w}$ que separe as duas classes de amostras. Entretanto, essa função tem infinitas soluções, isso é, diferentes hiperplanos que podem resolvê-la. Para solucionar este problema Vapnik, com o objetivo de generalizar o modelo por meio de uma função que visa minimizar o erro nos dados de teste, procurou maximizar a distância entre os hiperplanos de amostras de duas classes, nomeando de minimização do risco estrutural.[2] Nessa abordagem, determina-se o que chamamos de hiperplano de separação ótima (OSH, do inglês *Optimal Sepparting Hyperplane*). Todavia, nem todo modelo se separa de forma fácil, assim, para lidar com situações em que a separação das classes não é possível sem erros, introduziu uma variável de folga, ξi,[3] que admite que algumas amostras podem ter um erro associado à sua classificação. Dessa forma, chegamos na seguinte equação:

$$Min\frac{1}{2}||\mathbf{w}||^2 + \mathrm{C}\sum_{i=1}^{N}\xi_i \quad \text{..... Equação A.2}$$

Sujeito a $\begin{cases} \mathbf{w} \cdot \mathbf{x} + b \geq +1 - \xi_i \\ \mathbf{w} \cdot \mathbf{x} + b \leq -1 + \xi_i \\ \xi_i \geq 0 \end{cases}$ ..... **Equação A.3**

Onde C é uma constante que limita a variável de folga, determinando a margem suave da máquina de vetores, necessitando ser otimizada durante a construção do modelo. Estas equações explicam, de forma simplificada, o funcionamento do SVM, que não é o foco deste estudo, contudo existe na literatura diversas explicações mais detalhadas e completas para o mesmo.[4]

O SVM, ao empregar a variável de folga e a margem tem a tendencia ao sobreajuste, caracterizado pelo ajuste excessivo aos dados de calibração e perda de generalização. Portanto, é comum empregar a validação cruzada para otimizar os parâmetros do modelo, visando alcançar uma configuração robusta. Além disso, a utilização de kernels mais simples é uma estratégia eficaz para evitar ajustes excessivos.

- Adaptação para problemas de regressão

Como já dito anteriormente, o SVM pode ser adaptado para resolver problemas de regressão. Para isso, o espaço amostral é adaptado da seguinte forma: as amostras são multiplicadas com a modificação de adicionar e subtrair uma quantidade "d" do seu valor de interesse $\mathbf{y_j}$ para cada amostra $\mathbf{x_i}$, como demonstrado na **Figura A1**. Com isso, cria-se duas classes, uma positiva e outra negativa, e no cálculo do hiperplano será encontrado um OSH que passará justamente entre os valores de originais, $\mathbf{y_j}$.

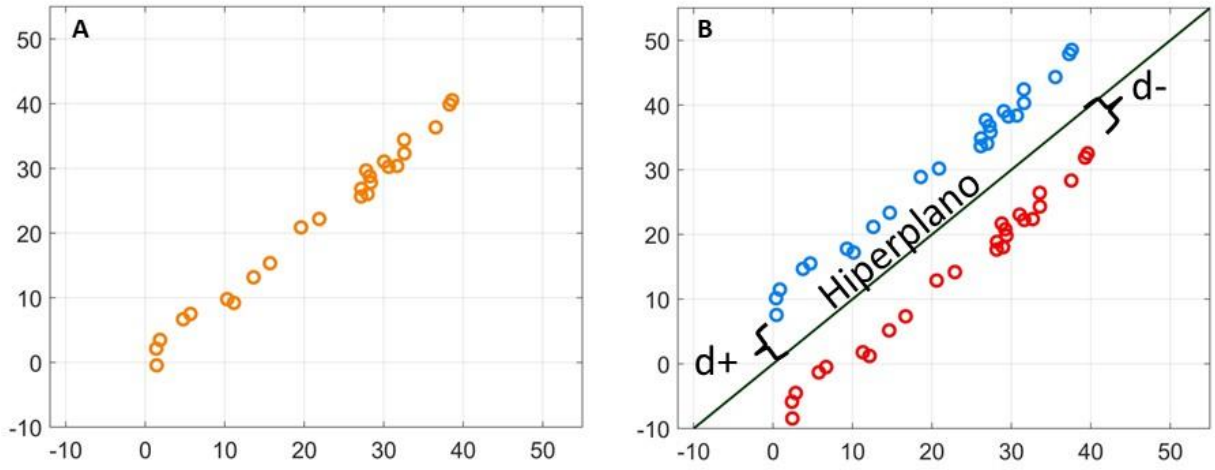


Figura A1. Classificador binário aplicado em regressão.

Assim, aplica-se a **Equação A.4** e **A.5**,[5] derivada da **Equação A.2** e **A.3**;

$$Min\,\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad \textbf{Equação A.4}$$

$$\text{Sujeito a} \begin{cases} y_i - \mathbf{w}\cdot\mathbf{x} \leq \varepsilon + \xi_i^* \\ \mathbf{w}\cdot\mathbf{x} - y_i \leq \varepsilon + \xi_i \\ \xi_i, \xi* \geq 0\,, i = 1, \ldots, N \end{cases} \quad \textbf{Equação A.5}$$

A constante C pondera os erros da função ($\xi i, \xi^* \geq 0$) que são delimitados por uma tolerância "ε" que também precisar ser otimizada. Dessa forma, uma separação binária pode ser modificada para uma regressão.

- Mapeamento Kernel

O SVM ficou conhecido pela sua capacidade de resolver problemas não lineares. Isso ocorre devido à aplicação do mapeamento kernel, ou truque kernel, que modifica o espaço de entrada original para um espaço de alta dimensão, chamado de espaço de características (*Feature Space*),[6] possibilitando que o problema seja solucionado com uma função linear, como fica demonstrada na **Figura A2**. Onde inicialmente temos uma classificação binaria impossível de separar utilizando uma função de reta. Ao aplicar a função kernel, o espaço amostral é modificado formando uma nova dimensão, onde é possível separar as classes utilizando uma função linear.

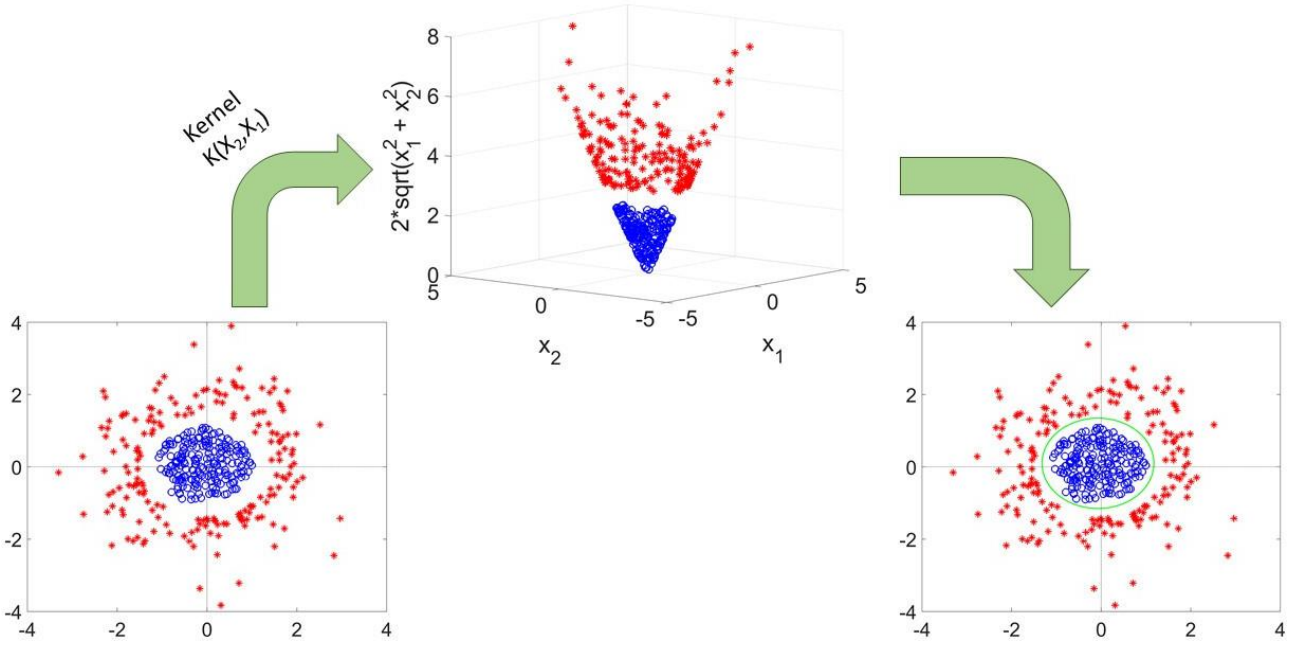


Figura A2. Exemplo de aplicação da Função Kernel.

A transformação kernel pode ser simbolizada matematicamente da seguinte forma:

$$K(\mathbf{x}_i, \mathbf{x}_\mathrm{j}) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \quad \textbf{Equação A.6}$$

Onde $\mathbf{x}_i$ e $\mathbf{x}_j$ são dois pontos do espaço de entrada e o $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ o novo ponto de alta dimensão. Existem diferentes tipos de kernel, entre eles os principais são radial (RBF, do inglês *Radial-Basis Function*), também chamado de gaussiano, linear, sigmoidal e polinomial.[3,7] Vale salientar que a função Kernel, independentemente do tipo, deve ter a constante γ otimizada.

Contudo, com a aplicação do kernel, a correlação entre o modelo SVM e a informação de entrada é perdida, impossibilitando que sejam identificadas as variáveis mais importantes para o desenvolvimento do modelo SVM.[8] Esse problema é popularmente conhecido como caixa preta do SVM. A solução da **Equação 4** gera uma função com produto escalar entre os vetores amostras, assim a função, $\phi(\mathbf{x}_i)$ pode substituir este produto escalar.

$$Min \frac{1}{2} \|\mathrm{w}\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \quad \textbf{Equação A.7}$$

$$\text{Sujeito a} \begin{cases} \boldsymbol{y_i} - f(\phi(\mathbf{x}_i), \mathbf{w}) \leq \varepsilon + \xi_i^* \\ f(\phi(\mathbf{x}_i), \mathbf{w}) - \boldsymbol{y}_i \leq \varepsilon + \xi_i \\ \xi_i, \xi* \geq 0 \, , i = 1, \ldots, n \end{cases} \quad \textbf{Equação A.8}$$

Para conseguir resolver a **Equação A.7**, aplica-se o multiplicador de Lagrange[1] resultando na **Equação A.8**.

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \quad \textbf{Equação A.8}$$

Em que $K(\mathbf{x}_i, \mathbf{x})$ representa a função kernel aplicada aos dados de entrada e α_i os vetores de suporte.

- Otimização

Como mencionado anteriormente, o SVR tem parâmetros que requerem ser determinados, incluindo; a constante C, que ajusta a suavidade da margem, a tolerância ε que estabelece o limite que uma amostra pode ser, ou não, erroneamente classificada; e constante γ, que determinar padrões internos do kernel selecionado pelo usuário. Ao contrário de métodos mais simples, como o PLS, o SVR demanda a otimização de dois, ou mais, hiperparâmetros que, na aprendizagem de máquina, são elementos que controlam o processo de aprendizagem do algoritmo, o que levanta o desafio de otimizar estes com o menor esforço computacional possível. Diante dessa questão, diversos métodos têm sido propostos como alternativas para a otimização do SVR.

Grade de Pesquisa, do inglês *Grid Search*, consiste em uma técnica de programação utilizada para otimizar hiperparâmetros,[9] sendo a mais lógica e utilizada, não só na quimiometria, mas em diversas áreas da ciência de dados. Neste método, é construído um número g^p de combinações de parâmetros, onde p é o número de parâmetros a ser otimizado e g é o tamanho da grade de pesquisa. Essas combinações são utilizadas para construir modelos e, posteriormente, avaliar a melhor combinação de parâmetros.[10]

No caso de uma matriz com 2 parâmetros e uma grade de tamanho 8, tem-se

um total de 64 combinações, como mostra a **Figura A3**. Conforme o tamanho da grade de pesquisa aumenta, o parâmetro de avaliação encontrado melhora, entretanto, aumenta também o tempo de processamento necessário para a otimização. Dessa forma, deve-se selecionar o valor de g de forma cautelosa.[11]

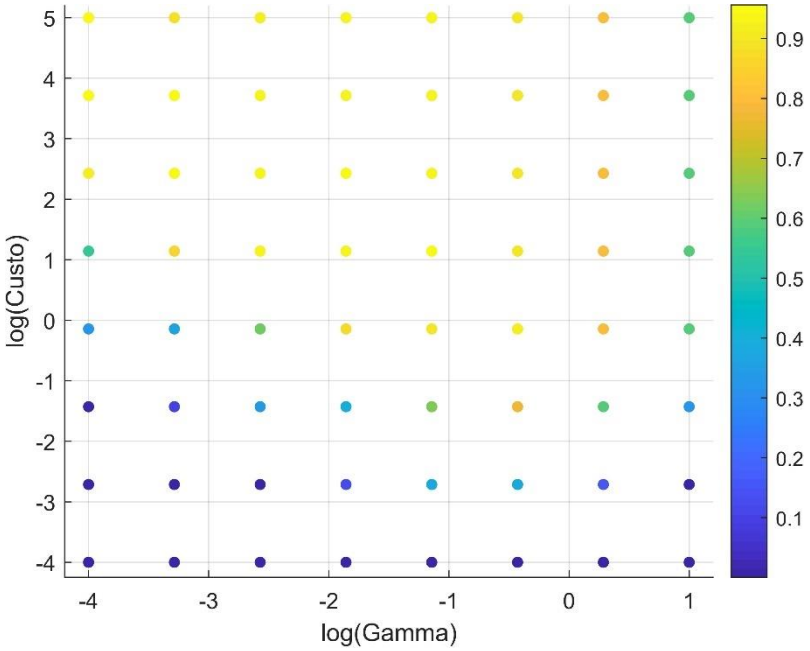


Figura A3. Exemplo de uma grade de pesquisa de 2 parâmetros e 8 variações, com o R^2 do modelo obtido.

Outra forma de otimizar o SVR consiste em utilizar o algoritmo genético (GA, do inglês *genetic algorithm*) que, por meio de uma otimização probabilística, busca obter os melhores parâmetros para o modelo. O GA utiliza conceitos de teoria da evolução para encontrar os melhores hiperparâmetros, onde se aplica a ideia de seleção natural, em diversos subconjuntos e ciclos, para conseguir otimizar o modelo.[9] A vantagem do GA perante a Grade de Pesquisa, está na possibilidade de se encontrar modelos melhores e em menor tempo, a depender das configurações de cada método. Contudo, a Grade de Pesquisa tem a vantagem de sempre encontrar o mesmo modelo ao final do processo, diferente do GA devido a sua natureza aleatória, além de não sofrer problema com mínimos locais, pontos numa função que são menores que os pontos próximos, contundo, não são o menor ponto, mínimo global.

Também é possível otimizar os hiperparâmetros do SVR utilizando o método enxame de partículas (PSO, do inglês *Particle Swarm Optimization*) que simula o comportamento de um grupo de indivíduos através de seu movimento num espaço

dimensional dos hiperparâmetros. Inicia-se colocando pequenos indivíduos nesse espaço que ao longo dos ciclos se movem semelhante a um enxame de vespas. Como também tem caráter aleatório, o PSO sofre do mesmo problema que o GAs, contudo com a vantagem de cair menos em mínimos locais.

O melhor método de otimização pode variar conforme a distribuição do conjunto amostral e sua propriedade.[12] Além disso, existem outras técnicas de otimização vem para o SVR,[12,13] cabendo ao usuário definir o melhor método de otimização para seu estudo.

REFERÊNCIAS – ANEXO A.

1. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
2. CHRISTOPHER J.C. BURGES. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **2**, 121–167 (1998).
3. Scholkopf, B. *et al.* Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* **45**, 2758–2765 (1997).
4. Zareef, M. *et al.* An Overview on the Applications of Typical Non-linear Algorithms Coupled With NIR Spectroscopy in Food Analysis. *Food Eng. Rev.* **12**, 173–190 (2020).
5. Ghorbani, M., Zargar, G. & Jazayeri-Rad, H. Prediction of asphaltene precipitation using support vector regression tuned with genetic algorithms. *Petroleum* **2**, 301–306 (2016).
6. Li, H., Liang, Y. & Xu, Q. Support vector machines and its applications in chemistry. *Chemom. Intell. Lab. Syst.* **95**, 188–198 (2009).
7. Lorena, A. C. & De Carvalho, A. C. P. L. F. Uma Introdução às Support Vector Machines. *Rev. Informática Teórica e Apl.* **14**, 43–67 (2007).
8. Üstün, B., Melssen, W. J. & Buydens, L. M. C. Visualisation and interpretation of Support Vector Regression models. *Anal. Chim. Acta* **595**, 299–309 (2007).
9. Liashchynskyi, P. & Liashchynskyi, P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. 1–11 (2019).
10. Liu, C., Yin, S. Q., Zhang, M., Zeng, Y. & Liu, J. Y. An Improved Grid

Search Algorithm for Parameters Optimization on SVM. *Appl. Mech. Mater.* **644**–**650**, 2216–2219 (2014).

11. Singh, A. Outliers and robust procedures in some chemometric applications. *Chemom. Intell. Lab. Syst.* **33**, 75–100 (1996).

12. Sun, Y., Ding, S., Zhang, Z. & Jia, W. An improved grid search algorithm to optimize SVR for prediction. *Soft Comput.* **25**, 5633–5644 (2021).

13. Tang, X., Zhuang, L., Cai, J. & Li, C. Multi-fault classification based on support vector machine trained by chaos particle swarm optimization. *Knowledge-Based Syst.* **23**, 486–490 (2010).

ANEXO B – ARTIGOS CIENTÍFICOS QUE ENVOLVEM SVR

Tabela B1. Artigos científicos que envolvem SVR.

Aplicação	Matriz	Fonte Analítica (X)	Observação	Referencia
Quantificação de adulterantes	Leite em pó	NIR	Comparou LS-SVM e PLS na identificação de adulterantes, primeiro artigo sobre.	Ferrão M. *et al.* (2006)[1]
Determinação de Hidroxi valor	Óleo de Soja	HATR/FT-IR	Utilização de dados de ordem superior em LS-SVM.	Ferrão M. *et al.* (2007)[2]
Quantificação de Lactobacillus	Leite Fermentado	Imagens de placa de petre	Utilização das bandas de vermelho, azul, verde e outras na quantificação.	Borin A. *et al.* (2007)[3]
Previsão do teor de sólidos solúveis na fruta de jabuticaba	Jaboticaba	NIR	Aplicou LS-SVM conseguindo produzir com capacidade de lidar a relação não linear do espectro NIR de sólidos.	Mariani, N. C. T. *et al.* (2014)[4]
Quantificação de escória utilizando LIBs	Escória	Espectroscopia de ruptura induzida por laser (LIBS)	PLS e SVR foram comparados na determinação de óxidos metálicos, devido à natureza de alto-absorção do LIBs no plasma, o que caracteriza não linearidade, o SVR conseguiu os melhores resultados.	Zhang, T. *et al.* (2015)[5]
Utilização de três métodos não lineares na predição de inibidores	Estrutura Molecular	Estrutura Molecular	---	Yuan, J. *et al.* (2016)[6]

Quantificação de biodiesel em diesel utilizando NIR	Diesel e Biodiesel	NIR	Os modelos obtiveram RMSEP abaixo do exigido pela metodologia analítica atual.	Alves J. C. L., Poppi R. J. (2016)[7]
Determinação de SAP utilizando NMR de ^{13}C e seleção de variáveis.	Petróleo	NMR de ^{13}C	Aplicação de GA em SVR em NMR de ^{13}C como fonte analítica.	Filgueiras, P. R. *et al.* (2016)[8]
Determinação de proteína total e glúten em farinha de trigo	Trigo	NIR	---	Chen, J. Zhu, S. Zhao, G. (2017)[9]
Seleção de variáveis em SVR para quantificar substâncias farmacológicas.	Fármacos	UV	Aplicação do algoritmo dos vagalumes conseguindo selecionar poucas variáveis.	Attia K. A.M. *et al.* (2017)[10]
Utilização de VIS-NIR na predição da densidade de raiz de arroz	Arroz	Vis-NIR	Aplicação de diversas seleções de variáveis (CARS, GAS, SPrA e UVE) em SVR.	Xu, S. *et al.* (2017)[11]
Determinação de stibnita em amostras de minerio	Minério	Raman	---	Cai Y. *et al.* (2018)[12]
Predição utilizando seis parâmetros atmosféricos	Ozônio Poluente	Parâmetros Atmosféricos	---	Mehdipour,V. Memarianfard, M.(2019)[13]

Determinação de tensão interfacial do petróleo	Petróleo	Dados Físico-Químico	---	Amar M. N., *et al.* (2019)[14]
Predição de aminoácidos livres em chá amarelo	Chá	Imagens hiperespectrais	Aplicação de fusão de baixo nível para melhorar a predição	Yang, B. *et al.* (2019)[15]
Comparação de três espectroscopias na terminação de Octanagem	Gasolina	NIR, Raman e NMR de ^{1}H	Comparou SVR e outros métodos, em diversas fontes analíticas, sendo o melhor resultado com SVR no espectro de NMR de ^{1}H.	Voigt, M. *et al.* (2019)[16]
Estimativa da radiação solar direta utilizando a qualidade do ar como parâmetro.	Índice de Qualidade do Ar (AQI)	Parâmetros climatológicos e AQI	O SVR conseguiu lidar com dados de fontes distintas, podendo prever a radiação solar.	Ma, Minglu *et al.* (2019)[17]
Predizer qualidade do ar utilizando a concentração de poluentes	Índice de Qualidade do Ar (AQI)	Concentração de Poluentes	---	Leong, W.C., Ahmad Z.K., (2020)[18]
Determinação de Número de Cetano utilizando diversos métodos de otimização	Biodiesel	Éster metílico de ácido graxo (FAME)	---	Bemani A. *et al.* (2020)[19]
Combinação de duas seleções de variáveis e detecção de outlier	Petróleo	MIR, NIR e NMR de ^{1}H	Aplicação de Teste de Grubbs para detectar outlier e uso de ASA-VIF na detecção de variáveis importantes em três espectros	Folli, G. S. *et al.* (2020)[20]

Determinação de chumbo em folhas de alface	Folhas de Alface	NIR/Vis-NIR	Transformação Wavelet e codificadores automáticos empilhados (WT-SAE) em SVR	Zhou, X. *et al.* (2020)[21]
Quantificação de vários adulterantes em Leite	Leite	MIR	Detecção simultânea de diversos adulterantes utilizando SVM, LS-SVM, ANN e PLS2	Amsaraj R. *et al.* (2021)[22]
Quantificação de propriedades FQ utilizando seleção de variáveis	Petróleo	MIR	Seleção de variáveis utilizando algoritmo genético em infravermelho	Mohammadi, M. *et al.* (2021)[23]
Determinação de qualidade de alimento	Ovo de Codorna	NIR Portátil	---	Brasil, Y.L. *et al.* (2022)[24]
Determinação de qualidade	Morango	Canal de Cores (RGB, HSV e HSL)	Uso de Imagens para modelagem	Basak J. K. *et al.* (2022)[25]
Uso de suscetibilidade magnética na determinação de poluente	Poeira Urbana	Suscetibilidade Magnética	Uso de suscetibilidade magnética para determinar concentração de metal.	Salazar-Rojas, T. *et al.* (2022)[26]
Detecção de Bactéria em Leite.	Leite	Raman	---	Du Y., *et al.* (2022)[27]
Predição de tempo de retenção	Oligonucleotídeos	Cromatografia iônica de gradiente	---	Enmark M., *et al* (2022)[28]
Predição de metal pesado em solo	Solo	Quebra induzida por laser com polarização	---	Zhao, H.-L., (2023). [29]

Determiação de elementos em alimentos.	Grão verde de Café	Imagem hiperespectral no infravermelho próximo	---	Sim, J. *et al*. (2023)[30]

REFERÊNCIAS – ANEXO B.

1. Ferrão, M. F., Mello, C., Borin, A., Maretto, D. A. & Poppi, R. J. LS-SVM: uma nova ferramenta quimiométrica para regressão multivariada. Comparação de modelos de regressão LS-SVM e PLS na quantificação de adulterantes em leite em pó empregando NIR. *Quim. Nova* **30**, 852–859 (2007).
2. Ferrão, M. F. *et al.* Non-destructive method for determination of hydroxyl value of soybean polyol by LS-SVM using HATR/FT-IR. *Anal. Chim. Acta* **595**, 114–119 (2007).
3. Borin, A. *et al.* Quantification of Lactobacillus in fermented milk by multivariate image analysis with least-squares support-vector machines. *Anal. Bioanal. Chem.* **387**, 1105–1112 (2007).
4. Mariani, N. C. T. *et al.* Predicting soluble solid content in intact jaboticaba [Myrciaria jaboticaba (Vell.) O. Berg] fruit using near-infrared spectroscopy and chemometrics. *Food Chem.* **159**, 458–462 (2014).
5. Zhang, T. *et al.* Quantitative and classification analysis of slag samples by laser induced breakdown spectroscopy (LIBS) coupled with support vector machine (SVM) and partial least square (PLS) methods. *J. Anal. At. Spectrom.* **30**, 368–374 (2015).
6. Yuan, J. *et al.* Predicting the biological activities of triazole derivatives as SGLT2 inhibitors using multilayer perceptron neural network, support vector machine, and projection pursuit regression models. *Chemom. Intell. Lab. Syst.* **156**, 166–173 (2016).
7. Alves, J. C. L. & Poppi, R. J. Quantification of conventional and advanced biofuels contents in diesel fuel blends using near-infrared spectroscopy and multivariate calibration. *Fuel* **165**, 379–388 (2016).
8. Filgueiras, P. R. *et al.* Determination of Saturates, Aromatics, and Polars in Crude Oil by 13C NMR and Support Vector Regression with Variable Selection by Genetic Algorithm. *Energy and Fuels* **30**, 1972–1978 (2016).
9. Chen, J., Zhu, S. & Zhao, G. Rapid determination of total protein and wet gluten in commercial wheat flour using siSVR-NIR. *Food Chem.* **221**, 1939–1946 (2017).
10. Attia, K. A. M., Nassar, M. W. I., El-Zeiny, M. B. & Serag, A. Firefly algorithm versus genetic algorithm as powerful variable selection tools and their effect on different multivariate calibration models in spectroscopy: A comparative study. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **170**, 117–123 (2017).
11. Xu, S., Zhao, Y., Wang, M. & Shi, X. Determination of rice root density from Vis–NIR spectroscopy by support vector machine regression and spectral variable selection techniques. *Catena* **157**, 12–23 (2017).
12. Cai, Y., Yang, C., Xu, D. & Gui, W. Quantitative analysis of stibnite content in raw ore by Raman spectroscopy and chemometric tools. *J. Raman Spectrosc.* **50**, 454–464 (2018).
13. Mehdipour, V. & Memarianfard, M. Ground-level O3 sensitivity analysis using support vector machine with radial basis function. *Int. J. Environ. Sci. Technol.* **16**, 2745–2754 (2019).
14. Nait Amar, M., Shateri, M., Hemmati-Sarapardeh, A. & Alamatsaz, A. Modeling oil-brine interfacial tension at high pressure and high salinity conditions. *J. Pet. Sci. Eng.* **183**, 106413 (2019).
15. Yang, B. *et al.* Rapid prediction of yellow tea free amino acids with hyperspectral images. *PLoS One* **14**, e0210084 (2019).

16. Voigt, M. *et al.* Using fieldable spectrometers and chemometric methods to determine RON of gasoline from petrol stations: A comparison of low-field 1H NMR@80 MHz, handheld RAMAN and benchtop NIR. *Fuel* **236**, 829–835 (2019).
17. Ma, M. *et al.* Estimation of horizontal direct solar radiation considering air quality index in China. *Energy Procedia* **158**, 424–430 (2019).
18. Leong, W. C., Kelani, R. O. & Ahmad, Z. Prediction of air pollution index (API) using support vector machine (SVM). *J. Environ. Chem. Eng.* **8**, 103208 (2020).
19. Bemani, A. *et al.* Modeling of cetane number of biodiesel from fatty acid methyl ester (FAME) information using GA-, PSO-, and HGAPSO- LSSVM models. *Renew. Energy* **150**, 924–934 (2020).
20. Folli, G. S. *et al.* Variable selection in support vector regression using angular search algorithm and variance inflation factor. *J. Chemom.* **34**, 1–16 (2020).
21. Zhou, X. *et al.* Development of deep learning method for lead content prediction of lettuce leaf using hyperspectral images. *Int. J. Remote Sens.* **41**, 2263–2276 (2020).
22. Amsaraj, R., Ambade, N. D. & Mutturi, S. Variable selection coupled to PLS2, ANN and SVM for simultaneous detection of multiple adulterants in milk using spectral data. *Int. Dairy J.* **123**, 105172 (2021).
23. Mohammadi, M. *et al.* Genetic algorithm based support vector machine regression for prediction of SARA analysis in crude oil samples using ATR-FTIR spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **245**, 118945 (2021).
24. Brasil, Y. L., Cruz-Tirado, J. P. P. & Barbin, D. F. Fast online estimation of quail eggs freshness using portable NIR spectrometer and machine learning. *Food Control* **131**, 108418 (2022).
25. Basak, J. K. ;, Madhavi, B. G. K., Paudel, B. & Kim, N. E. Prediction of Total Soluble Solids and pH of Strawberry Fruits Using RGB, HSV and HSL Colour Spaces and Machine Learning Models. *Foods* **10**, (2022).
26. Salazar-Rojas, T., Cejudo-Ruiz, F. R. & Calvo-Brenes, G. Comparison between machine linear regression (MLR) and support vector machine (SVM) as model generators for heavy metal assessment captured in biomonitors and road dust. *Environ. Pollut.* **314**, 120227 (2022).
27. Du, Y., Huang, H., Peng, Y., Wang, J. & Gao, Z. Rapid determination of <scp> *Staphylococcus aureus* </scp> enterotoxin B in milk using Raman spectroscopy and chemometric methods. *J. Raman Spectrosc.* **53**, 709–714 (2022).
28. Enmark, M., Häggström, J., Samuelsson, J. & Fornstedt, T. Building machine-learning-based models for retention time and resolution predictions in ion pair chromatography of oligonucleotides. *J. Chromatogr. A* **1671**, 462999 (2022).
29. ZHAO, H.-L., CAI, L.-L. & WU, G. On polarization resolved laser induced breakdown spectroscopy combined with support-vector regression to improve the accuracy of soil heavy-metal (Cd) detection. *Chinese J. Anal. Chem.* **51**, 100176 (2023).
30. Sim, J. *et al.* Support vector regression for prediction of stable isotopes and trace elements using hyperspectral imaging on coffee for origin verification. *Food Res. Int.* **174**, 113518 (2023).

ANEXO C –TESTE MANN-WHITNEY

O Test U, ou teste de Mann-Whitney, consiste em um teste que determina se dois conjuntos amostrais são estatisticamente iguais. Sua principal vantagem, em comparação ao test T, é que as amostras não precisam ter uma distribuição normal.[1] Considera-se dois conjuntos de amostras, A1 sendo x_1,x_2...x_m e A2 sendo y_1, y_2... y_n, sendo que m ≥ n. Afirmamos F e G são funções de distribuição das amostras A1 e A2, respectivamente; então considera-se a Hipótese nula:

$$H_0 : F(t) = G(t)\ para\ todo\ t. \qquad \textbf{Equação C.1}$$

A hipótese alternativa consiste em considerar o y maior ou menor que x. Pode-se descrevê-la da seguinte forma:

$$H_1 : F(t) = G(t - \Delta)\ para\ todo\ t.. \qquad \textbf{Equação C.2}$$

Uma forma diferente de interpretação é considerar a média de x e y.

$$\Delta = E(x) - E(y)\ . \qquad \textbf{Equação C.3}$$

Neste sentido, tem-se a seguinte hipótese nula e alternativa:

$$\begin{cases} H_0 : \Delta = 0 \\ H_1 : \Delta \neq 0 \end{cases}. \qquad \textbf{Equação C.4}$$

Após isso, os valores de x e y são organizados em ordem crescente e coloca-se os postos associados, sendo Sm e Sn as somas dos postos de x e y respectivamente. Com esses valores, calcula-se a estatística U.[2]

$$\begin{cases} U_m = Sm - \frac{1}{2}m(m+1) \\ U_n = Sn - \frac{1}{2}n(n+1) \end{cases}. \qquad \textbf{Equação C.5}$$

Sabendo que Sm + Sn, é a somatória de todos postos, junta-se as duas equações

e após uns ajustes tem-se:

$$U_m = m.n - U_n. \qquad \textbf{Equação C.6}$$

No teste de Mann-Whitney o termo estatístico W é considerado igual ao U_n então:

$$Z_{obs} = \frac{W - \frac{1}{2}m.n}{\sqrt{\frac{m.n(m+n+1)}{12}}}. \qquad \textbf{Equação C.7}$$

Neste trabalho, como foi utilizado o teste bilateral, foi calculado os valores críticos de $Z_{\propto/2}$ e $-Z_{\propto/2}$ tal que $P[Z > Z_{\propto/2}] = P[Z < -Z_{\propto/2} = {\propto}/{2}$. Então, para a hipótese nula ser aceita, tem-se que $Z_{obs} > Z_{\propto/2}$ ou $Z_{obs} < -Z_{\propto/2}$. Caso essa condição não seja obedecida, a hipótese alternativa deve ser aceita.[3]

Além de utilizar a hipótese nula do teste U, também foi calculado o p valor, utilizando a seguinte formula:

$$P - valor = P(|Z| > |Z_{obs}||H_0|) = 2P(Z > |Z_{obs}||H_0|). \qquad \textbf{Equação C.8}$$

REFERÊNCIAS – ANEXO C

1. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **18**, 50–60 (1947).
2. Perkins, G. Mann–Whitney U Test. *Key Top. Clin. Res.* 128–132 (2002) doi:10.3109/9780203450307-26.
3. ORCAN, F. Parametric or Non-parametric: Skewness to Test Normality for Mean Comparison. *Int. J. Assess. Tools Educ.* **7**, 255–265 (2020).