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

# Fusão de dados espectrais em quimiometria: Uma abordagem em química do petróleo

**Spectral data fusion in chemometric: An approach in petroleum chemistry**

**Mariana Kuster Moro**

**Tese de Doutorado em Química**

**Vitória**
**2022**

Mariana Kuster Moro

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**
**2022**

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

---

Moro, Mariana Kuster, 1989-
M867f Fusão de dados espectrais em quimiometria: uma abordagem em química do petróleo / Mariana Kuster Moro. - 2022.
163 f.

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

1. Petróleo. 2. Quimiometria. 3. Espectroscopia de ressonância magnética nuclear. 4. Espectroscopia de infravermelho. I. Filgueiras, Paulo Roberto Filgueiras. II. Universidade Federal do Espírito Santo. Centro de Ciências Exatas. III. Título.

CDU: 54

---

# Fusão de dados espectrais em quimiometria: Uma abordagem em química do petróleo

Mariana Kuster Moro

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 18/02/2022 por:

_____
Prof. Dr. Paulo Roberto Filgueiras[1]
Universidade Federal do Espírito Santo
Orientador(a)

_____
Prof. Dr. Lucas Mattos Duarte[1]
Universidade Federal Fluminense

_____
Prof. Dr. Waldomiro Borges Neto[1]
Universidade Federal de Uberlândia

_____
Prof. Dr. Wanderson Romão[1]
Instituto Federal do Espírito Santo

_____
Prof. Dr. Álvaro Cunha Neto[1]
Universidade Federal do Espírito Santo

[1] O documento será assinado eletronicamente em conformidade com as normas prescritas na Portaria Normativa PRPPG/ UFES nº 08/2021.

Universidade Federal do Espírito Santo
Vitória, fevereiro de 2022

# AGRADECIMENTOS

Agradeço, primeiramente, a Deus por essa conquista.

Agradeço ao professor Paulo Roberto Filgueiras pela orientação e prontidão na ajuda e pela confiança depositada.

Agradeço aos queridos amigos do Laboratório de Quimiometria pelo convívio diário e aprendizado.

Aos professores do Programa de Pós-graduação em Química da UFES por compartilharem o conhecimento.

Ao Laboratório de pesquisa e desenvolvimento de metodologias para análises de petróleo da UFES (LabPetro) e ao Centro de Pesquisas e Desenvolvimento Leopoldo Américo Miguez de Mello (Cenpes/Petrobras) pelo fornecimento e análises das amostras de petróleo.

Em especial, ao Olímpio Muniz Fiuza pelo amor e companherismo e a minha família, alicerce de todos os meus passos.

“Quanto mais aumenta nosso conhecimento, mais evidente fica nossa ignorância.”

John Fitzgerald Kennedy

# 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*)

ANP: Agência Nacional do Petróleo, Gás Natural e Biocombustíveis

API: *American Petroleum Institute*

ARO: teor de componentes aromáticos

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

ASTM: *ASTM International*

CarsPLS: amostragem adaptativa competitiva reponderada com mínimos quadrados parciais (*competitive adaptive reweighted sampling-partial least squares*)

DNNE: conjuntos de redes neurais decorrelacionadas (*decorrelated neural networks ensembles*)

EI: *Energy Institute*

ER: Erro de predição total

ERNN: conjunto de redes neurais com pesos aleatórios (*ensemble neural network with random weights*)

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

GA: algoritmo genético (*genetic algorithm*)

GC: cromatografia gasosa (*gas chromatography*)

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

IR: espectroscopia no infravermelho (*infrared spectroscopy*)

ISO: *International Organization for Standardization*

$K_{UOP}$: fator de caracterização

LabPetro: Laboratório de pesquisa e desenvolvimento de metodologias para análises de petróleo

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

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*)

NAT: número de acidez total

NB: teor de nitrogênio básico

NIPALS: mínimos quadrados parciais iterativos não lineares (*nonlinear iterative partial least square*)

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

RMN: ressonância magnética nuclear (*nuclear magnetic resonance*)

NT: teor de nitrogênio total

nVL: número de variáveis latentes

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

OPS: seleção de preditores ordenados (*ordered predictors selection*)

OSC: correção de sinal ortogonal (*orthogonal signal correction*)

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

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

PLS: mínimo quadrados parciais (*partial least squares*)

POL: teor de componentes polares

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

SARA: saturados, aromáticos, resinas e asfaltenos

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

SNV: variação normal padrão (standard normal variate)

SUVF: fluorescência ultravioleta síncrona (*synchronous ultraviolet fluorescence*)

$R^2$: coeficiente de determinação

$RMSE_C$: raiz quadrada do erro médio quadrático de calibração (*root mean square error of calibration*)

$RMSE_{CV}$: raiz quadrada do erro médio quadrático de validação cruzada (*root mean square error of cross validation*)

$RMSE_P$: raiz quadrada do erro médio quadrático de previsão (*root mean square error of prevision*)

RNA: rede neurais artificiais (*artificial neural network*)

S: teor de enxofre

SAP: saturados, aromáticos e polares

SAT: teor de componentes saturados

SimDis: destilação simulada (*simulated distillation*)

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

TBP: ponto de ebulição vedadeiro (*true boiling point*)

UOP: *Universal Oil Products*

# LISTA DE SÍMBOLOS

Acc – Acurácia do modelo

S - Desvio padrão amostral dos resíduos

$Eff_i$ – Eficiência do modelo em relação à classe i

$SPE_i$ – Especificidade do modelo em relação à classe i

$Erro_i$ – Erro da classe i

ER - Erro de predição total do modelo

$B_{(m,p)}$ – Matriz de coeficientes

P – Matriz de loadings

Q – Matriz de loadings

T – Matriz de scores

U – Matriz de scores

E – Matriz de resíduos

F – Matriz de resíduos

$Y_{(n,p)}$ – Matriz de variáveis dependentes

$X_{(n,m)}$ – Matriz de variáveis independentes

$\mu_R$ = Média dos resíduos

n – Número de amostras

C – Número de classes

m – Número de comprimentos de onda

$FN_i$ – Número de falsos negativos da classe i

$FP_i$ – Número de falsos positivos da classe i

p – Número de propriedades inferidas

$TN_i$ – Número de verdadeiros negativos da classe i

$TP_i$ – Número de verdadeiros positivos da classe i

$r_i$ – Resíduo da amostra i

$SEN_i$ – Sensibilidade do modelo em relação à classe i

$y_{prev,i}$ – Valor da propriedade predito pelo modelo

$y_{ref,i}$ – Valor experimental da propriedade

$\bar{y}$ – Valor experimental médio da propriedade

$\bar{y}_{predito}$ – Valor médio da propriedade predito pelo modelo

# RESUMO

A fusão de dados de diferentes fontes analíticas pode ser uma forma viável de estimar as propriedades físico-químicas do petróleo quando comparada ao uso de uma única técnica analítica. Isso ocorre porque saídas de diferentes fontes instrumentais podem transportar informações complementares e agir sinergicamente durante a calibração. Neste trabalho, foi investigado o potencial das estratégias de fusão de dados para estimar sete propriedades do petróleo bruto: teor de enxofre (S), teor de nitrogênio total (NT), teor de nitrogênio básico (NB), número de ácido total (NAT), frações de saturados (SAT), aromáticos (ARO) e polares (POL). Foram disponibilizadas 125 amostras de petróleo bruto, divididas em 70% para calibração e 30% para previsão. As propriedades físico-químicas das amostras foram determinadas por meio de métodos experimentais padronizados. Modelos de regressão por mínimos quadrados parciais (PLS) foram construídos a partir dos dados, previamente pré-tratados, de espectroscopia no infravermelho médio (MIR) com transformada de Fourier e de espectroscopia de ressonância magnética nuclear (RMN) de $^{1}H$ e $^{13}C$, individuais e fundidos a nível baixo, médio por PCA e PLS e alto em diferentes combinações. O número ótimo de variáveis latentes foi determinado por validação cruzada, pelo método Venetian Blind. Os resultados mostraram que, enquanto a fusão de nível médio por PLS aumentou a exatidão de alguns dos modelos, a fusão de nível baixo, médio por PCA e alto proporcionou melhorias insignificantes. Aplicando a fusão de nível médio por PLS, os teores de S, NT, NB, NAT, SAT, ARO e POL foram estimados, respectivamente, com $RMSE_P$ iguais a 0,057 m/m%, 0,041 m/m%, 0,0067 m/m%, 0,16 $mgKOHg^{-1}$, 4,73 m/m%, 3,66 m/m% e 6,50 m/m% e coeficientes de determinação iguais a 0,90, 0,83, 0,98, 0,91, 0,84, 0,67 e 0,66, para o conjunto de amostras teste. A fusão de nível médio por PLS demonstrou, para estes casos, ser a melhor estratégia fusão entre as demais, melhorando a exatidão dos modelos. A fusão de dados a nível baixo e médio por PCA também foi aplicada em dados de espectroscopia no infravermelho próximo (NIR) e MIR para discriminação de amostras de petróleo bruto em doce/azedo, pobre/rico em nitrogênio, não ácido/ácido/altamente ácido e leve/médio/pesado. Os modelos de classificação foram obtidos por meio do método de análise discriminante por mínimos quadrados parciais (PLS-DA). Os modelos de classificação baseados nos espectros individuais e nos espectros fundidos a nível baixo e médio foram comparados utilizando os parâmetros de desempenho sensibilidade, especificidade, eficiência, acurácia e erro de predição total. Em comparação com os modelos individuais, a fusão de dados a nível baixo e/ou médio proporcionou uma maior acurácia e menor taxa de erro de predição total em todas as quatro classificações propostas neste estudo. Os modelos de classificação construídos a partir da fusão de dados apresentaram alta acurácia, atingindo valores iguais ou superiores a 92%. Esses resultados demonstraram que as técnicas espectroscópicas propostas podem se complementar, por meio da estratégia de fusão de dados, gerando maior efeito sinérgico e, conseqüentemente, produzindo modelos com maior capacidade de previsão.



**Palavras-chave:** Fusão de dados. Petróleo bruto. Espectroscopia no infravermelho. RMN. PLS.

# ABSTRACT

Data fusion from different analytical sources can be a viable way to estimate physicochemical properties ou classificate crude oil samples, when compared to using a single analytical technique. This is because outputs from different instrumental techniques can carry additional information and act synergistically during calibration. In this work, the potential of data fusion strategies was investigated to estimate seven properties of crude oil: sulfur content (S), total nitrogen content (NT), basic nitrogen content (NB), total acid number (NAT), saturated (SAT), aromatic (ARO) and polar (POL) fractions. A total of 125 crude oil samples were available, divided into 70% for calibration and 30% for prediction. The physicochemical properties of the samples were determined using standardized experimental methods. Partial least squares (PLS) regression models were constructed from pre-treated mid-infrared spectroscopy (MIR) data and $^{1}H$ and $^{13}C$ nuclear magnetic resonance (RMN) spectroscopy data, using the data in individual form and also fused at low-level, mid-level with PCA and PLS and high-level in different combinations. The optimal number of latent variables was determined by cross-validation, using the Venetian Blind method. The results showed that, while mid-level data fusion by PLS increased the accuracy of some of the models, low, mid by PCA and high-level fusion provided negligible improvements. Using mid-level by PLS data fusion, the contents of S, NT, NB, NAT, SAT, ARO and POL were estimated, respectively, with $RMSE_P$ equal to 0,057 m/m%, 0,041 m/m%, 0,0067 m/m%, 0,16 $mgKOHg^{-1}$, 4,73 m/m%, 3,66 m/m% and 6,50 m/m% and coefficient of determination equal to 0,90, 0,83, 0,98, 0,91, 0,84, 0,67 and 0,66. Low and mid-level by PCA data fusion was also applied to near infrared (NIR) and MIR spectroscopy data for discrimination of crude oil samples into sweet/sour, poor/rich in nitrogen, non-acidic/acidic/highly acidic and light/ medium/heavy. The classification models were obtained through the method of discriminant analysis by partial least squares (PLS-DA). Classification models based on individual spectra and on the low and mid-level fused spectra were compared using the performance parameters sensitivity, specificity, efficiency, accuracy and total prediction error. Compared to the individual models, low and/or med-level data fusion provided greater accuracy and lower total prediction error rate in all four classifications proposed in this study. The classification models built from data fusion showed high accuracy, reaching values equal to or greater than 92%. These results demonstrate that the proposed spectroscopic techniques can complement each other, through the data fusion strategy, generating a greater synergistic effect and, consequently, producing models with greater predictive capacity.



**Keywords:** Data fusion. Crude oil. Infrared espectrocopy. NMR. PLS.

**SUMÁRIO**

# CAPÍTULO 1 INTRODUÇÃO E JUSTIFICATIVA

O petróleo apresenta uma matriz altamente complexa e sua composição e características variam de um reservatório para outro (THOMAS, 2001). Seu valor econômico depende do tipo, qualidade e quantidade de subprodutos que ele é capaz de gerar nas refinarias. Desta forma, é imprescindível a caracterização do petróleo para determinar as suas propriedades físicas e químicas de maior interesse, com vistas a precificá-lo para sua posterior comercialização. Além disso, para evitar problemas operacionais e garantir a qualidade do produto final, é importante monitorar e controlar alguns parâmetros do petróleo durante seu processamento e refino (VEMPATAPU e KANAUJIA, 2017).

A ASTM (*American Society for Testing and Materials*) padronizou diversos métodos para determinação de propriedades físico-químicas do petróleo e seus derivados como, por exemplo, octanagem, densidade, ponto de fulgor, viscosidade, lubricidade, entre outras. Entretanto, caracterizar o petróleo utilizando todos os métodos padronizados é um processo bastante trabalhoso que envolve diversas análises, geralmente, de alto custo. Ademais, estas análises, em conjunto, demandam um longo tempo (maior que 4 meses de ensaio), impossibilitando o ajuste dos parâmetros em tempo real para otimização das condições de operação e consequente controle da qualidade do produto final. Uma alternativa consiste em inferir estas propriedades por meio de técnicas espectroscópicas que, em geral, são rápidas, não destrutivas e permitem a determinação simultânea de várias propriedades (KHANMOHAMMADI *et al*., 2012).

A espectroscopia na região do infravermelho por transformada de Fourier (FTIR, do inglês *Fourier transform infrared*) tem sido a técnica analítica mais usada para predição de propriedades do petróleo e derivados (DOUGLAS *et al*., 2018, RODRIGUES *et al*., 2017, KHANMOHAMMADI *et al*., 2012, LOZANO *et al*., 2017). É uma técnica conveniente, pois os espectros na região do infravermelho se correlacionam bem com várias propriedades físico-químicas do petróleo (Baird 2016). Muitas calibrações desenvolvidas a partir dessa técnica têm sido utilizadas em linha para a otimização das condições operacionais de refino em tempo real (ALVES e POPPI, 2012, LAMBERT *et al*., 1995).

Outra técnica espectroscópica que tem se mostrado promissora e vem sendo

muito utilizada na caracterização de petróleo e seus derivados é a ressonância magnética nuclear (RMN) (VEMPATAPU e KANAUJIA, 2017). DUARTE *et al*. (2016) obtiveram resultados satisfatórios utilizando RMN $^{1}H$ associada a métodos quimiométricos de calibração multivariada para inferir várias propriedades do petróleo bruto.

A determinação das propriedades físico-químicas do petróleo por meio de informações espectrais é realizada com o auxílio da quimiometria, por meio de um processo chamado calibração multivariada. Os espectros, cujos sinais representam as variáveis independentes, são correlacionados, por meio de modelos matemáticos, com as propriedades físico-químicas, as quais representam as variáveis dependentes. Devido ao grande número de amostras e de variáveis, tanto dependentes como independentes, os cálculos envolvidos para construção dos modelos matemáticos se tornam complexos, sendo imprescindível o uso de ferramentas computacionais e o conhecimento de álgebra e estatística para o processamento dos dados. Nesse contexto de se trabalhar com muitas varáveis e amostras ao mesmo tempo, surgiu a Quimiometria, uma área da química cuja finalidade é agregar conhecimentos e técnicas necessárias para resolução de tais cálculos (BRERETON, 2000).

A quimiometria é a ciência de processar informações químicas para diferentes objetivos como, por exemplo, a discriminação de classes que busca dividir os dados em diferentes grupos de acordo com a similaridade entre eles. Uma das principais aplicações da quimiometria é a calibração que utiliza uma ou mais medidas analíticas para prever algum parâmetro ou propriedade de interesse que esteja correlacionado com essas medidas (BRERETON, 2007). Quando a informação analítica envolve duas ou mais variáveis, o processo é chamado de calibração multivariada (ADAMS, 1995). Com o objetivo de prever propriedades ou composição, a calibração multivariada pode ser realizada por diferentes métodos de regressão. O método mais popular é a regressão por mínimos quadrados parciais (PLS, do inglês *Partial Least Squares*) (BRERETON, 2007, KHANMOHAMMADI *et al*., 2012).

A PLS tem a vantagem de, no lugar dos dados brutos, utilizar combinações lineares das variáveis latentes para a construção do modelo. A variáveis latentes são combinações lineares das variáveis originais. Além de serem em número reduzido, são ortogonais entre si, o que reduz a elevada colinearidade das variáveis originais. Neste processo, as variáveis originais recebem “pesos”, de forma que quanto maior a

sua correlação com a variável resposta, maior será seu peso, aumentando sua influência no modelo construído. Desta forma, as variáveis mais importantes contribuem mais significativamente na previsão das variáveis dependentes (MILLER J. N. e MILLER J. C., 2010).

Na calibração multivariada objetiva-se encontrar um modelo matemático que relacione as propriedades de interesse com sinais analíticos, geralmente, de determinada técnica espectroscópica. Para cada técnica analítica ou diferentes conjuntos de dados é possível obter um modelo matemático diferente (MENDES *et al*., 2018). No entanto, medidas de uma única técnica analítica, algumas vezes, podem não ser suficientes para correlacioná-las com várias propriedades de interesse e gerar modelos de qualidade.

Neste contexto, aparece uma nova estratégia de modelagem chamada fusão de dados que consiste, basicamente, em fundir dados de diferentes fontes analíticas e, a partir deles, obter um único modelo matemático para inferência das variáveis dependentes ou classificação de amostras (CASTANEDO *et al*., 2013).

Um dos pioneiros na aplicação da estratégia de fusão de dados foi DOWNEY *et al*. (1997). Espectros no infravermelho médio e próximo foram agrupados em uma matriz e técnicas quimiométricas foram aplicadas para discriminação de dois tipos diferentes de café em um conjunto de amostras. Os autores compararam o modelo resultante da fusão com modelos construídos a partir dos dados de apenas uma técnica analítica. Foi observado que a fusão de dados produziu um modelo mais robusto, uma vez que foi necessária uma menor quantidade de componentes principais para uma discriminação eficaz das amostras. Segundo os autores, esses resultados eram um indicativo de que a fusão de dados provenientes de diferentes fontes analíticas era capaz de produzir modelos com melhor capacidade de discriminação ou de previsão de propriedades (DOWNEY *et al*., 1997).

Alguns trabalhos exploraram o impacto do nível da fusão de dados na capacidade preditiva ou discriminativa dos modelos. LI *et al*. (2018) fundiu dados de espectros na região do infravermelho médio e próximo a níveis baixo, médio e alto e comparou os modelos para a determinação da origem geográfica da erva medicinal *Panax*. Os dados espectrais tratados individualmente (sem fusão) não foram capazes de classificar as amostras quanto a sua origem de forma satisfatória. A fusão de nível alto gerou modelos de melhor classificação, com exatidão de 98-100% contra 95-98%

e 93-96% para nível médio e baixo, respectivamente. Segundo os autores, a fusão de nível baixo tem menor eficácia, pois os dados brutos podem conter informações redundantes, ruído ou interferência que obstruem o efeito sinérgico entre as informações (LI *et al*., 2018).

Nos últimos anos, uma extensa produção cientifica utilizando a fusão de dados de diversas fontes analíticas foi publicada de forma crescente ao longo do tempo, na área de química, como mostra a Figura 1.1. O gráfico da figura é resultado de uma análise bibliométrica utilizando o *software RStudio* com a ferramenta *Bibliometrix*. A pesquisa foi feita na base de dados *Scopus*, no período de 1980 e 2020, utilizando as palavras-chave "*data fusion*" e "*chemometrics*". De acordo com o gráfico, o número de pesquisas científicas envolvendo a fusão de dados na quimiometrica cresceu muito, principalmente, nos últimos anos.

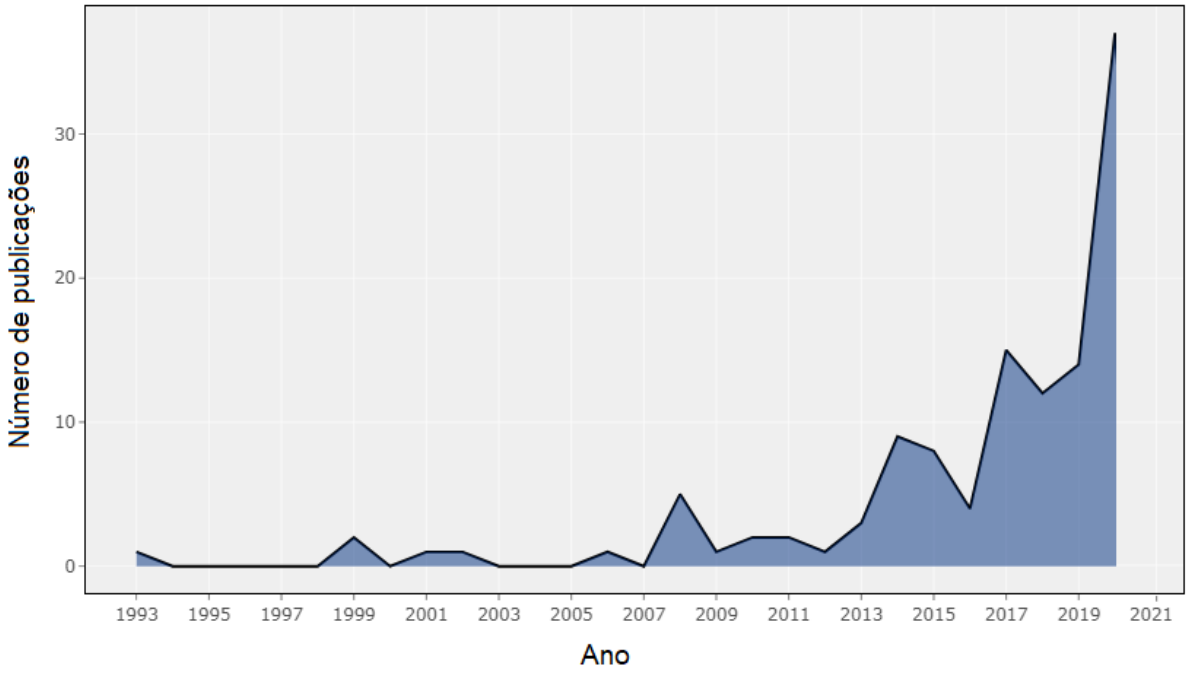



**Figura 1.1.** Número de publicações, na base de dados *Scopus*, com as palavras-chave "*data fusion*" e "*chemometrics*" ao longo dos anos.

BORRAS *et al*. (2015), em sua revisão bibliográfica, cita um total 77 artigos em que os autores aplicam a fusão em dados, utilizando diversos alimentos como objeto de estudo. Dentre estes, um número de 43, 25 e 9 artigos empregaram a fusão, respectivamente, a nível baixo, médio e alto. Uma grande parte destes trabalhos se

propôs a classificar\identificar a origem de alimentos a partir de análises quimiométricas (BORRAS *et al*., 2015). BAJOUB *et al*. (2016) empregaram fusão em níveis baixo e médio em dados de cromatografia líquida de alta eficiência para discriminar amostras de azeite de oliva extra-virgem de acordo com sua origem.

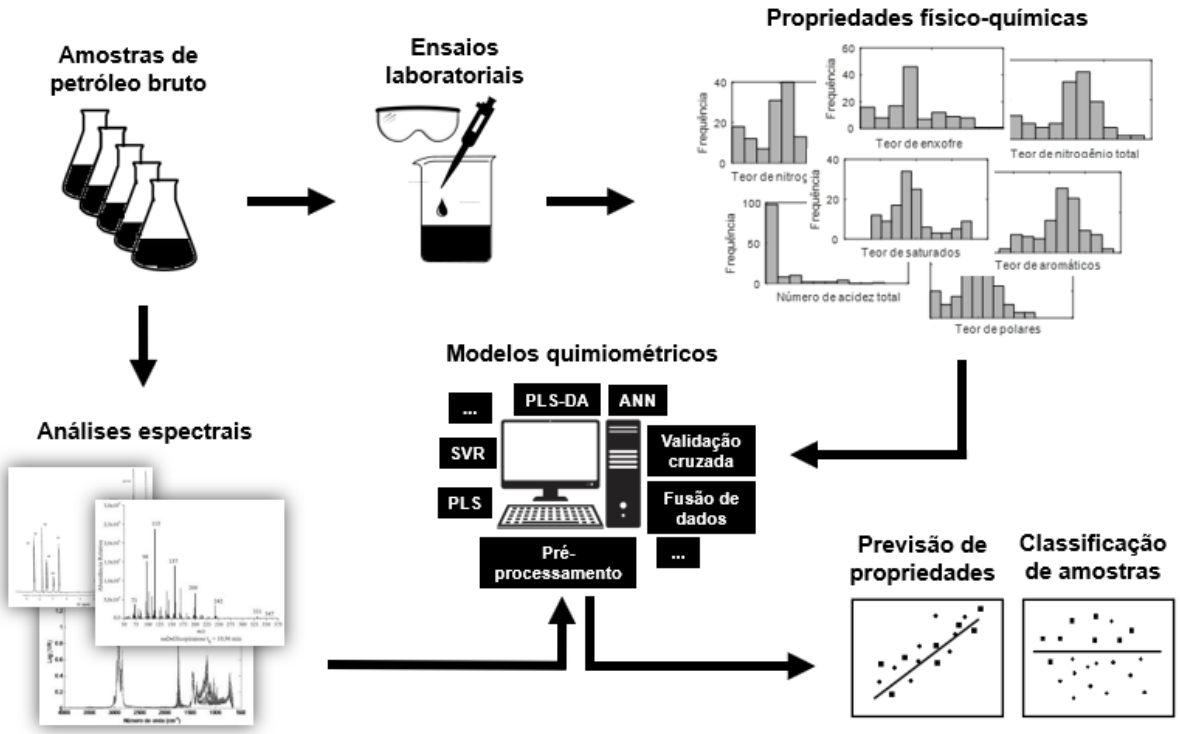
A estratégia de fusão de dados também foi aplicada para a classificação de vários outros alimentos: óleo de palma (OBSESAN *et al.*, 2017), suco de laranja (CUEVAS *et al*., 2017), cerveja (TAN *et al*., 2015), vinho tinto (APETREI *et al*., 2012), entre outros.

A inferência de propriedades também foi foco de estudo na área de alimentos. ZUDE *et al*. (2006), MENDONZZA *et al*. (2012) e XIAOBO *et al*. (2013) agruparam dados de diferentes técnicas analíticas para predição de propriedades de maçãs. A fusão de dados associada a quimiometria proporcionou modelos com boa capacidade preditiva nos três trabalhos. Como se observa, a fusão de dados tem sido bastante estudada na área de alimentos, entretanto é uma abordagem ainda muito pouco explorada quando se trata de outros objetos de estudo. A grande maioria dos trabalhos publicados se limita apenas à área de alimentos, enquanto que alguns poucos são encontrados no campo da medicina e biologia (SMOLINSKA *et al*., 2012).

Há poucos registros na literatura da aplicação da fusão de dados associada com os demais métodos quimiométricas em matrizes oleosas, seja para classificar ou caracterizar o objeto de estudo (MUHAMMAD e AZEREDO, 2014, PEINDER *et al*., 2009, DEARING *et al*., 2011, MORO *et al*., 2019). A inferência de propriedades físico-químicas do petróleo por meio de modelos construídos a partir da fusão de dados pode se apresentar como bastante promissora visto que a fusão de dados contribui para o aperfeiçoamento dos modelos e as técnicas espectrais trazem simplicidade e rapidez às análises estatísticas. Isso torna possível o controle de variáveis importantes em linha, o que permite a otimização do processo e garante a qualidade do produto final.

Este estudo objetiva explorar o potencial da estratégia de fusão de dados na construção de modelos de calibração para inferência de propriedades físico-químicas e classificação de amostras de petróleo bruto. Para tanto, a calibração multivariada será aplicada a espectros de NIR, MIR e RMN, fundidos a nível baixo, médio e alto.

# CAPÍTULO 2 FUNDAMENTAÇÃO TEÓRICA




- Este capítulo fornece uma visão geral da aplicação de técnicas de MIR e NIR de RMN $^{1}H$ e $^{13}C$, associadas à quimiometria, para caracterização de petróleo bruto.

- As espectroscopias de RMN e IR, aliadas à quimiometria, foram aplicadas com sucesso em diversos estudos para previsão de propriedades físico-químicas do petróleo.

- O uso de métodos de regressão não linear para previsão de propriedades do petróleo vem crescendo nos últimos anos.

- A fusão de dados multivariados, suas diferentes abordagens e aplicações na quiomiometria são revisadas.

## 2.1 Caracterização do petróleo

As reservas de petróleo foram formadas pelo acúmulo de matéria orgânica em bacias sedimentares e posterior decomposição ao longo de milhares de anos, sob condições particulares de temperatura, pressão e outras variáveis, de acordo com o reservatório. Como consequência, existe uma grande diversidade de tipos de petróleo ao redor do mundo (WAPLES, 1985). Dada a complexidade e variabilidade de sua composição química, o petróleo bruto precisa ser caracterizado (propriedades físico-químicas) por um conjunto de ensaios que compõe o portfólio de análise desse tipo de amostra. É a partir da resposta desse conjunto de características que os óleos são classificados e têm sua variabilidade determinada. A partir da caracterização, é possível avaliar a qualidade do óleo e, consequentemente, seu valor comercial, bem como o potencial do poço produtor e, além disso, ajustar as condições de produção e refino (MERDRIGNAC e ESPINAT, 2007). A Tabela 2.1 lista algumas das propriedades físico-químicas mais importantes a serem monitoradas no petróleo bruto para caracterização e avaliação da qualidade (RIAZI, 2005, SPEGHT, 2015).

Os procedimentos convencionais de caracterização, padronizados por organizações como a *ASTM International* (ASTM), o *Energy Institute* (EI) e o *American Petroleum Institute* (API), consistem em uma série de ensaios analíticos realizados em laboratórios químicos (RIAZI, 2005). Os métodos padrão usuais para determinar as diferentes propriedades físico-químicas do óleo também estão incluídos na Tabela 2.1. Alguns deles (por exemplo, densidade, curva de destilação, teor de nitrogênio e teor de enxofre) são de importância primordial, pois influenciam fortemente a qualidade e o preço do petróleo. Sua caracterização completa permite um conhecimento mais amplo do perfil químico do óleo, tornando as decisões mais seguras e adequadas por parte dos produtores e das refinarias (GAUTAM *et al*., 1998).

Uma análise bastante útil é aquela que separa os óleos brutos em quatro classes diferentes: saturados, aromáticos, resinas e asfaltenos (SARA). A determinação do SARA é de extrema importância, pois, em combinação com outras informações, pode indicar o custo do processamento do petróleo bruto para adequar o produto final às especificações do consumidor e à legislação ambiental (MULLINS *et al*., 2007, TIANGUANG *et al*., 2002). Os componentes saturados são moléculas não polares formadas por alcanos e cicloalcanos; enquanto os aromáticos são formados por

hidrocarbonetos mono e/ou poliaromáticos. Resinas e asfaltenos são considerados compostos polares do petróleo com estruturas poliaromáticas; ambos são polares, com alto peso molecular, geralmente na faixa de 2000 ± 500 para asfaltenos (SPEIGHT, 2004) e entre 500 e 1000 para resinas (WOODS *et al*., 2008) e alta relação carbono/hidrogênio, podendo ainda conter heteroátomos de nitrogênio, oxigênio e enxofre, bem como alguns metais (RIAZI, 2005).

**Tabela 2.1.** Propriedades físico-químicas e método de análise do petróleo bruto monitoradas para sua caracterização.

| **Propriedades** | **Método padrão** | **Propriedades** | **Método padrão** |
|---|---|---|---|
| Densidade (gravidade API) | ISO 12185 | Enxofre mercaptídico | UOP 163 |
| Viscosidade cinemática | ASTM D7042 | Pressão de vapor | ASTM D323 |
| Viscosidade dinâmica | ASTM D7042-21 | Ponto de fulgor | ISO 13736 |
| Ponto de fluidez | ASTM D7346 | Evolução de $H_2S$ | ASTM D5623 |
| Micro resíduo de carbono | ASTM D4530 | Teor de cinzas | ASTM D482 |
| Fator de caracterização ($K_{UOP}$) | UOP 375 | Teor de sal | ASTM D6470 |
| Hidrocarbonetos: saturados, aromaticos resinas and asfaltenos | ASTM D2549-02 | Metais (Ni, V, Pb, Cu, Fe, Zn, Ba, Ca, P, Mn, Si, Cd, Cr and Hg) | N2440 / ATSM D4927 and UOP 938 |
| Poder calorífico superior | ASTM D4809 | Tensão interfacial óleo/sal e água | ASTM D971<br>ASTM D2285 |
| Enxofre total | ASTM D4294-16 | Estabilidade e compatibilidade | ASTM D7112-09 |
| Nitrogênio total | ASTM D4629-17 | Curva de distribuição de acidez | ASTM D664 |
| Nitrogênio básico | UOP 269 | | |
| Número de acidez total | ASTM D664 | | |
| Curva de destilação simulada | ASTM D7169-20e1 | | |
| Curva ponto de ebulição verdadeiro | ASTM D2892-20 | | |
| Ponto de ebulição inicial | ASTM D7798-20 | | |
| Teor de água | ASTM D4377 | | |
| Análise elementar (C, H, N, O) | ASTM D5291<br>ASTM D5622 | | |

A caracterização do petróleo também inclui a quantificação de outros elementos, como enxofre, nitrogênio, oxigênio e alguns metais que se apresentam na forma de

compostos organometálicos ou como sais inorgânicos dissolvidos na água associada ao petróleo (KHUHAWAR *et al*., 2012). Todos esses não hidrocarbonetos são considerados impurezas (RIAZI, 2005).

O enxofre, embora presente em pequenas frações, é o terceiro elemento mais abundante no petróleo bruto e concentra-se preferencialmente na fração mais pesada. Os óleos com alto teor de enxofre são classificados como ácidos (ou azedos), visto que parte deste elemento está presente na forma de $H_2S$ contribuindo com a acidez e consequente corrosão dos equipamentos e tubulações metálicas; enquanto que aqueles com menor teor de enxofre, podem ser chamados de petróleo doce. O enxofre é indesejável, pois causa problemas como contaminação de catalisadores, corrosão de equipamentos e tubulações e produção de gases tóxicos de enxofre durante a combustão dos derivados de petróleo. O nitrogênio também é encontrado preferencialmente nas frações mais pesadas e é responsável por alguns problemas operacionais como formação de goma e inibição e desativação de catalisadores ácidos (PRADO *et al*., 2017, SPEIGHT, 2015).

Embora importante, caracterizar o petróleo por meio de todos os experimentos padronizados é um processo muito trabalhoso que envolve vários ensaios, geralmente de alto custo e que demanda muito tempo (LOVATTI *et al*., 2019). A Tabela 2.1 lista um total de 27 análises para uma caracterização completa do petróleo. Uma alternativa a essas análises seria inferir as propriedades utilizando técnicas espectroscópicas que geralmente são rápidas, não destrutivas e permitem a determinação simultânea de várias propriedades (KHANMOHAMMADI *et al*., 2012).

Entretanto, antes do surgimento da análise multivariada, a utilidade da espectroscopia era mais limitada, uma vez que a forma mais comum de análise dos espectros de óleos brutos e derivados era por meio da inspeção visual. Métodos estatísticos multivariados passaram a ser usados para analisar amostras de óleo apenas na década de 1980. As primeiras aplicações de análise multivariada em óleos brutos e derivados restringiam-se a metodologias de análise exploratória ou reconhecimento de padrões, como a modelagem independente e flexível por analogia de classes (SIMCA, do inglês *Soft Independent Modeling of Class Analogy*) em dados de cromatografia (OYGARD *et al*., 1984) e a análise de componentes principais (PCA, do inglês *Principal Component Analysis*) em dados de RMN $^{13}C$ (KVALHEIM *et al*., 1985 e BREKKE *et al*., 1989), de cromatografia gasosa (TELNAES *et al*., 1987) e em

dados de distribuições de saturados C6 e C7 (KVALHEIM, 1987). O método SIMCA também foi utilizado com sucesso para distinguir amostras de petróleos brutos advindas da mesma região geográfica, utilizando 47 parâmetros físico-químicos (URDAL *et al*., 1986).

Já os métodos de regressão foram aplicados em petróleo somente em 1988, quando BRAKSTAD *et al*. (1988) previu o peso molecular e a densidade de frações da destilação utilizando o método PLS e, como dados de entrada, a cromatografia gasosa e a espectroscopia de massas. A metodologia mostrou-se adequada para reduzir o tamanho do conjunto de dados e prever rapidamente as propriedades propostas. Poucos anos depois, BURG *et al*. (1995) comparam o desempenho dos métodos de regressão por componentes principal (PCR, do inglês *Principal Components Regression*), regressão linear múltipla (MLR, do inglês *Multiple Linear Regression*) e PLS para prever a viscosidade de óleos brutos. O PLS também foi aplicado para prever diversos parâmetros físico-químicos de amostras de gasolina, utilizando dados de perfis cromatográficos (FLUMIGNAN *et al*., 2008).

O número de artigos utilizando regressão multivariada para prever propriedades do petróleo e derivados foi inexpressivo até 2010. A partir deste ano, um grande número de artigos foi publicado usando métodos de regressão, como o PLS, em dados de diversas fontes analíticas, como cromatografia gasosa bidimensional (GODOY *et al*., 2011), espectroscopia no infravermelho com transformada de Fourier utilizando um sensor de reflectância total atenuada (ABBAS *et al*., 2012 e MULLER *et al*., 2012), espectroscopia de fluorescência ultravioleta síncrona (ABBAS *et al*., 2012), espectroscopia no infravermelho próximo (LAXALDE *et al*., 2011), espectrometria de massas por ionização e dessorção a laser (TERRA *et al*., 2015), entre outras. Alguns trabalhos utilizaram também métodos de seleção de variáveis (GODOY *et al*., 2011, MULLER *et al*., 2012, TERRA *et al*., 2015).

Métodos quimiométricos de regressão não lineares também têm sido aplicados a petróleo e derivados de maneira mais significativa nos últimos anos. Floresta aleatória foi aplicada em dados espectrais na região do infravermelho próximo para determinar propriedades de amostras de gasolina e nafta (LEE *et al*., 2013). Regressão por vetores de suporte foi usada com RMN $^{1}$H para prever temperaturas de destilação de petróleo bruto (FILGUEIRAS *et al*., 2015).

Enquanto isso, modelos de classificação continuaram a ser desenvolvidos ao longo dos anos para diferentes fins, como a discriminação de tipos de petróleo, utilizando a análise de fatores paralelos em dados de cromatografia gasosa acoplada à espectrometria de massas (EBRAHIMI *et al*., 2007), utilizando PCA em espectros na região do infravermelho e espectros de fluorescência ultravioleta-visível (ABBAS *et al*., 2008) e em cromatografia gasosa bidimensional abrangente (MISPELAAR *et al*., 2005 e VENTURA *et al*., 2011), entre outros vários estudos.

Dentre as técnicas espectroscópicas, a FTIR é uma das mais utilizadas para prever propriedades de petróleo e combustíveis, sendo uma das mais disponíveis em laboratórios. Seus espectros apresentam informações sobre diferentes grupos funcionais que possibilitam a correlação com diferentes propriedades físico-químicas de petróleo e combustíveis (BAIRD e OIA, 2016). Muitos modelos, desenvolvidos a partir de dados FTIR, têm sido usados em linha para controlar e otimizar as condições operacionais de refino de petróleo em tempo real (ALVES e POPPI, 2012 e LAMBERT *et al*., 1995).

Outra técnica espectroscópica que se mostrou muito promissora e amplamente utilizada para caracterizar petróleo e derivados é a ressonância magnética nuclear. Vários estudos na quimiometria têm relatado resultados satisfatórios ao aplicar RMN em petróleo e derivados (FILGUEIRAS *et al*., 2015; MOLINA *et al*., 2007; RAMOS *et al*., 2009; MASILI *et al*., 2012; MUHAMMAD e AZEREDO, 2014; MEJIA-MIRANDA *et al*., 2015; BARBOSA *et al*., 2016; DUARTE *et al*., 2016; FILGUEIRAS et al., 2016, LU *et al*., 2018 e VIEIRA *et al*., 2019).

## 2.2 Espectroscopia na região do infravermelho

A absorção da radiação na região do infravermelho induz a transições vibracionais moleculares cuja frequência depende das massas relativas e das ligações dos átomos envolvidos no modo de vibração (PAVIA *et al*., 2008). Na região do infravermelho médio, onde ocorre a transição fundamental, os modos de vibração são freqüentemente localizados em determinados grupos funcionais que então aparecem na mesma frequência ou próximo a ela, permitindo assim sua identificação em uma mistura (PASQUINI E BUENO, 2007). Entretanto, a situação é mais complexa na região do infravermelho próximo onde a absorção se deve a combinações ou

sobretons de modos de vibração, que não permitem a identificação, mas ainda possibilitam a quantificação, em particular com auxílio da análise multivariada de dados (PASQUINI, 2018).

Extrair informações quantitativas não é uma tarefa simples, sendo necessária a aplicação de métodos de calibração multivariada, principalmente para a análise de misturas complexas, como o petróleo. No entanto, para que a espectroscopia de infravermelho seja usada na calibração multivariada, as informações contidas nos espectros devem estar intimamente relacionadas à propriedade a ser estimada. Embora as amostras de petróleo geralmente exibam perfis espectrais muito semelhantes, em algumas regiões, é possível verificar pequenas diferenças entre os espectros de amostras com composições muito diferentes (PABÓN E SOUZA FILHO, 2019).

As principais vantagens da determinação das propriedades físico-químicas por meio de espectros, em comparação aos métodos tradicionais, são rapidez e facilidade de aquisição de espectros, exigência de pequena quantidade de amostra, não destrutível e determinação direta com mínimo ou nenhum preparo de amostra (BORRAS *et al*., 2015). A simplicidade operacional, rapidez e praticidade têm feito da espectroscopia no infravermelho uma alternativa para monitorar em linha as propriedades do petróleo nas refinarias, o que permite as tomadas de decisão em tempo real pelos operadores (KALLEVIK *et al*., 2000). A maioria dos estudos relatados na literatura empregou espectroscopia no infravermelho, seja na região do médio ou do próximo, para determinar propriedades de derivados de petróleo. O número de artigos que aplicam esta técnica em petróleo bruto é mais limitado (FALLA *et al*., 2006).

### 2.2.1 Aplicação da espectroscopia no infravermelho em petróleo

A Tabela 2.2 apresenta, em ordem cronológica, os artigos que usaram espectros NIR ou MIR como dados de entrada para construção de modelos quimiométricos que objetivaram prever as propriedades físico-químicas do petróleo bruto. Para cada estudo, foram identificadas as propriedades previstas, os pré-processamentos e os métodos de regressão aplicados.

WILT e WELCH (1998) foram um dos primeiros a utilizar espectros na região do infravermelho para estimar propriedades físico-químicas petróleo. Os autores

construíram modelos PLS para determinar o teor de asfaltenos em 50 amostras, utilizando a região do infravermelho médio. Além disso, construíram também modelos a partir da seleção de variáveis e da combinação de diferentes regiões espectrais. Os modelos construídos a partir da região espectral cujas absorbâncias são decorrentes dos compostos asfaltênicos apresentaram os melhores resultados, com erro padrão de 0,95% em peso e $R^2$ de 0,99, para o conjunto de predição.

Alguns anos depois, ASKE *et al*. (2001) também construíram modelos PLS, mas para predição de SARA, utilizando espectros MIR e NIR, separadamente. Os resultados mostraram que ambas as fontes de dados forneceram modelos com capacidade de predição semelhantes. Os modelos lineares puderam determinar SARA de forma satisfatória, com erros de predição iguais a 2,45, 2,20, 1,37 e 1,29% em peso, respectivamente, para saturados, aromáticos, resinas e asfaltenos, a partir de dados MIR.

As amostras, em ASKE *et al*. (2001), apresentaram perfis espectrais visualmente muito semelhantes; no entanto, ao sobrepor os espectros de algumas amostras com valores extremos de SARA, pode-se ver claramente as diferenças entre as bandas de absorção dessas amostras. Isso prova que o perfil espectral está diretamente relacionado à estrutura molecular dos componentes dos óleos que, por sua vez, reflete os teores de SARA. A região entre 1750 e 650 $cm^{-1}$ apresentou a maior diferença entre os espectros das amostras. Esta região apresenta bandas devido à deformação das ligações C – H em cadeias alifáticas e deformações angulares simétricas e assimétricas no plano do grupo $CH_2$; além disso, apresenta a região conhecida como impressão digital, que contém uma mistura de sinais relacionados às vibrações de cadeias alifáticas e aromáticas e a alguns heteroatomos (PAVIA *et al*., 2008). Apesar dos bons resultados, apenas 18 amostras foram utilizadas para a construção dos modelos, o que representa a principal desvantagem desta pesquisa.

Essa estreita relação entre espectro e propriedade permite que uma ampla gama de propriedades físico-químicas seja prevista com segurança utilizando a quimiometria. Nos anos seguintes, vários outros estudos surgiram propondo a construção de modelos, a partir de espectros no infravermelho, para estimar diversas propriedades do petróleo bruto, principalmente °API (ABBAS *et al*., 2012, PABÓN e SOUZA FILHO, 2019, DEARING *et al*., 2011, FILGUEIRAS *et al*., 2014, RODRIGUES *et al*., 2017, RAINHA *et al*., 2019, LONG *et al*., 2019). FILGUEIRAS *et al.* (2014)

alcançaram um modelo de alta exatidão com um erro quadrático médio da predição ($RMSE_P$) de 0,25 e um $R_P^2$ de 0,97 para a °API, usando dados de espectros no infravermelho.

Usando modelos lineares, alguns autores comparam o desempenho dos espectros MIR e NIR na calibração multivariada para prever propriedades (PABÓN e SOUZA FILHO, 2019, RAINHA *et al*., 2019, FOLLI *et al*., 2020). Em RAINHA *et al*. (2019), enquanto para °API e teor de nitrogênio total, a melhor fonte de dados foi NIR com um $RMSE_P$ igual a 0,9 e 0,0275 m/m%, respectivamente; para o teor de nitrogênio básico, a melhor foi a fonte de dados foi MIR com um $RMSE_P$ igual a 0,0134 m/m%.

**Tabela 2.2.** Propriedades do petróleo bruto previstas por quimiometria a partir de espectros de MIR e NIR.

| Referência | Ano | Dados | °API | Propriedades estimadas | Pré-tratamento | Método de regressão |
|---|---|---|---|---|---|---|
| WILT e WELCH | 1998 | MIR | não informado | teor de asfaltenos | centralização na média | PLS |
| HIDAJAT e CHONG | 2000 | NIR | 25.1 – 45.8 | Densidade<br>curva PEV | centralização na média e MSC | PLS |
| ASKE *et al*. | 2001 | MIR<br>NIR | 18.2 – 46.3 | SARA | Nenhum<br>primeira derivada | PCA e PLS |
| FALLA *et al*. | 2006 | NIR | não informado | curva PEV | centralização na média e primeira derivada | RNA[a] |
| PASQUINI e BUENO | 2007 | NIR | 13.2 – 49.6 | °API<br>curva PEV | correção da linha de base e normalização | RNA[a] e PLS |
| PEINDER *et al*. | 2009 | MIR,<br>RMN $^{1}H$ e<br>$^{13}C$ | 10.6 - 31.9 | rendimento de petróleo<br>Densidade<br>Viscosidade<br>teor de enxofre<br>ponto de fluidez<br>teor de asfaltenos<br>resíduo de carbono | primeira derivada e MSC | fusão de dados e PLS |
| DEARING *et al*. | 2011 | Raman,<br>NIR e<br>RMN $^{1}H$ | não informado | °API<br>rendimento por peso<br>rendimento por volume<br>teor de $H_2$ | normalização e centralização na média | fusão de dados e PLS |
| JINGYAN *et al*. | 2012 | MIR | 8.7 – 52.5 | número de acidez total | centralização na média, primeira e segunda derivada | seleção de variáveis e PLS |
| ABBAS *et al*. | 2012 | MIR e<br>EFS[b] | 21.3 – 45.5 | °API<br>alifáticos/aromáticos | SNV e primeira derivada | PLS |

| | | | | | | |
|---|---|---|---|---|---|---|
| MELENDEZ *et al*. | 2012 | MIR | não informado | SARA | normalização e primeira derivada | PLS |
| FILGUEIRAS *et al*. | 2014 | MIR | 16.0 – 23.0 | °API<br>teor de água<br>viscosidade cinemática | centralização na média e MSC | PCR, PLS e SVM[c] |
| RODRIGUES *et al*. | 2017 | MIR | 11.2 – 54.0 | °API | centralização na média | PLS e OPLS[d] |
| PABÓN e SOUZA FILHO | 2019 | NIR, MIR e FIR | 13.2 – 47.2 | SARA<br>°API | SNV e correção da linha de base | PLS |
| RAINHA *et al*. | 2019 | NIR e MIR | 11.4 – 57.5 | °API<br>número de acidez total<br>teor de nitrogênio básico | air-PLS e SNV | PLS, OPLS[d], iPLS, siPLS e carsPLS |
| LONG *et al*. | 2019 | NIR | 20.0 – 43.0 | °API<br>teor de enxofre | primeira derivada | PLS |
| MORO *et al*. | 2019 | MIR, RMN $^{1}H$ e $^{13}C$ | 11.4 – 54.0 | teor de enxofre<br>número de acidez total<br>teor de nitrogênio básico<br>número de acidez total<br>SAP[e] | MSC, SNV, icoshift e autoescalonamento | fusão de dados e PLS |
| MOHAMMADI *et al*. | 2020 | MIR | 18.5 – 54.07 | °API | correção da linha de base, SNV e OSC | PLS, SVM[c] e PLS-DA |
| BARRERA *et al*. | 2020 | MIR | não informado | número de acidez total | airPLS e normalização | PCR, PLS |
| FOLLI *et al*. | 2020 | NIR, MIR e RMN $^{1}H$ | não informado | teor de enxofre<br>teor de nitrogênio básico<br>número de acidez total | icoshift, centralização na média, SNV, MSC, primeira derivada e airPLS | ASA-VIF-SVR[f], Kernel-PLS e PLS |

[a]redes neurais artificiais, [b]espectroscopia de fluorescência síncrona, [c]máquinas de vetores de suporte (*support vector machine*), [d]projeções ortogonais a mínimos quadrados parciais, [e]teor de compostos saturados, aromáticos e polares, [f]algoritmo de busca angular e fator de inflação de variância com regressão de vetores de suporte (*angular search algorithm and variance inflation factor with support vector regression*).

De acordo com a Tabela 2.2, o MIR foi aplicado com uma frequência um pouco maior que a espectroscopia NIR como dados de entrada na calibração multivariada de petróleo bruto. No total, 14 artigos utilizaram MIR, enquanto 9 utilizaram NIR.

Quase todos os estudos da Tabela 2.2 aplicaram a regressão linear PLS. Como pode ser visto, os métodos lineares foram muito bem explorados nos primeiros anos, quando a quimiometria foi introduzida como uma ferramenta analítica para o petróleo e seus derivados. Posteriormente, com o desenvolvimento da área, métodos de regressão não lineares também passaram a ser investigados para a previsão das propriedades do óleo.

Alguns dos estudos compararam o modelo PLS com modelos não lineares (PASQUINI e BUENO, 2007; FALLA *et al*., 2006; RODRIGUES *et al*., 2017 e FOLLI *et al*., 2020). Seus resultados indicam que, de forma geral, não há um método de regressão mais adequado que os demais para dados de espectros no infravermelho. Em alguns estudos, métodos lineares forneceram melhores resultados, enquanto que em outros, os não lineares foram mais vantajosos. PASQUINI *et al*. (2007) alcançou melhores previsões com PLS que com RNA para gravidade API e a curva PEV. No entanto, FILGUEIRAS *et al*. (2014) mostraram que o modelo SVR produziu melhores resultados do que o PLS para determinação de gravidade API, mas resultados equivalentes para viscosidade cinemática e teor de água.

Apesar da aplicação de novos métodos quimiométricos como RNA e SVR, muitos trabalhos atuais (LONG *et al*., 2019; MORO *et al*., 2019; MOHAMMADI *et al*., 2020 e BARRERA *et al*., 2020) continuaram a aplicar o método linear PLS para determinar propriedades físico-químicas devido à sua simplicidade e eficácia para dados com comportamento linear. Não é justificável o uso de ferramentas/algoritmos não lineares ao invés de um simples PLS quando não há ganho satisfatório de exatidão que compense o aumento da complexidade da modelagem.

Uma análise visual comparativa entre espectro e propriedade estimada também foi feita por alguns autores. FALLA *et al*. (2006) estimou a curva SimDis de amostras de petróleo bruto com base na espectroscopia NIR, usando a metodologia de redes neurais. Percebeu-se que as similaridades e diferenças na curva SimDis entre as amostras também foram observadas nos espectros NIR. PASQUINI *et al*. (2007) alcançou a mesma observação. Seus resultados mostraram espectros mais próximos e semelhantes para amostras com PEV semelhantes.

Enquanto isso, BAIRD *et al*. (2016) revisaram métodos quimiométricos e técnicas analíticas para prever propriedades físico-químicas, mas abrangendo diferentes tipos de petróleo e derivados, não apenas petróleo bruto. Segundo os autores, a espectroscopia NIR tem sido a técnica analítica mais utilizada e o PLS o método quimiométrico mais aplicado.

## 2.3 Espectroscopia de ressonância magnética nuclear

A espectroscopia de RMN faz uso da interação entre a radiação eletromagnética e o núcleo de átomos que possuem a propriedade spin nuclear. Na área de petróleo, os núcleos mais estudados são $^{1}H$ e $^{13}C$, os quais permitem obter, por meio dessa técnica, informações sobre o esqueleto carbono-hidrogênio de moleculas orgânicas. Estes núcleos possuem um momento magnético próprio que interage com um campo magnético externo aplicado, se comportando de forma semelhante a pequenos imãs.

O fenômeno de RMN ocorre quando os núcleos, alinhados a um campo externo aplicado, são induzidos a absorver energia de uma radiação de radiofrequência incidente, mudando de orientação de spin em relação ao campo aplicado. Cada tipo de núcleo tem uma frequência de absorção específica, cuja intensidade da absorção é proporcional à concentração do mesmo. Após a retirada de energia, ocorre a relaxação do núcleo que retorna ao seu estado original, por meio da perda da energia absorvida durante a exitação. Neste momento, os núcleos emitem sinais eletromagnéticos que são detectados pelo espectrômetro de RMN.

Nem todos os núcleos ressonam na mesma frequência. O ambiente ao seu redor, como as ligações químicas e os átomos vizinhos, altera a frequência de absorção de energia pelos núcleos, o que contribui para elucidar a estrutura eletrônica, fornecendo informações detalhadas sobre o arranjo atômico, estado de reação, ambiente químico e grupos funcionais individuais de uma molécula (PAVIA *et al*., 2008).

As estruturas moleculares que compõem uma certa amostra de petróleo podem estar diretamente relacionadas aos sinais gerados em seus espectros de RMN. MOLINA *et al*. (2007) explora essa relação dividindo o espectro de hidrogênio em 12 regiões e associando cada uma delas a um respectivo grupo estrutural, como aromáticos, parafinas e naftalenos. Assim, é possível caracterizar as diferentes

frações que compõem os óleos, além de quantificar outros grupos de hidrocarbonetos. HASAN *et al*. (1983) e POVEDA *et al*. (2012) atribuíram grupos estruturais a diferentes regiões dos espectros RMN $^{1}H$ e $^{13}C$, de acordo com seus deslocamentos químicos. HASAN *et al*. (1983) dividiu os espectros de RMN $^{1}H$ e $^{13}C$ em cinco (de 0 a 9 ppm) e seis (de 0 a 160 ppm) regiões, respectivamente. Já POVEDA *et al*. (2012) dividiu em oito regiões (de 0 a 12 ppm) os espectros de RMN $^{1}H$ e 15 regiões (3 a 220 ppm) os espectros de RMN $^{13}C$. Essas regiões são amplamente utilizadas para elucidar as variáveis mais importantes na construção de modelos quimiométricos de petróleo bruto.

A RMN é uma das principais ferramentas para elucidação estrutural de hidrocarbonetos e tem sido amplamente aplicada em petróleo bruto e seus derivados nos últimos anos. Isso se deve à capacidade da RMN de fornecer informações detalhadas sobre a estrutura química, especialmente para matrizes complexas como o petróleo (BRERETON, 2000).

### 2.3.1 Aplicação da espectroscopia de RMN em petróleo

A RMN pode ser empregada para fins qualitativos ou quantitativos, permitindo identificar e quantificar diferentes tipos de hidrocarbonetos. Em petróleo, a espectroscopia de RMN foi inicialmente utilizada para caracterização estrutural sem o uso de qualquer ferramenta quimiométrica. Como exemplo, ALI *et al*. (1990), a partir de espectros de RMN $^{1}H$ e $^{13}C$, propuseram modelos estruturais hipotéticos para compostos asfaltênicos. URDAL *et al*. (1986) determinaram os principais constituintes e as estruturas médias de componentes orgânicos/inorgânicos de óleos brutos usando apenas espectros de RMN $^{13}C$.

Com a popularização das ferramentas quimiométricas, os espectros de RMN passaram a ser empregados também associados à regressão multivariada para estimar propriedades do petróleo. Em uma pesquisa desenvolvida por MOLINA *et al*. (2007), modelos para predição de nove propriedades físico-químicas foram desenvolvidos a partir de espectros de RMN $^{1}H$ e regressão por PLS. Foram obtidos modelos com altos coeficientes de determinação, mostrando a viabilidade da utilização desta espectroscopia e quimiometria em amostras de petróleo.

Ao longo dos anos, várias outras metodologias de RMN associadas à quimiometria foram desenvolvidas para determinar diversas propriedades do petróleo e seus derivados. A Tabela 2.3 mostra artigos, organizados por data, que utilizaram espectros de RMN $^{1}H$ e $^{13}C$ como dados de entrada em modelos quimiométricos para prever propriedades do petróleo bruto. Também foram identificadas as propriedades previstas, os métodos de regressão e os pré-processamentos aplicados.

Utilizando PLS, MASILI *et al*. (2012), a partir de espectros de RMN $^{1}H$, determinaram propriedades do petróleo bruto que são essencialmente importantes para fins de monitoramento em plantas de refino. Os autores alcançaram um erro de predição igual a 0,17 mg $KOHg^{-1}$ para o número de acidez total e 0,25% em peso para o teor de enxofre. Foi demonstrada a viabilidade do método linear em dados espectrais, mesmo para um conjunto de amostras de grande variabilidade na composição, compreendendo óleos leves, médios e pesados.

Vale ressaltar também que a maioria dos estudos listados na Tabela 2.3 utilizou dados de RMN $^{1}H$ para a calibração multivariada e poucos utilizaram dados de RMN $^{13}C$. Isso pode ser devido a algumas desvantagens da RMN $^{13}C$ em relação a $^{1}H$, como o alto custo da análise e a necessidade de maior quantidade de amostra, uma vez que a abundância isotópica do carbono 13 é muito menor que a do hidrogênio. Além disso, os espectros de RMN $^{1}H$ apresentam pouco mais de 20 mil variáveis, enquanto que os espectros de RMN $^{13}C$ têm um número muito maior de variáveis, geralmente mais de 60 mil, o que demanda maior poder computacional para as análises quimiométricas (FILGUEIRAS *et al*., 2016). Além disso, os espectros de RMN $^{1}H$ podem ser obtidos muito rapidamente, em poucos minutos, e o RMN $^{13}C$ leva mais tempo para ser adquirido, podendo chegar a duas horas.

No entanto, em alguns casos, a técnica de RMN $^{13}C$ pode produzir resultados superiores, uma vez que a RMN $^{1}H$ não é capaz obter sinais relacionados a carbonos internos que não estão diretamente ligados a átomos de hidrogênio. Isso ocorre especialmente quando se deseja determinar propriedades em estruturas poliaromáticas, como resinas e asfaltenos, que contêm vários átomos de carbonos quaternários, ou seja, sem a presença de hidrogênio vizinho a ser detectado em espectros de RMN $^{1}H$ (FILGUEIRAS *et al*., 2016).

O mesmo padrão identificado em pesquisas que empregaram espectroscopia no infravermelho pode ser observado nas que empregaram RMN. Como pode ser visto

na Tabela 2.3, as primeiras pesquisas publicadas utilizaram predominantemente o método linear PLS. Somente em 2015, FILGUEIRAS *et al*. (2015) utilizaram o método não linear SVR para prever as temperaturas de destilação de óleos a partir de espectros de RMN $^{1}$H. Um ano depois, o SVR foi novamente aplicado ao petróleo, mas, neste caso, em associação à seleção variáveis chamada algoritmo genético (FILGUEIRAS *et al*., 2016).

**Tabela 2.3.** Propriedades do petróleo bruto previstas por quimiometria a partir de espectros de RMN $^{1}H$ e $^{13}C$.

| Referência | Ano | Espectro | °API | Propriedade estimada | Pré-tratamento | Método de regressão |
|---|---|---|---|---|---|---|
| MOLINA *et al*. | 2007 | RMN $^{1}H$ | 12,3 – 45,1 | curva SimDis<br>PEI<br>teor de enxofre<br>teor de nitrogênio<br>teor de ceras<br>vanádio e níquel<br>MRC<br>insolubilidade em $nC_7$ | não informado | PLS |
| RAMOS *et al*. | 2009 | RMN $^{1}H$ de baixo campo | óleos leves a extrapesados | Viscosidade | normalização e centralização na média | PLS |
| PEINDER *et al*. | 2009 | MIR, RMN $^{1}H$ e $^{13}C$ | 10,6 – 31,9 | rendimento de petróleo<br>densidade<br>viscosidade<br>teor de enxofre<br>ponto de fluidez<br>teor de asfaltenos<br>resíduo de carbono | primeira derivada e MSC | fusão de dados e PLS |
| DEARING *et al*. | 2011 | Raman, NIR e RMN $^{1}H$ | não informado | °API<br>rendimento por peso<br>rendimento por volume<br>teor de $H_2$ | normalização e centralização na média | fusão de dados PLS |
| MASILI *et al*. | 2012 | RMN $^{1}H$ | 43,6 – 19,2 | $K_{UOP}$ | seleção de variáveis, | PLS |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | densidade<br>número de acidez total<br>curva PEV<br>teor de enxofre | normalização e centralização na média | |
| MUHAMMAD e AZEREDO | 2014 | RMN $^{1}H$ de baixo campo | 9,5 – 34,0 | °API<br>viscosidade | normalização e centralização na média | fusão de dados e PLS |
| MEJIA-MIRANDA | 2015 | RMN $^{1}H$ | 12,4 – 44,6 | taxa de corrosão | normalização | PLS |
| FILGUEIRAS *et al*. | 2015 | RMN $^{1}H$ | 17,0 – 54,0 | curva SimDis | alinhamento, SNV e normalização | ePLS e eSVR |
| BARBOSA *et al*. | 2016 | RMN $^{1}H$ de baixo campo | 16,9 – 28,9 | número de acidez total<br>teor de enxofre | não informado | MLR |
| DUARTE *et al*. | 2016 | RMN $^{1}H$ | 11,4 – 54,0 | °API<br>resíduo de carbono<br>WAT<br>teor de nitrogênio básico | alinhamento e SNV | PLS |
| FILGUEIRAS *et al*. | 2016 | RMN $^{13}C$ | 10,0 – 55,0 | SAP | alinhamento | GA[a]-SVR |
| DUARTE *et al.* | 2017 | RMN $^{1}H$ | 13,1 – 54,0 | curva de destilação | alinhamento e SNV | PLS |
| LU *et al*. | 2018 | RMN | não informado | resíduo de carbono<br>teor de asfaltenos | não informado | PLS, MLR, LS-SVM, DNNE[b], ERNN[c] |
| VIEIRA *et al*. | 2019 | RMN $^{1}H$ | não informado | $K_{UOP}$<br>teor de nitrogênio<br>parâmetro de solubilidade<br>ponto de fluidez<br>teor de enxofre | alinhamento e SNV | PLS, OPLS, iPLS, siPLS e MPA[d] |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | viscosidade cinemática | | |
| MORO *et al*. | 2019 | MIR, RMN $^{1}H$ e $^{13}C$ | 11,4 – 54,0 | teor de enxofre<br>teor de nitrogênio total<br>teor de nitrogênio básico<br>número de acidez total<br>SAP | MSC, SNV, icoshift e autoescalonamento | fusão de dados e PLS |
| LOVATTI *et al*. | 2019 | RMN $^{1}H$ e $^{13}C$ | 11,4 – 52,4 | ponto de fluidez máximo | Icoshift, autoescalonamento, SNV, normalização, centralização na média e 1ª derivada | floresta randômica |
| LOVATTI *et al*. | 2020 | RMN $^{1}H$ e $^{13}C$ | 11,4 – 52,4 | número de acidez total | Icoshift, autoescalonamento, SNV, normalização, centralização na média e primeira derivada | fusão de dados, PCA e floresta randômica |
| PAULO *et al*. | 2020 | RMN $^{1}H$ e $^{13}C$ | 11,4 – 54,0 | °API<br>número de acidez total<br>calor de combustão<br>viscosidade cinemática<br>SARA | icoshift, autoescalonamento, SNV | PSO[e]-PLS e OPS[f]-PLS |
| FOLLI *et al*. | 2020 | NIR, MIR e RMN $^{1}H$ | não informado | teor de enxofre<br>teor de nitrogênio básico<br>número de acidez total | icoshift, centralização na média, SNV, MSC, 1ª derivada e airPLS | ASA-VIF-SVR, Kernel-PLS and PLS |

[a]algorismo genético, [b]redes neurais com pesos aleatórios, [c] redes neurais recorrentes de Elman, [d]análise de modelo populacional, [e]método do enxame de partículas, [f]seleção dos preditores ordenados.

Já LU *et al*. (2018), no lugar de aplicar apenas um método, testou e comparou diferentes métodos não lineares - redes neurais não correlacionadas, redes neurais com pesos aleatórios e máquinas de vetores de suporte de mínimos quadrados - com métodos lineares - PLS e MLR - para construir um modelo de predição de resíduos de carbono e teor de asfaltenos em petróleo bruto, a partir de espectros de RMN $^{1}$H. As redes neurais com pesos aleatórios produziram o melhor modelo com um erro quadrático médio de 0,16% em peso e de 0,098% em peso para carbono residual e teor asfaltenos, respectivamente.

Os artigos mais recentes, para inferir propriedades físico-químicas do petróleo bruto, empregam métodos não lineares, uma vez que os lineares já estão bem consolidados na área. LOVATTI *et al*. (2019) e LOVATTI *et al*. (2020) utilizaram floresta randômica para discriminar amostras de petróleo bruto em relação ao seu ponto de fluidez máximo e a valores de NAT, respectivamente. FOLLI *et al*. (2020) empregaram o SVR com seleção variáveis para estimar NAT, o teor de nitrogênio básico e o teor de enxofre em petróleo bruto, a partir de dados de RMN $^{1}$H, MIR e NIR. Os melhores modelos apresentaram $RMSE_P$ iguais a 0,0071 m/m%, 0,0623 m/m% e 0,1426 $mgKOHg^{-1}$ para nitrogênio básico, enxofre e NAT, respectivamente.

Além da técnica de RMN de alto campo, amplamente difundida, também existem estudos que aplicaram a técnica de RMN de baixo campo (RAMOS *et al*., 2009, MUHAMMAD e AZEREDO, 2014, BARBOSA *et al*., 2016). Nesse caso, as resoluções das medições e a sensibilidade são bem menores. Sua grande vantagem é o baixo custo dos equipamentos e de manutenção (RUDSZUCK *et al*., 2019). Apesar disso, a técnica tem sido pouco explorada para predizer propriedades físico-químicas de petróleo bruto em comparação com a de alto campo, o que pode ser verificado em função do baixo número de estudos, conforme demonstrado na Tabela 2.3.

Nestes trabalhos, todos utilizaram regressão linear multivariada para prever algumas propriedades, como número de acidez total, teor de enxofre (BARBOSA *et al*., 2016), viscosidade (RAMOS *et al*., 2009, MUHAMMAD e AZEREDO, 2014) e densidade API (MUHAMMAD e AZEREDO, 2014) de amostras de petróleo bruto. Muhammad e Aezeredo também comparam a exatidão de modelos de RMN de baixo e alto campo. Os autores descobriram que, apesar de oferecer menos sensibilidade e resolução nas análises de dados, a relaxometria de baixo campo proporcionou modelos com desempenho superior ao RMN de alto campo. Foram observadas

diferenças significativas nos perfis de relaxação entre as amostras cujas propriedades estimadas diferiram bastante, o que mostra que tais propriedades estão bem correlacionadas com dados de RMN de baixo campo.

## 2.4 Estimativa das propriedades a partir de espectros de IR, RMN

Entre os 34 artigos apresentados nas Tabelas 2.2 e 2.3, um total de 14 estimaram a densidade ou o °API a partir de dados de IR ou RMN. Vários deles alcançaram bons resultados com altos coeficientes de determinação (maiores que 0,97) para amostras de predição (ABBAS *et al*., 2012, FILGUEIRAS *et al*., 2014, RODRIGUES *et al*., 2017 e RAINHA *et al*., 2019). Estes números mostram que as pesquisas relacionadas à estimativa da densidade/°API são as mais bem estabelecidas na literatura.

As publicações listadas nas Tabelas 2.2 e 2.3 apresentam modelos construídos a partir de dados de IR e RMN para estimar as propriedades físico-químicas listadas na primeira coluna da Tabela 2.1. Já as propriedades listadas na segunda coluna ainda não foram previstas por modelos quimiométricos, utilizando espectros de IR ou RMN. Algumas dessas propriedades não foram estimadas, visto que podem não ser tão relevantes quanto outras para a caracterização do petróleo. É preciso lembrar também que a quimiometria do petróleo é uma área relativamente nova e que ainda há muito a ser explorado.

Para ser estimada, a propriedade deve ter relação com o sinal analítico das técnicas espectroscópicas´ou seja, as informações fornecidas pelos espectros devem estar direta ou indiretamente relacionadas à propriedade estimada. Por exemplo, os teores de enxofre e água estão diretamente relacionados às informações contidas nos espectros de infravermelho, já que estes produzem bandas de absorção específicas nas regiões em torno de 2550 e 3300 $cm^{-1}$, respectivamente. Ou seja, caso o modelo apresente boa capacidade preditiva, a explicação estaria na alta correlação entre a propriedade prevista e as técnicas mencionadas. Um bom exemplo são os resultados de MORO *et al*. (2019) cujos modelos de previsão de teor de enxofre total a partir de dados de RMN apresentaram boa exatidão.

Para algumas propriedades como a densidade, a curva PEV, entre outras, a relação entre espectro e propriedade é indireta, pois não há uma banda de absorção específica para essas propriedades. Nestes casos, os espectros estão relacionados a

algum outro componente que, por sua vez, é proporcional à propriedade inferida.

Para espectroscopia de RMN, os espectros fornecem sinais diretamente relacionados aos átomos de carbono ou hidrogênio, portanto, para propriedades como teor de enxofre e água, por exemplo, as informações dentro dos espectros são indiretamente correlacionadas a essas propriedades por meio dos modelos de calibração. No entanto, no que se refere à determinação de, por exemplo, teor de componentes polares, a relação entre esta propriedade e os espectros de RMN $^{13}C$ pode ser considerada direta, pois os espectros provêem informações sobre as estruturas poliaromáticas que, por sua vez, estão diretamente relacionadas ao teor de polares.

Algumas propriedades, como teor de sal e teor de metal, não estão diretamente correlacionadas com os espectros e sua predição em petróleo por técnicas espectroscópicas não é conhecida. Não foram encontradas pesquisas para prever as propriedades listadas na segunda coluna da Tabela 2.1 em petróleo bruto, possivelmente devido à falta de correlação entre essas propriedades e os espectros no infravermelho e de RMN.

Metais, por exemplo, não apresentam sinais espectrais na região do infravermelho, sendo impossíveis de serem estimados diretamente por estes espectros. No entanto, os metais podem ser encontrados ligados à matéria orgânica devido à complexação do metal. Nesse caso, é possível estimar indiretamente a concentração de metais utilizando tais espectros e a quimiometria (SHI *et al*., 2014).

Existem vários estudos que estimam as concentrações de metais a partir de espectros de NIR em amostras de solo, mas não em petróleo. Esses estudos mostram o bom potencial do infravermelho e da quimiometria para a estimativa de várias concentrações de metais pesados no solo (SHI *et al*., 2014). Entretanto, mais estudos são necessários para analisar a viabilidade desse procedimento em petróleo.

Os teores de sal e cinzas também não foram estimados em petróleo e seus derivados; entretanto, eles já foram estimados, de forma satisfatória, em produtos comerciais de carne processada (MARCHI *et al*., 2017) e resíduos de alimentos (RAMBO *et al*., 2020), respectivamente, utilizando a espectroscopia NIR. Uma vez que sal e cinzas não possuem uma banda de absorção específica, não podem ser medidos diretamente pelos espectros NIR. A previsão de sal foi feita indiretamente com base na mudança do espectro da água (MARCHI *et al*., 2017). De acordo com

POJIĆ *et al*. (2010) as cinzas podem ser previstas pela correlação com a quantidade total de compostos orgânicos e de água.

A maioria das pesquisas listadas nas Tabelas 2.2 e 2.3 informou a faixa de °API das amostras. Fornecer esses dados é extremamente importante para indicar a variabilidade composicional das amostras utilizadas. Um conjunto de amostras com grande faixa de °API inclui óleos leves, médios e pesados. Modelos construídos a partir de amostras com menor variabilidade de °API tendem a apresentar bons resultados de parâmetros de desempenho, mas apresentam menor robustez. Porém, modelos construídos, a partir de amostras com alta variabilidade composicional, podem ser considerados mais robustos, uma vez que podem ser aplicados a uma maior variabilidade de amostras externas.

## 2.5 Quimiometria

A quimiometria utiliza de ferramentas matemáticas, estatísticas e computacionais para extrair informações relevantes a partir de um conjunto de dados químicos para facilitar a compreensão, principalmente, de fontes de dados complexas (PASQUINI, 2018). Muitos autores fazem uso da quimiometria em diferentes áreas da ciência para planejar e otimizar experimentos, inferir propriedades, classificar ou determinar a origem de amostras, entre outras finalidades (BORRÀS *et al*., 2015).

De forma geral, a quimiometria pode ser aplicada em três vertentes: planejamento de experimentos, métodos qualitativos e quantitativos. O planejamento das condições experimentais tem o objetivo de otimizar os pocedimentos, de forma a realizar o menor número possível de ensaios, obtendo o máximo de informações relevantes sobre o sistema. Já os métodos qualitativos consistem em uma análise exploratória de um conjunto de dados para reconhecimento de padrões como, por exemplo, detecção de padrões ou anomalias, correlações entre variáveis e fatores agrupadores ou diferenciadores. Por fim, os métodos quantitativos utilizam a calibração para quantificar informações, estabelecendo uma relação matemática entre os dados químicos e a resposta de interesse (MILLER J. N. e MILLER J. C., 2010).

Dentre os métodos quimiométricos, a calibração multivariada tem sido amplamente empregada na área de petróleo para prever propriedades de interesse a partir de medições de espectroscopia (RODRIGUES *et al*., 2018, DOUGLAS *et al*.,

2018, RODRIGUES *et al*., 2017, LOZANO *et al*., 2017).

O processo de calibração multivariada inicia-se com a aquisição de dados analíticos. Os dados são organizados em duas matrizes diferentes. Uma matriz é formada pelas variáveis dependentes que, geralmente, são propriedades que se desejam inferir. A outra é formada pelas variáveis independentes que, geralmente, são dados de absorbância, medidos em diferentes comprimentos de onda. Em ambas as matrizes, cada coluna corresponde a uma variável e cada linha a uma amostra (WOLD *et al*., 2001).

A matriz de variáveis independentes, geralmente, é formada por dados de uma única técnica analítica específica. Entretanto é possível reunir dados de duas ou mais técnicas analíticas diferentes nesta única matriz. Essa estratégia é chamada de fusão de dados e objetiva agrupar informações complementares para construir modelos mais exatos (CASTANEDO, 2013).

A próxima etapa, denominada pré-processamento ou pré-tratamento, consiste em modificar o conjunto de dados da matriz de variáveis independentes, para torná-lo adequado à calibração. Sua função não é criar uma nova informação, mas sim remover fontes indesejadas de variabilidade como, por exemplo, o espalhamento elástico quando os fótons que entram em contato com o amostrador e ruídos que prejudicam a capacidade de previsão dos modelos (PASQUINI, 2018).

Os dados tratados são, em seguida, divididos em um conjunto de calibração e outro de previsão. Ao primeiro conjunto, é aplicado o algoritmo de calibração multivariada para construção de equações matemáticas que relacionem as variáveis dependentes com as independentes. Na sequência, o conjunto de previsão é utilizado para avaliar os modelos quanto a sua capacidade preditiva (BRERETON, 2000).

Uma consideração importante ao construir modelos quimiométrico é o número de amostras dos conjuntos de calibração e previsão. Enquanto que alguns estudos utilizaram um grande número de amostras, outros utilizaram um número pequeno. O número mínimo de amostras para compor um conjunto de calibração deve levar em consideração a variabilidade do problema a ser modelado. O modelo de calibração é treinado para reconhecer informações que estão presentes no conjunto de calibração, portanto, qualquer nova amostra cuja informação distoe muito do conjunto de calibração será identificada como uma amostra anômala pelo modelo. Alguns artigos nas Tabelas 2.2 e 2.3 (DUARTE *et al*., 2016, PASQUINI e BUENO, 2007, HIDAJAT e

CHONG, 2000, JINGYAN *et al*., 2012, MORO *et al*. 2019) utilizaram, no conjunto de calibração, uma série de amostras representativas, compreendendo vários tipos de petróleo, originários de diferentes campos de exploração ou diferentes países. Desta forma, foi possível obter um conjunto de várias amostras com propriedades extremas e construir um modelo com aplicação mais abrangente. Portanto, a representatividade é alcançada não apenas com um bom número de amostras, mas também com uma alta variabilidade nos dados e com a diversidade das amostras.

Amostras menos representativas não necessariamente produzirão modelos com baixa capacidade preditiva, mas modelos com aplicabilidade limitada a apenas aquele tipo específico de óleo usado para construir o modelo.

Mesmo usando amostras representativas, o modelo de calibração deve ser validado contra outro conjunto independente de amostras. Esta etapa permite avaliar se o modelo é satisfatório ou não. Dessa forma, na maioria dos artigos citados, o número total de amostras foi dividido em dois conjuntos: de calibração e previsão, como discutido anteriormente. Um conjunto é aplicado para construir o modelo matemático usando um método de regressão, seguido de sua otimização usando a validação cruzada. O outro conjunto de amostras, denominado amostras externas, ou seja, amostras que não foram utilizadas para construir o modelo, é aplicado para avaliar sua exatidão na validação externa. Esta avaliação com amostras externas é essencial, uma vez que um modelo bem ajustado para amostras internas não implica que seja preciso para amostras externas, que é o objetivo principal de qualquer modelo quimiométrico. Geralmente, de 20 a 30% do total de amostras são usados no conjunto de validação externa.

### 2.5.1 Pré-processamento

Mesmo quando a obtenção dos espectros de RMN é feita de forma cuidadosa e padronizada, podem ocorrer desalinhamentos nos deslocamentos químicos, devido às interferências e perturbações espectrais, trazendo prejuízos à modelagem. Inicialmente, algumas abordagens foram propostas para resolver este problema, mas nenhuma delas foi amplamente adotada devido ao baixo desempenho de alinhamento e alto custo computacional (SAVORANI *et al*., 2010). Somente em 2010, SAVORANI *et al*. (2010) desenvolveram uma ferramenta para remover o desalinhamento dos

espectros de RMN chamada *Icoshift*.

O *Icoshift* alinha os espectros de cada amostra de forma independente usando como base um espectro de referência, o qual pode ser a média ou mediana de todos os espectros ou um espectro real. Os espectros são divididos em intervalos definidos pelo usuário e o alinhamento ocorre dentro de cada intervalo por meio de uma ferramenta matemática que maximiza a correlação cruzada local e desloca todo o espectro para direita ou esquerda.

Depois de desenvolvido, o *Icoshift* foi aplicado com sucesso em diversos estudos como os de FILGUEIRAS *et al.* (2015), DUARTE *et al*. (2016), DUARTE *et al*. (2017), FILGUEIRAS *et al*. (2016), VIEIRA *et al*. (2019) e PAULO *et al*. (2020) para alinhar espectros de petróleo bruto antes da construção dos modelos de calibração.

Para obter bons modelos, também é necessário realizar pré-tratamentos cujo objetivo é reduzir variações indesejáveis nos dados que podem influenciar nos resultados finais. Além do alinhamento, os espectros de RMN e de infravermelho precisam ser pré-tratados para remover variações sistemáticas não informativas, reduzir o ruído e eliminar mudanças de linha de base que não estão associadas aos objetivos dos modelos. FALLA *et al*. (2006) destaca a importância do pré-processamento de dados que pode melhorar a discriminação da informação espectral.

Nos estudos listados nas Tabelas 2.2 e 2.3, o pré-tratamento mais aplicado foi a centralização na média. Este método centraliza os dados de cada variável em torno de sua média. Para centrar os dados na média calcula-se a média das intensidades em cada comprimento de onda e, em seguida, subtrai-se cada intensidade pelo seu respectivo valor médio. Sua função é evitar que os pontos mais distantes do centro tenham maior influência do que os mais próximos, evitando que a magnitude das variáveis afete os cálculos. É altamente recomendável centrar os dados na média para variáveis contínuas, como as dos espectros, uma vez que torna mais evidente as diferenças relativas das intensidades de sinais entre as variáveis (ADAMS, 1995).

Para os espectros na região do infravermelho, o pré-tratamento mais aplicado foi a primeira derivada. Em geral, este pré-tratamento é amplamente aplicado em espectros na região do infravermelho porque atua removendo mudanças de linha de base e aumenta a seletividade (LIRA *et al*., 2010).

Outro pré-processamento muito utilizado é o escalonamento, cuja função é ponderar os valores de cada variável pela sua variância. É importante escalonar os

dados quando a matriz é composta por variáveis com diferentes ordens de grandeza, caso contrário, é possível que uma variável se sobreponha à outra, impedindo que aquelas que tenham pequena ordem de grandeza contribuam com a modelagem. O escalonamento iguala o impacto das variáveis na contrução do modelo, dando a mesma importância a todas elas. Entretanto, tem como desvantagem aumentar a importância de variáveis que contem ruído, o que pode reduzir a acurácia dos modelos (ADAMS, 1995).

Dois outros pré-processamento importantes e comumente utilizados em espectros, como pode ser visto nas Tabelas 2.2 e 2.3, são a correção de dispersão multiplicativa (MSC) e a variável normal padrão (SNV). Ambos removem desvios de linha de base nos espectros causados por espalhamento multiplicativo de luz. Desta forma, todas as amostras passam a ficar na mesma linha de base, evitando que sua magnitude afete os cálculos.

### 2.5.2 Regressão por mínimos quadrados parciais

A calibração tem o objetivo de correlacionar uma propriedade de interesse com medidas de um instrumento analítico. O resultado do procedimento é um modelo matemático capaz de quantificar a propriedade de interesse em novas amostras onde ela é desconhecida.

Correlacionar, por exemplo, uma propriedade com valores de absorbância em apenas um comprimento de onda constitui uma calibração univariada. Na grande maioria das vezes, a calibração univariada não fornece um modelo adequado para previsão de propriedades. Isso acontece principalmente em amostras reais e complexas, porque a informação útil geralmente está dispersa por todo o espectro, sendo necessária a tomada de dados de absorbância em diversos comprimentos de onda. Cada comprimento de onda representa uma variável e, neste caso, o processo é chamado de calibração multivariada de dados (PASQUINI, 2018).

Existem diversos métodos de calibração linear multivariada. Em todos eles, o objetivo é encontrar uma equação linear que relacione as variáveis preditoras e preditas. Considera-se, por exemplo, que a análise espectroscópica de n amostras em *m* comprimentos de onda diferentes forma a matriz de variáveis independentes $\mathbf{X}_{(n,m)}$. Considera-se também que a determinação de P propriedades em cada amostra

constitui a matriz de variáveis dependentes $\mathbf{Y}_{(n,p)}$. As duas matrizes podem ser relacionadas por meio da equação linear (ANDERSSON, 2009):

$$\mathbf{Y}_{(n,p)} = \mathbf{X}_{(n,m)}.\mathbf{B}_{(m,p)} \qquad \text{Equação (2.1)}$$

A regressão linear multivariada mais empregada é a regressão por mínimos quadrados parciais (PLS). Trata-se de um método relativamente simples, com resposta rápida e, em muitos casos, bons resultados (MENDES *et al*., 2018). A PLS visa construir um modelo matemático linear que relacione variáveis independentes com uma ou mais variáveis dependentes. Para isso, ela reduz o número de variáveis originais a um conjunto menor de componentes não correlacionados para posteriormente realizar a regressão sobre esses componentes. Isso reduz a dimensionalidade e a multicolinearidade entre as variáveis antes da regressão (BRERETON, 2000).

O método decompõe, não somente a matriz **X**, mas também a matriz **Y**, em matrizes de scores (**T** e **U**) e de loadings (**P** e **Q**), acrescido da matriz de resíduos (**E** e **F**) de acordo com as equações (BRERETON, 2000):

$$\mathbf{X} = \mathbf{T}.\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad \text{Equação (2.2)}$$

$$\mathbf{Y} = \mathbf{U}.\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \qquad \text{Equação (2.3)}$$

As matrizes das componentes principais, **T** e **U**, estão em direções onde a variância dos dados originais são máximas. Entretanto, como essas direções não estão necessariamente correlacionadas, as componentes principais são ligeiramente rotacionadas de forma a maximar a covariância entre **T** e **U**. Após a rotação, uma relação linear é estabelecida entre os scores para cada i variável latente, da seguinte forma (BRERETON, 2000):

$$\mathbf{U} = \boldsymbol{B}\mathbf{T} \qquad \text{Equação (2.4)}$$

Os coeficientes de **B** são, então, calculados pela equação:

$$\mathbf{B} = \mathbf{U}^{\mathrm{T}}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1} \qquad \text{Equação (2.5)}$$

Com os valores de absorbância e os coeficientes calculados a partir da Equação 2.5, é possível estimar as propriedades de novas amostras $\widehat{\boldsymbol{Y}}$ com a Equação 2.6 (BRERETON, 2000):

$$\widehat{\mathbf{Y}} = \mathbf{B}\mathbf{T}\boldsymbol{Q}^{T} + \mathbf{F} \qquad \text{Equação (2.6)}$$

A ideia do PLS é encontrar uma relação linear ótima entre os scores **U** e **T**. Para que isso aconteça, as componentes principais das variáveis dependentes e independentes sofrem uma pequena rotação até que o ângulo entre elas seja zero, obtendo, com isso, uma correlação máxima entre as variáveis. Após rotacionadas, as componentes principais são chamadas de variáveis latentes (WOLD, 2001).

É um passo crítico determinar o número ótimo de variáveis latentes para compor o modelo. Deve-se selecionar um número mínimo necessário procurando obter a maior correlação possível, uma vez que muitas variáveis latentes podem introduzir ruído à calibração (WOLD, 2001).

### 2.5.3 Validação cruzada e parâmetros de desempenho

A validação cruzada tem o objetivo de determinar o número ótimo de variáveis latentes do modelo. É considerada uma validação interna, pois não utiliza amostras externas, ou seja, é conduzida apenas com amostras que foram utilizadas para a construção do modelo. Um método de validação cruzada muito utilizado, chamado *leave one out* (deixe um de fora), é conduzido da seguinte forma:

- A regressão é aplicada ao conjunto de calibração contendo todas as amostras exceto uma que é deixada de fora para ser utilizada como amostra de teste. Com apenas a primeira variável latente do modelo estima-se a variável da amostra que foi deixada de fora e calcula-se o erro de previsão. Em seguida, esta amostra volta para o conjunto de calibração e uma nova amostra é deixada de fora e todo este

processo é realizado novamente.

- Este processo é realizado até que todas as amostras, uma por vez, tenham sido deixadas de fora do conjunto de calibração para serem utilizadas como teste. Ao fim, calcula-se a média do erro de previsão de todas as amotras. Este valor, chamado de raiz do erro quadrático médio de validação cruzada ($RMSE_{CV}$ - do inglês, r*oot mean square error of cross validation*), refere-se ao erro de previsão quando se utiliza apenas a primeira variável latente do modelo.
- Para se determinar o $RMSE_{CV}$ quando se utiliza duas ou mais variáveis latentes, todo este processo é realizado da mesma forma, mas as amostras de teste são estimadas utilizando-se duas ou mais variáveis latentes ao invés de apenas uma. Com isso, é possível determinar o número ótimo de variáveis latentes que proporciona o menor valor de $RMSE_{CV}$.

Outro método de validação cruzada chamado *k-fold*, também muito utilizado, divide as amostras em k subconjuntos de mesmo tamanho. Um subconjunto é utilizado para teste e os demais para calibração e, a partir daí, calcula-se o $RMSE_{CV}$. Em seguida, outro subconjunto é utilizado para teste e todo este processo é realizado novamente até que todos os subconjuntos, um por vez, tenham sido deixados de fora do conjunto de calibração para serem utilizados como teste.

Os k subconjuntos podem ser formados de diferentes formas. Quando as amostras são separadas em blocos o método é denominado *contiguos block*, quando são separadas aleatoriamente é denominado *random block*, quando é separada de forma ordenada, mas não em blocos, é denominado *venetian blind*. A Figura 2.1 mostra um esquema que representa a seleção de amostras de teste para os métodos de validação cruzada citados. Cada retângulo representa uma amostra e as amostras de teste selecionadas estão representadas em amarelo.

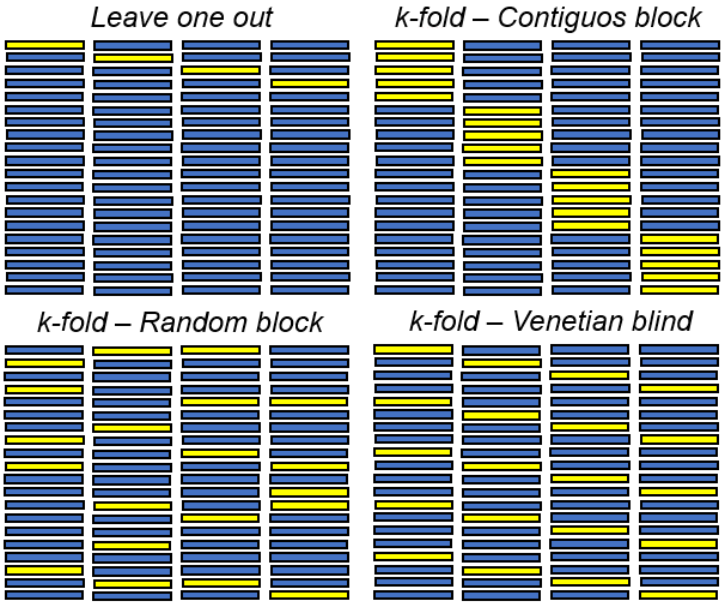



**Figura 2.1.** Representação da seleção de amostras de teste nos métodos de validação cruzada.

Após a validação cruzada e a escolha do número ótimo de variáveis latentes, o desempenho do modelo deve ser testado. Dois parâmetros de desempenho são amplamente utilizados para avaliar a adequação do modelo aos dados experimentais, são eles: raiz quadrada média do erro de calibração ($RMSE_C$) e de predição ($RMSE_P$). Quanto mais baixos forem seus valores, mais acurado será o modelo. Eles são uma boa medida da exatidão e permitem comparar a capacidade preditiva de diferentes modelos. $RMSE_P$ é especialmente importante, pois mostra a capacidade de prever amostras externas, que é o principal objetivo dos modelos quimiométricos. Quase todos os artigos das Tabelas 2.2 e 2.3 apresentaram o valor de $RMSE_P$.

Além das figuras de mérito mais comuns, $RMSE_C$, $RMSE_P$ e $R^2$, alguns estudos também apresentaram outras para validar seus modelos. PASQUINI e BUENO (2007) apresentaram a repetibilidade; ABBAS *et al*. (2012), MEJIA-MIRANDA *et al*. (2015), DUARTE *et al*. (2016), FILGUEIRAS *et al*. (2014) e MORO *et al*. (2019) apresentaram o erro relativo de predição; PABÓN e SOUZA FILHO (2019) e JINGYAN *et al*. (2012) a relação entre desempenho e desvio; DUARTE *et al*. (2016), FOLLI *et al*. (2020), MORO *et al*. (2019), RODRIGUES *et al*. (2017) e RAINHA *et al*. (2019) o teste de viés; DUARTE *et al*. (2016), MORO *et al*. (2019) e RODRIGUES *et al.* (2017) o teste de permutação. FILGUEIRAS *et al.* (2014) também apresentou sensibilidade, seletividade e limite de detecção. Todas essas figuras de mérito contribuem para

validar os modelos e aumentar a confiança em suas previsões.

### 2.5.4 Classificação por mínimos quadrados parciais

A classificação de amostras em grupos distintos, baseada em medidas multivariadas, também é uma das subáreas da quimiometria. Denomina-se classificação supervisionada os métodos de classificação que utilizam as informações das classes das amostras de treinamento na construção dos modelos. A análise discriminante por mínimos quadrados parciais (PLS-DA, do ingês *partial least squares discriminant analysis*) é um dos métodos de classificação superviosionada mais utilizados. Seu algoritmo trata-se de uma modificação do PLS, adaptado para resolução de problemas de classificação. O procedimento é o mesmo utilizado pelo PLS, no entanto, enquanto que no PLS as variáveis dependentes são quantitativas, no PLS-DA as variáveis são qualitativas, representando as diferentes classes (SUN, 2009; BARKER e RAYENS, 2003).

Existem duas abordagens para aplicação deste método: PLS1-DA e PLS2-DA. Na primeira, as amostras são classificadas como pertencentes ou não a uma única determinada classe, sendo atribuído a cada amostra o código 1 (pertencente à classe) ou 0 (não pertencente à classe). Caso existam três classes, por exemplo, devem ser construídos três modelos. Já no PLS2-DA, as amostras são classificadas como pertencentes a duas ou mais classes, sendo atribuído os códigos 1 para pertencente à classe 1, o código 2 para pertencente à classe 2, e assim sucessivamente, de acordo com o número existente de classes. Dessa forma, com o PLS2-DA, um único modelo é desenvolvido para as todas as classes (TRYGG e WOLD, 2002).

Assim como o PLS seleciona um número ideal de variáveis latentes que minimiza o valor do $RMSE_{CV}$, o PLS-DA também determina um número ideal de variáveis latentes que minimiza a taxa de erro de classificação, por meio de um método de validação cruzada. Da mesma forma, após a construção do modelo, é necessário um conjunto de amostras teste para determinação das figuras de mérito e avaliação da qualidade do modelo.

O modelo deve prever a classe a qual pertence uma determinada amostra. Entretanto, na prática, ele não fornece como resultado variáveis discretas, mas sim contínuas. Em uma classificação binária, por exemplo, os resultados não são

exatamente 0 ou 1, mas são, geralmente, próximos a estes valores. Dessa forma, para se atribuir uma classe a uma amostra a partir do resultado do modelo, deve ser estabelecido um valor limite, chamado de limiar ou *threshold*, entre as classes. Para resultados acima deste limite é atribuído uma classe e para resultados abaixo é atribuído outra classe. Este valor é comumente estimado pela teoria bayesiana (JAMES *et al*., 2013).

## 2.6 Fusão de dados

A fusão de dados é uma estratégia que combina dados provenientes de diferentes fontes de aquisição com o objetivo de fazer inferências mais exatas e precisas sobre um objeto de estudo. Algumas vezes, apenas uma fonte de aquisição de dados pode ser insuficiente para fornecer toda informação que se deseja sobre um objeto de estudo, dessa forma, os dados quando fundidos podem fornecer informações complementares (LAHAT *et al*., 2015).

O conceito fusão de dados surgiu no início da década de 70 para aplicações no campo militar. Utilizavam-se diversas fontes de informação em conjunto, para identificar alvos como mísseis, aeronaves e formações militares, e seu potencial ameaçador (KHALEGHI *et al*., 2013). Atualmente, a fusão de dados é considerada uma ferramenta multidisciplinar e vem sendo aplicada em diversas áreas como, por exemplo, robótica, medicina, monitoramento meteorológico, sensoriamento remoto, reconhecimento de padrões, química analítica, entre outras áreas (LAHAT *et al*., 2015).

Com o foco voltado para Química, a fusão de dados pode ser aplicada também em análises quimiométricas, como para classificação ou calibração multivariada de dados. A calibração multivariada tem por objetivo construir modelos matemáticos de regressão em função da resposta de uma determinada técnica analítica (MENDES *et al*., 2018). No entanto, medidas de uma única técnica, algumas vezes, podem não ser suficientes para correlacioná-las com certas propriedades de interesse e gerar modelos de qualidade. Os resultados obtidos por TAN *et al*. (2015) são um bom exemplo da ineficácia do uso de técnicas analíticas de forma isolada. A partir da espectroscopia UV-Vis ou autores obtiveram um espectro na região ultravioleta (240-400 nm) e outro na região do visível (380 a 700 nm). Com esses espectros separados

e também dados de fluorescência, contruíram três modelos cujo objetivo era classificar amostras de cervejas, de acordo com o fabricante. Os três modelos, criados a partir de cada espectro separado, não puderam classificar de forma satisfatória as amostras de cerveja. A acurácia da validação cruzada foi de 42,2-47,4%, 70,4% e 58,5% para fluorescência, ultravioleta e visível, respectivamente (TAN *et al*., 2015).

Nesses casos, é conveniente que seja feita a fusão de dados já que, na maioria das vezes, as diferentes técnicas analíticas podem ter informações complementares. Isso permite uma interpretação e extração de informações mais acuradas para um determinado conjunto de amostras (MENDES *et al*., 2018). Matrizes, com baixo poder preditivo individual, podem produzir modelos com boa exatidão, após uma fusão de dados adequada, porque a relação multivariada entre blocos concatenados é considerada. Vale salientar que as matrizes fundidas consistem em diferentes conjuntos de variáveis que foram medidas para o mesmo conjunto de amostras (SMOLINSKA *et al*., 2019).

Nos últimos anos, a fusão de dados tem sido utilizada na área de petróleo para estimar propriedades físico-químicas, como viscosidade, densidade, teor de enxofre (PEINDER *et al*., 2009), teor de nitrogênio e número de acidez total (MORO *et al*., 2019), entre outras propriedades. Em todos os casos, é aplicada a regressão linear PLS, entretanto a estratégia de fusão de dados também pode ser aplicada a metodologias não lineares, como o SVR. Associada à regressão não linear SVR, a fusão de dados é encontrada para outras finalidades como, por exemplo, estimar parâmetros bioquímicos da cultura de soja (MAIMAITIJIANG *et al*., 2017), calibrar sensores que medem a poluição do ar (FERRER-CID *et al*., 2020), prever a concentração de aminoácidos livres do chá amarelo (YANG *et al*., 2019), entre outras.

Tanto a fusão de dados quanto o SVR trazem vantagens intrínsecas a sua aplicação. Enquanto a primeira consegue reunir informações que se complementam e atuam sinergicamente, o segundo é capaz de modelar sistemas com comportamento não linear, trazendo resultados melhores que métodos lineares, como o PLS. Tanto o SVR quanto a fusão de dados vêm ganhando importância na área de petróleo, produzindo resultados satisfatórios (MORO *et al*., 2019; FOLLI *et al*., 2020), entretanto, a combinação dos dois métodos carece de estudos nessa área. Portanto, é particularmente promissor estudar a aplicação do SVR para obter alta exatidão no processo de fusão de dados de matrizes oleosas.

Em 1999, STEIMENTZ *et al*. (1999), propuseram uma metodologia que foi seguida pela maioria dos autores que aplicaram fusão de dados para avaliação da qualidade e autenticação de alimentos e bebidas. De acordo com a metodologia, os dados podem ser fundidos a três níveis diferentes: baixo, médio e alto (Figura 2.2). A simples concatenação dos dados brutos ou pré-processados antes de qualquer modelagem, como citado anteriormente no trabalho de DOWNEY *et al*. (1997), caracteriza a fusão de nível baixo. Quando os dados são combinados após uma redução de sua dimensão ou extração de recursos, caracteriza-se a fusão de nível médio. Por fim, a fusão de nível alto é realizada combinando-se as respostas dos modelos advindos dos dados individuais.

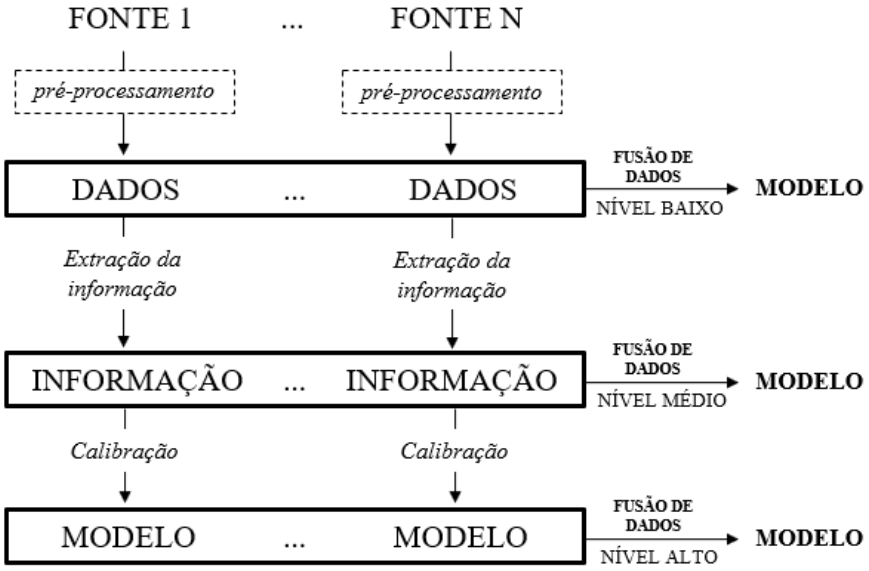



**Figura 2.2.** Representação da fusão de dados a nível baixo, médio e alto. Adaptado de BORRÀS *et al*., 2015.

### 2.6.1 Fusão de nível baixo

A fusão de nível baixo, também conhecida como *measurement level*, é a forma mais simples de concatenação das informações. Pode ser feita com duas ou mais matrizes. A Figura 2.3 apresenta um exemplo de fusão a nível baixo com espectros de MIR e de RMN $^{1}$H.

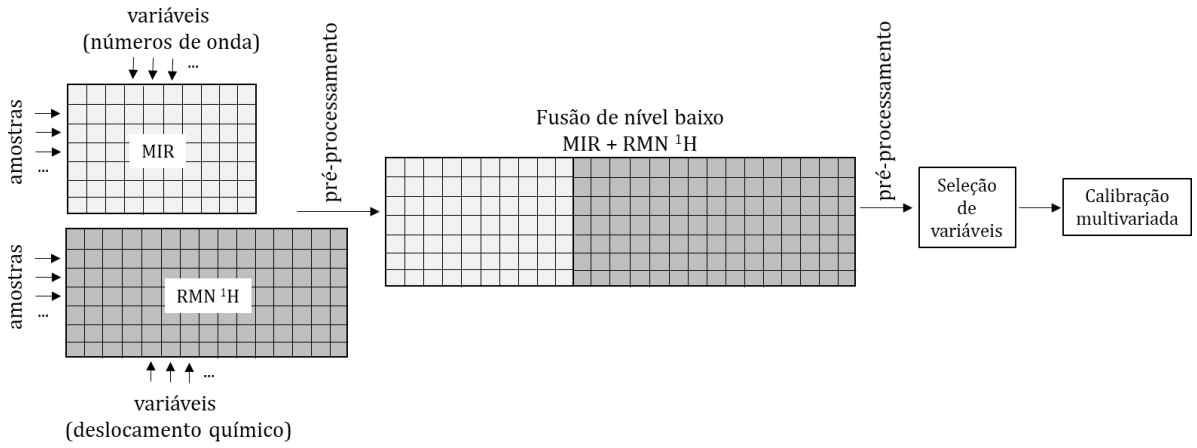



**Figura 2.3.** Representação da fusão de dados a nível baixo com dados de MIR e RMN $^{1}$H

O pré-processamento das matrizes pode ser realizado antes e/ou após a fusão. Também podem ser aplicados métodos de seleção de variáveis após a fusão e antes da calibração multivariada. A aplicação da seleção de variáveis antes da fusão não configura a estratégia de nível baixo, uma vez que fusão não é mais dos dados brutos. A matriz fundida pode ser utilizada para diversas finalidades como classificação de amostras, predição de variáveis dependentes e análise exploratória (SMOLINSKA *et al*., 2019).

Essa estratégia, geralmente, só é vantajosa quando os dados fundidos são semelhantes em variância, caso contrário pode haver predominância de uma fonte de dados sobre a outra. Por exemplo, unindo-se uma matriz A a uma matriz B, cuja variância é muito inferior a A, é provável que o modelo consiga capturar bem a variância de A e despreze, em parte, a variância de B, impedindo que ela contribua com a formação do modelo. Com isso, a variação conjunta entre as matrizes não será capturada pelo modelo e a fusão de dados perderá a sua função. Uma forma de contornar esse problema é aplicando o pré-processamento escalonamento, após a fusão das matrizes. O resultado serão todas as variáveis apresentando variância igual a um, de forma que todas contribuam com equidade para formação do modelo (SMOLINSKA *et al*., 2019).

O problema de predominância de uma fonte de dados sobre a outra pode ocorrer também quando as dimensões das duas matrizes são muito diferentes, ou seja, quando o número de variáveis de uma matriz é muito maior que da outra. Além disso, a fusão de nível baixo requer uma alta capacidade de processamento computacional,

devido ao elevado volume de informação. Uma forma de contornar esses dois últimos problemas citados consiste em aplicar a fusão a nível médio que reduz a dimensão das matrizes antes da concatenação (BORRÀS *et al*., 2015).

### 2.6.2 Fusão de nível médio

Na fusão de nível médio, também conhecida como "*feature level*" ou "*characteristic level*", primeiramente são extraídas informações relevantes, como variáveis selecionadas, componentes principais ou variáveis latentes, de cada fonte de dados, separadamente, por meio de métodos como a seleção de variáveis, análise de componentes principais (PCA) ou regressão por mínimos quadrados parciais (PLS). Essa etapa é chamada de extração de recursos. As matrizes, agora com menor dimensão, são fundidas para posterior calibração multivariada. Por exemplo, o PLS é aplicado em cada uma das matrizes separadamente e, na sequencia, as variáveis latentes dos dois modelos são concatenadas em uma única matriz e um novo modelo PLS é construído para previsão das propriedades. Não existe um número pré-determinado de variáveis latentes a serem selecionadas em cada um dos modelos para fusão. Geralmente, escolhe-se a quantidade de variáveis latentes que deverá fornecer o modelo mais exato após a fusão, ou seja, a otimização é feita levando em conta o modelo final. Também não é necessário que sejam selecionadas as mesmas quantidades de variáveis latentes em ambos os modelos (SMOLINSKA *et al*., 2019).

Outra possibilidade, por exemplo, é aplicar um método para selecionar, em ambas as matrizes, as variáveis que estejam mais correlacionadas com a variável que se deseja inferir. Após a seleção de variáveis e consequente redução da dimensão, as duas matrizes são fundidas para posterior calibração multivariada (SMOLINSKA *et al*., 2019).

Além de proporcionar uma redução significativa do número de variáveis (BORRÀS *et al*., 2015), a fusão de nível médio tem maior capacidade de eliminar ruídos e redundância na informação, quando comparada com a fusão de nível baixo (BEVILACQUA *et al*., 2017). A Figura 2.4 apresenta um exemplo de fusão a nível médio com espectros de MIR e de RMN $^{1}$H.

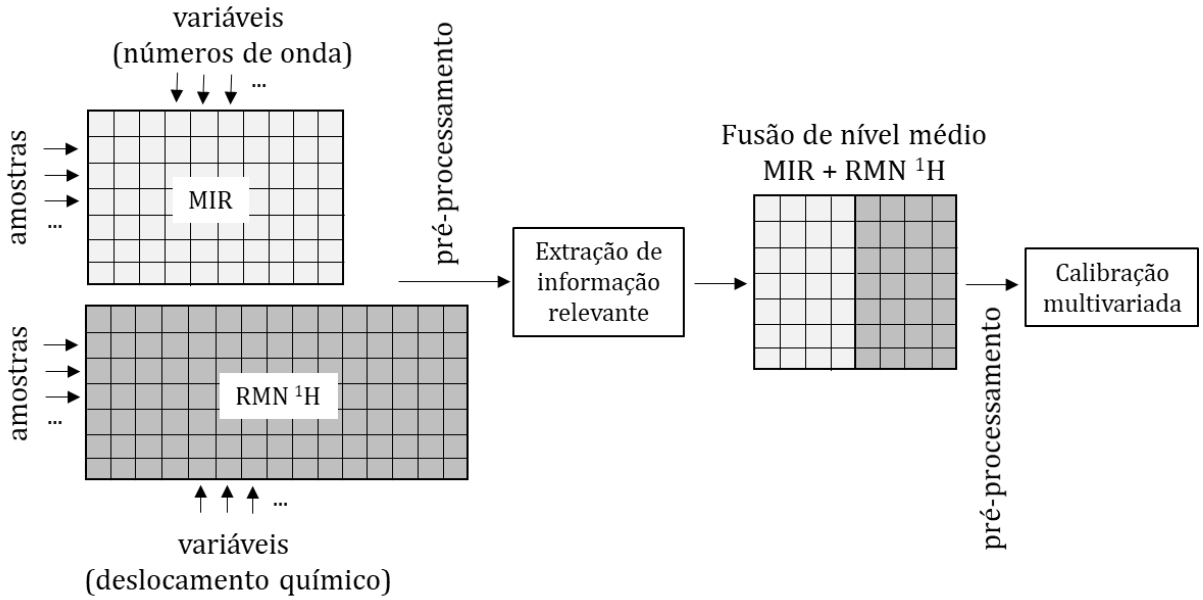



**Figura 2.4.** Representação da fusão de dados a nível médio com dados de MIR e RMN $^{1}$H

Geralmente, após a extração de recursos e consequente de redução da dimensão, as matrizes apresentam um número de variáveis muito semelhante, o que elimina o problema da predominância de uma fonte de dados sobre a outra em decorrência da discrepância entre as dimensões das matrizes (SMOLINSKA *et al*., 2019). No entanto, como muitas combinações de métodos de extração e pré-processamento são possíveis, a otimização e validação dos modelos se tornam trabalhosas em relação à fusão de nível baixo (BORRÀS *et al*., 2015). Outro problema associado, é o risco de eliminação de variáveis, durante a extração de recursos, que poderiam contribuir para a modelagem quando concatenadas com as demais fontes de dados (SMOLINSKA *et al*., 2019). Entretanto, uma estratégia de "reciclagem", proposta por GEURTS *et al*. (2016), pode ser implementada na tentativa de reavaliar se as variáveis que foram eliminadas contêm informação relevante, quando associadas às demais fontes de dados.

Assim como na fusão de nível baixo, pode haver predominância de uma fonte de dados sobre a outra caso, após a extração de recursos, as matrizes a serem concatenadas tenham variâncias muito diferentes. Para impedir que isso aconteça, pode-se aplicar o pré-processamento escalonamento, após a fusão das matrizes, para que todas as variáveis apresentem variância igual a um e contribuam com equidade para formação do modelo final.

Outra observação importante consiste na interpretação do modelo. É importante analisar quais das variáveis originais, ou seja, quais regiões dos espectros foram mais importantes para previsão das propriedades. Como na fusão de nível médio o modelo final é construído a partir dos recursos extraídos, e não das variáveis originais, fica inviável a análise das variáveis mais importantes a partir do modelo final. Entretanto, os recursos são extraídos a partir das variáveis originais, dessa forma, é possível verificar neste procedimento quais delas apresentam maiores "*loadings*" e, portanto, são mais importantes.

### 2.6.3 Fusão de nível alto

No procedimento de fusão a nível alto, também conhecida como "*decision fusion*", os modelos são construídos para cada conjunto de dados analíticos, separadamente. Na sequência, os resultados dos modelos individuais, que podem ser de classificação ou de previsão, são fundidos para a obtenção de um resultado final de previsão. Diferente dos casos anteriores, o que é fundido são as previsões dos modelos individuais e não os dados brutos ou os modelos em si. Em decorrência disso, podem ser utilizados diferentes métodos de calibração para os diferentes conjuntos de dados (BALLABIO *et al*., 2019). Cada modelo individual é construído de forma independente, dessa forma, podem ser aplicados, em um ou em ambos os modelos, seleção de variáveis, regressão linear e/ou não linear, entre outras alternativas. A Figura 2.5 apresenta um exemplo de fusão a nível alto com espectros na região do infravermelho e espectros de RMN $^{1}$H.

Uma das alternativas para a combinação das respostas consiste simplesmente em obter uma média das previsões dos modelos individuais para produzir uma previsão final. Desta forma, cada modelo individual terá igual importância na média final. Outra alternativa consiste em atribuir pesos diferentes às previsões individuais, dando mais importância para uma fonte de dados em relação às outras na previsão final. O critério para escolha das alternativas e/ou dos pesos atribuídos deve ser a otimização da exatidão da previsão final (BALLABIO *et al*., 2019).

Quando a diferença entre os resultados dos modelos individuais é muito grande, é mais adequado atribuir um peso maior à resposta cuja exatidão é melhor. Entretanto, é essencial que a fusão das respostas forneça um resultado final mais acurado que

as respostas individuais. Sobretudo para a fusão de nível alto, é necessário processar os dados com acurácia, para que nenhuma informação seja perdida (BORRÀS et al., 2015). Segundo FLOUDAS *et al*. (2007), na medida que os dados são processados de forma individual, informações que poderiam atuar de forma sinérgica entre os blocos são perdidas até o momento final em que ocorre a fusão. Já os problemas associados à fusão de nível baixo, como predominância de uma fonte sobre outra e alta demanda de processamento, são contornados aqui, visto que os dados brutos são modelados para posterior fusão de suas previsões (FLOUDAS *et al*., 2007).

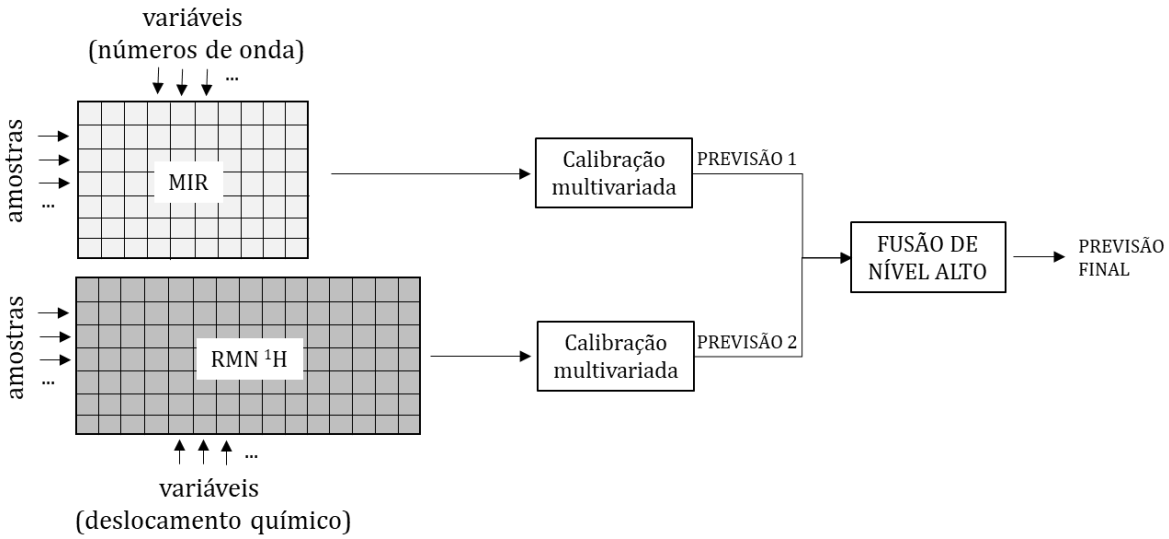



**Figura 2.5.** Representação da fusão de dados a nível alto com dados de MIR e RMN $^{1}$H

A fusão de nível alto tem a proposta muito semelhante ao método chamado modelo de análise populacional (MPA, do inglês *Model population analysis*), originalmente prosposto por Li *et al*. (2010). O MPA é conduzido da seguinte forma: por meio de uma amostragem aleatória, são construídos subconjuntos de variáveis e, a partir deles, sobmodelos. Os submodelos são analisados com base em suas figuras de mérito e o resultado final de predição é uma média entre os resultados dos modelos mais acurados. Assim como a fusão de nível alto, o MPA chega a um resultado final a partir de uma combinação entre os resultados de modelos individuais. A diferença reside no fato de que para a fusão, os modelos individuais são construídos a partir de dados de diferentes fontes analíticas, enquanto que para o MPA, os modelos individuais são construídos a partir de subconjuntos de dados de apenas uma fonte.

# CAPÍTULO 3 OBJETIVOS

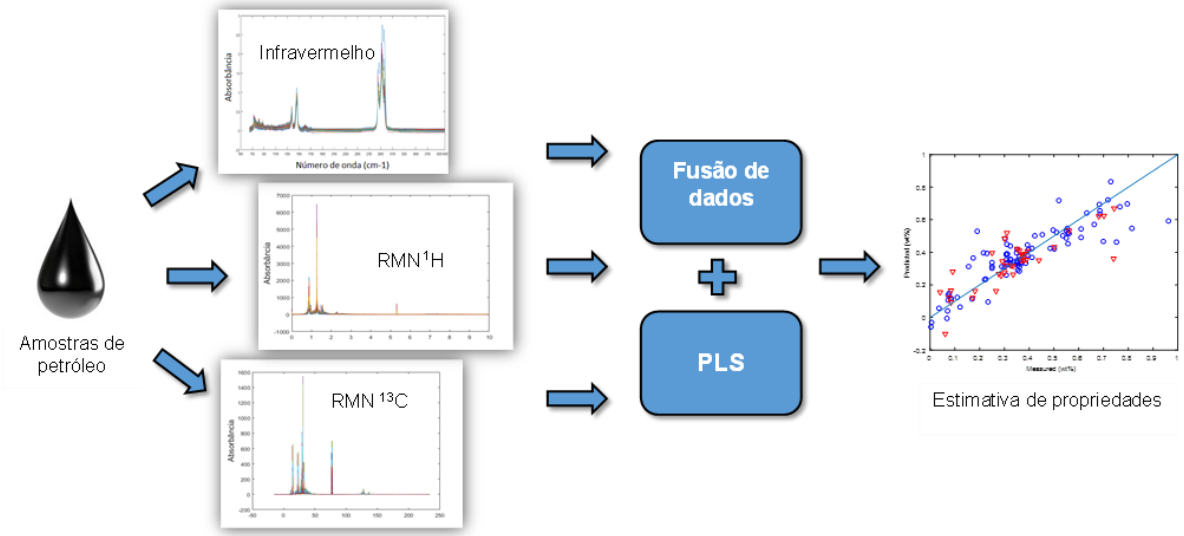
## 3.1 Objetivo geral

Desenvolver metodologias de fusão de dados, associadas a análises quimiométricas, para determinar propriedades físico-químicas e classificar amostras de petróleo a partir de dados de IR e RMN, de modo a investigar a influência da fusão de dados, em seus diferentes níveis, na melhoria de desempenho de modelos quimiométricos.

## 3.2 Objetivos específicos

- Adquirir espetros na região do infravermelho médio e próximo por transformada de Fourier com reflectância total atenuada de amostras de petróleo;
- Adquirir espectros de ressonância magnética nuclear de $^{1}H$ e $^{13}C$ de amostras de petróleo;
- Determinar por métodos experimentais padronizados as propriedades físico-químicas do petróleo: °API, teor de nitrogênio total, nitrogênio básico, enxofre total, número de acidez total e frações de saturados, aromáticos e polares;
- Pré-processar e fundir, a nível baixo, médio e alto, os dados dos espectros de NIR, MIR, RMN $^{1}H$ e $^{13}C$;
- Estimar o teor de nitrogênio total, nitrogênio básico e enxofre, o número de acidez total e as frações de saturados, aromáticos e polares das amostras de petróleo, por meio de modelos PLS a partir dos espectros de MIR, RMN $^{1}H$ e $^{13}C$ isolados e fundidos;
- Construir modelos de classificação para discriminar amostras de petróleo bruto entre as classes doce/azedo, pobre/rico em nitrogênio; não ácidos/ácidos/altamente ácidos e leves/médios/pesados, utilizando o algoritmo PLS-DA e os espectros de NIR e MIR.
- Comparar a capacidade de previsão de modelos regressão e classificação obtidos com e sem fusão de dados.

# CAPÍTULO 4 FUSÃO DE DADOS DE MIR, RMN $^{1}H$ E $^{13}C$ PARA PREDIÇÃO DE PROPRIEDADES DO PETRÓLEO BRUTO




- Estratégias de fusão de dados foram aplicadas para prever sete propriedades físico-químicas do petróleo.

- Modelos de regressão PLS foram construídos a partir das fusões de nível baixo, médio e alto dos espectros de MIR, RMN $^{1}H$ e $^{13}C$.

- As fusões de nível baixo, médio por PCA e alto não proporcionaram modelos com maior exatidão, em comparação com os modelos individuais.

- A fusão de nível médio por PLS forneceu modelos com menores valores de $RMSE_P$ para todas as propriedades

## 4.1 Introdução

O conhecimento das propriedades físico-químicas do petróleo bruto após a sua exploração é fundamental para determinação de seu valor econômico e para o ajuste das condições operacionais durante seu processamento. Os experimentos clássicos de determinação dessas propriedades podem ser substituídos por técnicas como a espectroscopia na região do infravermelho médio, com reflectância total atenuada e a espectroscopia de ressonância magnética nuclear que têm se mostrado alternativas eficazes, práticas e não destrutivas para a caracterização do petróleo (VEMPATAPU e KANAUJIA, 2017, KHANMOHAMMADI *et al*., 2012). Isso só é possível por meio do uso da quimiometria que permite que as propriedades físico-químicas do petróleo sejam estimadas a partir de modelos matemáticos que fazem uma correlação entre os dados espectrais e as propriedades (BRERETON, 2000).

Para construir uma melhor correlação entre os espectros e as variáveis de interesse, é possível a união de dados de duas ou mais fontes analíticas cujas informações sejam complementares entre si, ao ponto de produzirem modelos matemáticos mais satisfatórios. Essa abordagem, chamada de fusão de dados, também conhecida como fusão de sensores ou fusão de informações, é uma tecnologia que permite explorar todas as informações contidas nos dados e a sinergia entre elas, tanto para análises qualitativas quanto quantitativas (CASTANEDO *et al*., 2013).

A fusão de dados tem sido amplamente estudada na área de alimentos, embora seja muito pouco explorada na área de petróleo. Diante desse contexto, este estudo teve como objetivo construir metodologias analíticas para determinação de algumas propriedades físico-químicas de petróleo bruto, utilizando análises quimiométricas associadas aos métodos de fusão de dados. Os dados fornecidos por três técnicas espectroscópicas diferentes (MIR, RMN $^{1}H$ e $^{13}C$) foram modelados individualmente. Para utilizar a sinergia e fornecer informações complementares, os dados também foram modelados após fundidos nos níveis baixo e médio, em várias combinações diferentes.

## 4.2 Metodologia

### 4.2.1 Amostras e ensaios físico-químicos

Um total de 125 amostras de petróleo bruto provenientes de diferentes campos de petróleo da costa brasileira foram fornecidas pelo Cenpes/Petrobras. As amostras foram armazenadas em frascos âmbar para evitar a degradação de compostos fotossensíveis, com batoque para não perder compostos voláteis, e em temperatura adequada para análises posteriores. Os espectros MIR e NIR de todas as amostras foram adquiridos e fornecidos pelo LabPetro/UFES. As análises físico-quimicas foram realizadas pelo Cenpes/Petrobras.

As amotras apresentaram API° variando de 11.4 a 54.0, compreendendo exemplos de óleos leves a pesados, de acordo com a classificação da *American Petroleum Institute* (API, 2021). Foram determinadas um total de 7 propriedades físico-químicas, seguindo os procedimentos padronizados de análise da *American Standard Testing Material* (ASTM) e da *Universal Oil Product* (UOP). As propriedades quantificadas foram enxofre total, nitrogênio total, nitrogênio básico, número de acidez total e as frações dos constituintes saturados, aromáticos e polares. Para uma amostragem adequada, antes das análises, as amostras foram aquecidas a 45º por 15 min para homogeneização e dissolução de compostos cristalizados.

A determinação do teor total de enxofre (S) foi feita seguindo a norma ASTM D4294-16 (2016). A espectrometria de fluorescência de raios X por energia dispersiva foi aplicada e a excitação característica de cada amostra foi comparada com amostras padrão de enxofre.

O teor total de nitrogênio (NT) foi determinado pelo método padrão ASTM D4629-17 (2017). O ensaio consiste na combustão das amostras, seguida da determinação do nitrogênio por quimioluminescência, utilizando uma curva de calibração padrão.

O teor de nitrogênio orgânico básico (NB) foi determinado pelo método UOP – 269 (2010), por titulação potenciométrica, utilizando um eletrodo de pH e ácidos perclórico e acético como titulante e solvente, respectivamente.

Para determinar o número de acidez total (NAT) das amostras de óleo, foi realizada uma titulação potenciométrica utilizando o eletrodo seletivo *Solvotrode easy*

*clean*, conforme a norma ASTM D664 (2004). Os resultados foram expressos em miligramas de hidróxido de potássio necessários para neutralizar um grama de amostra.

Os teores de saturados (SAT), aromáticos (ARO) e polares (POL) foram determinados pela norma ASTM D2549-02 (2017) modificada, variando o tamanho da partícula de sílica e a quantidade de solventes. Os teores de resina e asfalteno foram determinados como uma propriedade única, denominada polar. Uma coluna cromatográfica de 25 cm com eluições sucessivas de 200 mL de hexano, 200 mL de hexano: diclorometano (1:1) e 200 mL de metanol separaram, respectivamente, as frações saturada, aromática e polar de 0,2 g de petróleo acomodado no topo da coluna. Um total de 20 g de sílica gel, com partículas variando de 230 a 400 mesh, ativada a 120 °C durante 12 horas, foi utilizada como fase estacionária.

### 4.2.2 Ensaios espectrais

Espectros na região do infravermelho médio por transformada de Fourier de todas as amostras foram obtidos em um espectrômetro PerkinElmer, equipado com um acessório de reflexão total atenuada, com cristal de seleneto de zinco. Os espectros foram medidos na região entre 4000 e 650 $cm^{-1}$ com uma resolução de 4 $cm^{-1}$ e 32 leituras por amostra.

**Tabela 4.1.** Condições experimentais de obtenção dos espectros de RMN de $^{1}H$ e $^{13}C$.

| | $^{1}H$ | $^{13}C$ |
|---|---|---|
| Fequência (MHz) | 399,73 | 100,51 |
| Largura espectral (Hz) | 6410,3 | 25510,2 |
| Tempo de aquisição (s) | 2,556 | 1,285 |
| Atraso de relaxação (s) | 1 | 7 |
| Ângulo de pulso (°) | 90 | 90 |
| Número de transientes | 64 | 1000 |

Também foram obtidos espectros de ressonâcia magnética nuclear de hidrogênio e carbono em um espectro da marca Varian/Agilent, modelo VNMRS400,

operando a 9,4 T e utilizando uma sonda Broadband 1H/X/D de 5 mm. As condições instrumentais para os espectros de hidrogênio e carbono estão dispostas na Tabela 4.1. O solvente utilizado na dissolução das amostras foi o clorofórmio deuterado, $CDCl_3$, com acetilacetonato de cromo III, $Cr(acac)_3$, 0,05 $mol.L^{-1}$.

### 4.2.3 Calibração multivariada

Todo o tratamento dos dados, a construção e a validação dos modelos, bem como qualquer outra análise quimiométrica empregada neste trabalho, foram realizados no *software* MATLAB®, versão 7.1 (The Mathworks, Natik, USA).

Os procedimentos descritos nesta seção foram empregados para a construção de modelos de calibração para cada uma das sete propriedades avaliadas: teor de enxofre, nitrogênio total, nitrogênio básico, número de acidez total e teores das frações de saturados, aromáticos e polares. Para cada propriedade, modelos de predição PLS foram construídos a partir dos conjuntos individuais de MIR, RMN $^{1}H$ e RMN $^{13}C$ e também a partir das quatro seguintes combinações de fusão de espectros: MIR + RMN $^{1}H$, MIR + RMN $^{13}C$, RMN $^{1}H$ + RMN $^{13}C$ e MIR + RMN $^{1}H$ + RMN $^{13}C$, fundidos a nível baixo, médio e alto.

Primeiramente, os espectros de MIR, RMN $^{1}H$ e RMN $^{13}C$ foram, cada um, organizados em uma matriz em que cada linha corresponde a uma amostra e cada coluna corresponde a um número de onda da varredura espectral ou deslocamento químico. Os dados de cada uma das sete propriedades estimadas foram organizados em vetores, de forma que cada linha corresponde a uma amostra.

Usando o algoritmo Rank-KS (LIU *et al*., 2014), o conjunto de amostras foi dividido em duas frações, de forma que 70% foram destinadas para calibração e 30% para previsão. As amostras do conjunto de calibração foram usadas para construir os modelos que mais tarde foram aplicados às amostras do conjunto de previsão. O Rank-KS foi escolhido, pois consegue separar as amostras em conjuntos com boa representatividade, uma vez que a separação é baseada não somente no espaço amostral das variáveis independentes (espectros) mas também da variável dependente (propriedade). O método utiliza com critério de seleção a maior distância entre uma amostra já selecionada e as demais amostras, buscando uma distribuição uniforme durante o processo.

Após a separação em amostras de calibração e previsão, as etapas subsequentes se diferenciam na construção dos modelos individuais e fundidos, portanto, estão descritas a seguir de acordo com cada objetivo.

#### *4.2.3.1 Construção dos modelos individuais*

Para construção dos modelos individuais, após a separação em amostras de calibração e teste, os espectros foram pré-processados utilizando três técnicas diferentes: MSC, SNV e primeira derivada. O pré-processamento que forneceu o melhor resultado foi selecionado para os testes posteriores. Além disso, os espectros de RMN foram alinhados usando o algoritmo *Icoshift* (Interval Correlation Optimized Shifting) (TOMASI *et al*., 2012).

A calibração multivariada foi feita utilizando a Regressão por Mínimos Quadrados Parciais, mais precisamente, com algoritmo NIPALS (*Non-Linear Interative Partial Least Squares*), desenvolvido por Herman Wold (WOLD, 1975), aplicando concomitantemente o autoescalolamento das matrizes. O algoritmo NIPALS obtem as componentes das matrizes dos espectros e das propriedades de forma iterativa. A cada iteração é produzido um modelo adicionando-se um componente principal. Os primeiros componentes são calculados a partir das matrizes originais e os demais a partir de seus resíduos. As iterações terminam quando se alcança o número máximo de componentes ou quando a matriz de resíduos é igual a zero.

Para previsão de cada uma das sete propriedades, S, NB, NT, NAT, SAT, ARO e POL, foram construídos modelos a partir de cada um dos três espectros, pré-processados com os diferentes métodos mencionados anteriormente.

O número ótimo de variáveis latentes foi determinado por validação cruzada, pelo método Venetian Blind, k-fold, com k igual a 5. Para reduzir as chances de superajuste, o número de variáveis latentes para os modelos individuais foi limitado a um máximo de 10. Posteriormente, foram determinadas as principais figuras de mérito para avaliar a qualidade dos modelos. A sequência das etapas descritas está ilustrada na Figura 4.1.

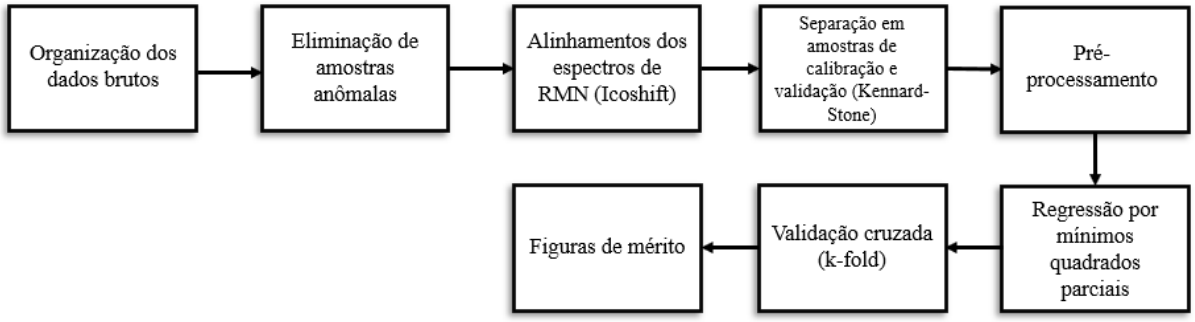



**Figura 4.1.** Etapas da calibração multivariada sem fusão de dados.

### *4.2.3.2 Construção dos modelos fundidos*

Para fusão de dados a nível baixo, médio e alto, as quatro seguintes combinações de espectros foram testadas: MIR e RMN $^{1}$H, MIR e RMN $^{13}$C, RMN $^{1}$H e RMN $^{13}$C, e MIR, RMN $^{1}$H e RMN $^{13}$C.

Após a separação em amostras de calibração e teste, o espectro de MIR foi pré-processado com MSC e os espectros de RMN com SNV. As escolhas de pré-processamentos foram baseadas em resultados preliminares.

Na fusão de baixo nível, os dados pré-processados dos espectros individuais foram concatenados em uma matriz única. A nova matriz foi novamente pré-processada com SNV e, em seguida, o PLS e a validação cruzada foram executados sequencialmente. Para reduzir as chances de superajuste, o número de variáveis latentes foi limitado a um máximo de 10. A sequência de etapas está ilustrada na Figura 4.2.

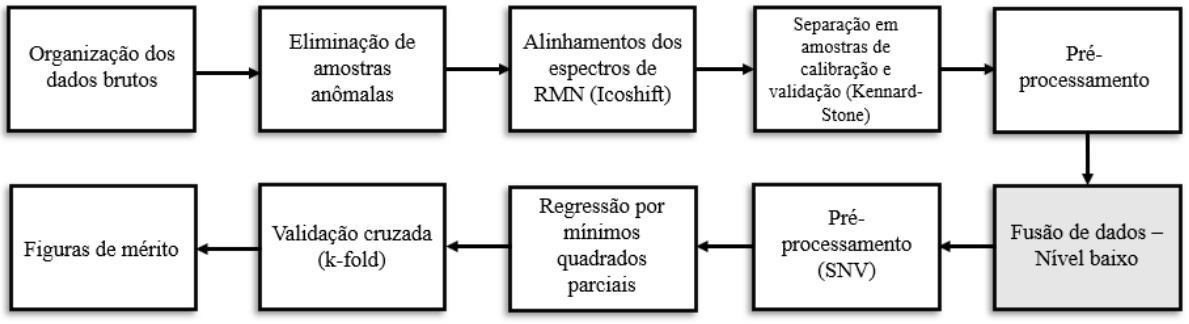



**Figura 4.2.** Etapas da calibração multivariada com fusão de dados a nível baixo.

Na fusão de nível médio, após a etapa de pré-processamento, foi aplicada a

Análise de Componentes Principais (PCA) nas matrizes separadas. Em seguida, os 10 primeiros componentes principais de cada conjunto de dados foram fundidos em uma única matriz e procedeu-se com a calibração multivariada, como mostra a Figura 4.3. Para reduzir as chances de superajuste, o número de variáveis latentes foi limitado a um máximo de 10. Essa abordagem foi chamada de fusão de dados de nível médio PCA.

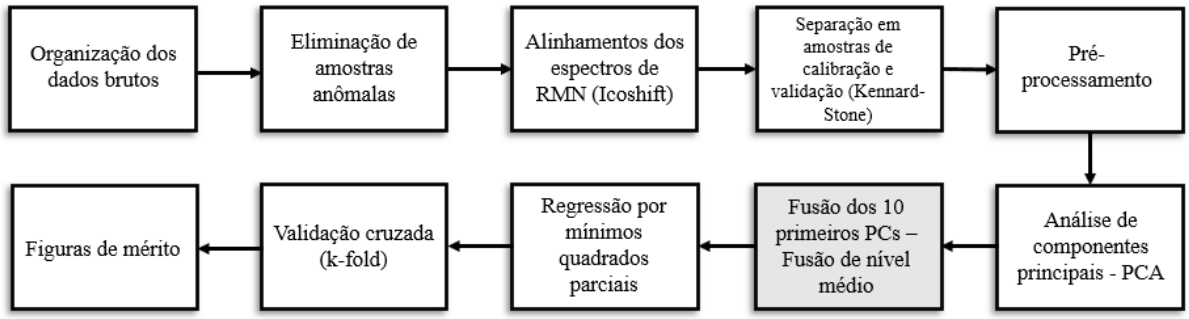



**Figura 4.3.** Etapas da calibração multivariada com fusão de dados a nível médio PCA.

Em uma outra estratégia de fusão de dados de nível médio, o PLS foi aplicado separadamente nos três espectros individuais pré-processados. Depois disso, uma determinada quantidade deescores dos modelos PLS foram fundidos, formando uma única matriz. O número de scores foi escolhido de forma a otimizar os modelos finais. Um novo PLS e validação cruzada foram executados sequencialmente nesta nova matriz. Para reduzir as chances de superajuste, o número de variáveis latentes do modelo final foi limitado a 10. Essa abordagem foi chamada de fusão de dados de nível médio PLS. A sequência de etapas está ilustrada na Figura 4.4.

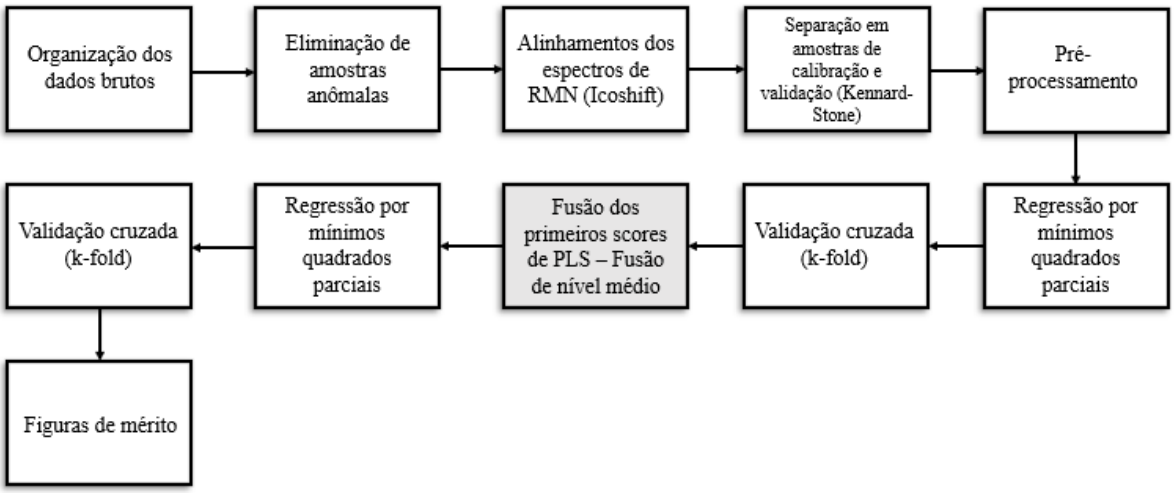



**Figura 4.4.** Etapas da calibração multivariada com fusão de dados a nível médio PLS.

Na fusão de nível alto, após a etapa de pré-processamento, o PLS e a validação cruzada foram aplicados sequencialmente nas matrizes separadas. Em seguida, os valores previstos pelos modelos individuais foram combinados para gerar uma única resposta de previsão para as amostras de calibração. A sequencia de etapas está ilustrada na Figura 4.5.

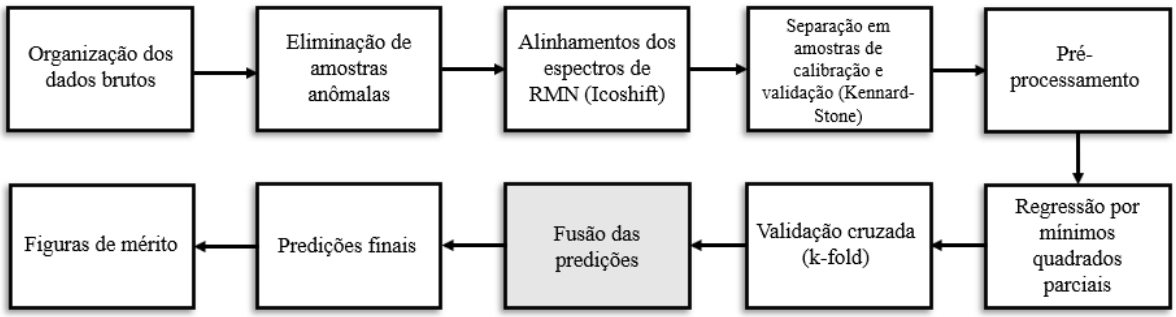



**Figura 4.5.** Etapas da calibração multivariada com fusão de dados a nível alto.

A Figura 4.6 mostra um esquema da metodologia de fusão de dados utilizada para construção dos modelos fundidos a nível baixo, médio e alto.

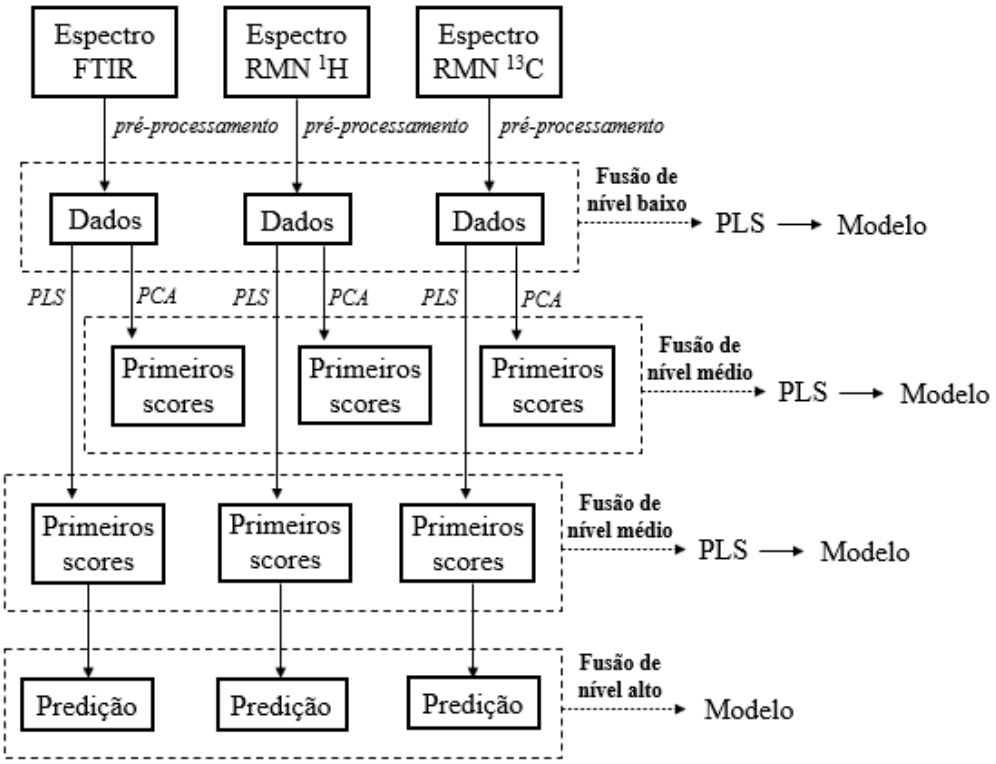



**Figura 4.6.** Esquema de metodologia de fusão de dados a níveis baixo, médio e alto.

### 4.2.4 Figuras de mérito e análise de variáveis

Com o objetivo de avaliar a qualidade dos modelos construídos, foram determinadas as principais figuras de mérito, sendo elas: coeficiente de determinação ($R^2$), exatidão e erro médio percentual. O coeficiente de determinação, calculado pela Equação 4.1, é uma medida do ajuste linear entre os dados experimentais e os previstos pelo modelo. Seu valor varia de zero a um e quanto mais próximo de um, mais explicativo é o modelo e, portanto, melhor ele se ajusta aos dados reais.

$$R^2 = 1 - \frac{\sum_i (y_{ref,i} - y_{prev,i})^2}{\sum_i (y_{ref,i} - \bar{y})^2} \qquad \text{Equação (4.1)}$$

Em que $y_{ref}$ é o valor experimental da propriedade, $y_{prev}$ é o valor previsto pelo modelo e $\bar{y}$ é o valor médio da propriedade. O coeficiente de determinação pode ser calculado utilizando o conjunto de amostras de calibração ($R_C^2$) ou de previsão ($R_P^2$).

A exatidão foi calculada por meio do parâmetro RMSE (*Root Mean Squares*

*Error*), que calcula uma média da diferença entre o valor real e o predito (Equação 4.2).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{ref,i} - y_{prev,i}\right)^2}{n}} \qquad \text{Equação (4.2)}$$

Em que $y_{ref,i}$ e $y_{prev,i}$ são, respectivamente, o valor experimental e o valor predito pelo modelo e n é o número de amostras. RMSE pode ser calculado utilizando o conjunto de amostras de calibração ($RMSE_C$) ou de previsão ($RMSE_P$). Para uma melhor interpretação do erro, foi calculado também o erro médio percentual que mede a porcentagem do erro em relação à média da propriedade, de acordo com a Equação 4.3 (FILGUEIRAS *et al*., 2014a).

$$\%\mathrm{RMSE_P} = \frac{\mathrm{RMSE_P}}{\bar{y}_{predito}} \mathrm{x}100 \qquad \text{Equação (4.3)}$$

Em que $\bar{y}_{predito}$ é o valor médio da propriedade predito pelo modelo. O valor de $RMSE_P$ foi utilizado como principal parâmetro para comparação entre os modelos, de forma que quanto menor o seu valor, menor o erro de previsão das amostras externas e, consequentemente, melhor a capacidade preditiva do modelo. A comparação dos valores de $RMSE_P$ entre os modelos foi realizada aplicando o teste randômico (VOET, 1994), unicaudal, a um nível de significância de 5%. O objetivo do teste é verificar se há diferença significativa entre os valores de $RMSE_P$ de dois modelos hipotéticos A e B. Para isso, são estabelecidas a seguintes hipóteses:

Hipótese nula $H_0$: $\mathrm{RMSEP_A^2} = \mathrm{RMSEP_B^2}$
Hipótese alternativa $H_a$: $\mathrm{RMSEP_A^2} > \mathrm{RMSEP_B^2}$

Deve ser calculado a diferença real entre as exatidões da seguinte forma:

$$\mathrm{d_{real}} = \mathrm{RMSEP_A^2} - \mathrm{RMSEP_B^2} \qquad \text{Equação (4.4)}$$

$$\mathrm{d_{real}} = \frac{\sum_{i=1}^{n}\left(e_{i,A}^2 - e_{i,B}^2\right)}{n} \qquad \text{Equação (4.5)}$$

Em seguida, o vetor $\mathbf{e}_{\mathbf{A}}^{2}$ é permutado randomicamente e a diferença $d_{real}$ é calculada novamente. Este último passo é repetido diversas vezes para se construir um gráfico de frequências de $d_{real}$. Caso $d_{real}$ seja estatisticamente significativo ao nível de significância de 0,05, a hipótese nula é rejeitada e constata-se uma diferença entre as duas exatidões.

Também foi verificado a presença de erros sistemáticos e tendências nos resíduos deixados pelos modelos, utilizando, respectivamente, o teste de viés, ou teste de *bias*, e o teste de permutação não paramétrico, descrito por FILGUEIRAS *et al*. (2014b), ambos com um intevalo de confiança de 95%.

O teste de viés verifica se há presença de erros sistemáticos nos resíduos. Caso haja, isso significa que, tendenciosamente, os resultados das previsões estão abaixo ou acima dos valores de referência. É realizado o Teste-T (t-student) bilateral, no qual se verifica se a média dos resíduos ($\mu_R$), dada pela Equação 4.6, é significativamente diferente de zero ou se a diferença é simplesmente resultado de erros aleatórios.

$$\mu_R = \frac{\sum_{i=1}^{n}(y_{ref,i} - y_{prev,i})}{n} \qquad \text{Equação (4.6)}$$

Para isso, são estabelecidas a seguintes hipóteses:

Hipótese nula $H_0$: $\mu_R = 0$
Hipótese alternativa $H_a$: $\mu_R \neq 0$

Aceita-se a hipótese nula caso $-t_{tab} \leq t_{calc} \leq +t_{tab}$ e considera-se então que não há erros sistemáticos. O valor de $t_{tab}$ é encontrado a partir da tabela da distribuição t-student, considerando um nível de significância de 0,05 e n-1 graus de liberdade. Um nível de significância igual a 0,05 significa que há 5% de chance de a hipótese nula ser rejeitada quando, na verdade, ela é verdadeira. O valor de $t_{calc}$ é calculado da seguinte forma:

$$t_{calc} = \frac{\mu_R}{S/\sqrt{n}} \qquad \text{Equação (4.7)}$$

Em que n é o número de amostras e S é o desvio padrão amostral dos resíduos,

calculado pela Equação 4.8.

$$s = \frac{\sum_{i=1}^{n}\left((y_{ref,i} - y_{prev,i}) - \mu_R\right)^2}{n-1} \qquad \text{Equação (4.8)}$$

O gráfico a seguir representa a função de densidade de probabilidade para um teste t-student, com um nível de significância de 0,05 e as demarcações das regiões de aceitação e rejeição da hipótese nula.

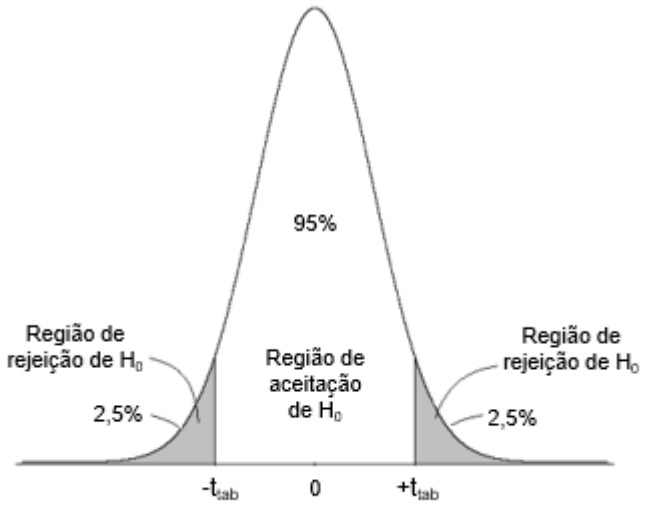



**Figura 4.7.** Teste de hipótese e área de rejeição e aceitação da hipótese nula.

Já o teste de permutação não paramétrico verifica a presença de tendências nos resíduos, ou seja, verifica se há uma relação entre os resíduos ($y_{ref} - y_{prev}$) e a variável dependente ($y_{ref}$) e se essa relação é significativa ou devida ao acaso. Para isso, são estabelecidas a seguintes hipóteses:

Hipótese nula $H_0$: os resíduos, $r_i = y_{ref,i} - y_{prev,i}$, são independentes de $y_{ref,i}$
Hipótese alternativa $H_a$: os resíduos são dependentes de $y_{ref,i}$, conforme a equação:

$$r_i = f(y_{ref,i}) + e_i \qquad \text{Equação (4.9)}$$

Em que $e_i$ é um erro aleatório não dependente de $y_{ref,i}$ e $f(y_{ref,i})$ é uma função polinomial que relaciona os erros $r_i$ com os valores de referência $y_{ref,i}$. Caso o maior

coeficiente da função $f(y_{ref,i})$ seja estatisticamente significativo ao nível de significância de 0,05, a hipótese nula é rejeitada e constata-se a presença de tendência nos resíduos. Neste estudo, foram testadas as funções polinomiais de grau um e dois.

Por fim, após a confecção dos modelos finais, a matriz de *loadings*, comumente referida como matriz de pesos, foi avaliada para investigar as variáveis mais importantes, ou seja, aquelas com maior contribuição nos modelos.

## 4.3 Resultados e discussão

### 4.3.1 Propriedades físico-químicas

A Figura 4.8 mostra histogramas da distribuição das propriedades físico-químicas das amostras de óleo, determinadas pelos experimentos. Os valores dessas propriedades para cada uma das amostras estão dispostos na Tabela A.1 do Apêndice A. O teor total de enxofre é uma das características mais importantes que determinam a qualidade do petróleo bruto e, conseqüentemente, seu valor comercial. O enxofre pode causar corrosão de equipamentos metálicos, contaminação de catalisadores e produção de gases poluentes durante a queima do combustível, por esses motivos, seu controle é importante. Em geral, quanto menor o teor de enxofre no óleo, melhor a sua qualidade.

As amostras utilizadas apresentaram teor total de enxofre variando de 0,0044 a 0,96% em massa. De acordo com a classificação de RIAZI (2005), a maioria dessas amostras são classificadas como óleos doces, uma vez que o teor de enxofre é inferior a 0,5% em massa. As demais amostras são classificadas como óleo azedo; no entanto, o teor máximo de enxofre nessas amostras (1% em massa) é baixo em comparação com certos óleos pesados, cujo teor pode chegar a 6% em massa.

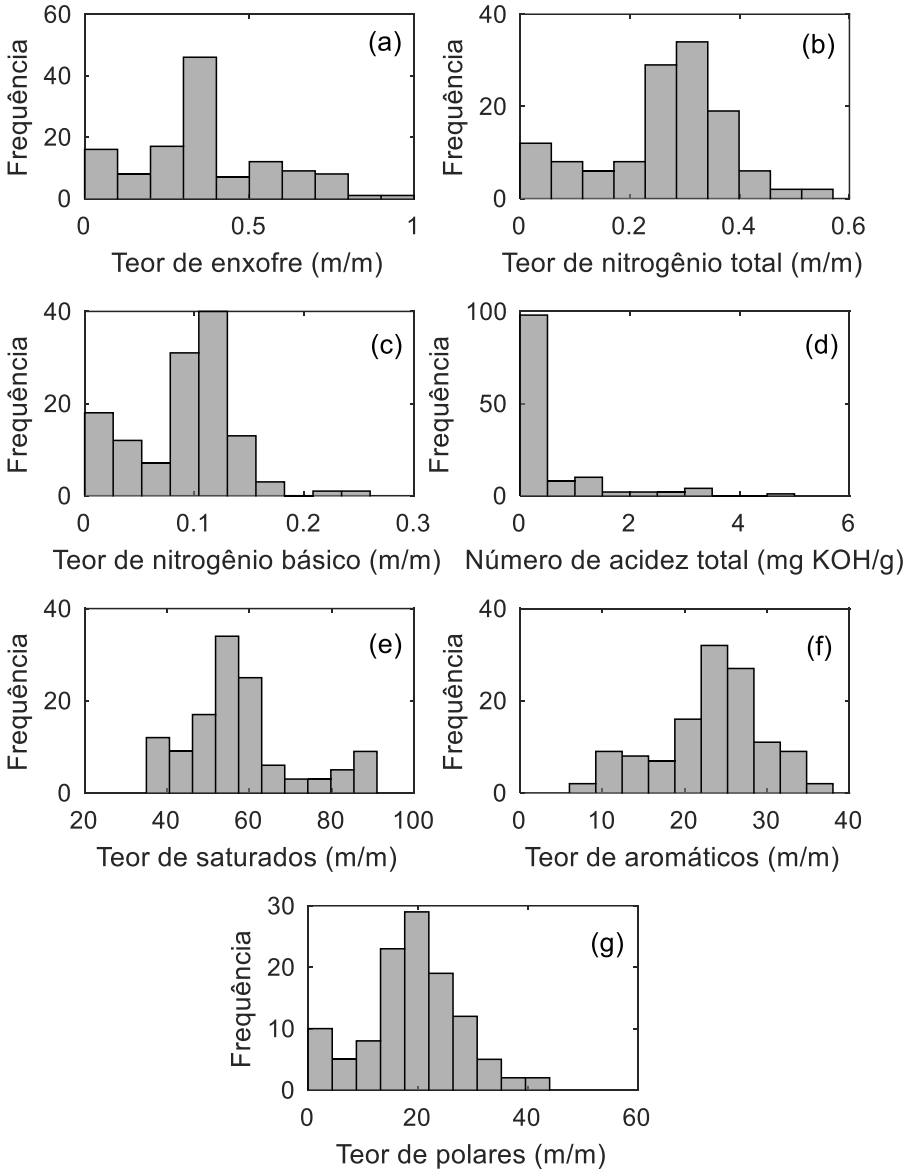



**Figura 4.8.** Histograma das amostras de petróleo para (a) enxofre, (b) nitrogênio total, (c) nitrogênio básico, (d) número de acidez total, (e) saturados, (f) aromáticos e (g) polares.

Além do enxofre, compostos nitrogenados trazem vários efeitos deletérios no processamento do petróleo. Em especial, os compostos nitrogenados básicos podem reagir e desativar os sítios ácidos dos catalisadores. O teor de nitrogênio total é geralmente menor que o teor de enxofre no petróleo bruto, variando de 0,01 a 0,9% em massa (RIAZI, 2005). Segundo PRADO *et al*. (2017), quase 90% do petróleo bruto

possui teor de nitrogênio total inferior a 0,25% em massa, sendo classificados como pobres em nitrogênio. Óleos com teor acima deste são classificados como ricos em nitrogênio. Como mostra a Figura 4.8b, as amostras de óleo utilizadas apresentaram teores totais de nitrogênio variando de 0,00087 a 0,57% em massa, sendo a maioria delas classificadas como ricas em nitrogênio. Além do nitrogênio total, o nitrogênio básico também foi medido e apresentou uma concentração variando de 0,0005 a 0,21% em massa, de modo que a maioria das amostras apresentou um valor médio igual a 0,1% (Figura 4.8c).

Para evitar problemas de corrosão no processo de refino, recomenda-se que o número total de acidez seja inferior a 0,5 mg $KOHg^{-1}$. De acordo com a Figura 4.8d, o número total de acidez variou de 0,030 a 4,96 mg $KOHg^{-1}$ e a maioria das amostras apresentou baixos valores dessa propriedade, sendo consideradas óleos não ácidos. As demais amostras são classificadas como óleos ácidos, uma vez que o NAT é superior a 0,5 mg $KOHg^{-1}$.

De acordo com WOODS *et al*. (2008), óleos convencionais (óleos não pesados) têm maiores quantidades de compostos saturados e uma fração polar pequena. As amostras utilizadas neste trabalho apresentaram uma fração saturada variando de 36,6 a 90,6%, fração aromática de 7,0 a 37,2% e fração polar de 0,0 a 43,5%, em massa. As Figuras 4.8e, 16f e 16g mostram que as amostras de óleo são constituídas principalmente por compostos saturados que são formados por cadeias simples. O segundo maior constituínte são os compostos aromáticos e, em menor escala, os polares. Para algumas amostras, a fração polar chega a zero.

### 4.3.2 Espectros de MIR, RMN $^{1}H$ e $^{13}C$

A Figura 4.9a mostra uma sobreposição dos espectros de MIR de todas as amostras de petróleo. Os picos próximos a 2900 e 1450 $cm^{-1}$ são decorrentes, respectivamente, do estiramento e da deformação da ligação C – H das cadeias alifáticas. O sinal de intensidade média, próximo à região de 1375 $cm^{-1}$, corresponde às vibrações dos grupos metila, ligados aos átomos de carbono. A região entre 1200 e 650 $cm^{-1}$, conhecida como região de impressão digital, apresenta uma mistura de sinais relacionados às vibrações de cadeias alifáticas e aromáticas, além de sinais de alguns heteroátomos como enxofre, nitrogênio, flúor e cloro (PAVIA *et al*., 2008).

Os espectros sobrepostos de RMN $^{1}H$ de todas as amostras de petróleo são mostrados na Figura 4.9b. Segundo MASILI *et al*. (2012), os sinais entre 0,5 e 1,0 ppm referem-se ao grupo $CH_3$ em cadeias alifáticas. Já o pico de alta intensidade na região 1,2 - 1,3 ppm é devido ao grupo $CH_2$, também em cadeias alifáticas. Sinais entre 1,5 - 1,6 ppm referem-se ao $CH_2$ β-aromático. A região destacada, entre 2,0 - 2,5 ppm, refere-se ao $CH_3$ α-aromáticos e entre 7,2 - 7,4 ppm, às cadeias de di-aromáticos. O solvente, $CDCl_3$, tem um pico na região de 5,3 ppm, onde não há sinal de amostra e, portanto, nenhum problema de sobreposição. O sinal do solvente pode ser usado para verificar o alinhamento dos picos entre os espectros das amostras.

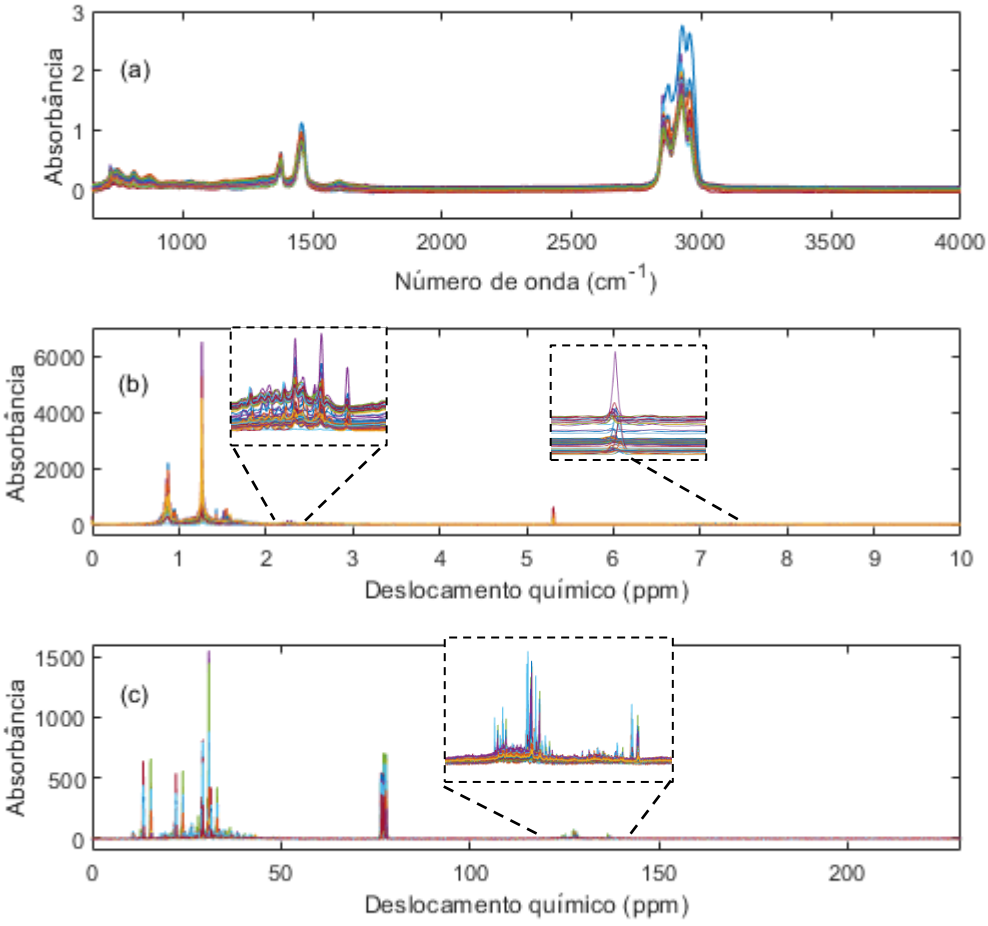



**Figura 4.9.** Sobreposição de espectros (a) de MIR (b) RMN $^{1}H$ e (c) RMN $^{13}C$ das amostras de petróleo.

A Figura 4.9c mostra uma sobreposição dos espectros de RMN $^{13}C$ de todas as amostras de petróleo. O pico ligeiramente abaixo de 80 ppm corresponde ao sinal

do solvente $CDCl_3$, utilizado para a análise. Nesta região, não há sinal correspondente às amostras, portanto, não há problemas de sobreposição entre o sinal do solvente e das amostras. A região abaixo de 40 ppm corresponde principalmente a carbonos de cadeias alifáticas e a região de destaque entre 120 e 140 ppm corresponde, principalmente, a carbonos aromáticos e carbonos ligados a heteroátomos.

### 4.3.3 Pré-processamento dos dados

A etapa de pré-processamento é fundamental para bons resultados na análise multivariada. O pré-processamento utilizado para os espectros no infravermelho foi o MSC, cuja função foi corrigir o deslocamento da linha de base. Isso é essencial para calibração multivariada com dados na região do infravermelho, uma vez que seus espectros apresentam alta variação da linha de base, prejudicando a regressão. A Figura 4.10 mostra os espectros no infravermelho médio de todas as amostras antes e após o tratamento com MSC.

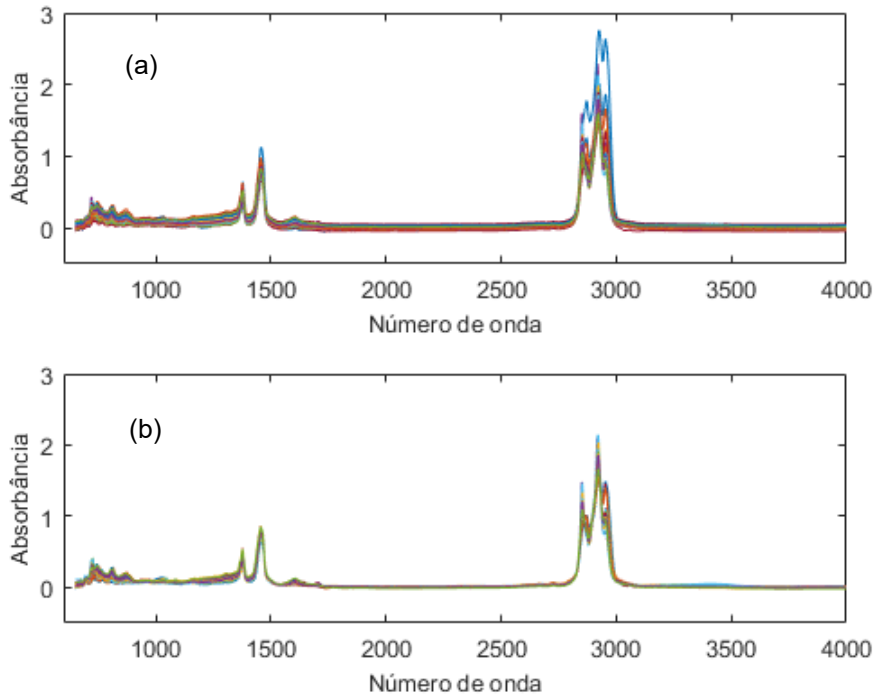



**Figura 4.10.** Espectros no infravermelho médio (a) antes e (b) após o pré-processamento MSC.

Para os espectros de RMN, o pré-processamento utilizado foi o SNV, pois apresentou melhores resultados em relação aos demais, de acordo com testes

preliminares. O SNV possui um efeito muito similar ao MSC, porém sua correção é através da normalização de cada espectro individualmente (mesma amostra), ao contrário do MSC que se baseia na normalização dos espectros contra um espectro de referência, que geralmente é o espectro médio das amostras. A grande vantagem do SNV aplicado sobre dados de RMN é a sua capacidade de padronizar as intensidades dos picos entre uma amostra e outra.

Além de pré-processados, os espectros de RMN $^{1}H$ e $^{13}C$ também foram previamente alinhados com o algoritmo *icoshift* para corrigir problemas de desvios entre uma amostra e outra. A Figura 4.11 mostra os espectros de RMN $^{1}H$ de todas as amostras, antes e após o alinhamento e o tratamento com SNV. A região entre o deslocamento químico 1 e 2 foi ampliada para melhor vizalização do efeito do alinhamento.

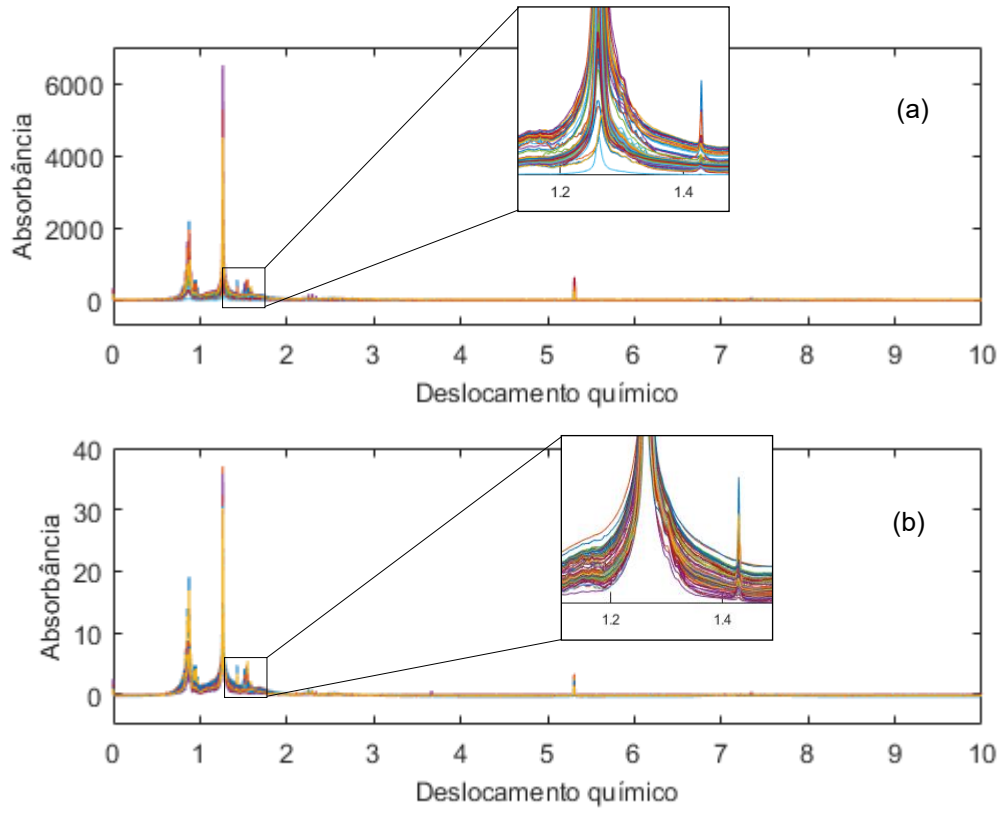



**Figura 4.11.** Espectros de RMN $^{1}H$ (a) antes e (b) após alinhamento e pré-processamento por SNV.

Após os pré-processamentos, procedeu-se com a construção dos modelos, utilizando a calibração linear PLS. A Figura 4.12 mostra a variância das matrizes MIR,

RMN $^{1}H$ e RMN $^{13}C$ explicada pelas variáveis latentes após a aplicação do PLS. A variância explicada pela primeira variável latente é maior que pela segunda. A variância explicada pela segunda é maior que pela terceira e assim sucessivamente até atingir um patamar próximo a 1 (ou 100%) para MIR e RMN $^{1}H$ e próximo a 0,9 (ou 90%) para RMN $^{13}C$, como mostra a figura.

Deve-se utilizar um método de validação cruzada para se determinar o número ótimo de variáveis latentes, de forma que este não seja muito grande para que não ocorra superajuste aos dados de calibração e nem muito pequeno para que o máximo de variância consiga ser explicada pelo modelo.

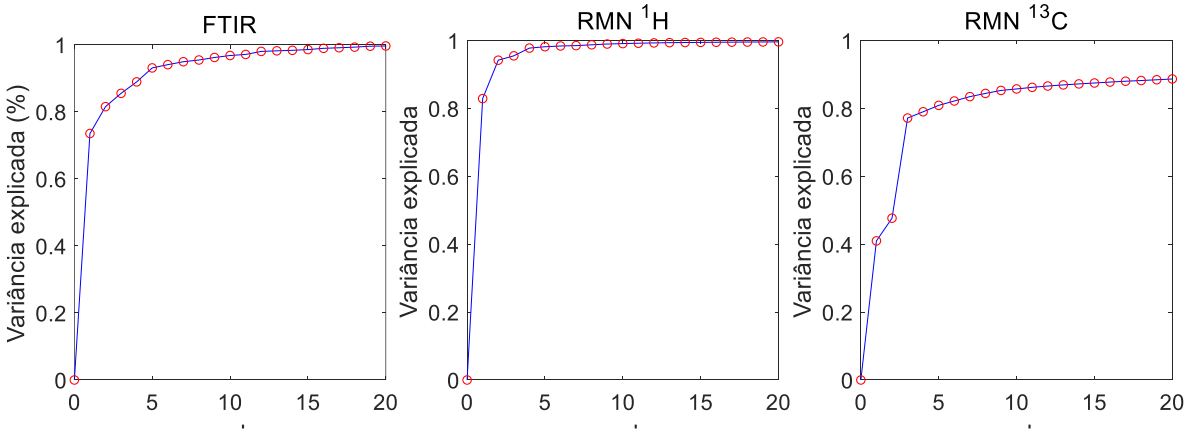



**Figura 4.12.** Variância explicada das matrizes de MIR, RMN $^{1}H$ e RMN $^{13}C$ pelas variáveis latentes do PLS.

### 4.3.4 Modelos a partir dos espectros individuais

Os primeiros modelos construídos foram aqueles a partir dos espectros separados (não-fundidos) e seus principais resultados estão dispostos na Tabela 4.2. Estes modelos foram chamados de “modelos individuais”.

Aparentemente, as três técnicas analíticas, MIR, RMN $^{1}H$ e $^{13}C$ apresentaram desempenho semelhante de previsão, para cada propriedade. O teste de randomização, descrito por VOET (1994), foi aplicado no conjunto de dados de teste para comparar os valores de $RMSE_P$ desses três modelos, para cada propriedade. A Tabela 4.3 apresenta os resultados dos testes (p-valor). Considera-se que a diferença entre as exatidões é estatisticamente significativa, e não atribuída ao acaso, quando p-valor é menor que 0,05. O teste não indicou diferenças significativas na exatidão

dos modelos em todas as propriedades, exceto no caso do nitrogênio básico, cujo valor do $RMSE_P$ do modelo de RMN $^{13}C$ diferiu significativamente do modelo MIR (p-valor = 0,008) e de RMN $^{1}H$ (p-valor = 0,0001), apresentando desempenho inferior. Para previsão de enxofre, também houve diferença significativa entre a exatidão do modelo RMN $^{13}C$ e a do modelo MIR (p-valor = 0,049) que apresentou desempenho inferior. Pode-se afirmar, portanto, que, de forma geral, as três fontes de dados apresentam desempenhos semelhantes para calibração multivariada sem fusão de dados.

A Figura 4.13 apresenta os gráficos oriundos da validação cruzada, somente para os modelos individuais que apresentaram os menores valores de $RMSE_P$ (destacados em negrito na Tabela 4.2), para cada propriedade. Tais gráficos mostram os valores de $RMSE_{CV}$ em função do número de variáveis latentes que compõe o modelo. A análise destes gráficos permitiu escolher o número ideal de variáveis latentes que minimiza o erro de previsão dos modelos. A Tabela 4.2 apresenta o nvl escolhido para compor cada um dos modelos individuais.

Para todos os modelos em negrito, os valores $RMSE_C$ e $RMSE_P$ são relativamente próximos, portanto, não há indicativo de superajuste aos dados de calibração. O superajuste ocorre quando o modelo tem uma acurácia elevada em relação aos dados de calibração, apresentando um alto valor de $RMSE_C$, porém é pouco acurado para amostras externas, apresentando $RMSE_P$ relativamente baixo. Não existe um limite pré-definido para classificação de modelos superajustados ou não, dessa forma, para este trabalho, baseado em análises visuais dos gráficos de dados medidos versus dados previstos, foi considerado superajuste aos dados de calibração quando $RMSE_P > 3.RMSE_C$.

Para os mesmos modelos, destacados em negrito na Tabela 4.2, a Figura 4.14 apresenta gráficos dos dados obtidos pelos experimentospadronizados *versus* os dados previstos pelos modelos. Os resultados de S, NT, NB, NAT e SAT apresentam uma boa linearidade para amostras de teste, expressa por valores mais altos do coeficiente de determinação ($R_P^2$). Para ARO e POL, valores mais baixos de coeficiente de determinação foram alcançados para as amostras teste (menores que 0,6) e, portanto, uma menor linearidade pode ser vista nos gráficos.

**Tabela 4.2.** Principais parâmetros estatísticos dos modelos individuais (sem fusão).

| | | nvl | [a]$RMSE_C$ | [a]$RMSE_P$ | $RMSE_P$% | $R_C^2$ | $R_P^2$ | $T_{tab,test}$ | $T_{cal,test}$ | p-valor (1º) | p-valor (2º) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MIR | 4 | 0,105 | 0,103 | 30,40 | 0,758 | 0.673 | 2,028 | 0,191 | 0,29 | 0,092 |
| S | RMN $^{1}$H | 4 | 0,196 | 0,089 | 26,42 | 0,753 | 0,755 | 2,028 | 0,826 | 0,45 | 0,18 |
| | **RMN $^{13}$C** | **6** | **0,055** | **0,081** | **24,03** | **0,936** | **0,797** | **2,028** | **0,657** | **0,36** | **0,50** |
| | MIR | 5 | 0,031 | 0,052 | 19,05 | 0,948 | 0,760 | 2,028 | 0,387 | 0,010 | 0,016 |
| NT | **RMN $^{1}$H** | **8** | **0,030** | **0,049** | **17,77** | **0,951** | **0,773** | **2,028** | **0,021** | **0,049** | **0,042** |
| | RMN $^{13}$C | 6 | 0,022 | 0,073 | 26,81 | 0,974 | 0,585 | 2,028 | 0,904 | 0,002 | 0,006 |
| | MIR | 6 | 0,010 | 0,012 | 12,78 | 0,952 | 0,933 | 2,030 | 0,120 | 0,005 | 0,40 |
| NB | **RMN $^{1}$H** | **6** | **0,011** | **0,010** | **10,34** | **0,948** | **0,947** | **2,030** | **0,363** | **0,12** | **0,24** |
| | RMN $^{13}$C | 6 | 0,010 | 0,020 | 21,57 | 0,959 | 0,770 | 2,030 | 0,966 | 0,21 | 0,056 |
| | **MIR** | **5** | **0,279** | **0,255** | **58,64** | **0,910** | **0,799** | **2,028** | **0,120** | **0,003** | **0,000** |
| NAT | RMN $^{1}$H | 4 | 0,307 | 0,287 | 66,01 | 0,889 | 0,725 | 2,028 | 1,806 | 0,29 | 0,000 |
| | RMN $^{13}$C | 6 | 0,210 | 0,295 | 67,69 | 0,949 | 0,683 | 2,028 | 0,594 | 0,29 | 0,000 |
| | MIR | 3 | 6,116 | 5,827 | 10,37 | 0,813 | 0,762 | 2,032 | 0,628 | 0,15 | 0,116 |
| SAT | **RMN $^{1}$H** | **4** | **6,250** | **5,522** | **9,83** | **0,807** | **0,788** | **2,032** | **0,502** | **0,11** | **0,391** |
| | RMN $^{13}$C | 6 | 3,220 | 6,687 | 11,90 | 0,950 | 0,673 | 2,032 | 0,207 | 0,39 | 0,017 |
| | MIR | 4 | 3,190 | 4,323 | 18,84 | 0,775 | 0,574 | 2,028 | 0,261 | 0,014 | 0,038 |
| ARO | RMN $^{1}$H | 4 | 2,964 | 4,651 | 20,27 | 0,806 | 0,507 | 2,028 | 1,218 | 0,032 | 0,23 |
| | **RMN $^{13}$C** | **6** | **1,805** | **4,255** | **18,55** | **0,930** | **0,589** | **2,028** | **0,142** | **0,011** | **0,007** |
| | MIR | 4 | 4,529 | 7,420 | 37,28 | 0,725 | 0,541 | 2,037 | 0,086 | 0,21 | 0,062 |
| POL | RMN $^{1}$H | 3 | 5,365 | 7,850 | 39,43 | 0,610 | 0,520 | 2,037 | 1,000 | 0,086 | 0,048 |
| | **RMN $^{13}$C** | **5** | **3,505** | **7,337** | **36,86** | **0,838** | **0,555** | **2,037** | **1,003** | **0,46** | **0,12** |

[a]Unidade m/m% para as propriedades S, NT, NB, SAT, ARO e POL e unidade $mgKOHg^{-1}$ para NAT.

**Tabela 4.3.** P-valor do teste de randomização para comparação entre os modelos individuais.

| | | MIR | RMN $^{1}$H | RMN $^{13}$C |
|---|---|---|---|---|
| **Enxofre** | MIR | 1,0 | 0,11 | 0,049 |
| | RMN $^{1}$H | 0,11 | 1,0 | 0,11 |
| | RMN $^{13}$C | 0,049 | 0,11 | 1,0 |
| | | MIR | RMN $^{1}$H | RMN $^{13}$C |
| **Nitrogênio total** | MIR | 1,0 | 0,29 | 0,11 |
| | RMN $^{1}$H | 0,29 | 1,0 | 0,065 |
| | RMN $^{13}$C | 0,11 | 0,065 | 1,0 |
| | | MIR | RMN $^{1}$H | RMN $^{13}$C |
| **Nitrogênio básico** | MIR | 1,0 | 0,08 | 0,008 |
| | RMN $^{1}$H | 0,08 | 1,0 | 0,0001 |
| | RMN $^{13}$C | 0,008 | 0,0001 | 1,0 |
| | | MIR | RMN $^{1}$H | RMN $^{13}$C |
| **Número de acidez total** | MIR | 1,0 | 0,35 | 0,28 |
| | RMN $^{1}$H | 0,35 | 1,0 | 0,4 |
| | RMN $^{13}$C | 0,28 | 0,4 | 1,0 |
| | | MIR | RMN $^{1}$H | RMN $^{13}$C |
| **Teor de saturados** | MIR | 1,0 | 0,38 | 0,13 |
| | RMN $^{1}$H | 0,38 | 1,0 | 0,12 |
| | RMN $^{13}$C | 0,13 | 0,12 | 1,0 |
| | | MIR | RMN $^{1}$H | RMN $^{13}$C |
| **Teor de aromáticos** | MIR | 1,0 | 0,2 | 0,39 |
| | RMN $^{1}$H | 0,2 | 1,0 | 0,18 |
| | RMN $^{13}$C | 0,39 | 0,18 | 1,0 |
| | | MIR | RMN $^{1}$H | RMN $^{13}$C |
| **Teor de polares** | MIR | 1,0 | 0,36 | 0,49 |
| | RMN $^{1}$H | 0,36 | 1,0 | 0,14 |
| | RMN $^{13}$C | 0,49 | 0,14 | 1,0 |

A Figura 4.15 apresenta os gráficos de erro *versus leverage* para identificação de possíveis amostras anômalas, também chamadas de *outliers*. Amostras com alto *leverege* tem um peso significativo na construção do modelo de regressão, uma vez que diferem bastante do perfil médio das demais amostras. Alto valor de *leverage* associado a um erro alto é um indicativo de *outlier*. Os gráficos também apresentam uma linha vertical e outra horizontal que delimitam a fronteira para altos valores de *leverage* e de erro, respectivamente. Observa-se que para NAT a amostra 99 cruza a duas fronteiras, portanto possui simultaneamente *leverage* e erro altos, indicando a possibilidade de que seja uma amostra anômala. Dessa forma, essa amostra foi

eliminada do conjunto de dados e a regressão foi novamente realizada. O novo modelo não apresentou erro significativamente menor, portanto, a amostras 99 não foi considerada anômala e permaneceu no conjunto de dados.

Observa-se também que para previsão de nitrogênio básico, a amostra 21, apesar de não possuir *leverage* acima do limite, possui erro muito superior às demais amostras indicando a possibilidade de que seja um *outlier*. A amostra foi eliminada e a regressão foi novamente realizada. O novo modelo conseguiu prever nitrogênio básico com exatidão significativamente superior, com $RMSE_P$ igual a 0,0096 m/m% contra um $RMSE_P$ igual a 0,0547 m/m% quando a amostra 21 é utilizada. Dessa forma, essa amostra foi considerada anômala e excluída definitivamente do conjunto de dados para previsão de NB. Os resultados apresentados na Tabela 4.2 e nas Figuras 4.13 e 4.14 para nitrogênio básico são referentes à modelos construídos já com a devida eliminação da amostra 21.

De acordo com a análise, a amostra 21 foi a única considerada *outlier* e, portanto, foi eliminada do conjunto de dados para calibração e previsão somente da propriedade NB, tanto para os modelos individuais, quanto para todos os demais modelos construídos ao longo desta pesquisa.

Após a etapa de calibração e predição, os resíduos deixados pelos modelos foram avaliados por meio de testes estatísticos. Os resíduos, mostrados nos gráficos da Figura 4.16, foram avaliados quanto à presença de erros sistemáticos e tendências. Observando os gráficos, os resíduos parecem não apresentar tendência, entretanto, esta é uma análise visual muito subjetiva. Para tornar a análise mais confiável, foi realizado o teste de permutação não paramétrico para as amostras de teste e os resultados do p-valor para funções de primeiro e segundo grau estão dispostos na Tabela 4.2. Para os modelos destacados em negrito, o teste indicou a presença de tendência linear e quadrática nos resíduos de previsão das amostras teste das propriedades NT, NAT e ARO, visto que seu p-valor foi menor que 0,05. Isso indica que o conjunto de dados pode estar correlacionado com estas propriedades de forma linear e não linear. Isso mostra que o modelo não consegue capturar toda relação existente entre as variáveis independentes e a dependente, ficando parte dessa relação refletida nos resíduos. É possível ainda que um método de calibração não linear seja mais adequado para capturar toda a variância dos dados, relacioná-la com

estas propriedades e, consequentemente, aumentar o coeficiente de determinação e reduzir o $RMSE_P$.

Pela análise visual dos gráficos também é muito subjetiva a afirmação de presença ou ausência de erros sistemáticos. Para uma avaliação mais precisa, foi realizado o teste de viés (ou teste de *bias*) que indicou que não há presença de erros sistemáticos significativos para as amostras teste, pois o valor obtido foi menor que o tabelado ($t_{cal,teste} < t_{tab,teste}$), não somente para os modelos destacados em negrito, mas para todos eles.

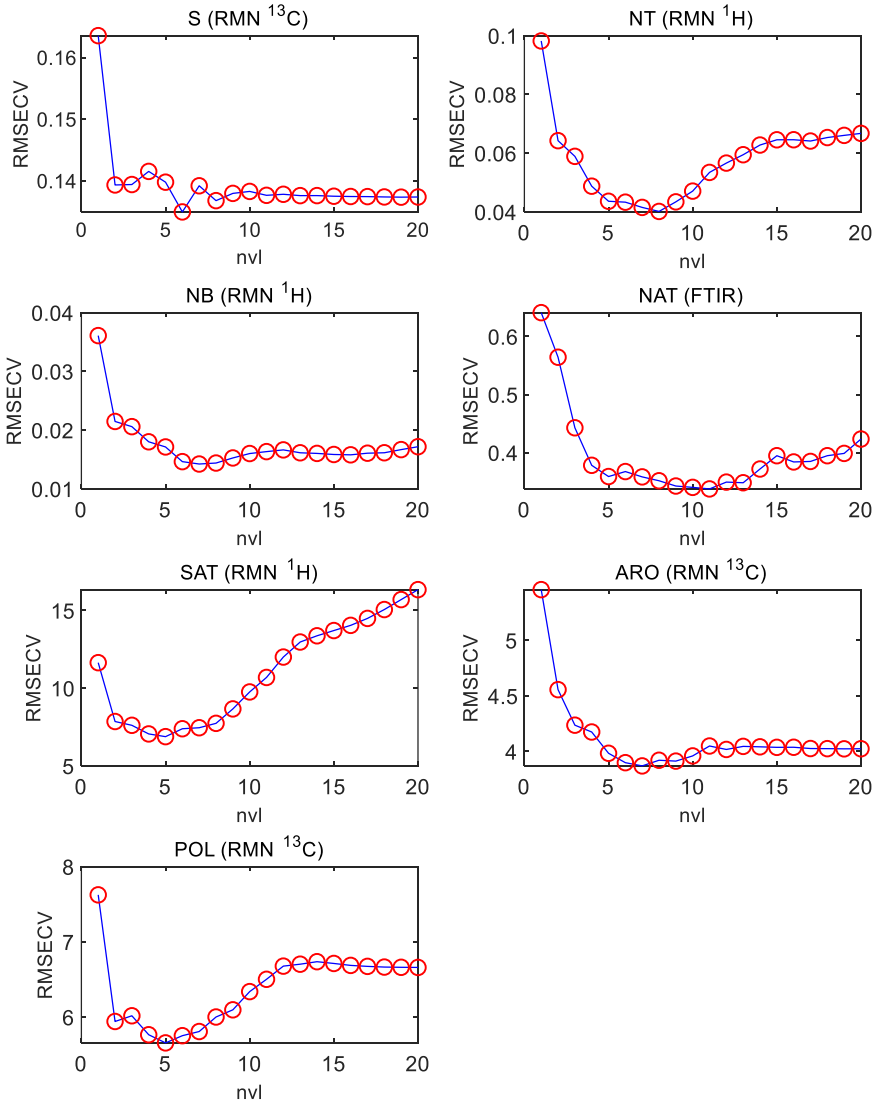



**Figura 4.13.** $RMSE_{CV}$ *versus* número de variáveis latentes para os modelos individuais.

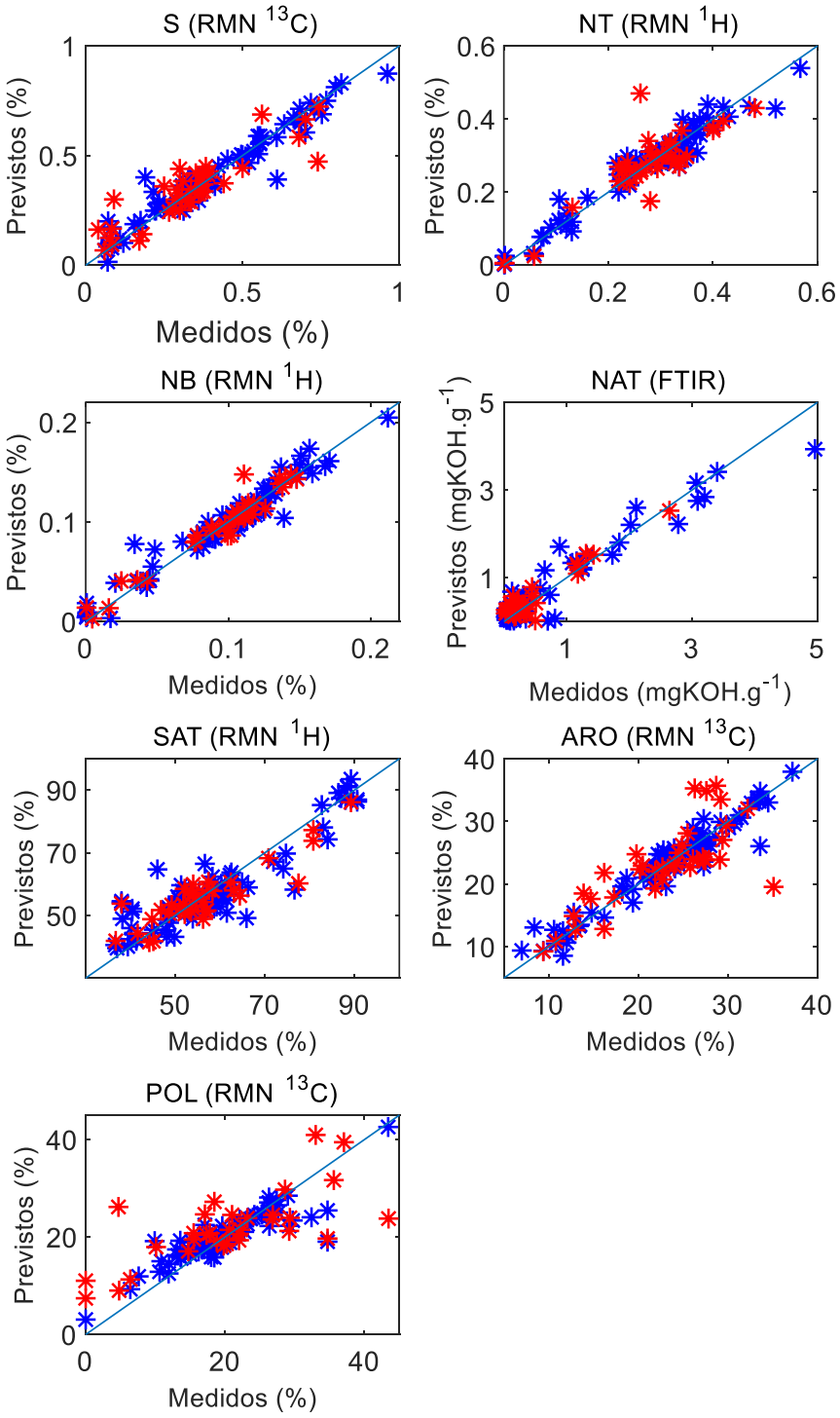



**Figura 4.14.** Gráficos de correlação entre os valores de referência experimentais e os valores preditos pelos modelos individuais. Legenda: azul - amostras de calibração, vermelho - amostras teste.

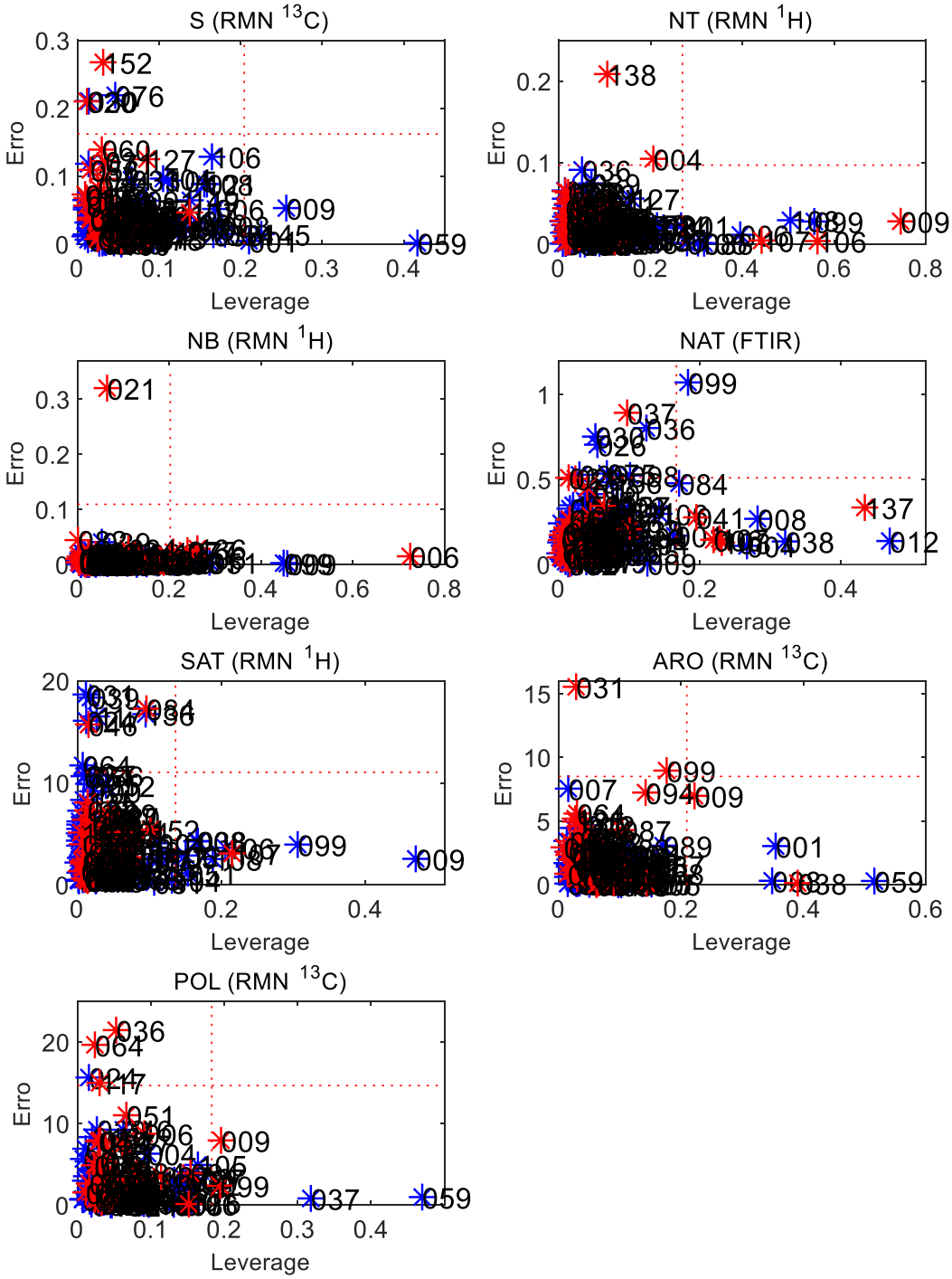



**Figura 4.15.** Erro de previsão *versus leverage*. Legenda: azul - amostras de calibração, vermelho - amostras teste.

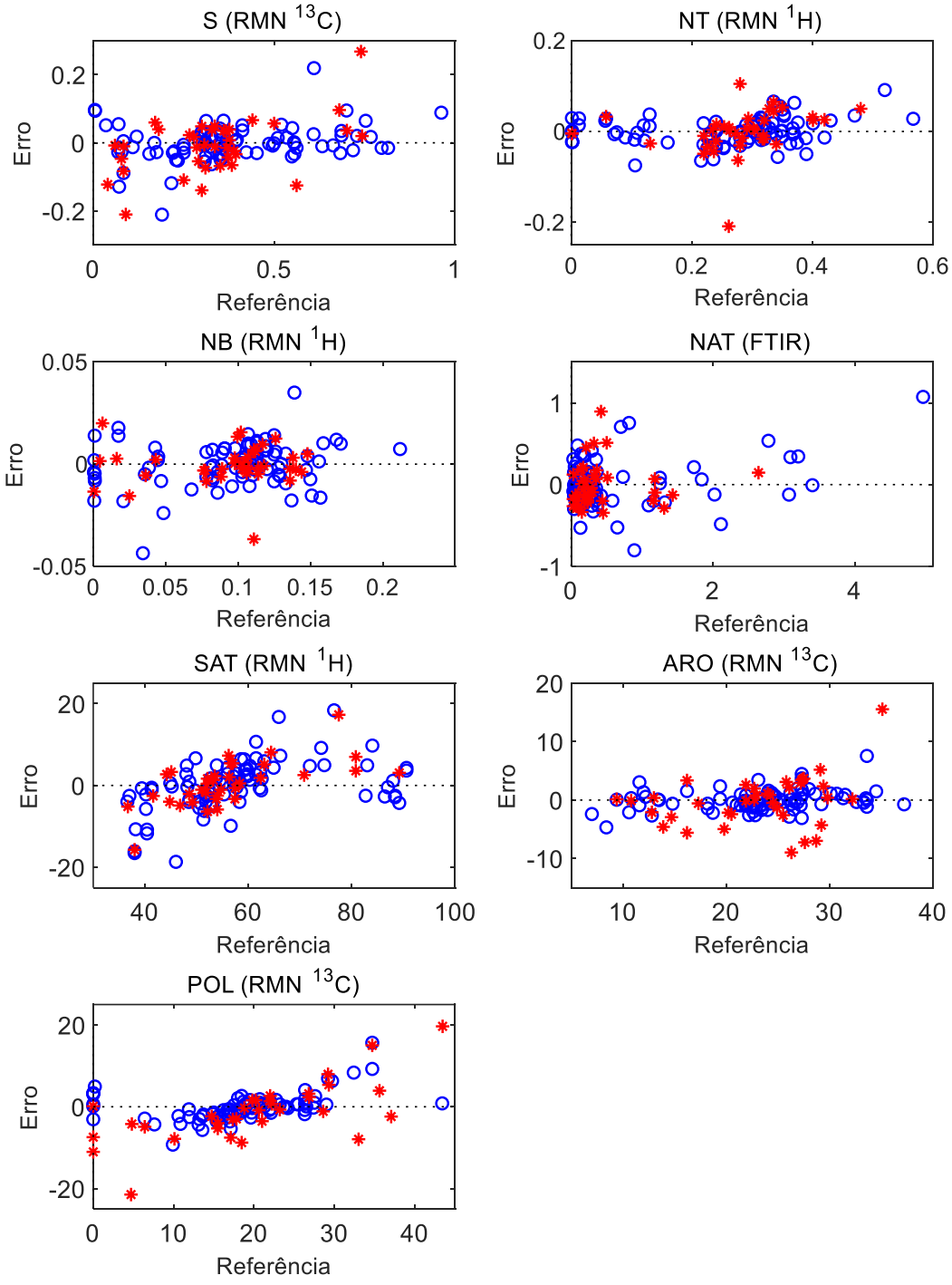



**Figura 4.16.** Distribuição dos resíduos dos modelos individuais. Legenda: azul - amostras de calibração, vermelho - amostras teste.

### 4.3.5 Modelos a partir da fusão de nível baixo

A fusão a nível baixo foi feita por meio da simples concatenação das matrizes de dados das diferentes fontes analíticas. Foram feitas todas as combinações possíveis dos espectros, resultando nas quatro fusões: MIR + RMN $^{1}$H, MIR + RMN $^{13}$C, RMN $^{1}$H + RMN $^{13}$C e MIR + RMN $^{1}$H + RMN $^{13}$C. Os resultados da modelagem de nível baixo estão apresentados na Tabela 4.4, com os principais parâmetros dos modelos para cada propriedade físico-química. Os modelos que proporcionaram os menores valores de $RMSE_P$ estão destacados em negrito. Para facilitar a comparação, foi também inserido na tabela (destacado em cinza) o melhor resultado dos modelos individuais. Comparando-se os resultados com os modelos individuais, verifica-se que a fusão de nível baixo não trouxe melhoria significativa na exatidão, exceto para as propriedades SAT e ARO que apresentaram um $RMSE_P$ ligeiramente inferior, com uma redução de 5,52 para 5,43 m/m% e de 4,26 para 3,87 m/m%, respectivamente.

Para todos os modelos destacados em negrito, os valores $RMSE_C$ e $RMSE_P$ são relativamente próximos, portanto, não há indicativo de superajuste aos dados de calibração. Quanto aos resíduos, nenhum dos modelos destacados apresentou erros sistemáticos, de acordo com o teste de viés. Já os modelos de NT, ARO e POL apresentaram tendências linear e quadrática nos resíduos, o modelo de NB tendência linear, e NAT tendência quadrática, visto que o p-valor, resultante dos testes de permutação não paramétricos, foi menor que 0,05, como mostra a Tabela 4.4. Isso mostra que os modelos não conseguiram capturar toda relação existente entre as variáveis dependentes e independentes, ficando essa relação refletida nos resíduos.

O baixo desempenho da fusão de baixo nível pode ser decorrente das desvantagens dessa estratégia. Segundo ROUSSEL *et al*. (2003), a simples concatenação de dados ruidosos e correlacionados pode gerar resultados piores, em vez de melhorá-los. Além disso, segundo BORRÀS *et al*. (2015), quando os conjuntos de dados a serem fundidos apresentam dimensões e variâncias muito diferentes, pode haver o predomínio de uma fonte de dados sobre a outra. As matrizes dos espectros de MIR, RMN $^{1}$H e RMN $^{13}$C utilizadas neste trabalho apresentaram 3.351, 20.433 e 65.533 variáveis, respectivamente. Essa discrepância entre as dimensões das matrizes pode ter contribuído para o fracasso da fusão de nível baixo.

**Tabela 4.4.** Principais parâmetros estatísticos dos modelos da fusão a nível baixo.

| | | nvl | [a]$RMSE_C$ | [a]$RMSE_P$ | $RMSE_P$% | $R_C^2$ | $R_P^2$ | $T_{tab,test}$ | $T_{cal,test}$ | p-valor (1º) | p-valor (2º) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| S | RMN $^{13}C$ | 6 | 0,055 | 0,081 | 24,03 | 0,936 | 0,797 | 2,028 | 0,657 | 0,36 | 0,50 |
| | MIR + RMN $^{1}H$ | 4 | 0,102 | 0,096 | 28,31 | 0,770 | 0,717 | 2,028 | 0,426 | 0,31 | 0,44 |
| | MIR + RMN $^{13}C$ | 6 | 0,051 | 0,092 | 27,26 | 0,945 | 0,739 | 2,028 | 0,451 | 0,28 | 0,094 |
| | **RMN $^{1}H$ + RMN $^{13}C$** | **7** | **0,050** | **0,083** | **24,63** | **0,947** | **0,786** | **2,028** | **0,610** | **0,48** | **0,40** |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 7 | 0,052 | 0,090 | 26,52 | 0,943 | 0,751 | 2,028 | 0,125 | 0,31 | 0,22 |
| NT | **RMN $^{1}H$** | **8** | **0,030** | **0,049** | **17,77** | **0,951** | **0,773** | **2,028** | **0,021** | **0,049** | **0,042** |
| | **MIR + RMN $^{1}H$** | **6** | **0,028** | **0,049** | **17,94** | **0,957** | **0,769** | **2,028** | **0,299** | **0,047** | **0,018** |
| | MIR + RMN $^{13}C$ | 6 | 0,020 | 0,066 | 24,11 | 0,977 | 0,653 | 2,028 | 0,778 | 0,0024 | 0,000 |
| | RMN $^{1}H$ + RMN $^{13}C$ | 8 | 0,014 | 0,057 | 20,82 | 0,990 | 0,696 | 2,028 | 0,761 | 0,046 | 0,003 |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 6 | 0,026 | 0,061 | 22,40 | 0,964 | 0,695 | 2,028 | 0,624 | 0,0033 | 0,010 |
| NB | **RMN $^{1}H$** | **6** | **0,011** | **0,010** | **10,34** | **0,948** | **0,947** | **2,030** | **0,363** | **0,12** | **0,24** |
| | **MIR + RMN $^{1}H$** | **7** | **0,008** | **0,009** | **9,54** | **0,969** | **0,959** | **2,030** | **0,175** | **0,012** | **0,23** |
| | MIR + RMN $^{13}C$ | 6 | 0,009 | 0,016 | 17,64 | 0,967 | 0,840 | 2,030 | 0,205 | 0,35 | 0,019 |
| | RMN $^{1}H$ + RMN $^{13}C$ | 6 | 0,011 | 0,021 | 22,61 | 0,951 | 0,753 | 2,030 | 1,235 | 0,16 | 0,005 |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 6 | 0,010 | 0,017 | 18,74 | 0,956 | 0,822 | 2,030 | 0,784 | 0,39 | 0,010 |
| NAT | MIR | 5 | 0,279 | 0,255 | 58,64 | 0,910 | 0,799 | 2,028 | 1,120 | 0,003 | 0,000 |
| | MIR + RMN $^{1}H$ | 4 | 0,254 | 0,270 | 61,98 | 0,924 | 0,781 | 2,028 | 1,885 | 0,025 | 0,000 |
| | MIR + RMN $^{13}C$ | 6 | 0,192 | 0,258 | 59,25 | 0,958 | 0,765 | 2,028 | 0,539 | 0,15 | 0,000 |
| | RMN $^{1}H$ + RMN $^{13}C$ | 7 | 0,166 | 0,269 | 61,68 | 0,969 | 0,751 | 2,028 | 1,357 | 0,28 | 0,000 |
| | **MIR + RMN $^{1}H$ + RMN $^{13}C$** | **7** | **0,143** | **0,250** | **57,42** | **0,977** | **0,792** | **2,028** | **1,126** | **0,064** | **0,000** |
| SAT | RMN $^{1}H$ | 4 | 6,250 | 5,522 | 9,83 | 0,807 | 0,788 | 2,032 | 0,502 | 0,11 | 0,391 |
| | MIR + RMN $^{1}H$ | 2 | 7,396 | 5,937 | 10,57 | 0,724 | 0,756 | 2,032 | 0,855 | 0,13 | 0,47 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MIR + RMN $^{13}$C | 6 | 3,115 | 6,131 | 10,91 | 0,953 | 0,725 | 2,032 | 0,154 | 0,39 | 0,006 |
| | RMN $^{1}$H + RMN $^{13}$C | 6 | 3,491 | 5,966 | 10,62 | 0,941 | 0,741 | 2,032 | 0,457 | 0,44 | 0,49 |
| | **MIR + RMN $^{1}$H + RMN $^{13}$C** | **6** | **3,478** | **5,431** | **9,67** | **0,942** | **0,789** | **2,032** | **0,3259** | **0,20** | **0,43** |
| | RMN $^{13}$C | 6 | 1,805 | 4,255 | 18,55 | 0,930 | 0,589 | 2,028 | 0,142 | 0,011 | 0,007 |
| | MIR + RMN $^{1}$H | 6 | 2,803 | 4,314 | 18,80 | 0,831 | 0,564 | 2,028 | 1,2433 | 0,059 | 0,22 |
| ARO | MIR + RMN $^{13}$C | 6 | 1,743 | 4,125 | 17,98 | 0,935 | 0,596 | 2,028 | 0,2000 | 0,033 | 0,005 |
| | RMN $^{1}$H + RMN $^{13}$C | 7 | 1,704 | 4,039 | 17,60 | 0,938 | 0,617 | 2,028 | 0,2805 | 0,017 | 0,017 |
| | **MIR + RMN $^{1}$H + RMN $^{13}$C** | **7** | **1,717** | **3,879** | **16,91** | **0,937** | **0,632** | **2,028** | **0,2749** | **0,026** | **0,020** |
| | **RMN $^{13}$C** | 5 | 3,505 | 7,337 | 36,86 | 0,838 | 0,555 | 2,037 | 1,003 | 0,46 | 0,12 |
| | MIR + RMN $^{1}$H | 3 | 5,059 | 8,273 | 41,56 | 0,653 | 0,507 | 2,037 | 1,0831 | 0,039 | 0,038 |
| POL | MIR + RMN $^{13}$C | 2 | 5,502 | 7,708 | 38,72 | 0,584 | 0,500 | 2,037 | 0,6126 | 0,049 | 0,055 |
| | **RMN $^{1}$H + RMN $^{13}$C** | **5** | **3,853** | **7,564** | **38,00** | **0,804** | **0,565** | **2,037** | **1,3405** | **0,023** | **0,029** |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 5 | 3,882 | 7,537 | 37,86 | 0,801 | 0,568 | 2,037 | 1,1642 | 0,038 | 0,052 |

[a]Unidade m/m% para as propriedades S, NT, NB, SAT, ARO e POL e unidade mgKOHg$^{-1}$ para NAT.

### 4.3.6 Modelos a partir da fusão de nível médio por PCA

A fusão de nível médio foi aplicada por meio de duas estratégias diferentes: pela concatenação dos scores de modelos PLS e pela concatenação dos componentes principais de modelos PCA. PCA é um procedimento numérico que transforma as variáveis originais, geralmente fortemente correlacionadas, em um conjunto de variáveis não correlacionadas chamadas de componentes principais. Na fusão a nível médio, os componentes principais dos dados das diferentes técnicas analíticas são fundidos em uma única matriz.

Neste trabalho, optou-se por fundir os dez primeiros componentes principais de cada PCA que explicam 99,99%, 99,96% e 99,45% da variância dos dados de MIR, RMN $^{1}H$ e $^{13}C$, respectivamente. Os resultados estão apresentados na Tabela 4.5, com os principais parâmetros para cada propriedade físico-química.

Os modelos que proporcionaram os menores valores de $RMSE_P$ estão destacados em negrito na tabela. Para todos eles, os valores $RMSE_C$ e $RMSE_P$ são relativamente próximos, portanto, não há indicativo de superajuste aos dados de calibração. Quanto aos resíduos, nenhum dos modelos destacados apresentou erros sistemáticos, de acordo com o teste de viés. Já os modelos de NT e NAT apresentaram tendência quadrática e os modelos SAT e NAT tendênica linear nos resíduos uma vez que o p-valor, resultante do teste de permutação não paramétrico, foi menor que 0,05, como mostra a Tabela 4.5.

Em comparação com os modelos individuais, verifica-se que a fusão reduziu os valores de $RMSE_P$ de 0,255 para 0,209 $mgKOHg^{-1}$; de 4,26 para 3,59 m/m% e de 7,34 para 6,53 m/m% nos modelos de predição de NAT, ARO e POL, respectivamente. Para as demais propriedades, a fusão a baixo nível apresentou pior desempenho comparada aos modelos dos espectros individuais.

**Tabela 4.5.** Principais parâmetros estatísticos dos modelos provenientes da fusão a nível médio por PCA.

| | | nvl | [a]$RMSE_C$ | [a]$RMSE_P$ | $RMSE_P$% | $R_C^2$ | $R_P^2$ | $T_{tab,test}$ | $T_{cal,test}$ | p-valor (1º) | p-valor (2º) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| S | RMN $^{13}C$ | 6 | 0,055 | 0,081 | 24,03 | 0,936 | 0,797 | 2,028 | 0,657 | 0,36 | 0,50 |
| | MIR + RMN $^{1}H$ | 3 | 0,099 | 0,097 | 28,72 | 0,784 | 0,725 | 2,028 | 1,369 | 0,22 | 0,34 |
| | **MIR + RMN $^{13}C$** | **2** | **0,112** | **0,089** | **26,51** | **0,720** | **0,753** | **2,028** | **1,231** | **0,45** | **0,204** |
| | RMN $^{1}H$ + RMN $^{13}C$ | 8 | 0,107 | 0,110 | 32,50 | 0,762 | 0,647 | 2,028 | 1,565 | 0,25 | 0,259 |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 2 | 0,104 | 0,096 | 28,34 | 0,756 | 0,727 | 2,028 | 1,020 | 0,20 | 0,131 |
| NT | **RMN $^{1}H$** | 8 | 0,030 | 0,049 | 17,77 | 0,951 | 0,773 | 2,028 | 0,021 | 0,049 | 0,042 |
| | MIR + RMN $^{1}H$ | 3 | 0,040 | 0,060 | 21,89 | 0,910 | 0,682 | 2,028 | 1,068 | 0,012 | 0,002 |
| | MIR + RMN $^{13}C$ | 3 | 0,047 | 0,073 | 26,86 | 0,871 | 0,668 | 2,028 | 1,630 | 0,0 | 0,036 |
| | **RMN $^{1}H$ + RMN $^{13}C$** | **9** | **0,045** | **0,054** | **19,85** | **0,893** | **0,716** | **2,028** | **0,406** | **0,071** | **0,003** |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 3 | 0,040 | 0,059 | 21,68 | 0,909 | 0,710 | 2,028 | 1,432 | 0,003 | 0,003 |
| NB | RMN $^{1}H$ | 6 | 0,011 | 0,010 | 10,34 | 0,948 | 0,947 | 2,030 | 0,363 | 0,12 | 0,24 |
| | **MIR + RMN $^{1}H$** | **3** | **0,015** | **0,015** | **16,29** | **0,892** | **0,866** | **2,030** | **0,763** | **0,42** | **0,30** |
| | MIR + RMN $^{13}C$ | 2 | 0,019 | 0,019 | 20,09 | 0,823 | 0,792 | 2,030 | 1,230 | 0,36 | 0,002 |
| | RMN $^{1}H$ + RMN $^{13}C$ | 5 | 0,015 | 0,021 | 22,74 | 0,894 | 0,752 | 2,030 | 0,010 | 0,056 | 0,42 |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 4 | 0,013 | 0,018 | 19,48 | 0,919 | 0,821 | 2,030 | 0,163 | 0,19 | 0,49 |
| NAT | MIR | 5 | 0,279 | 0,255 | 58,64 | 0,910 | 0,799 | 2,028 | 1,120 | 0,003 | 0,000 |
| | **MIR + RMN $^{1}H$** | **10** | **0,288** | **0,209** | **47,96** | **0,910** | **0,913** | **2,028** | **0,341** | **0,0** | **0,002** |
| | MIR + RMN $^{13}C$ | 2 | 0,453 | 0,425 | 97,62 | 0,754 | 0,453 | 2,028 | 1,176 | 0,006 | 0,010 |
| | RMN $^{1}H$ + RMN $^{13}C$ | 9 | 0,338 | 0,244 | 56,05 | 0,874 | 0,835 | 2,028 | 1,426 | 0,002 | 0,000 |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 2 | 0,338 | 0,246 | 56,43 | 0,863 | 0,800 | 2,028 | 0,245 | 0,035 | 0,003 |
| SAT | RMN $^{1}H$ | 4 | 6,250 | 5,522 | 9,83 | 0,807 | 0,788 | 2,032 | 0,502 | 0,11 | 0,391 |
| | **MIR + RMN $^{1}H$** | **2** | **6,544** | **5,600** | **9,97** | **0,784** | **0,808** | **2,032** | **1,087** | **0,016** | **0,47** |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MIR + RMN $^{13}$C | 2 | 6,690 | 6,215 | 11,06 | 0,774 | 0,740 | 2,032 | 0,795 | 0,069 | 0,48 |
| | RMN $^{1}$H + RMN $^{13}$C | 2 | 6,788 | 5,938 | 10,57 | 0,767 | 0,783 | 2,032 | 1,077 | 0,011 | 0,42 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 2 | 6,582 | 5,687 | 10,12 | 0,781 | 0,805 | 2,032 | 0,876 | 0,008 | 0,45 |
| | RMN $^{13}$C | 6 | 1,805 | 4,255 | 18,55 | 0,930 | 0,589 | 2,028 | 0,142 | 0,011 | 0,007 |
| | MIR + RMN $^{1}$H | 3 | 3,028 | 3,933 | 17,14 | 0,795 | 0,617 | 2,028 | 0,264 | 0,11 | 0,109 |
| ARO | **MIR + RMN $^{13}$C** | **11** | **3,272** | **3,595** | **15,67** | **0,784** | **0,674** | **2,028** | **0,190** | **0,17** | **0,20** |
| | RMN $^{1}$H + RMN $^{13}$C | 3 | 3,502 | 4,311 | 18,79 | 0,725 | 0,552 | 2,028 | 1,191 | 0,052 | 0,29 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 3 | 2,971 | 3,823 | 16,66 | 0,802 | 0,630 | 2,028 | 1,039 | 0,16 | 0,27 |
| | **RMN $^{13}$C** | 5 | 3,505 | 7,337 | 36,86 | 0,838 | 0,555 | 2,037 | 1,003 | 0,46 | 0,12 |
| | MIR + RMN $^{1}$H | 2 | 4,465 | 7,321 | 36,78 | 0,736 | 0,565 | 2,037 | 0,082 | 0,11 | 0,11 |
| POL | MIR + RMN $^{13}$C | 3 | 4,591 | 7,217 | 36,25 | 0,714 | 0,558 | 2,037 | 0,774 | 0,36 | 0,45 |
| | **RMN $^{1}$H + RMN $^{13}$C** | **4** | **4,546** | **6,538** | **32,84** | **0,723** | **0,649** | **2,037** | **0,446** | **0,18** | **0,25** |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 10 | 4,041 | 7,254 | 36,44 | 0,799 | 0,558 | 2,037 | 0,123 | 0,26 | 0,45 |

[a]Unidade m/m% para as propriedades S, NT, NB, SAT, ARO e POL e unidade mgKOHg$^{-1}$ para NAT.

### 4.3.7 Modelos a partir da fusão de nível médio por PLS

O algoritmo PLS reduz as variáveis originais a um conjunto menor de variáveis não correlacionados, chamadas de variáveis latentes. Na fusão a nível médio, os scores dos modelos provenientes das diferentes fontes analíticas são fundidos em uma única matriz. Neste trabalho, foram feitas todas as combinações possíveis dos espectros, resultando nas quatro fusões: MIR + RMN $^{1}$H, MIR + RMN $^{13}$C, RMN $^{1}$H + RMN $^{13}$C e MIR + RMN $^{1}$H + RMN $^{13}$C.

A Tabela 4.6 mostra o número de variáveis latentes (nvl) utilizadas na fusão dos scores (nível médio PLS) para posterior construção dos modelos. Foram escolhidas as quantidades de variáveis latentes capazes de gerar os modelos mais bem ajustados.

**Tabela 4.6.** Número de variáveis latentes utilizadas na fusão dos modelos individuais.

| | **S** | **NT** | **NB** | **NAT** | **SAT** | **ARO** | **POL** |
|---|---|---|---|---|---|---|---|
| **MIR** | 12 | 8 | 10 | 20 | 6 | 7 | 10 |
| **RMN $^{1}$H** | 12 | 5 | 10 | 20 | 7 | 7 | 5 |
| **RMN $^{13}$C** | 12 | 5 | 10 | 20 | 7 | 7 | 5 |

Os resultados da modelagem de nível médio por PLS estão apresentados na Tabela 4.7, com os principais parâmetros dos modelos para cada propriedade físico-química. Os modelos que proporcionaram os menores valores de $RMSE_P$ estão destacados em negrito na tabela. Comparando estes resultados com os dos modelos individuais, verifica-se que esta estratégia reduziu o erro de predição dos modelos de todas as sete propriedades.

A Figura 4.17 apresenta os gráficos oriundos da validação cruzada, somente para os modelos individuais que apresentaram os menores valores de $RMSE_P$ (destacados em negrito na Tabela 4.7). Tais gráficos mostram os valores de $RMSE_{CV}$ em função do número de variáveis latentes que compõe o modelo. A análise destes gráficos permitiu escolher o número ideal de variáveis latentes que minimiza o erro de

previsão dos modelos. A Tabela 4.7 apresenta o nvl escolhido para compor cada um dos modelos fundidos.

Para os modelos destacados em negrito na Tabela 4.7, os valores $RMSE_C$ e $RMSE_P$ são relativamente próximos, obedecendo o critério $RMSE_P < 3.RMSE_C$, portanto, não há indicativo de superajuste aos dados de calibração, exceto para previsão de NAT cujo $RMSE_P$ foi 3,8 vezes maior que o $RMSE_C$.

Para estes mesmos modelos, a Figura 4.18 apresenta gráficos dos dados obtidos pelos experimentos padronizados *versus* os dados previstos pelos modelos. Os resultados de S, NB, NT, NAT e SAT apresentam uma boa linearidade, expressa por valores mais altos do coeficiente de determinação. Já para ARO e POL, valores mais baixos de coeficiente de determinação foram alcançados, respectivamente iguais a 0,67 e 0,65, e, portanto, uma menor linearidade pode ser vista nos gráficos.

Para os resíduos, mostrados nos gráficos da Figura 4.19, nenhum dos modelos apresentou erros sistemáticos, de acordo com o teste de viés. Também foi realizado o teste de permutação não paramétriconos nos residuos e os resultados do p-valor estão dispostos na Tabela 4.7. Os modelos de NB, NAT e ARO, em negrito, apresentaram tendência quadrática nos resíduos, visto que seu p-valor foi menor que 0,05. Além disso, o modelo de NB apresentou também tendência linear.

**Tabela 4.7.** Principais parâmetros estatísticos dos modelos provenientes da fusão a nível médio por PLS.

| | | nvl | [a]$RMSE_C$ | [a]$RMSE_P$ | $RMSE_P$% | $R_C^2$ | $R_P^2$ | $T_{tab,test}$ | $T_{cal,test}$ | p-valor (1º) | p-valor (2º) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| S | RMN $^{13}$C | 6 | 0,055 | 0,081 | 24,03 | 0,936 | 0,797 | 2,028 | 0,657 | 0,36 | 0,50 |
| | **MIR + RMN $^{1}$H** | **2** | **0,0537** | **0,0565** | **16,76** | **0,936** | **0,902** | **2,028** | **0,417** | **0,46** | **0,37** |
| | MIR + RMN $^{13}$C | 4 | 0,0229 | 0,0765 | 22,67 | 0,988 | 0,817 | 2,028 | 0,507 | 0,44 | 0,10 |
| | RMN $^{1}$H + RMN $^{13}$C | 4 | 0,0195 | 0,0756 | 22,40 | 0,992 | 0,825 | 2,028 | 0,385 | 0,26 | 0,464 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 4 | 0,0258 | 0,0608 | 18,03 | 0,985 | 0,887 | 2,028 | 0,3746 | 0,22 | 0,307 |
| NT | RMN $^{1}$H | 8 | 0,030 | 0,049 | 17,77 | 0,951 | 0,773 | 2,028 | 0,021 | 0,049 | 0,042 |
| | **MIR + RMN $^{1}$H** | **5** | **0,0231** | **0,0414** | **15,15** | **0,970** | **0,825** | **2,028** | **0,218** | **0,41** | **0,12** |
| | MIR + RMN $^{13}$C | 2 | 0,0187 | 0,0474 | 17,36 | 0,980 | 0,778 | 2,028 | 2,315 | 0,13 | 0,018 |
| | RMN $^{1}$H + RMN $^{13}$C | 2 | 0,0272 | 0,0543 | 19,90 | 0,957 | 0,722 | 2,028 | 2,248 | 0,037 | 0,046 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 2 | 0,0203 | 0,0436 | 15,98 | 0,976 | 0,804 | 2,028 | 2,060 | 0,37 | 0,035 |
| NB | RMN $^{1}$H | 6 | 0,011 | 0,010 | 10,34 | 0,948 | 0,947 | 2,030 | 0,363 | 0,12 | 0,24 |
| | **MIR + RMN $^{1}$H** | **3** | **0,0059** | **0,0067** | **7,18** | **0,984** | **0,977** | **2,030** | **0,117** | **0,018** | **0,026** |
| | MIR + RMN $^{13}$C | 6 | 0,0032 | 0,0161 | 17,34 | 0,996 | 0,853 | 2,030 | 0,578 | 0,31 | 0,043 |
| | RMN $^{1}$H + RMN $^{13}$C | 3 | 0,0033 | 0,0117 | 12,58 | 0,995 | 0,921 | 2,030 | 0,931 | 0,36 | 0,14 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 3 | 0,0037 | 0,0093 | 10,07 | 0,994 | 0,950 | 2,030 | 0,168 | 0,29 | 0,084 |
| NAT | MIR | 5 | 0,279 | 0,255 | 58,64 | 0,910 | 0,799 | 2,028 | 1,120 | 0,003 | 0,000 |
| | MIR + RMN $^{1}$H | 5 | 0,0647 | 0,220 | 50,58 | 0,995 | 0,825 | 2,028 | 1,910 | 0,26 | 0,0 |
| | **MIR + RMN $^{13}$C** | **5** | **0,0420** | **0,160** | **36,67** | **0,998** | **0,907** | **2,028** | **1,022** | **0,49** | **0,0012** |
| | RMN $^{1}$H + RMN $^{13}$C | 6 | 0,0292 | 0,258 | 59,29 | 0,999 | 0,767 | 2,028 | 0,078 | 0,31 | 0,025 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 9 | 0,0308 | 0,222 | 50,96 | 1,000 | 0,827 | 2,028 | 0,070 | 0,30 | 0,0 |
| SAT | RMN $^{1}$H | 4 | 6,250 | 5,522 | 9,83 | 0,807 | 0,788 | 2,032 | 0,502 | 0,11 | 0,391 |
| | **MIR + RMN $^{1}$H** | **2** | **4,782** | **4,730** | **8,42** | **0,885** | **0,839** | **2,032** | **0,195** | **0,20** | **0,20** |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MIR + RMN $^{13}$C | 8 | 2,363 | 5,967 | 10,62 | 0,974 | 0,740 | 2,032 | 0,464 | 0,34 | 0,49 |
| | RMN $^{1}$H + RMN $^{13}$C | 2 | 2,891 | 5,073 | 9,03 | 0,958 | 0,819 | 2,032 | 0,500 | 0,14 | 0,52 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 2 | 2,984 | 4,805 | 8,55 | 0,955 | 0,840 | 2,032 | 0,354 | 0,079 | 0,27 |
| | RMN $^{13}$C | 6 | 1,805 | 4,255 | 18,55 | 0,930 | 0,589 | 2,028 | 0,142 | 0,011 | 0,007 |
| | MIR + RMN $^{1}$H | 2 | 2,359 | 4,064 | 17,71 | 0,874 | 0,604 | 2,028 | 0,067 | 0,071 | 0,030 |
| ARO | **MIR + RMN $^{13}$C** | **2** | **1,444** | **3,658** | **15,94** | **0,953** | **0,671** | **2,028** | **0,047** | **0,099** | **0,031** |
| | RMN $^{1}$H + RMN $^{13}$C | 3 | 1,355 | 4,045 | 17,63 | 0,959 | 0,620 | 2,028 | 0,492 | 0,020 | 0,002 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 3 | 1,429 | 3,954 | 17,23 | 0,954 | 0,632 | 2,028 | 0,516 | 0,029 | 0,005 |
| | RMN $^{13}$C | 5 | 3,505 | 7,337 | 36,86 | 0,838 | 0,555 | 2,037 | 1,003 | 0,46 | 0,12 |
| | MIR + RMN $^{1}$H | 3 | 3,648 | 6,541 | 32,86 | 0,820 | 0,639 | 2,037 | 1,185 | 0,48 | 0,49 |
| POL | MIR + RMN $^{13}$C | 3 | 3,099 | 6,880 | 34,56 | 0,870 | 0,613 | 2,037 | 0,022 | 0,20 | 0,29 |
| | RMN $^{1}$H + RMN $^{13}$C | 2 | 3,477 | 6,577 | 33,04 | 0,834 | 0,634 | 2,037 | 0,494 | 0,34 | 0,31 |
| | **MIR + RMN $^{1}$H + RMN $^{13}$C** | **3** | **3,112** | **6,500** | **32,66** | **0,869** | **0,657** | **2,037** | **0,469** | **0,13** | **0,43** |

[a]Unidade m/m% para as propriedades S, NT, NB, SAT, ARO e POL e unidade mgKOHg$^{-1}$ para NAT.

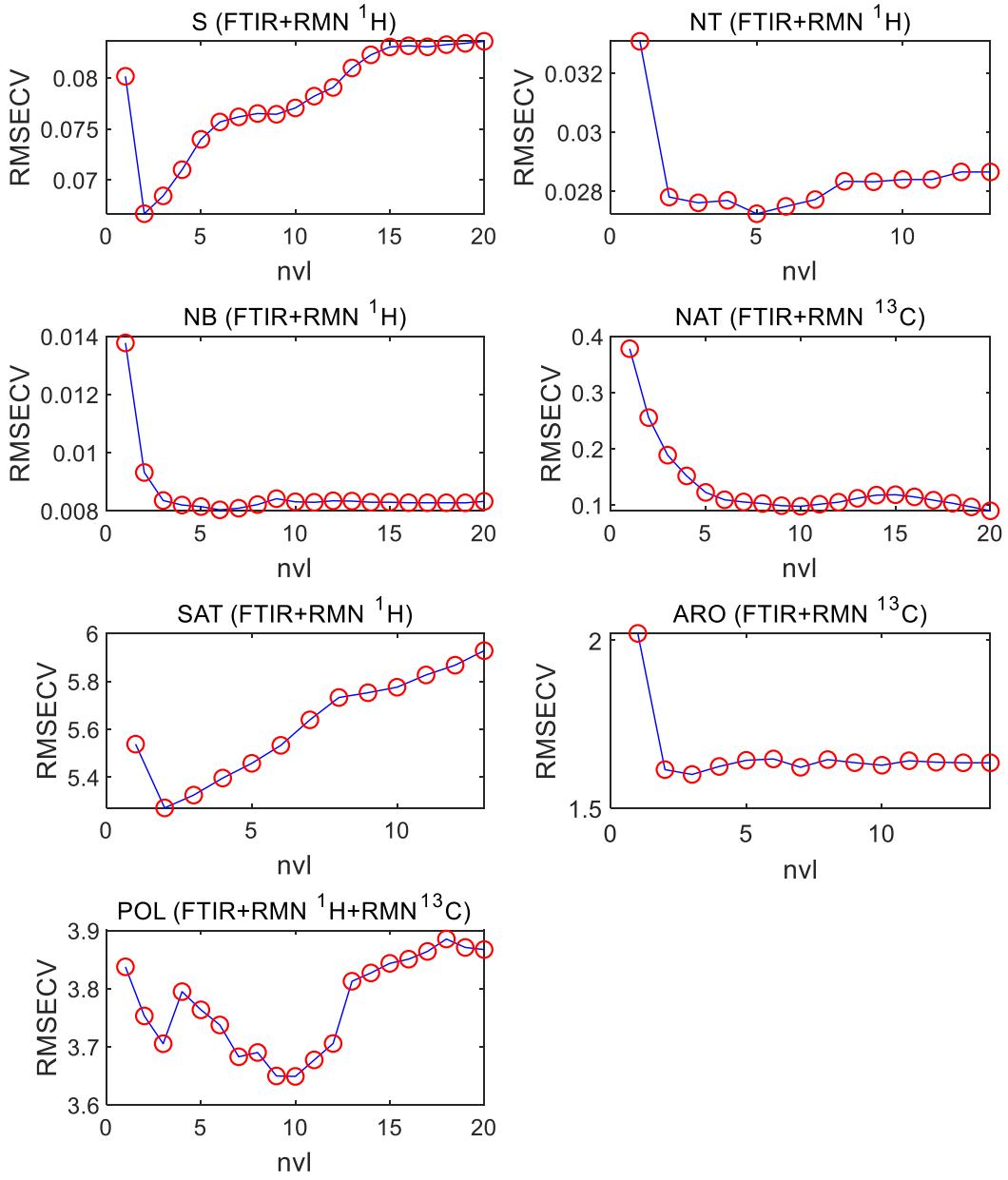



**Figura 4.17.** $RMSE_{CV}$ *versus* número de variáveis latentes para os modelos fundidos a nível médio com PLS. Unidade de $RMSE_{CV}$ m/m% para as propriedades S, NT, NB, SAT, ARO e POL e $mgKOHg^{-1}$ para NAT.

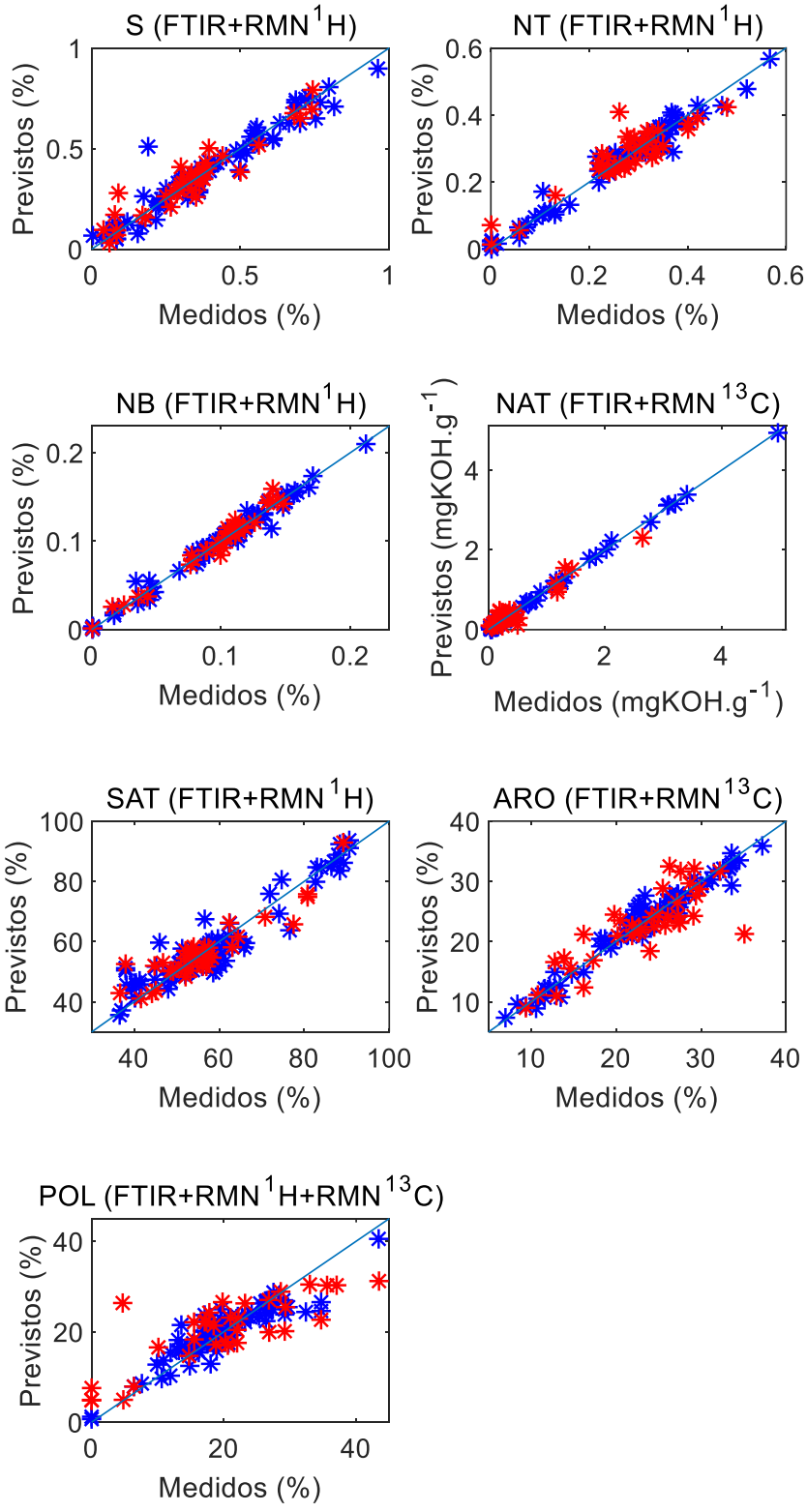



**Figura 4.18.** Gráficos de correlação entre os valores de referência experimentais e os valores preditos pelos modelos fundidos a nível médio com PLS. Legenda: azul - amostras de calibração, vermelho - amostras teste.

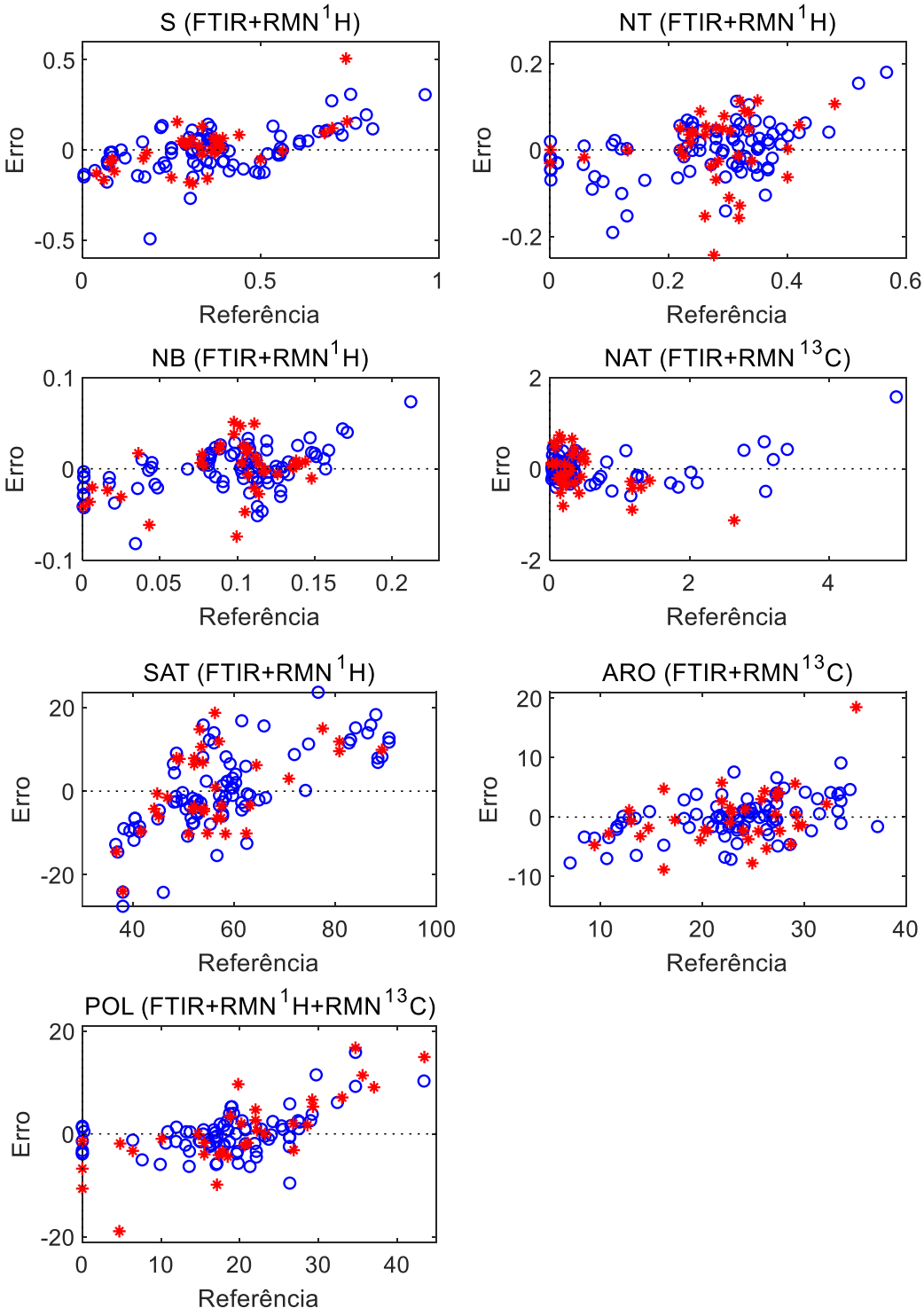



**Figura 4.19.** Distribuição dos resíduos dos modelos fundidos a nível médio com PLS. Legenda: azul - amostras de calibração, vermelho - amostras teste.

### 4.3.8 Modelos a partir da fusão de nível alto

Na fusão de nível alto, as previsões geradas pelos modelos individuais foram combinadas por meio de uma média simples para gerar uma previsão final para as amostras de calibração. Neste trabalho, foram feitas todas as combinações possíveis, resultando nas quatro fusões: MIR + RMN $^{1}$H, MIR + RMN $^{13}$C, RMN $^{1}$H + RMN $^{13}$C e MIR + RMN $^{1}$H + RMN $^{13}$C.

Os resultados da modelagem de nível alto estão apresentados na Tabela 4.8, com os principais parâmetros dos modelos para cada propriedade físico-química. Os modelos que proporcionaram os menores valores de $RMSE_P$ estão destacados em negrito. Em comparação com os modelos individuais, verifica-se que esta estratégia reduziu ligeiramente o $RMSE_P$ de 0,0096 para 0,0091 m/m%, para NB; de 0,26 para 0,20 m/m%, para NAT; de 5,52 para 5,29 $mgKOHg^{-1}$, para SAT; de 4,26 para 3,99 m/m%, para ARO e de 7,34 para 7,07 m/m%, para POL.

Para os modelos destacados na Tabela 4.8, os valores $RMSE_C$ e $RMSE_P$ são relativamente próximos, obedecendo o critério $RMSE_P < 3.RMSE_C$, portanto, não há superajuste aos dados de calibração para nenhuma das propriedades. Quanto aos resíduos, nenhum dos modelos destacados apresentou erros sistemáticos, de acordo com o teste de viés. Já os modelos de NB e SAT apresentaram tendência linear nos resíduos, o modelo de NAT apresentou tendência quadrática e o modelo de NT tendência linear e quadrática, uma vez que o p-valor, resultante do teste de permutação não paramétrico, foi menor que 0,05 nestes casos, como mostra a Tabela 4.8.

**Tabela 4.8.** Principais parâmetros estatísticos dos modelos provenientes da fusão a nível alto.

| | | [a]$RMSE_C$ | [a]$RMSE_P$ | $RMSE_P$% | $R_C^2$ | $R_P^2$ | $T_{tab,test}$ | $T_{cal,test}$ | p-valor (1º) | p-valor (2º) |
|---|---|---|---|---|---|---|---|---|---|---|
| S | RMN $^{13}C$ | 6 | 0,055 | 0,081 | 24,03 | 0,936 | 0,797 | 2,028 | 0,657 | 0,36 |
| | MIR + RMN $^{1}H$ | 0,0982 | 0,0910 | 26,97 | 0,775 | 0,741 | 2,028 | 0,280 | 0,45 | 0,39 |
| | MIR + RMN $^{13}C$ | 0,0676 | 0,0874 | 25,92 | 0,896 | 0,761 | 2,028 | 0,1767 | 0,40 | 0,17 |
| | **RMN $^{1}H$ + RMN $^{13}C$** | **0,0668** | **0,0807** | **23,91** | **0,899** | **0,802** | **2,028** | **0,787** | **0,26** | **0,33** |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 0,0753 | 0,0848 | 25,13 | 0,871 | 0,777 | 2,028 | 0,409 | 0,32 | 0,38 |
| NT | RMN $^{1}H$ | 8 | 0,030 | 0,049 | 17,77 | 0,951 | 0,773 | 2,028 | 0,021 | 0,049 |
| | MIR + RMN $^{1}H$ | 0,0249 | 0,0476 | 17,43 | 0,963 | 0,785 | 2,028 | 0,189 | 0,032 | 0,021 |
| | MIR + RMN $^{13}C$ | 0,0211 | 0,0567 | 20,76 | 0,973 | 0,715 | 2,028 | 0,410 | 0,011 | 0,004 |
| | RMN $^{1}H$ + RMN $^{13}C$ | 0,0211 | 0,0538 | 19,71 | 0,973 | 0,731 | 2,028 | 0,621 | 0,031 | 0,005 |
| | **MIR + RMN $^{1}H$ + RMN $^{13}C$** | **0,0209** | **0,0506** | **18,55** | **0,974** | **0,759** | **2,028** | **0,313** | **0,030** | **0,006** |
| NB | RMN $^{1}H$ | 6 | 0,011 | 0,010 | 10,34 | 0,948 | 0,947 | 2,030 | 0,363 | 0,12 |
| | **MIR + RMN $^{1}H$** | **0,0086** | **0,0091** | **9,80** | **0,965** | **0,955** | **2,030** | **0,044** | **0,026** | **0,25** |
| | MIR + RMN $^{13}C$ | 0,0075 | 0,0144 | 15,55 | 0,973 | 0,880 | 2,030 | 0,571 | 0,13 | 0,18 |
| | RMN $^{1}H$ + RMN $^{13}C$ | 0,0077 | 0,0132 | 14,19 | 0,972 | 0,898 | 2,030 | 0,865 | 0,42 | 0,096 |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 0,0072 | 0,0116 | 12,57 | 0,976 | 0,921 | 2,030 | 0,571 | 0,20 | 0,16 |
| NAT | MIR | 5 | 0,279 | 0,255 | 58,64 | 0,910 | 0,799 | 2,028 | 1,120 | 0,003 |
| | **MIR + RMN $^{1}H$** | **0,234** | **0,198** | **45,41** | **0,933** | **0,863** | **2,028** | **1,171** | **0,27** | **0,0** |
| | MIR + RMN $^{13}C$ | 0,185 | 0,215 | 49,47 | 0,959 | 0,832 | 2,028 | 0,306 | 0,29 | 0,0 |
| | RMN $^{1}H$ + RMN $^{13}C$ | 0,215 | 0,275 | 63,12 | 0,943 | 0,735 | 2,028 | 1,248 | 0,44 | 0,0 |
| | MIR + RMN $^{1}H$ + RMN $^{13}C$ | 0,194 | 0,213 | 48,98 | 0,955 | 0,839 | 2,028 | 0,997 | 0,29 | 0,0 |
| SAT | RMN $^{1}H$ | 4 | 6,250 | 5,522 | 9,83 | 0,807 | 0,788 | 2,032 | 0,502 | 0,11 |
| | **MIR + RMN $^{1}H$** | **5,750** | **5,294** | **9,43** | **0,827** | **0,812** | **2,032** | **0,128** | **0,048** | **0,44** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MIR + RMN $^{13}$C | 3,920 | 5,620 | 10,01 | 0,921 | 0,779 | 2,032 | 0,243 | 0,11 | 0,14 |
| | RMN $^{1}$H + RMN $^{13}$C | 4,084 | 5,657 | 10,07 | 0,914 | 0,774 | 2,032 | 0,368 | 0,15 | 0,15 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 4,452 | 5,347 | 9,52 | 0,898 | 0,806 | 2,032 | 0,0021 | 0,055 | 0,29 |
| | RMN $^{13}$C | 6 | 1,805 | 4,255 | 18,55 | 0,930 | 0,589 | 2,028 | 0,142 | 0,011 |
| | MIR + RMN $^{1}$H | 2,848 | 4,135 | 18,03 | 0,810 | 0,588 | 2,028 | 0,770 | 0,066 | 0,13 |
| ARO | MIR + RMN $^{13}$C | 2,050 | 4,215 | 18,38 | 0,904 | 0,591 | 2,028 | 0,0191 | 0,016 | 0,012 |
| | **RMN $^{1}$H + RMN $^{13}$C** | **1,975** | **3,993** | **17,41** | **0,910** | **0,609** | **2,028** | **0,623** | **0,085** | **0,076** |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 2,211 | 4,011 | 17,49 | 0,888 | 0,608 | 2,028 | 0,475 | 0,063 | 0,052 |
| | RMN $^{13}$C | 5 | 3,505 | 7,337 | 36,86 | 0,838 | 0,555 | 2,037 | 1,003 | 0,46 |
| | MIR + RMN $^{1}$H | 4,529 | 7,240 | 36,37 | 0,710 | 0,564 | 2,037 | 0,504 | 0,23 | 0,078 |
| POL | **MIR + RMN $^{13}$C** | **3,588** | **7,072** | **35,53** | **0,818** | **0,577** | **2,037** | **0,482** | **0,48** | **0,11** |
| | RMN $^{1}$H + RMN $^{13}$C | 3,960 | 7,425 | 37,31 | 0,781 | 0,549 | 2,037 | 1,025 | 0,31 | 0,072 |
| | MIR + RMN $^{1}$H + RMN $^{13}$C | 3,919 | 7,147 | 35,91 | 0,786 | 0,572 | 2,037 | 0,682 | 0,37 | 0,085 |

[a]Unidade m/m% para as propriedades S, NT, NB, SAT, ARO e POL e unidade mgKOHg$^{-1}$ para NAT.

### 4.3.9 Comparação entre os modelos

A Tabela 4.9 resume os principais resultados daqueles modelos que apresentaram os menores valores de $RMSE_P$, para cada propriedade. Estes modelos foram destacados em negrito nas Tabelas 4.2, 4.4, 4.5, 4.7 e 4.8 e já foram discutidos nos tópicos anteriores. O objetivo deste tópico é fazer uma comparação mais clara dos modelos individuais com as estratégias de fusão.

Comparando-se com os modelos individuais, a fusão de nível baixo conseguiu reduzir o $RMSE_P$ apenas para as propriedades NB, SAT e ARO. As demais propriedades obtiveram valores de $RMSE_P$ aproximadamente iguais ou maiores com a aplicação da fusão. O teste randômico (VOET, 1994), foi aplicado para comparar os valores de $RMSE_P$ e verificar se essas diferenças são estatisticamente significativas. Os resultados do p-valor dos testes foram 0,22, 0,44 e 0,052, respectivamente, para NB, SAT e ARO, indicando que não há diferenças significativas entre a exatidão do modelo individual e do fundido, para as três propriedades.

Comparando-se com os modelos individuais, a fusão de nível médio por PCA conseguiu reduzir o $RMSE_P$ apenas para as propriedades NAT, ARO e POL. As demais propriedades obtiveram valores de $RMSE_P$ aproximadamente iguais ou maiores com a aplicação da fusão. O teste randômico (VOET, 1994), foi aplicado para comparar os valores de $RMSE_P$ e verificar se essas diferenças são estatisticamente significativas. Os resultados do p-valor dos testes foram 0,23, 0,032 e 0,054, respectivamente, para NAT, ARO e POL, indicando que há diferença significativa apenas entre a exatidão do modelo individual e do fundido para a propriedade ARO. Dessa forma, a fusão de nível médio por PCA conseguiu reduzir o $RMSE_P$ de forma estatisticamente significativa apenas para a propriedade ARO.

Tendo em vista os resultados, pode-se afirmar que, de forma geral, assim como a fusão de nível baixo, a fusão de nível médio por PCA não foi capaz de capturar a complementaridade e o sinergismo entre os dados multi-espectrais. PEINDER *et al.* (2009) obtiveram conclusões semelhantes ao usar infravermelho, RMN $^{1}H$ e $^{13}C$ fundidos a nível baixo e médio para estimar sete propriedades do petróleo bruto. Seus modelos baseados em espectros fundidos não levaram a melhorias significativas, em comparação com os espectros individuais.

**Tabela 4.9.** Principais resultados do modelo ótimo de cada propriedade.

| Prop. | Estratégia | Espectro | [a]$RMSE_P$ | $R_P^2$ | Tendência linear | Tendência quadrática | Erro sistemático |
|---|---|---|---|---|---|---|---|
| S | Sem fusão | RMN $^{13}C$ | 0,081 | 0,80 | Não | Não | Não |
| | Fusão baixa | RMN $^{1}H$+$^{13}C$ | 0,083 | 0,80 | Não | Não | Não |
| | Fusão média PCA | MIR+RMN $^{13}C$ | 0,089 | 0,75 | Não | Não | Não |
| | Fusão média PLS | MIR+RMN $^{1}H$ | 0,057 | 0,90 | Não | Não | Não |
| | Fusão alta | RMN $^{1}H$+$^{13}C$ | 0,081 | 0,80 | Não | Não | Não |
| NT | Sem fusão | RMN $^{1}H$ | 0,049 | 0,77 | Sim | Sim | Não |
| | Fusão baixa | MIR+RMN $^{1}H$ | 0,049 | 0,77 | Sim | Sim | Não |
| | Fusão média PCA | RMN $^{1}H$+$^{13}C$ | 0,054 | 0,72 | Não | Sim | Não |
| | Fusão média PLS | MIR+RMN $^{1}H$ | 0,041 | 0,83 | Não | Não | Não |
| | Fusão alta | MIR+RMN $^{1}H$+$^{13}C$ | 0,051 | 0,76 | Sim | Sim | Não |
| NB | Sem fusão | RMN $^{1}H$ | 0,0096 | 0,95 | Não | Não | Não |
| | Fusão baixa | MIR+RMN $^{1}H$ | 0,0090 | 0,96 | Sim | Não | Não |
| | Fusão média PCA | MIR+RMN $^{1}H$ | 0,0150 | 0,87 | Não | Não | Não |
| | Fusão média PLS | MIR+RMN $^{1}H$ | 0,0067 | 0,98 | Sim | Sim | Não |
| | Fusão alta | MIR+RMN $^{1}H$ | 0,0091 | 0,96 | Sim | Não | Não |
| NAT | Sem fusão | MIR | 0,26 | 0,80 | Sim | Sim | Não |
| | Fusão baixa | MIR+RMN $^{1}H$+$^{13}C$ | 0,25 | 0,79 | Não | Sim | Não |
| | Fusão média PCA | MIR+RMN $^{1}H$ | 0,21 | 0,91 | Sim | Sim | Não |
| | Fusão média PLS | MIR+RMN $^{13}C$ | 0,16 | 0,91 | Não | Sim | Não |
| | Fusão alta | MIR+RMN $^{1}H$ | 0,20 | 0,86 | Não | Sim | Não |
| SAT | Sem fusão | RMN $^{1}H$ | 5,52 | 0,79 | Não | Não | Não |
| | Fusão baixa | MIR+RMN $^{1}H$+$^{13}C$ | 5,43 | 0,79 | Não | Não | Não |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Fusão média PCA | MIR+RMN $^{1}H$ | 5,60 | 0,81 | Sim | Não | Não |
| | Fusão média PLS | MIR+RMN $^{1}H$ | 4,73 | 0,84 | Não | Não | Não |
| | Fusão alta | MIR+RMN $^{1}H$ | 5,29 | 0,81 | Sim | Não | Não |
| | Sem fusão | RMN $^{13}C$ | 4,26 | 0,59 | Sim | Sim | Não |
| | Fusão baixa | MIR+RMN $^{1}H$+$^{13}C$ | 3,88 | 0,63 | Sim | Sim | Não |
| ARO | Fusão média PCA | MIR+RMN $^{13}C$ | 3,59 | 0,67 | Não | Não | Não |
| | Fusão média PLS | MIR+RMN $^{13}C$ | 3,66 | 0,67 | Sim | Não | Não |
| | Fusão alta | RMN $^{1}H$+$^{13}C$ | 3,99 | 0,61 | Não | Não | Não |
| | Sem fusão | RMN $^{13}C$ | 7,34 | 0,56 | Não | Não | Não |
| | Fusão baixa | RMN $^{1}H$+$^{13}C$ | 7,56 | 0,56 | Sim | Sim | Não |
| POL | Fusão média PCA | RMN $^{1}H$+$^{13}C$ | 6,54 | 0,65 | Não | Não | Não |
| | Fusão média PLS | MIR+RMN $^{1}H$+$^{13}C$ | 6,50 | 0,66 | Não | Não | Não |
| | Fusão alta | MIR+RMN $^{13}C$ | 7,07 | 0,58 | Não | Não | Não |

[a]Unidade m/m% para as propriedades S, NT, NB, SAT, ARO e POL e unidade $mgKOHg^{-1}$ para NAT.

O mesmo não aconteceu no trabalho de CASALE *et al*. (2010). Para testar a autenticidade de amostras de azeite de oliva, os autores construíram modelos através da fusão de scores da PCA dos espectros de massas, de UV-Vísível e de NIR. Os resultados do modelo fundido foram comparados com os dos modelos individuais e o potencial da fusão de nível médio por PCA foi demonstrado.

VERA *et al*. (2011) também conseguiram melhores resultados, não somente com a fusão de nível médio por PCA, como também com a fusão de nível baixo. Foram construídos modelos para classificar amostras de cerveja da mesma marca, de acordo com o local de fabricação, utilizando dados fundidos e separados de espectrometria de massas – nariz eletrônico (MS e-nose), infravermelho médio – língua óptica (IR - *optical-tongue*) e UV–visível. A porcentagem de classificação correta com a fusão de nível baixo e médio foi maior em comparação com resultados obtidos a partir das técnicas individuais.

Já os modelos de fusão de nível médio por PLS, comparando com todas as demais estratégias, apresentam maior linearidade e exatidão, com os maiores coeficientes de determinação ($R_P^2$) e os menores erros de previsão ($RMSE_P$) para todas as propriedades, exceto para aromáticos cujo melhor modelo foi alcançado por meio da fusão a nível médio por PCA. Comparando-se os modelos da fusão de nível médio por PLS com os modelos individuais, o $RMSE_P$ foi reduzido cerca de 29,6; 16,3; 30,2; 38,4; 14,3; 14,1 e 11,4% para as propriedades S, NT, NB, NAT, SAT, ARO e POL, respectivamente.

O teste randômico (VOET, 1994) foi aplicado para comparar as exatidões entre os modelos individuais e os modelos da fusão a nível médio por PLS e verificar se essas diferenças são estatisticamente significativas. Os resultados do p-valor dos testes foram 0,024, 0,23, 0,11, 0,024, 0,06, 0,035 e 0,11, respectivamente, para S, NT, NB, NAT, SAT, ARO e POL, indicando que há diferença significativa entre a exatidão do modelo individual e do fundido para as propriedades S, NAT e ARO. Apesar da considerável redução percentual (30,2%) na estimativa de NB, o teste não considerou estastisticamente diferentes os valores de $RMSE_P$ para o modelo individual e fundido a nível médio.

Em relação à fusão de nível alto, comparando-se com os modelos individuais, houve redução dos valores de $RMSE_P$ para as propriedades NB, NAT, SAT, ARO e

POL. As demais propriedades obtiveram valores de $RMSE_P$ aproximadamente iguais ou maiores com a aplicação da fusão. O teste randômico (VOET, 1994), foi aplicado para comparar estes valores de $RMSE_P$ e verificar se essas diferenças são estatisticamente significativas. Os resultados do p-valor dos testes foram 0,25, 0,13, 0,36, 0,22 e 0,36, respectivamente, para NB, NAT, SAT, ARO e POL, indicando que não há diferenças significativas entre a exatidão do modelo individual e do fundido, para as cinco propriedades.

Os resultados mostram que as fusões de nível baixo, nível médio por PCA e nível alto não trouxeram melhorias significativa nas exatidões dos modelos. Já a fusão de nível médio por PLS reduziu, para todas as propriedades, o valor de $RMSE_P$, entretanto somente para S, NAT e ARO o teste randômico acusou diferenças significativas entre as exatidões dos modelos individuais e fundidos.

Além de melhorar a capacidade preditiva, a fusão de nível médio por PLS reduziu o número de variáveis das matrizes originais para um número pequeno, variando de apenas 2 a 5 variáveis latentes, conforme mostrado na Tabela 4.7, produzindo modelos com relativa simplicidade. Enquanto isso, os modelos não fundidos utilizaram de 4 a 8 variáveis latentes (Tabela 4.2).

A fusão a nível baixo aumentou consideravelmente o número de variáveis e, com isso, o tempo de processamento dos dados, principalmente quando incluído dados de RMN $^{13}C$. As fusões de nível médio e alto, por outro lado, se mostraram extremamente mais rápidas.

De forma geral, exceto para aromáticos e polares, foi possível obter, para todas as propriedades, altos coeficientes de determinação, com valores de $R_P^2$ acima de 0,8, com destaque para nitrogênio básico que apresentou um $R_P^2$ igual a 0,98.

As propriedades que apresentaram maiores valores de $\%RMSE_P$ foram NAT e POL, cujos melhores modelos tiveram valores iguais a 36,67 e 32,66%, respectivamente. O alto erro médio percentual de previsão de teor de componentes polares pode ser decorrente do erro associado ao procedimento experimental. O ensaio para determinação do teor de polares consiste no fracionamento do petróleo por cromatografia líquida. Essa metodologia é sujeita a perda de asfaltenos que são adsorvidos na coluna de sílica gel. Já o alto erro médio percentual de previsão de número de acidez total pode estar associado a não linearidade entre a relação da

propriedade preditora e as variáveis independentes, uma vez que os resíduos de todos os modelos NAT da Tabela 4.9 apresentaram tendência quadrática.

Após as etapas de calibração e predição, os resíduos de predição foram avaliados por testes estatísticos. Para nenhum dos modelos verifica-se a presença de erros sistemáticos nos resíduos, entretanto tendência é detectada em vários casos. As únicas propriedades que não apresentaram tendência quadrática nos resíduos dos modelos da Tabela 4.9 foram S e SAT. Enquanto isso, NAT apresentou tendência quadrática em todos os casos, indicando que uma regressão não linear pode proporcionar um melhor ajuste aos dados e com isso uma redução do erro percentual de previsão que, neste estudo, foi o maior entre todas as propriedades. A única propriedade que não apresentou tendência linear nos resíduos de todos os modelos da Tabela 4.9 foi S. Para as demais propriedades, alguns modelos apresentaram tendência linear nos resíduos, outros não.

A tendência, linear ou quadrática, mostra que o modelo não foi capaz de explicar toda a variância contida nos dados experimentais. Por consequência, essa variância não explicada acaba sendo refletida nos resíduos. Isso ocorre geralmente quando a relação entre a variável preditora e os dados espectrais é não linear, mas utiliza-se de um método de regressão linear devido a sua maior simplicidade, como no caso deste trabalho. Cabe salientar que um modelo com tendência nos resíduos não é necessariamente um modelo ruim, uma vez que sua capacidade de previsão pode ser relativamente alta e, portanto, satisfatória. A tendência quadrática ou linear nos resíduos indica a possibilidade de se alcançar modelos ainda melhores lançando mão de métodos não lineares ou aperfeiçoando os lineares, respectivamente.

Os modelos de fusão de nível médio por PLS para previsões de S, NAT, SAT, ARO e POL apresentaram maior exatidão que os modelos encontrados na literatura. Para estimar o teor de enxofre, PEINDER *et al.,* (2009) obteve um $RMSE_P$ igual a 0,25 m/m% utilizando espectros de MIR e regressão com PLS, enquanto o mínimo alcançado neste trabalho foi de 0,057 m/m% (Tabela 4.9). BARBOSA *et al*. (2016) também previu o teor total de S com um $RMSE_P$ igual a 0,51 m/m%, usando RMN de baixo campo associada à regressão linear múltipla.

Para a determinação de NAT, um valor muito menor de $RMSE_P$ foi alcançado neste trabalho (0,16 $mgKOH \cdot g^{-1}$), em comparação com os resultados de BARBOSA *et al*. (2016) cujo $RMSE_P$ mais baixo foi 0,3 $mgKOH \cdot g^{-1}$.

Enquanto foram obtidos valores de $RMSE_P$ iguais a 4,73 e 6,50 m/m% para saturados e polares (Tabela 4.9), respectivamente, aplicando a fusão de dados, FILGUEIRAS *et al*. (2016) obtiveram 4,4 e 3,7 m/m%, respectivamente, utilizando RMN $^{13}C$. Porém, para ARO, foi obtido um modelo mais exato com $RMSE_P$ igual a 3,59 (Tabela 4.9) enquanto FILGUEIRAS *et al*. (2016) obtiveram 4,3 m/m%. RODRIGUES *et al*. (2017) também obtiveram modelos menos exatos com valores de $RMSE_P$ iguais a 6,7 e 4,05 m/m% para previsão de SAT e ARO, respectivamente, a partir de dados de cromatografia gasosa.

Não foram encontrados estudos que estimaram o teor de nitrogênio total em petróleo bruto, na literatura. Para prever NB, o menor valor de $RMSE_P$ alcançado neste estudo foi igual a 0,0067 m/m%. Enquanto isso, DUARTE *et al*. (2016) conseguiu um $RMSE_P$ maior, igual a 0,009 %m/m, utilizando RMN $^{1}H$ e PLS, para previsão de NB. TERRA *et al*. (2015), por sua vez, conseguiu um modelo mais exato, com $RMSE_P$ igual a 0,0012 m/m%, utilizando espectroscopia de massa, seleção variáveis e PLS. Os autores, entretanto, utilizaram 70 amostras com teor de NB variando de 0,0011 a 0,17 m/m%. Neste trabalho, os modelos foram desenvolvidos utilizando amostras com uma maior variedade de teor de NB (de 0,0005 a 0,26 m/m%) e um total de 126 amostras, o que o torna o modelo mais robusto.

Além da alta exatidão, vale destacar que os modelos foram obtidos com o algoritmo PLS, que é um método linear e, portanto, rápido e relativamente simples. Não foi necessário utilizar seleção de variáveis e/ou métodos não lineares que muitas vezes consomem tempo e requerem alto processamento computacional.

### 4.3.10 Análise de variáveis

Em um modelo linear, as variáveis com os maiores coeficientes apresentam maior contribuição para construção dos modelos e, portanto, maior correlação com a propriedade estimada. Já as variáveis com coeficientes próximos a zero não apresentam muita relevância para o modelo.

A análise das variáveis teve o objetivo de verificar a contribuição e a importância das variáveis originais dentro dos modelos para previsão das propriedades propostas. Para auxiliar na análise, as Figuras 4.20a, 4.21a e 4.22a mostram, respectivamente, os espectros de MIR, RMN $^{1}H$ e RMN $^{13}C$, de todas as amostras de petróleo. As

Figuras 4.20b, 4.20c, e 4.20d mostram uma sobreposição dos coeficientes de regressão das primeiras variáveis latentes (utilizadas na fusão) de modelos individuais de MIR, para predição do teor de saturados, aromáticos e polares, respectivamente. As Figuras 4.21b, 4.21c e 4.21d mostram o mesmo, mas para os modelos de RMN $^{1}H$ e as Figuras 4.22b, 4.22c e 4.22d mostram o mesmo, mas para os modelos de RMN $^{13}C$.

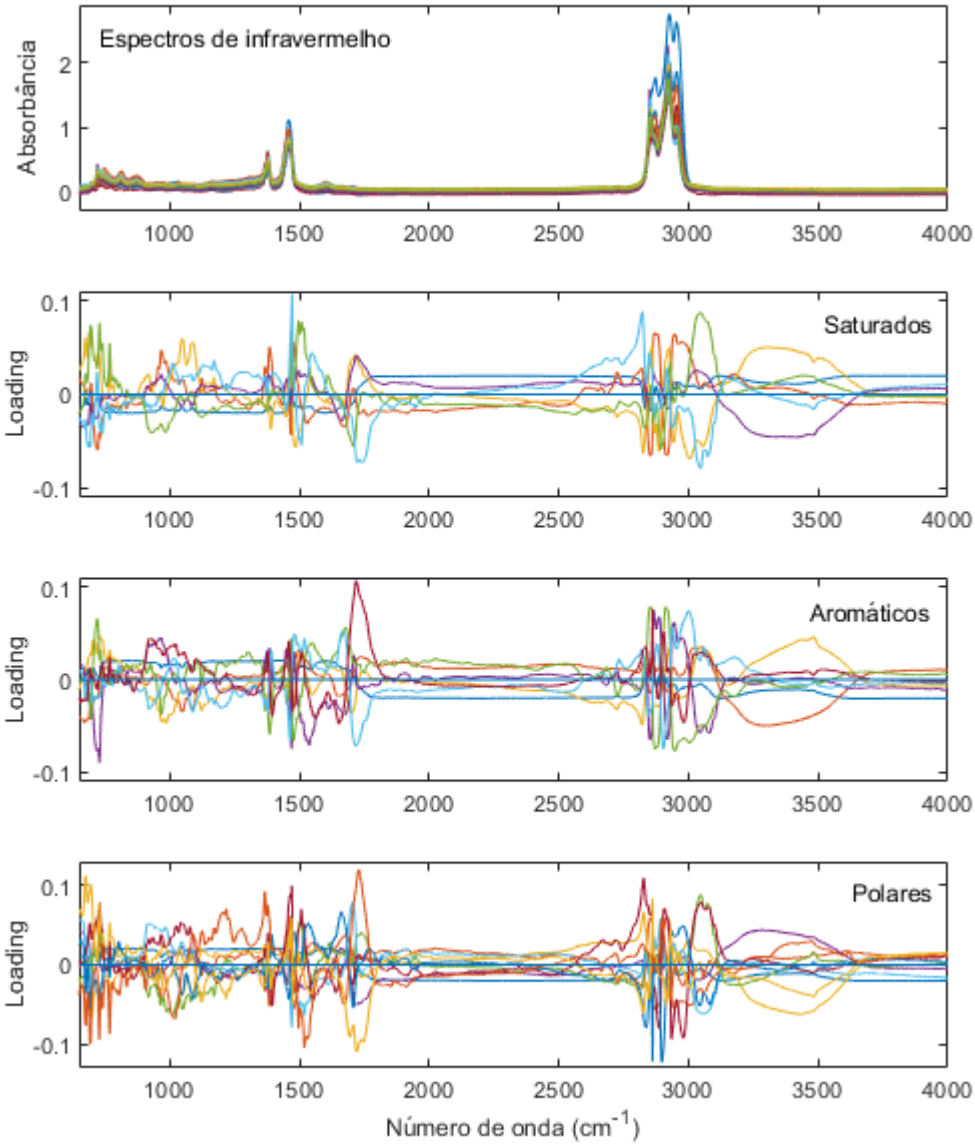



**Figura 4.20.** Espectros de MIR de todas as amostras de petróleo e coeficientes de regressão das primeiras variáveis latentes dos modelos individuais construídos a partir dos espectros de MIR para predição saturados, aromáticos e polares.

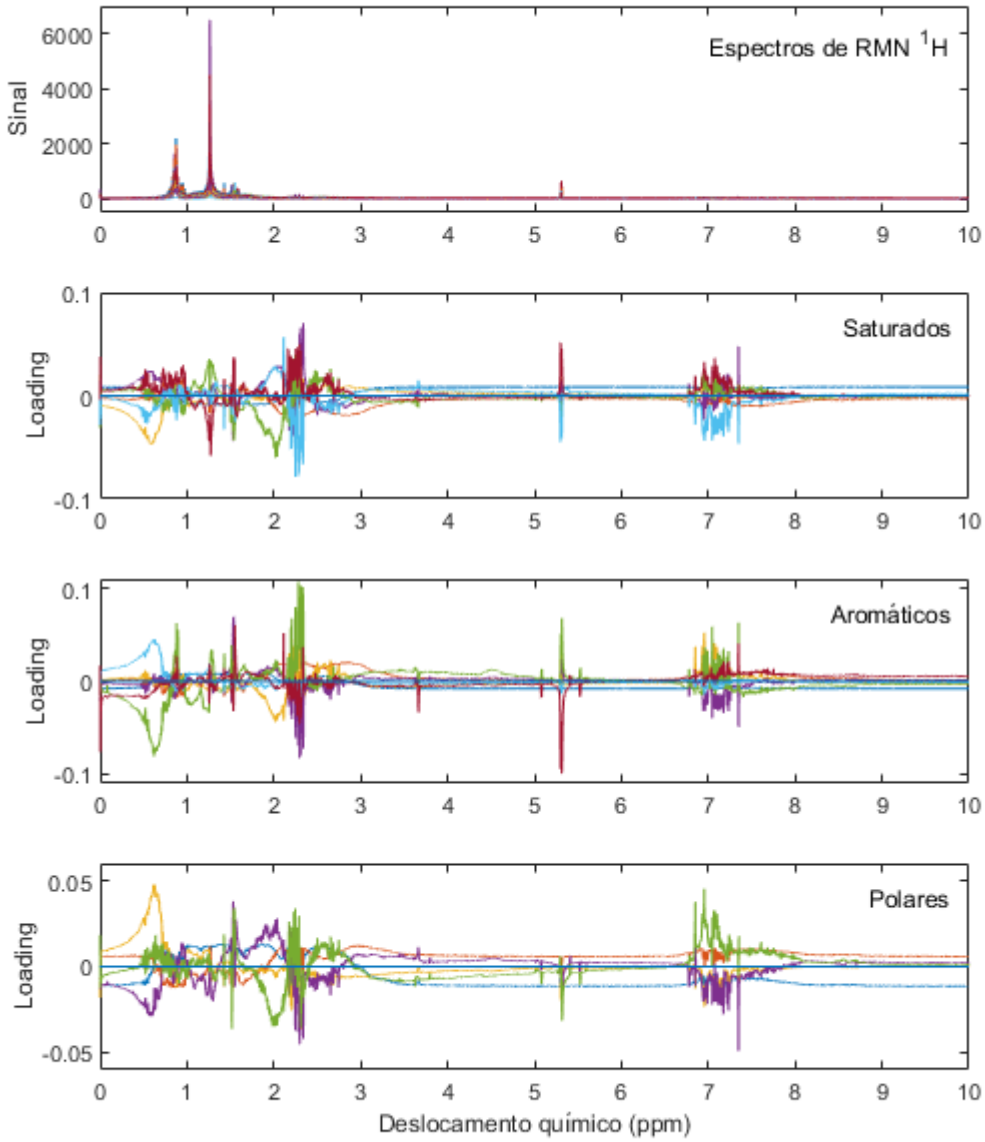



**Figura 4.21.** Espectros de RMN $^{1}$H de todas as amostras de petróleo e coeficientes de regressão das primeiras variáveis latentes dos modelos individuais construídos a partir dos espectros de RMN $^{1}$H para predição saturados, aromáticos e polares.

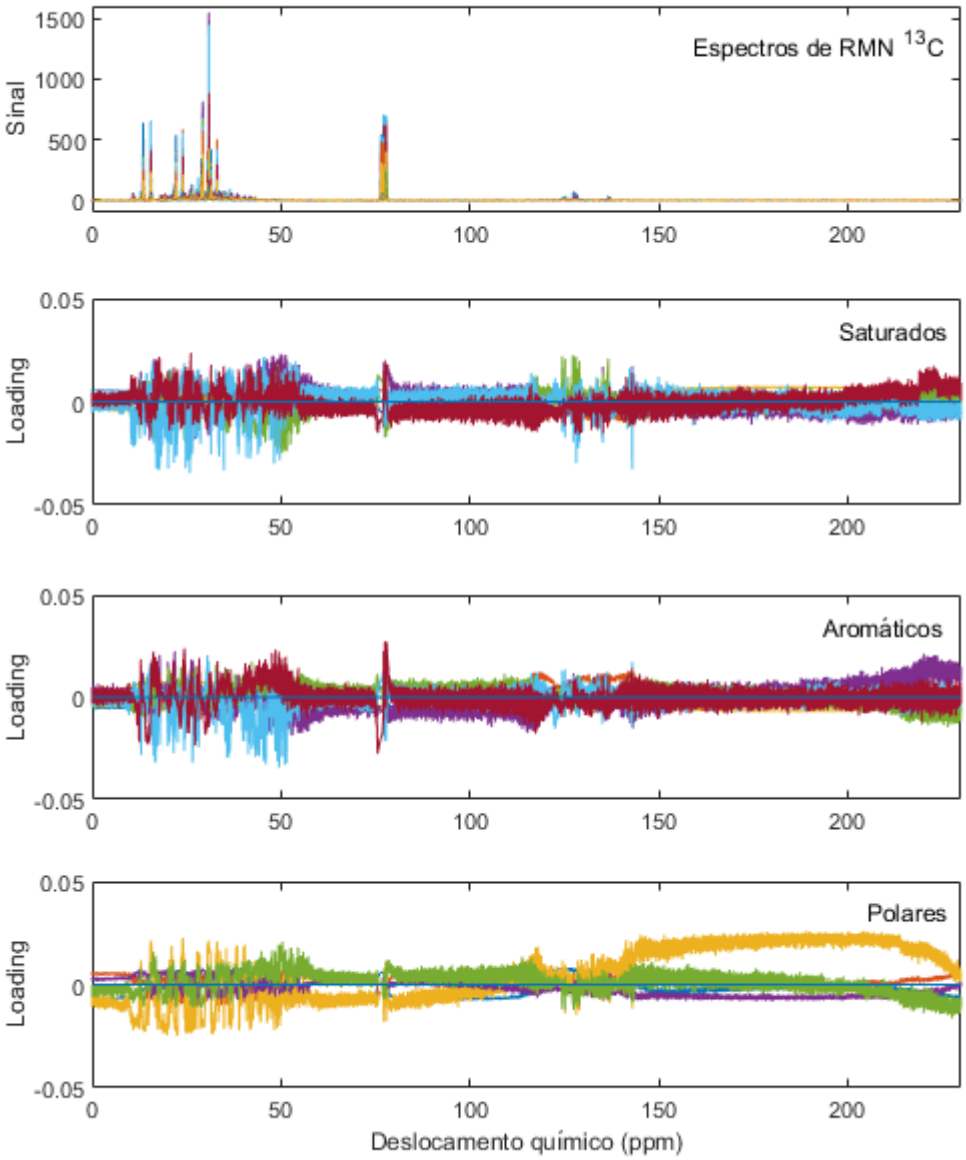



**Figura 4.22.** Espectros de RMN $^{13}C$ de todas as amostras de petróleo e coeficientes de regressão das primeiras variáveis latentes dos modelos individuais construídos a partir dos espectros de RMN $^{13}C$ para predição saturados, aromáticos e polares.

Para uma mesma fonte de dados (MIR, RMN $^{1}H$ ou RMN $^{13}C$), todos os três gráficos possuem perfis semelhantes, de modo que as variáveis importantes são basicamente as mesmas para os modelos de saturados, aromáticos e polares. Além disso, as variáveis mais importantes são de regiões espectrais com sinais analíticos

relevantes. Isso prova que essas propriedades estão bem correlacionadas com os três espectros e os modelos foram construídos com base nessa correlação.

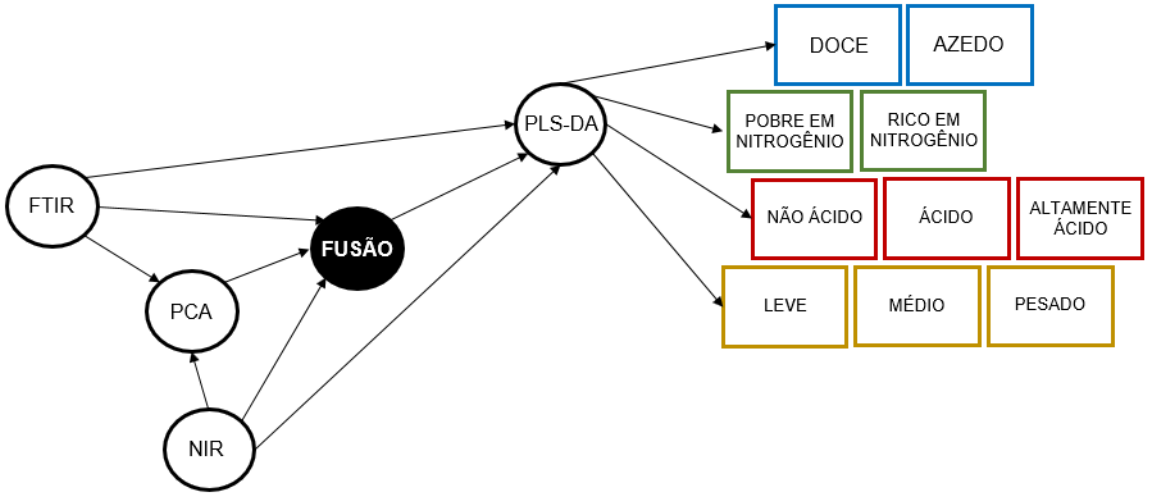
## 4.4 Conclusão

Neste estudo, as técnicas analíticas MIR, RMN $^{1}H$ e RMN $^{13}C$ foram aplicadas para prever sete propriedades de amostras de petróleo. A regressão por mínimos quadrados parciais, associado à fusão de dados a nível baixo, médio por PCA, médio por PLS e alto, foi aplicado para construir modelos de calibração. Posteriormente, os modelos foram comparados quanto a sua acurácia e linearidade.

As fusões de nível baixo, médio por PCA e alto não proporcionaram modelos com maior exatidão, em comparação com os modelos individuais (sem fusão); portanto, o uso destas estratégias não se justifica para este caso. Já a fusão de nível médio por PLS forneceu modelos com menores valores de $RMSE_P$ para todas as propriedades e, significativamente, com melhor desempenho de previsão para S, NAT e ARO. Dessa forma, a fusão de nível médio por PLS provou ser a melhor estratégia neste estudo.

Os resultados provaram que os dados multiespectrais podem se complementar e agir sinergicamente para estabelecer uma correlação entre os sinais espectrais e a propriedade estimada. A fusão de dados de MIR, RMN $^{1}H$ e $^{13}C$, associadas ao algoritmo de calibração PLS, pode ser considerada uma metodologia rápida, eficiente e acurada para estimar algumas propriedades físico-químicas do petróleo bruto.

# CAPÍTULO 5 FUSÃO DE DADOS APLICADA EM ESPECTROS NIR E MIR PARA CLASSIFICAÇÃO DE AMOSTRAS DE PETRÓLEO




- As fusões de dados MIR e NIR de baixo e médio nível são aplicadas para discriminação de amostras de petróleo bruto.

- O PLS-DA é utilizado para classificação das amostras baseado em seus espectros.

- A fusão de nível médio é aplicada após extração de recursos por meio de uma PCA.

- É viável a utilizar fusão de dados espectrais e PLS-DA para estabelecer um modelo rápido e com boa exatidão.

## 5.1 Introdução

O petróleo é a *commodity* mais importante do mundo, sendo a principal fonte de energia da atualidade e impactando fortemente a economia global. Suas características físico-químicas variam amplamente de um reservatório para outro e, em consequencia, existe uma grande variedade de petróleos em todo o mundo (WAPLES, 1985).

Existem algumas classificações que agrupam os óleos de acordo com suas semelhanças, levando em consideração alguns fatores específicos de interesse como, por exemplo, a classificação em óleos leves, médios ou pesados, de acordo com seu valor de densidade API. Esta propriedade é muito importante, pois determina o valor do produto no mercado global. Ao caracterizar/classificar o óleo de acordo com o grau API, é possível estimar as frações do destilado, evitar problemas no transporte, no processamento e armazenamento e proporcionar melhor desempenho na sua produção e refino (RIAZI, 2005, SPEIGHT, 2015).

Para realizar essas classificações, métodos alternativos, como a espectroscopia associada à quimiometria, podem ser aplicados. Vários estudos têm aplicado métodos de reconhecimento de padrões para discriminar amostras de petróleo e derivados. GALTIER *et al*. (2011) utilizaram diferentes métodos quimiométricos em espectros no infravermelho médio para classificar petróleo bruto de acordo com sua origem geográfica. LOVATTI *et al*. (2020) discriminaram amostras de petróleo bruto em diferentes grupos, de acordo com seu número de acidez total, aplicando o método de reconhecimento de padrões chamado floresta randômica em dados de RMN $^{1}H$ e $^{13}C$.

Nos últimos anos, tem crescido a produção científica com aplicação de quimiometria em problemas de classificação envolvendo petróleo bruto, conforme mostrado na Figura 5.1. O gráfico é resultado de uma pesquisa realizada nas bases de dados *Web of Science* (Figura 5.1a) e *Scopus* (Figura 5.1b), no período de 2010 a 2021, utilizando as palavras-chave “*crude oil*” e “*classification*” ou “PLS-DA” ou “*discriminant analysis*”.

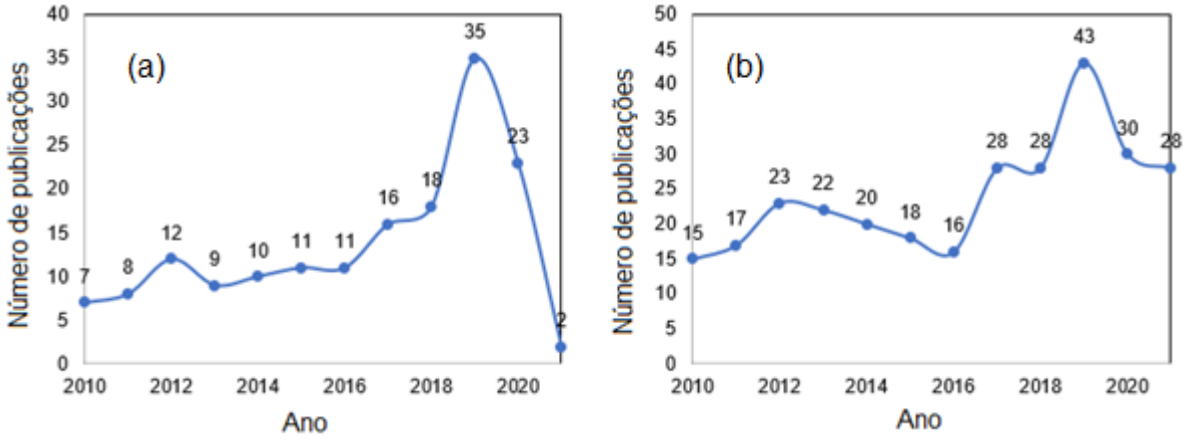



**Figura 5.1.** Número de publicações científicas relacionadas ao petróleo bruto e a análise discriminante. Dados coletados a partir da base de dados (a) *Web of Science* e (b) *Scopus*.

O algoritmo PLS-DA é um dos métodos discriminantes mais empregados na quimiometria, sendo utilizado para diversos fins como, por exemplo, detectar adulterantes ou contaminantes em misturas de biodiesel/diesel (MAZIVILA *et al*., 2015), detectar adulterante em amostras de petróleo bruto e classificar amostras combustíveis de acordo com seu tipo (nafta, gasolina, diesel e querosene) (SANTOS *et al*., 2021), detectar adulteração em amostras de gasolina (MABOOD *et al*., 2017, NESPECA *et al*., 2018), classificar amostras de gasolina de acordo com sua marca (BARNETTI e ZHANG, 2018), entre outras finalidades.

Para melhorar a capacidade de discriminação dos modelos, produzindo melhores inferências, com menores taxas de erro, a estratégia de fusão de dados pode ser aplicada. Especificamente para problemas de classificação envolvendo petróleo bruto, a fusão de dados tem sido utilizada para, por exemplo, discriminar entre amostras de petróleo bruto e óleo diesel (GUIDA *et al*., 2015), discriminar entre manchas de óleo e sósias (GUO *et al*., 2017), discriminar entre amostras de solo poluído e não poluído por derramamentos de óleo (OZIGIS *et al*., 2019) e para a detecção de defeitos em oleodutos de transporte de petróleo bruto (LI *et al*., 2019). A espectrometria de massas, a espectroscopia de infravermelho e de RMN foram amplamente utilizadas em estratégias de fusão de dados. Quanto aos métodos quimiométricos, o PLS e o PLS-DA são um dos mais aplicados para problemas de regressão e classificação, respectivamente (BORRÀS *et al*., 2015).

Este estudo teve como objetivo construir modelos de classificação para

diferenciar amostras de petróleo bruto entre as classes doce/azedo, pobre/rico em nitrogênio; não ácidos/ácidos/altamente ácidos e leves/médios/pesados, utilizando o algoritmo PLS-DA e seus espectros NIR e MIR. Para utilizar a sinergia entre as duas técnicas instrumentais, os dados NIR e MIR também foram modelados após serem fundidos a nível baixo e médio, na tentativa não apenas de obter modelos mais precisos, mas também de revelar até que ponto a fusão de dados pode melhorar o desempenho de discriminação.

## 5.2 Metodologia

### 5.2.1 Amostras e ensaios espectrais

Um total de 121 amostras de petróleo bruto advindas da bacia sedimentar costeira brasileira foram fornecidas pelo Cenpes/Petrobras. Estas amostras são as mesmas utilizadas no estudo apresentado no Capítulo 4. De acordo com a classificação API (API, 2021), as amostras variaram de óleos leves a pesados, apresentando °API entre 11,4 e 54,0.

Os espectros MIR e NIR de todas as amostras foram adquiridos e fornecidos pelo LabPetro/UFES. Os espectros MIR foram medidos em um espectrômetro PerkinElmer, equipado com acessório de reflexão total atenuada, com cristal de seleneto de zinco. Os espectros foram coletados na faixa de 4000 a 650 $cm^{-1}$, com resolução de 4 $cm^{-1}$ e 32 varreduras por amostra. Foram medidas três réplicas por amostra.

Os espectros NIR foram medidos em um espectrômetro PerkinElmer, Spectrum 400. Três réplicas foram medidas por amostra, coletadas no intervalo de 10000 - 4000 $cm^{-1}$, com resolução de 8 $cm^{-1}$ e 128 varreduras.

### 5.2.2 Ensaios físico-químicos e atribuição de classes

As análises físico-quimicas foram realizadas e seus resultados fornecidos pelo Cenpes/Petrobras. Às amostras foram atribuídas diferentes classes com base em algumas classificações usuais de petróleo bruto e com base nos resultados dos

experimentos físico-químicos, conforme mostra a Tabela 5.1.

Seguindo a classificação de RIAZI (2005), às 93 amostras com teor de enxofre abaixo de 0,5% em massa, foram atribuídas a classe "doce"; e, às 28 amostras com teor de enxofre acima ou igual a 0,5% em massa, foram atribuídas a classe “azedo”. A concentração de enxofre (S) nas amostras de petróleo foi determinada de acordo com a ASTM D4294-16, utilizando fluorescência de raios-X por energia dispersiva. As amostras foram irradiadas por um tubo de raios X e a radiação X excitada característica foi medida e comparada com medições anteriores de amostras padrão com concentrações conhecidas de enxofre.

**Tabela 5.1.** Valores limites para as classificações do petróleo bruto.

| Propriedade | Limites | Classe | Número da classe | Número de amostras |
|---|---|---|---|---|
| Teor de enxofre | < 0.5 m/m% | Doce | 1 | 93 |
| | ≥ 0.5 m/m% | Azedo | 2 | 28 |
| Teor de nitrogênio total | < 0.25 m/m% | Pobre em N | 1 | 43 |
| | ≥ 0.25 m/mt% | Rico em N | 2 | 78 |
| Número de acidez total | ≤ 0.5 $mgKOHg^{-1}$ | Não-ácido | 1 | 96 |
| | > 0.5 and ≤ 1 $mgKOHg^{-1}$ | Ácido | 2 | 8 |
| | > 1 $mgKOHg^{-1}$ | Altamente ácido | 3 | 17 |
| °API | ≥ 31 | Leve | 1 | 10 |
| | < 31 and ≥ 22 | Médio | 2 | 84 |
| | < 22 | Pesado | 3 | 27 |

Seguindo a classificação de PRADO *et al*. (2017), às 43 amostras com teor de nitrogênio total abaixo de 0,25% em massa foram atribuídas a classe “pobre em nitrogênio”, e às 78 amostras com teor de nitrogênio total acima ou igual a 0,25% em massa foram atribuídas a classe “rico em nitrogênio”. O teor de nitrogênio total (NT) foi determinado pelo método padrão ASTM D4629-17, por uma combustão oxidativa completa com detecção por quimioluminescência. O hidrocarboneto é oxidado em um fluxo rico de oxigênio puro na zona de combustão, onde todos os compostos ligados ao nitrogênio são convertidos em óxido de nitrogênio. Em seguida, o nitrogênio é quantificado por quimioluminescência, utilizando uma curva de calibração padrão.

Seguindo a classificação de ZAFAR *et al*. (2016), às 96 amostras com índice de

acidez total menor ou igual a 0,5 mgKOHg$^{-1}$ foram atribuídas a classe “não ácida”, às 8 amostras com índice de acidez total acima de 0,5 e abaixo ou igual a 1,0 mgKOHg$^{-1}$ foram atribuídas a classe “ácida”, e às 17 amostras com índice de acidez total acima de 1,0 mgKOHg$^{-1}$ foram atribuídas a classe “altamente ácida”. Para medir a acidez das amostras, foi aplicado o método padrão ASTM D664. A acidez foi calculada a partir da quantidade de solução padrão de hidróxido de potássio necessária para neutralizar a amostra de óleo em uma titulação potenciométrica, utilizando o eletrodo seletivo Solvotrode easyClean. Os resultados foram expressos em mg(KOH).g$^{-1}$(óleo).

Seguindo a classificação API (API, 2021), às 10 amostras com densidade API maior ou igual a 31, foi atribuída a classe “leve”; às 84 amostras com API inferior a 31 e superior ou igual a 22 foi atribuída a classe “médio”; e às 27 amostras com API abaixo de 22 foi atribuída a classe “pesado”. A densidade API foi determinada de acordo com a ISO 12185, usando um densímetro digital automático Anton Paar, modelo DMA 5000.

### 5.2.3 Fusão de dados e PLS-DA

Os procedimentos descritos nesta seção foram seguidos para a construção de modelos de classificação com base nos quatro diferentes parâmetros mostrados na Tabela 5.1.

Anteriormente à modelagem, por meio do algoritmo Kennard-Stone (KENNARD e STONE, 1969), as amostras foram divididas em dois conjuntos: 70% foram utilizadas para construir os modelos (amostras de treinamento) e os 30% restantes foram utilizadas para testar exatidão dos modelos na previsão de amostras externas (amostras de teste).

Em seguida, os espectros MIR e NIR foram pré-tratados de diferentes formas, de acordo com a fusão aplicada. Os espectros MIR e NIR, para modelagem individual e fusão de nível baixo, foram pré-tratados com MSC. Já o espectro MIR e NIR para fusão de nível médio foram pré-tratados com SNV.

Para cada classificação, modelos PLS-DA foram construídos a partir dos conjuntos individuais de MIR e NIR e a partir da fusão desses dois espectros (MIR + NIR) a nível baixo e médio. A fusão de nível baixo foi feita por uma simples concatenação, em uma única matriz, dos espectros MIR e NIR pré-processados. Após

a fusão, a nova matriz foi autoescalonada e, no caso específico de classificação em óleos não ácidos/ácidos/altamente ácidos foi pré-processada com MSC novamente.

Para a fusão de nível médio, a PCA foi aplicada separadamente aos dois espectros individuais pré-processados. Em seguida, as 10 primeiras componentes principais de cada PCA foram fundidas em uma única matriz. Depois disso, a nova matriz foi autoescalonada.

Modelos de classificação multivariada foram então construídos com dados fundidos ou individuais, utilizando o algoritmo PLS-DA e as amostras de treinamento. O número ideal de variáveis latentes foi encontrado aplicando a validação cruzada Venetian blind 5-fold. Para evitar superajuste, o número de variáveis latentes para todos os modelos foi limitado a 10. Para avaliar a habilidade de predição dos modelos em amostras externas, ou seja, amostras que não foram utilizadas na construção do modelo, seu desempenho foi testado com as amostras teste.

A qualidade de todos os modelos obtidos, tanto os individuais quanto os fundidos, foi determinada por meio de parâmetros de desempenho e, sem seguida, os modelos foram comparados entre si.

### 5.2.4 Parâmetros de avaliação dos modelos

Os parâmetros de avaliação dos modelos utilizados neste estudo estão descritos a seguir.

Sensibilidade:

$$\mathrm{SEN_i} = \frac{\mathrm{TP_i}}{(\mathrm{TP_i}+\mathrm{FN_i})} \qquad \text{Equação (5.1)}$$

Em que $TP_i$ (verdadeiro positivo) é o número de amostras corretamente classificadas como pertencentes à classe “i”, $FN_i$ (falso negativo) é o número de amostras que pertencem à classe “i”, mas foram classificadas pelo modelo como pertencentes a outra classe. $TP_i + FN_i$ representa o número de amostras que pertencem à classe "i". A sensibilidade (SEN) expressa a capacidade do modelo de prever corretamente os "verdadeiros" positivos.

Especificidade:

$$\mathrm{SPE_i} = \frac{\mathrm{TN_i}}{(\mathrm{TN_i}+\mathrm{FP_i})} \qquad \text{Equação (5.2)}$$

Em que $TN_i$ (verdadeiro negativo) é o número de amostras corretamente classificadas como não pertencentes à classe “i”, $FP_i$ (falso positivo) é o número de amostras que foram classificadas pelo modelo como pertencentes à classe “i”, mas pertencem a outra classe. TPi + FNi representa o número de amostras que não pertencem à classe "i". A especificidade (SPE) expressa a capacidade de prever corretamente os "verdadeiros" negativos.

Eficiência:

$$\mathrm{Eff_i} = \sqrt{\mathrm{SEN_i} \times \mathrm{SPE_i}} \qquad \text{Equação (5.3)}$$

A eficiência (Eff) avalia a eficácia da classificação do modelo de forma geral, para cada classe “i”.

Acurácia:

$$\mathrm{Acc} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Equação (5.4)}$$

A acurácia (Acc), também chamada de exatidão, é a razão entre o número de classificações corretas e o número total de casos, ou seja, representa a porcentagem de classificações corretas. Ele mede o desempenho do modelo como um todo, apresentando a taxa de previsões corretas. Ele varia de zero a um: um quando todas as previsões estão corretas e zero quando todas as previsões estão incorretas.

Erro de predição total:

$$\mathrm{ER} = \frac{\sum_{i=1}^{C} Error_i}{C} \qquad \text{Equação (5.5)}$$

Em que C é o número de classes. O erro de predição total (ER) representa uma média de erros de cada classe. O erro de uma classe “i” é calculado pela Equação 5.6 que representa a média de falsos positivos e negativos para a classe “i”.

$$\mathrm{Error_i} = \frac{(1-SEN_i)+(1-SPE_i)}{2} \qquad \text{Equação (5.6)}$$

A Figura 5.2 mostra um esquema da metodologia de construção e avaliação de modelos de classificação construídos a partir dos espectros individuais e fundidos a baixo nível.

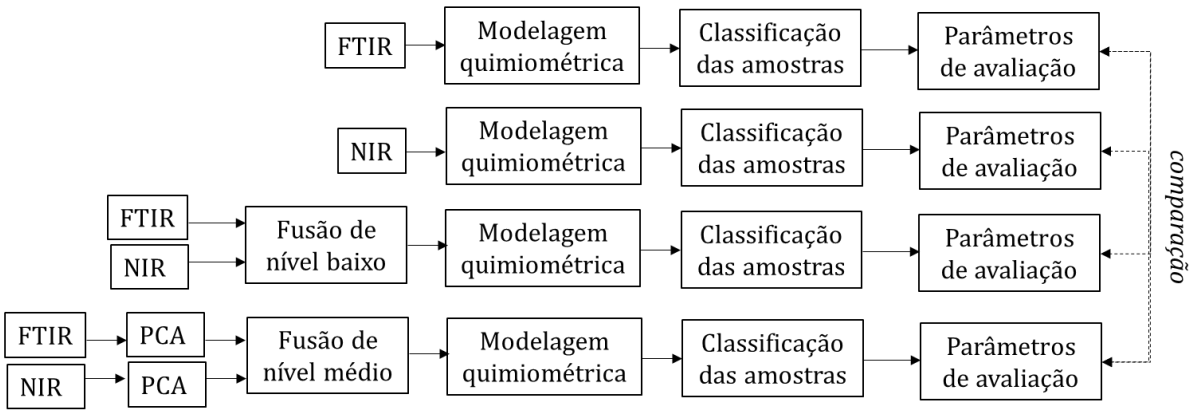



**Figura 5.2.** Esquema da metodologia de construção e avaliação de modelos de classificação.

## 5.3 Resultados e discussão

Os espectros MIR e NIR de todas as amostras de petróleo estão sobrepostos nos gráficos das Figuras 4.20a e 5.3a, respectivamente. Nos espectros NIR, as bandas mais intensas aparecem, aproximadamente, em quatro regiões: 4000-4500 e 7000-7700 $cm^{-1}$, devido às bandas de combinação de CH, e 5500-6250 e 8000-9000 $cm^{-1}$, devido ao primeiro e segundo sobretons de CH, respectivamente (PAVIA *et al*., 2008). Em ambos os espectros, o pré-tratamento MSC foi aplicado para correção de sinal relacionado ao espalhamento de luz. Os espectros pré-tratados são mostrados nas Figuras 4.20b e 5.3b.

Modelos de classificação PLS-DA foram construídos individualmente com os espectros MIR e NIR e para estes mesmos espectros fundidos a nível baixo e médio, todos previamente pré-tratados. O número de variáveis latentes foi determinado por validação cruzada com base no valor mínimo de ER. Os modelos discriminantes atribuíram às amostras uma classe, ambas as classes ou a nenhuma classe.

O desempenho de classificação dos modelos PLS-DA foi avaliado em termos de sensibilidade, especificidade, eficiência, acurácia e erro de predição total. Os

principais parâmetros estatísticos dos modelos obtidos estão dispostos nas Tabelas 5.2 a 5.5. O erro de predição total e a acurácia são parâmetros importantes para avaliar a capacidade de classificação do modelo. Quanto maior a Acc e a Eff e menor o ER, maior a capacidade do modelo de classificar as amostras corretamente.

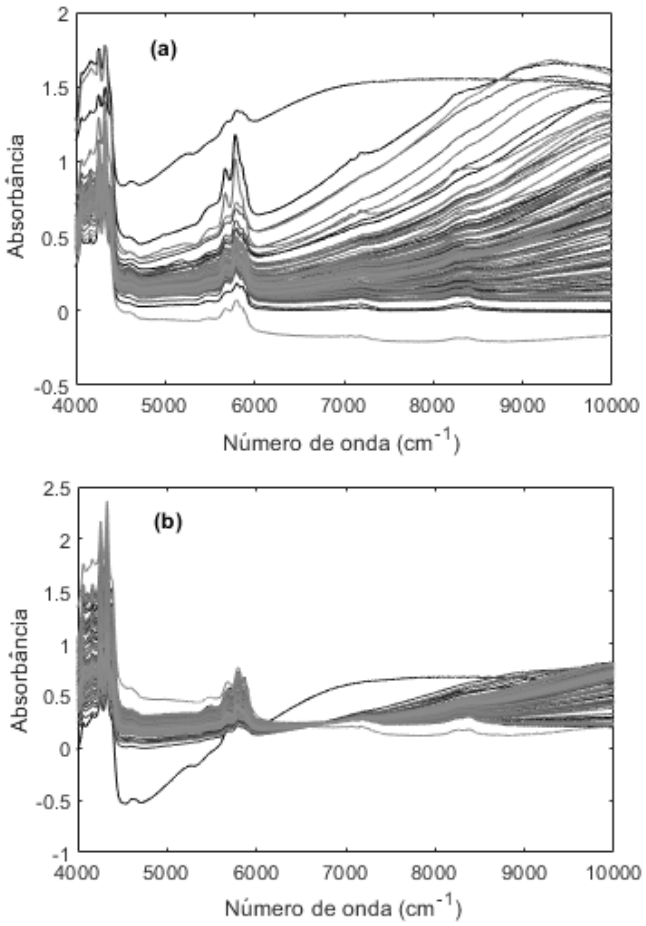



**Figura 5.3.** Sobreposição de espectros NIR (a) antes e (b) após pré-tratamento com MSC.

**Tabela 5.2.** Principais parâmetros estatísticos dos modelos PLS-DA a partir dos espectros individuais e fundidos para classificação em óleo doce/azedo.

| Fonte de dados | Classe | nlv | Treinamento | | | | | | | | Teste | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FP | FN | NAS | SEM | SPE | Eff | ER | Acc | FP | FN | NAS | SEN | SPE | Eff | ER | Acc |
| MIR | 1 | 6 | 3 | 4 | 0 | 0.95 | 0.85 | 0.89 | 0.11 | 0.92 | 1 | 1 | 0 | 0.96 | 0.88 | 0.91 | 0.08 | 0.94 |
| | 2 | | 4 | 3 | 0 | 0.95 | 0.94 | 0.94 | | | 1 | 1 | 0 | 0.88 | 0.96 | 0.91 | | |
| NIR | 1 | 4 | 3 | 13 | 0 | 0.80 | 0.85 | 0.82 | 0.17 | 0.81 | 1 | 1 | 0 | 0.96 | 0.88 | 0.91 | 0.08 | 0.94 |
| | 2 | | 13 | 3 | 0 | 0.85 | 0.80 | 0.82 | | | 1 | 1 | 0 | 0.88 | 0.96 | 0.91 | | |
| Fusão de nível baixo | 1 | 3 | 2 | 5 | 0 | 0.92 | 0.90 | 0.90 | 0.09 | 0.92 | 0 | 2 | 0 | 0.93 | 1.0 | 0.96 | 0.04 | 0.94 |
| | 2 | | 5 | 2 | 0 | 0.90 | 0.92 | 0.90 | | | 2 | 0 | 0 | 1.0 | 0.93 | 0.96 | | |
| Fusão de nível médio | 1 | 2 | 2 | 4 | 0 | 0,94 | 0,90 | 0,92 | 0,08 | 0,93 | 0 | 2 | 0 | 0,93 | 1,0 | 0,96 | 0,04 | 0,94 |
| | 2 | | 4 | 2 | 0 | 0,90 | 0,94 | 0,92 | | | 2 | 0 | 0 | 1,0 | 0,93 | 0,96 | | |

SEN = sensibilidade; SPE = especificidade; Eff = eficiência; Acc = acurácia; ER = erro de predição total; NAS = número de amostras não classificadas; FP = falso positivo; FN = falso negativo

**Tabela 5.3.** Principais parâmetros estatísticos dos modelos PLS-DA a partir dos espectros individuais e fundidos para classificação em óleo pobre/rico em nitrogênio.

| Fonte de dados | Classe | nlv | Treinamento | | | | | | | | Teste | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FP | FN | NAS | SEM | SPE | Eff | ER | Acc | FP | FN | NAS | SEN | SPE | Eff | ER | Acc |
| MIR | 1 | 2 | 4 | 4 | 0 | 0.87 | 0.93 | 0.89 | 0.10 | 0.91 | 1 | 3 | 0 | 0.77 | 0.96 | 0.85 | 0.14 | 0.89 |
| | 2 | | 4 | 4 | 0 | 0.93 | 0.87 | 0.89 | | | 3 | 1 | 0 | 0.96 | 0.77 | 0.85 | | |
| NIR | 1 | 7 | 5 | 8 | 0 | 0.73 | 0.91 | 0.81 | 0.18 | 0.85 | 1 | 3 | 0 | 0.77 | 0.96 | 0.85 | 0.14 | 0.89 |
| | 2 | | 8 | 5 | 0 | 0.91 | 0.73 | 0.81 | | | 3 | 1 | 0 | 0.96 | 0.77 | 0.85 | | |
| Fusão de nível baixo | 1 | 5 | 5 | 4 | 0 | 0.87 | 0.91 | 0.88 | 0.11 | 0.89 | 0 | 2 | 0 | 0.85 | 1.0 | 0.92 | 0.08 | 0.94 |
| | 2 | | 4 | 5 | 0 | 0.91 | 0.87 | 0.88 | | | 2 | 0 | 0 | 1.0 | 0.85 | 0.92 | | |
| Fusão de nível médio | 1 | 9 | 2 | 5 | 0 | 0.83 | 0.96 | 0.89 | 0.10 | 0.92 | 2 | 0 | 0 | 1.0 | 0.91 | 0.95 | 0.04 | 0.94 |
| | 2 | | 5 | 2 | 0 | 0.96 | 0.83 | 0.89 | | | 0 | 2 | 0 | 0.91 | 1.0 | 0.95 | | |

SEN = sensibilidade; SPE = especificidade; Eff = eficiência; Acc = acurácia; ER = erro de predição total; NAS = número de amostras não classificadas; FP = falso positivo; FN = falso negativo

**Tabela 5.4.** Principais parâmetros estatísticos dos modelos PLS-DA a partir dos espectros individuais e fundidos para classificação em óleo não ácido/ácido/altamente ácido.

| Fonte de dados | Classe | nvl | Treinamento | | | | | | | | Teste | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FP | FN | NAS | SEN | SPE | Eff | ER | Acc | FP | FN | NAS | SEN | SPE | Eff | ER | Acc |
| MIR | 1 | 9 | 2 | 1 | 11 | 0.98 | 0.88 | 0.92 | 0.09 | 0.96 | 1 | 0 | 3 | 1.0 | 0.75 | 0.86 | 0.21 | 0.97 |
| | 2 | | 1 | 2 | 1 | 0.60 | 0.99 | 0.77 | | | 0 | 1 | 1 | 0.0 | 1.0 | 0.0 | | |
| | 3 | | 0 | 0 | 0 | 1.0 | 1.0 | 1.0 | | | 0 | 0 | 2 | 1.0 | 1.0 | 1.0 | | |
| NIR | 1 | 10 | 3 | 1 | 7 | 0.98 | 0.82 | 0.89 | 0.18 | 0.92 | 3 | 0 | 1 | 1.0 | 0.57 | 0.75 | 0.27 | 0.91 |
| | 2 | | 0 | 5 | 0 | 0.17 | 1.0 | 0.41 | | | 0 | 2 | 0 | 0.0 | 1.0 | 0.0 | | |
| | 3 | | 3 | 0 | 1 | 1.0 | 0.95 | 0.97 | | | 0 | 1 | 0 | 0.8 | 1.0 | 0.89 | | |
| Fusão de nível baixo | 1 | 6 | 2 | 3 | 5 | 0.95 | 0.87 | 0.90 | 0.14 | 0.92 | 1 | 0 | 0 | 1.0 | 0.86 | 0.92 | 0.11 | 0.97 |
| | 2 | | 2 | 3 | 1 | 0.40 | 0.97 | 0.62 | | | 0 | 1 | 0 | 0.5 | 1.0 | 0.70 | | |
| | 3 | | 2 | 0 | 2 | 1.0 | 0.97 | 0.98 | | | 0 | 0 | 0 | 1.0 | 1.0 | 1.0 | | |
| Fusão de nível médio | 1 | 5 | 2 | 2 | 8 | 0.97 | 0.86 | 0.87 | 0.12 | 0.93 | 1 | 0 | 1 | 0.86 | 1.0 | 0.97 | 0.11 | 0.97 |
| | 2 | | 1 | 3 | 0 | 0.50 | 1.0 | 0.44 | | | 0 | 1 | 0 | 1.0 | 0.5 | 1.0 | | |
| | 3 | | 2 | 0 | 2 | 1.0 | 1.0 | 0.98 | | | 0 | 0 | 0 | 1.0 | 1.0 | 1.0 | | |

SEN = sensibilidade; SPE = especificidade; Eff = eficiência; Acc = acurácia; ER = erro de predição total; NAS = número de amostras não classificadas; FP = falso positivo; FN = falso negativo

**Tabela 5.5.** Principais parâmetros estatísticos dos modelos PLS-DA a partir dos espectros individuais e fundidos para classificação em óleo leve/médio/pesado.

| Fonte de dados | Classe | nvl | Treinamento | | | | | | | | Teste | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FP | FN | NAS | SEN | SPE | Eff | ER | Acc | FP | FN | NAS | SEN | SPE | Eff | ER | Acc |
| MIR | 1 | 7 | 1 | 0 | 1 | 1.0 | 0.99 | 0.99 | 0.05 | 0.95 | 0 | 0 | 1 | 1.0 | 1.0 | 1.0 | 0.04 | 0.97 |
| | 2 | | 2 | 2 | 3 | 0.96 | 0.91 | 0.93 | | | 1 | 0 | 1 | 1.0 | 0.9 | 0.94 | | |
| | 3 | | 1 | 2 | 2 | 0.88 | 0.98 | 0.92 | | | 0 | 1 | 0 | 0.88 | 1.0 | 0.93 | | |
| NIR | 1 | 6 | 1 | 3 | 0 | 0.57 | 0.99 | 0.75 | 0.14 | 0.91 | 0 | 1 | 2 | 0.0 | 1.0 | 0.0 | 0.25 | 0.90 |
| | 2 | | 5 | 2 | 3 | 0.96 | 0.79 | 0.87 | | | 2 | 1 | 2 | 0.96 | 0.75 | 0.84 | | |
| | 3 | | 1 | 2 | 2 | 0.88 | 0.98 | 0.92 | | | 1 | 1 | 1 | 0.86 | 0.96 | 0.90 | | |
| Fusão de nível baixo | 1 | 5 | 0 | 0 | 3 | 1.0 | 1.0 | 1.0 | 0.02 | 0.95 | 0 | 0 | 2 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 |
| | 2 | | 0 | 4 | 5 | 0.93 | 1.0 | 0.96 | | | 0 | 0 | 1 | 1.0 | 1.0 | 1.0 | | |
| | 3 | | 4 | 0 | 3 | 1.0 | 0.93 | 0.96 | | | 0 | 0 | 0 | 1.0 | 1.0 | 1.0 | | |
| Fusão de nível médio | 1 | 5 | 2 | 2 | 8 | 0.97 | 0.86 | 0.87 | 0.12 | 0.93 | 1 | 0 | 0 | 0.86 | 1.0 | 0.97 | 0.11 | 0.97 |
| | 2 | | 1 | 3 | 0 | 0.50 | 1.0 | 0.44 | | | 0 | 1 | 0 | 1.0 | 0.5 | 1.0 | | |
| | 3 | | 2 | 0 | 2 | 1.0 | 1.0 | 0.98 | | | 0 | 0 | 1 | 1.0 | 1.0 | 1.0 | | |

SEN = sensibilidade; SPE = especificidade; Eff = eficiência; Acc = acurácia; ER = erro de predição total; NAS = número de amostras não classificadas; FP = falso positivo; FN = falso negativo

Vale ressaltar que a fusão de nível baixo é geralmente vantajosa quando as dimensões das matrizes fundidas são muito semelhantes, ou seja, quando o número de variáveis é semelhante, caso contrário, pode haver predomínio de uma fonte de dados sobre a outra. As matrizes MIR e NIR utilizadas neste trabalho apresentaram 3351 e 3001 variáveis, respectivamente. Devido à similaridade do número das variáveis, não houveram problema de predominância de dados em decorrência de discrepância entre as dimensões das matrizes.

No entanto, esse problema também pode ocorrer quando as matrizes fundidas têm variâncias muito diferentes. Nesse caso, é provável que o modelo consiga capturar bem a variância de uma matriz e desconsiderar parcialmente a variância da outra, evitando que esta contribua para a construção do modelo. Dessa forma, a variação conjunta entre as matrizes não será capturada e a fusão dos dados será inútil.

O gráfico da Figura 5.4 mostra a variância contida em cada uma das variáveis das matrizes MIR e NIR. Fica claro que as variâncias das variáveis entre as duas matrizes são bastante diferentes. Os espectros NIR têm variância total igual a 0,0025, muito maior que a variância total dos espectros MIR, igual a 0,0000144. Esta discrepância pode causar sobreposição das informações contidas no NIR sobre o MIR.

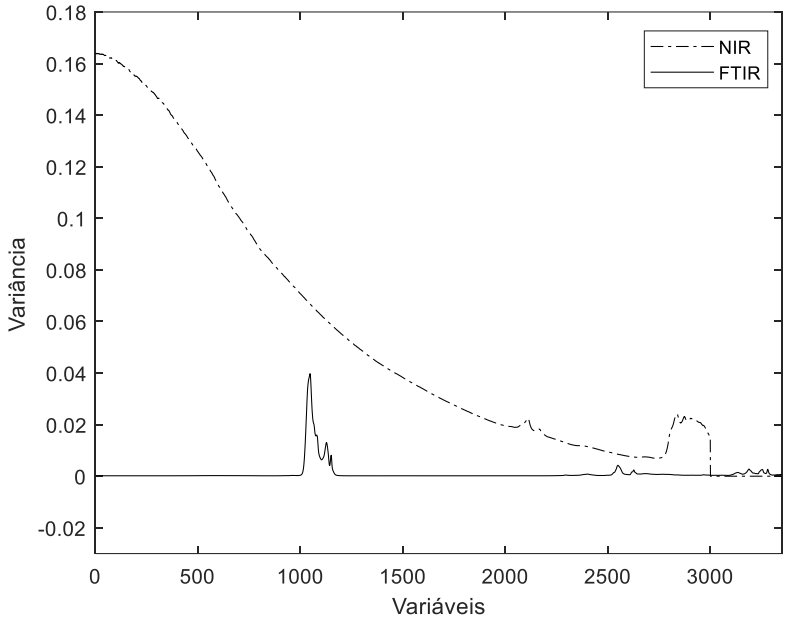



**Figura 5.4.** Variância contida nas variáveis dos espectros MIR e NIR.

Para resolver este problema, o pré-processamento auto escalonamento pode ser aplicado após a fusão das matrizes, como foi feito neste trabalho. Desta forma, todas as variáveis terão variância igual a um e, conseqüentemente, todas contribuirão com eqüidade para a formação do modelo.

Conforme mostram as tabelas, diferentes números de variáveis latentes foram necessários para os diferentes conjuntos de dados (MIR, NIR ou espectros fundidos) para se obter modelos com as menores taxas de erro possíveis.

Comparando os resultados entre as fontes de dados, para todas as classificações, os espectros NIR forneceram modelos com os piores desempenhos, tanto para amostras de treinamento quanto para amostras externas.

Em geral, os desempenhos de classificação foram satisfatórios. Para todas as classificações, os melhores modelos apresentaram acurácias superiores a 92,0% para treinamento e a 97,0% para previsão. Esses resultados são semelhantes a algumas classificações PLS-DA para petróleo bruto encontradas na literatura. GALTIER *et al*. (2011) utilizaram espectros no infravermelho médio de petróleo bruto para classificá-los de acordo com sua origem geográfica. O modelo PLS-DA alcançou uma acurácia igual a 100%. LOVATTI *et al*. (2019) utilizaram floresta randômica para discriminar amostras de óleo, de acordo com o ponto de fluidez máximo, obtendo resultados com exatidão de 90% e 71%, respectivamente, para espectros de RMN $^{1}$H e RMN $^{13}$C.

NESPECA *et al*. (2018) desenvolveram um modelo PLS-DA, a partir de dados de cromatografia gasosa, para detectar adulteração em gasolina. O modelo pôde discriminar corretamente 100% das amostras de treinamento e validação. MAZIVILA *et al*. (2015) também obtiveram previsões 100% corretas, utilizando dados MIR e PLS-DA, para determinar simultaneamente o óleo empregado e a rota (metílica ou etílica) na produção de amostras de biodiesel. No entanto, é importante ressaltar que gasolina e biodiesel são matrizes menos complexas que o petróleo bruto.

Para a classificação em doce/azedo, ambas os espectros (MIR e NIR) foneceram modelos equivalentes que apresentaram, para as amostras de teste, os mesmos valores de sensibilidade (0,96 em relação à classe 1) e especificidade (0,88 em relação à classe 1) e, como consequência, a mesma eficiência (0,91), acurácia (0,94) e erro de predição total (0,08). Para o conjunto de treinamento, os espectros MIR forneceram uma taxa de erro de predição menor (0,11) que os espectros NIR (0,17). A exatidão também foi melhor para MIR (0,92) em relação ao NIR (0,81).

As fusões de baixo e médio nível trouxeram uma pequena melhoria para o conjunto de treinamento, reduzindo o ER para valores iguais a 0,09 e 0,08, respectivamente, embora a acurácia tenha permanecido praticamente a mesma do modelo MIR individual, para o conjunto de treinamento. Nos quatro modelos, não houveram amostras não atribuídas.

Como pôde ser visto, o ER é utilizado como principal parâmetro para comparação dos modelos. É mais comum comparar os valores de ER das amostras de teste, entretanto em todos os modelos de classificação em óleo doce/azedo e em alguns modelos seguintes, o ER das amostras de treinamento foi maior que das amostras de teste. Neste caso, utiliza-se sempre o maior valor de ER como referência para comparação dos modelos, seja a partir das amostras de teste ou de treinamento.

Seguindo essa lógica, o modelo NIR apresentou o pior desempenho, com ER igual a 0,17, seguido do modelo MIR, com ER igual a 0,11. As fusões de nível baixo e médio trouxeram uma pequena melhoria, com ER igual a 0,09 e 0,08, respectivamente.

Para classificação em óleo pobre/rico em nitrogênio, o modelo MIR apresentou maior capacidade de reconhecer suas próprias amostras e amostras diferentes, obtendo maiores valores de sensibilidade (0,77 em relação à classe 1), especificidade (0,96 em relação à classe 1) e acurácia (0,89) e menor ER (0,14), em relação ao modelo NIR, cuja sensibilidade, especificidade, acurácia e ER foram iguais a 0,73 (em relação à classe 1), 0,91 (em relação à classe 1), 0,85 e 0,18, respectivamente.

A fusão de baixo nível melhorou ligeiramente a sensibilidade e especificidade, alcançando uma eficiência de 0,88. Também reduziu o ER para 0,11, embora a acurácia tenha permanecido a mesma do modelo MIR individual (0,89). Já a fusão de nível médio reduziu o ER para 0,10, aumentou a eficiência para 0,89 e a acurácia para 0,92. Nos quatro modelos, não houveram amostras não atribuídas.

Para classificação em óleo não ácido/ácido/altamente ácido, o conjunto de dados MIR forneceu um modelo com maior capacidade classficatória em comparação ao conjunto de NIR. Como pode ser visto na Tabela 5.4, o modelo MIR apresentou maior sensibilidade, especificidade, eficiência e acurácia e menor erro de previsão total tanto para amostras de treinamento quanto para de teste.

As fusões de nível baixo e médio produziram modelos com ER inferiores, respectivamente iguais a 0,14 e 0,12, ao modelo MIR, cujo ER foi igual a 0,21. Além

disso, houve mais amostras não atribuídas no modelo MIR do que nos modelos fundidos, tanto para o conjunto de treinamento quanto para o de teste. No entanto, o modelo MIR apresentou maior acurácia (0,97) que os modelos de fusão a nível baixo (0,92) e médio (0,93).

Para a classificação em óleo leve/médio/pesado, a técnica MIR produziu um modelo com maior capacidade classificatória do que técnica NIR, cujo erro de predição total (0,25) foi cinco vezes maior que o MIR (0,05). A fusão dos conjuntos MIR e NIR, a nível baixo, melhorou o desempenho do modelo, alcançando maior eficiência e menor ER. Como pode ser visto na Tabela 5.5, a fusão trouxe uma melhoria em todos os parâmetros estatísticos dos modelos. Apesar disso, houve algumas amostras que não foram atribuídas a nenhuma classe nos quatro modelos, tanto para o conjunto de treinamento quanto de teste.

MOHAMMADI *et al*. (2020), assim como feito neste estudo, utilizou dados MIR e PLS-DA para classificação de petróleo bruto em três classes, com base em seus valores de gravidade °API. O resultado da classificação apresentou 100% de acurácia para as amostras de treinamento e de teste. Sem fusão de dados, os autores obtiveram resultados equivalentes aos deste estudo, entretanto utilizaram apenas 20 amostras de petróleo, enquanto que aqui foram utilizadas 121 amostras com uma faixa de °API ligeiramente maior.

Em geral, os melhores resultados foram obtidos para os conjuntos de dados MIR e NIR combinados a nível médio. Em outras palavras, a fusão de dados levou a um refinamento ou aprimoramento dos resultados. Isso é importante, uma vez que a esta metodologia não é usada para complementar os resultados das técnicas individuais, mas sim para substituí-los. Portanto, deve apresentar resultados melhores do que as técnicas individuais e, somente assim, sua utilização se justifica.

## 5.4 Conclusão

Neste estudo, as espectroscopias MIR e NIR foram aplicados para discriminar amostras de petróleo bruto de acordo com quatro classificações. O método de análise discriminante de mínimos quadrados parciais, associado à fusão de dados de baixo nível, foi usado para construir modelos a partir dos espectros individuais e dos espectros fundidos. Posteriormente, os desempenhos de classificação de todos os

modelos foram comparados com base nos principais parâmetros estatísticos.

Diante dos resultados obtidos, a abordagem discriminante baseada na metodologia proposta mostrou-se uma forma promissora e eficiente de diferenciar amostras de petróleo bruto de acordo com as classificações apontadas. Os protocolos experimentais aqui apresentados, juntamente com os modelos construídos, mostraram que a fusão de dados de baixo e/ou médio nível melhoraram a capacidade discriminante de todos os modelos de classificação.

# CAPÍTULO 6 CONCLUSÕES GERAIS

A fusão de dados integra informações de diferentes fontes de aquisição sobre um mesmo conjunto de amostras. O método, portanto, tem a capacidade de prover novas informações que nenhuma fonte de dados, de maneira isolada, seria capaz de fornecer. Em problemas envolvendo matrizes complexas como o petróleo, é muito improvável que apenas uma técnica analítica consiga fornecer informações suficientes para explicar toda variância dos dados por meio de um modelo quimiométrico e assim conseguir uma boa exatidão em suas previsões. Dessa forma, neste trabalho, a fusão de dados foi aplicada para predição de propriedades e classificação de amostras de petróleo bruto.

Quatro estratégias de fusão de dados foram apresentadas e suas aplicações foram demonstradas sobre os diferentes dados espectrais. O principal objetivo foi verificar a viabilidade da fusão e comparar os resultados das diferentes estratégias. Um aspecto importante é a condução adequada das fusões, de forma que se escolham os pré-tratamentos e estratégias que evitem contribuições ambíguas e permitam que sejam capturadas as informações relevantes em todas as fontes de dados, independentemente de sua dimensão ou variância, no processo de análise multivariada.

Os resultados mostraram que, a partir da fusão, foi possível agregar infiormações diferentes advindas de duas ou mais fontes analíticas que, em conjunto, atuaram sinergicamente para contruir modelos de maior exatidão. Embora a fusão de nível médio por PLS tenha aumentado a exatidão dos modelos, as fusões de nível baixo, médio por PCA e alto não trouxeram melhorias significativas. Já os resultados de classificação mostraram que a fusão de baixo e médio nível melhoraram de forma sutil a capacidade discriminante em todos os casos.

O uso da fusão de dados só se justifica nos casos em que houve melhoria da capacidade preditiva e classificatória dos modelos em relação aos modelos individuais, visto que a inclusão de duas ou mais técnicas acrescenta custos e demanda de tempo ao processo.

As espectroscopias de MIR, NIR, RMN $^{1}H$ e $^{13}C$, utilizadas neste trabalho, se mostram técnicas simples, rápidas e não destrutivas, trazendo vantagens em relação

aos experimentos padronizados para determinação das propriedades físico-químicas do petróleo bruto. A fusão de dados e a regressão linear por PLS também se destacam pela sua simplicidade de aplicação e rapidez de processamento, não exigindo uma alta demanda computacional, como nos casos de seleção de variáveis ou método não lineares.

# CAPÍTULO 7 REFERÊNCIAS BIBLIOGRÁFICAS

1. ABBAS, O.; REBUFA, C.; DUPUY, N.; PERMANYER, A.; KISTER, J. Assessing petroleum oils biodegradation by chemometric analysis of spectroscopic data. *Talanta*, **75**, 857-871, 2008.

2. 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.

3. ADAMS, M. J. Chemometrics in analytical spectroscopy. Cambrigde: The Royal Society of Chemistry, 1995, 216 p.

4. ALI, L. H.; AL-GHANNAM, K. A.; AL-RAWI, J. M. Chemical structure of asphaltenes in heavy crude oils investigated by n.m.r. *Fuel*, **69**, 519–521, 1990.

5. ALVES, J. C. L.; POPPI, R. J. Diesel oil quality parameter determinations using support vector regression and near infrared Spectroscopy for hydrotreating feedstock monitoring. *J. Near Infrared. Spectrosc.*, **20**, 419-425, 2012.

6. ANDERSSON, M. A comparison of nine PLS1 algorithms. *J. Chemometrics*, **23**, 518-529, 2009.

7. APETREI, I. M.; RODRÍGUEZ-MÉNDEZ, M. L.; APETREI, C.; NEVARES, I.; ALAMO, M.; SAJA, J. A. Monitoring of evolution during red wine aging in oak barrels and alternative method by means of an electronic panel test. *Food Res. Int.*, **45**, 244-249, 2012.

8. API. American petroleum institute. Disponível em: https://www.api.org/. Acessado em 04/07/2021.

9. ASKE, N.; KALLEVIK, H.; SJOBLOM, J. Determination of saturate, aromatic, resin, and asphaltenic (sara) components in crude oils by means of infrared and near-infrared spectroscopy. *Energ. Fuels*., **15**, 1304-12, 2001.

10. ASTM. Annual Book of ASTM Standards. Standards practices for infrared, multivariate, quantitative analysis, E1655, vol 03.06. ASTM International, West Conshohocken, Pennsylvania, USA, 2000.

11. ASTM D2549-02, Standard Test Method for Separation of Representative Aromatics and Nonaromatics Fractions of High-Boiling Oils by Elution Chromatography, ASTM International, West Conshohocken, PA, 2017.

12. ASTM D4294-16, Standard Test Method for Sulfur in Petroleum and Petroleum Products by Energy Dispersive X-ray Fluorescence Spectrometry, ASTM International, West Conshohocken, PA, 2016.

13. ASTM D4629-17, Standard Test Method for Trace Nitrogen in Liquid Hydrocarbons by Syringe/Inlet Oxidative Combustion and Chemiluminescence Detection, ASTM International, West Conshohocken, PA, 2017.

14. ASTM D664-04, Standard test method for acid number of petroleum products by potentiometric titration; ASTM International, West Conshohocken, PA, 2004.

15. BAIRD, Z. S.; OJA, V. Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density. *Chemom. Intell. Lab. Syst.*, **158**, 41−77, 2016.

16. BAJOUB, A.; MEDINA-RODRÍGUEZ, S.; GÓMEZ-ROMERO, M.; AJAL, E.; BAGUR-GONZÁLEZ, M. G.; FERNÁNDEZ-GUTIÉRREZ, A.; CARRASCO-PANCORBO, A. Assessing the varietal origin of extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data fusion and chemometrics. *Food Chem.,* **215***,* 245-255, 2016.

17. BALLABIO, D., TODESCHINI, R., CONSONNI, V. Recent advances in high-level fusion methods to classify multiple analytical chemical data. *Data Fusion Methodology and Applications*, **31**, 129–155, 2019.

18. BARBOSA, L. L.; SAD, C. M. S.; MORGAN, V. G.; FILGUEIRAS, P. R.; CASTRO, E. R. V. Application of low field NMR as an alternative technique to quantification of total acid number and sulphur content in petroleum from Brazilian reservoirs. *Fuel*, **176,**146-152, 2016.

19. BARKER, M.; RAYENS, W. Partial least squares for discrimination. *J. Chemom*., **17**, 166–173, 2003.

20. BARNETT, I.; ZHANG, M. Discrimination of brands of gasoline by using DART-MS and chemometrics. *Forensic Chem*., 10, 58-66, 218.

21. BARRERA, D. R.; CHACÓN, H. R.; MOLINA, D. Prediction of the total acid number (TAN) of colombian crude oils via ATR–FTIR spectroscopy and chemometric methods. *Talanta*, **206**, 2020.

22. BEVILACQUA, M.; MARINI, F.; RINNAN, A.; RASMUSSEN, M. A.; SKOV, T. Recent chemometrics advances for Foodomics. *Trends Anal. Chem*., **96**, 42–51, 2017.

23. BIANCOLILLO, A.; BUCCI, R.; MAGRÌ, A. L.; MAGRÌ, A. D.; MARINI, F. Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication. *Anal. Chim. Acta*, **820**, 23-31, 2014.

24. BLANCHET, L.; SMOLINSKA, A. Data fusion in metabolomics and proteomics for biomarker discovery. *Statistical Analysis in Proteomics*, **1362**, 209-223, 2016.

25. BORRÀS, E.; FERRE, J.; BOQUE, R.; MESTRES, M.; ACENA, L.; BUSTO, O. Data fusion methodologies for food and beverage authentication and quality assessment - a review. *Anal. Chim. Acta*, **891**, 1-14, 2015.

26. BRAKSTAD, F.; KARSTANG, T. V.; SØRENSEN, J.; STEEN, A. Prediction of molecular weight and density of distillation fractions from gas chromatographic - mass spectrometric detection and multivariate calibration. *Chemometr. Intell. Lab*., **3**, 321-328, 1988.

27. BREKKE, T.; KVALHEIM, O. M.; SLETTEN, E. Assignment of $^{13}C$ nuclear magnetic resonance spectra of complex mixtures by multivariate analysis. *Chemometr. Intell. Lab.*, **7**, 101-17, 1989.

28. BRERETON, R. G. Introduction to multivariate calibration in analytical chemistry. *Analyst.*, **125**, 2125-2154, 2000.

29. BRERETON, R. G. Applied chemometrics for scientists. Chichester: John Wiley & Sons Ltd, 2007, 379 p.

30. BURG, P.; SELVES, J. L.; COLIN, J. P. Numerical simulation of crude oil behaviour from chromatographic data. *Anal. Chim. Acta*, **317**, 107-25, 1995.

31. CASALE, M.; CASOLINO, C.; OLIVERI, P.; FORINA, M. The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil, *Food Chem*., **118**, 163-170, 2010.

32. CASTANEDO, F. A Review of Data Fusion Techniques. The *Scientific World Journal*, 2013.

33. CUEVAS, F. J.; PEREIRA-CARO, G.; MORENO-ROJAS, J. M.; MUÑOZ-REDONDO, J. M.; RUIZ-MORENO, M. J. Assessment of premium organic orange juices authenticity using HPLC-HR-MS and HS-SPME-GC-MS combining data fusion and chemometrics. *Food Control*, **82**, 203-211, 2017.

34. DEARING, T.; THOMPSON, W.; RECHSTEINER, C. Characterization of crude oil products using data fusion of process raman, infrared, and nuclear magnetic resonance (NMR) spectra. *Appl. Spectrosc*., **65**, 181-6, 2011.

35. DOUGLAS, R. K.; NAWAR, S.; ALAMAR, M. C.; MOUAZEN, A. M.; COULON, F. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. *Sci. Total Environ*., **617**, 147-155, 2018.

36. DOWNEY, G.; BRIANDET, R.; WILSON, R. H.; KEMSLEY, E. K. Near- and mid-infrared spectroscopies in food authentication: coffee varietal identification. *J. Agric. Food Chem.*, **45**, 4357-4361, 1997.

37. DUARTE, L. M.; FILGUEIRAS, P. R.; SILVA, S. R. C.; DIAS, J. C. M.; OLIVEIRA, L. M. S. L.; CASTRO, E. V. R.; OLIVEIRA, M. A. L. Determination of some physicochemical properties in Brazilian crude oil by $^{1}$H NMR spectroscopy associated to chemometric approach. *Fuel*, **181**, 660-669, 2016.

38. DUARTE, L. M.; FILGUEIRAS, P. R.; DIAS, J. C. M.; OLIVEIRA, L M. S. L.; CASTRO, E. V. R.; OLIVEIRA, M. A. L. Study of distillation temperature curves from brazilian crude oil by $^{1}$H NMR-PLS. Energy Fuels, **31**, 3892-3897, 2017.

39. EBRAHIMI, D.; LI, J.; HIBBERT, D. B. Classification of weathered petroleum oils by multi-way analysis of gas chromatography–mass spectrometry data using PARAFAC2 parallel factor analysis. *J. Chromatogr. A.*, **1166**, 163–70, 2007.

40. FALLA, F. S.; LARINI, C.; ROUX, G. A.; QUINA, F. H.; MORO, L. F. L., NASCIMENTO, C. A. O. Characterization of crude petroleum by NIR. *J. Petrol. Sci. Eng*., **51**,127–37, 2006.

41. FERRER-CID, P.; BARCELO-ORDINAS, J. M.; GARCIA-VIDAL, J.; RIPOLL, A.; VIANA, M. Multi-sensor data fusion calibration in IoT air pollution platforms. *IEEE Internet of Things Journal*, **7**, 3124-3132, 2020.

42. FILGUEIRAS, P. R.; SAD, C. M. S.; LOUREIRO, A. R.; SANTOS, M. F. P.; CASTRO, E. V. R.; DIAS, J. C. M.; POPPI, R. J. Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. *Fuel*, **116**, 123-30, 2014a.

43. FILGUEIRAS, P. R.; ALVES, J. C. L.; SAD, C. M. S.; CASTRO, E. V. R.; DIAS, J. C. M.; POPPI, R. J. Evaluation of trends in residuals of multivariate calibration models by permutation test. *Chemom. Intell. Lab. Syst.*, **133**, 33-41, 2014b.

44. FILGUEIRAS, P. R.; TERRA, L. A.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; POPPI, R. J. Prediction of the distillation temperatures of crude oils

using $^{1}$H NMR and support vector regression with estimated confidence intervals. *Talanta*, **142**, 197-205, 2015.

45. FILGUEIRAS, P. R.; PORTELA, N. A.; SILVA, S. R. C.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; NETO, A. C.; ROMÃO, W.; POPPI, R. J. Determination of saturates, aromatics, and polars in crude oil by $^{13}$C NMR and support vector regression with variable selection by genetic algorithm. Fuel, **30**, 1972-1978, 2016.

46. FLOUDAS, N.; POLYCHRONOPOULOS, A.; AYCARD, O.; BURLET, J.; AHRHOLDT, M. High level sensor data fusion approaches for object recognition in road environment. In: IEEE Intelligent Vehicles Symposium, 2007, Istambul – Turquia. p. 136.

47. FLUMIGNAN, D. L.; FERREIRA, F. O.; TININIS, A. G.; OLIVEIRA, J. E. Multivariate calibrations in gas chromatographic profiles for prediction of several physicochemical parameters of Brazilian commercial gasoline. *Chemometr. Intell. Lab. Syst.*, **92**, 53-60, 2008.

48. FOLLI, G. S.; NASCIMENTO, M. H. C.; DE PAULO, E. H.; DA CUNHA, P. H. P.; ROMÃO, W.; FILGUEIRAS, P. R. Variable selection in support vector regression using angular search algorithm and variance inflation factor. *J. Chemom.*, **34**, e3282, 2020.

49. GALTIER, O.; ABBAS, O.; LE DRÉAU, Y.; REBUFA, C.; KISTER, J.; ARTAUD, J.; DUPUY, N. Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions. *Vib. Spectrosc*., **55**, 132-140, 2011.

50. GAUTAM, K.; JIN, X. H.; HANSEN, M. Review of spectrometric techniques for the characterization of crude oil and petroleum products. *Appl. Spectrosc. Rev.*, **33**, 427-443, 1998.

51. GEURTS, B. P.; ENGEL, J.; RAFII, B.; BLANCHET, L.; SUPPERS, A.; SZYMANSKA, E.; JANSEN, J. J.; BUYDENS, L. M. C. Improving high-dimensional data fusion by exploiting the multivariate advantage. *Chemometr. Intell. Lab. Syst.*, **156**, 231–240, 2016.

52. GODOY, L. A. F.; PEDROSO, M. P.; FERREIRA, E. C.; AUGUSTO, F.; POPPI, R.J. Prediction of the physicochemical properties of gasoline by comprehensive two-dimensional gas chromatography and multivariate data processing. *J. Chromatogr. A.*, **1218**, 1663-1667, 2011.

53. GUIDA, R.; Ng, S. W.; Lervolino, P. S- and X-band SAR data fusion. In: IEEE 5th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR), 2015, Cingapura. P. 578.

54. GUO, H; WU, D; AN, J. Discrimination of oil slicks and lookalikes in polarimetric SAR images using CNN. *Sensors*, **17**, 1837-1856, 2017.

55. HASAN, M.; ALI, M.; BUKHARI, A. Structural characterization of Saudi Arabian heavy crude oil by n.m.r. spectroscopy. *Fuel*, **62**, 518-23, 1983.

56. HAWARE, R. V.; WRIGHT, P. R.; MORRIS, K. R.; HAMAD, M. L. Data fusion of Fourier transform infrared spectra and powder X-ray diffraction patterns for pharmaceutical mixtures. *J. Pharm. Biomed. Anal*., **56**, 944–949, 2011.

57. HIDAJAT, K.; CHONG, S. M. Quality characterisation of crude oils by partial least square calibration of NIR spectral profiles. *J. Near Infrared Spectrosc*., **8**, 53-9, 2000.

58. ISO12185:1996. Crude petroleum and petroleum products-determination of density - Oscillating U-tube method. 1996.

59. JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R.; An introduction to statistical learning. Nova Iorque: Springer, 2013.

60. JINGYAN, L.; XIAOLI, C.; SONGBAI, T. Research on determination of total acid number of petroleum using mid-infrared attenuated total reflection spectroscopy. *Energ. Fuel*, **26**, 5633–5637, 2012.

61. KALLEVIK, H.; HANSEN, S. B.; SAETHER, O.; KVALHEIM, O. M.; SJÖBLOM, J. Crude oil model emulsion characterised by means of near infrared spectroscopy and multivariate techniques. *J. Disper. Sci. Technol*., **21**, 245–62, 2000.

62. KENNARD, R. W.; STONE, L.A. Computer-aided Design of Experiments. Technometrics, **11**, 137-148, 1969.

63. KHALEGHI, B.; KHAMIS, A.; KARRAY, F. O.; RAZAVI, S. N. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, **14**, 28–44, 2013.

64. KHANMOHAMMADI, M.; GARMARUDI, A. B.; GUARDIA, M. Characterization of Petroleum based products by infrared spectroscopy and chemometrics. *Trends Analyt. Chem.*, **35**, 135-149, 2012.

65. KHUHAWAR, M. Y.; MIRZA, M. A.; JAHANGIR, T. M. Determination of metal ions in crude oils. In: Crude oil emulsions- Composition, stability and characterization, 121-144, 2012.

66. KVALHEIM, O. M.; AKSNES, D. W.; BREKKE, T.; EIDE, M. O.; SLETTEN, E. Crude oil characterization and correlation by principal component analysis of $^{13}C$ nuclear magnetic resonance spectra. *Anal. Chem.*, **57**, 2858-64, 1985.

67. KVALHEIM, O. M. Oil-source correlation by the combined use of principal component modelling, analysis of variance and a coefficient of congruence. Chemometr. Intell. Lab., **2**, 127-36, 1987.

68. LAHAT, D.; ADALI, T.; JUTTEN, C. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, **103**, 1449–1477, 2015.

69. LAMBERT, D.; DESCALES, B.; LLINAS, J.; MARTENS, A. On-line NIR monitoring and optimization for refining and petrochemical processes. *Analysis Magazine*, 23, 9-13, 1995.

70. LAXALDE, J.; RUCKEBUSCH, C.; DEVOS, O.; CAILLOL, N.; WAHL, F.; DUPONCHEL, L. Characterisation of heavy oils using near-infrared spectroscopy: Optimisation of pre-processing methods and variable selection. *Anal. Chim. Acta*, **705**, 227–234, 2011.

71. LEE, S.; CHOI, H.; CHA, K.; CHUNG, H. Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. Microchem. J., **110**, 739-48, 2013.

72. LI, A.; YE, D.; LYU, E.; SONG, S.; MENG, M..; DE SILVA, C. W. RGB-Thermal Fusion Network for Leakage Detection of Crude Oil Transmission Pipes. In: International Conference on Robotics and Biomimetics (ROBIO), 2019, Dali - China.

73. LI, H.; LIANG, Y.; XU, Q.; CAO, D. Model population analysis for variable selection. J. Chemom., **24**, p. 418–423, 2010.

74. 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.

75. LIU, W.; ZHAO, Z.; YUAN, H.; SONG, C.; LI, X. An optimal selection method of samples of calibration set and validation set for spectral multivariate analysis. *Guang Pu Xue Yu Guang Pu Fen Xi*, **34**, 947-951, 214.

76. LIRA, L. F. B.; VASCONCELOS, F. V. C.; PEREIRA, C. F.; PAIM, A. P. S.; STRAGEVITCH, L.; PIMENTEL, M. F. Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration. *Fuel*, **89**, 405-9, 2010.

77. LONG, J.; WANG, K.; YANG, M.; ZHONG, W. Rapid crude oil analysis using near-infrared reflectance spectroscopy. *Petrol. Sci. Technol.*, **37**, 354-60, 2019.

78. LOVATTI, B. P. O.; NASCIMENTO, M. H. C.; NETO, A. C.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Use of random forest in the identification of important variables. Microchem. J., **145**, 1129-1134, 2019.

79. LOVATTI, B. P. O.; NASCIMENTO, M. H. C.; RAINHA, K. P.; OLIVEIRA, E. C. S.; NETO, A. C.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Different strategies for the use of random forest in NMR spectra. *J. Chemom.*, e3231, 2020.

80. LOZANO, D. C. P.; RUIZ, J. A. O.; HERNÁNDEZ, R. C.; GUERRERO, J. E.; OSPINO, E. M. APPI(+)-FTICR mass spectrometry coupled to partial least squares with genetic algorithm variable selection for prediction of API gravity and CCR of crude oil and vacuum residues. *Fuel*, **193**, 39-44, 2017.

81. LU, J.; DING, J.; LIU, C.; JIN, Y. Prediction of physical properties of crude oil based on ensemble random weights neural network. *International Federation of Automatic Control*, **51**, 655-660, 2018.

82. MABOOD, F.; GILANI, S. A.; ALBROUMI, M.; ALAMERI, S.; AL NABHANI, M. M. O.; JABEEN, F.; HUSSAIN, J.; AL-HARRASI, A.; BOQUÉ, R.; FAROOQ, S.; HAMAED, A. M.; NAUREEN, Z.; KHAN, A.; HUSSAIN, Z. Detection and estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with multivariate methods. Fuel, **197**, 388-396, 2017.

83. MAIMAITIJIANG, M.; GHULAM, A.; SIDIKE, P.; HARTLING, S.; MAIMAITIYIMING, M.; PETERSON, K.; SHAVERS, E.; FISHMAN, J.; PETERSON, J.; KADAM, S.; BURKEN, J.; FRITSCHI, F. Unmanned aerial system (UAS) - based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. *Journal of Photogrammetry and Remote Sensing*, **134**, 43–58, 2017.

84. MASILI, A.; PULIGHEDDU, S.; SASSU, L.; SCANOA, P.; LAI, A. Prediction of physical–chemical properties of crude oils by $^{1}$H NMR analysis of neat samples and chemometrics. *Magn. Reson. Chem*., **50**, 729-738, 2012.

85. MARCHI, M.; MANUELIAN, C. L.; TON, S.; MANFRIN, D.; MENEGHESSO, M.; CASSANDRO, M.; PENASA, M. Prediction of sodium content in commercial processed meat products using near infrared spectroscopy. *Meat Science*, **125**, 61-65, 2017.

86. MAZIVILA, S. J.; GONTIJO, L. C.; SANTANA, F. B.; DE MITSUTAKE, H.; SANTOS, D. Q.; BORGES NETO, W. Fast detection of adulterants/contaminants in biodiesel/diesel blend (B5) employing mid-infrared spectroscopy and PLS-DA. *Energy Fuels*, **29**, 227-232, 2014.

87. MEJIA-MIRANDA, C.; LAVERDE, D.; MOLINA, V. D. Correlation for predicting corrosivity of crude oils using proton nuclear magnetic resonance and chemometric methods. *Energy Fuels*, **29**, 7595-7600, 2015.

88. MELENDEZ, L.; LACHE, A.; RUIZ, O.; PACHO, Z.; OSPINO, E. M. Prediction of the SARA analysis of colombian crude oils using ATR–FTIR spectroscopy and chemometric methods. *J. Petrol. Sci. Eng*., **90-91**, 56–60, 2012.

89. MENDES, R. K.; DANTAS, M. V. C.; NOGUEIRA, A. B.; ETCHEGARAY, A.; FILGUEIRAS, P. R.; POPPI, R. J. Simultaneous determination of different phenolic compounds using electrochemical biosensor and multivariate calibration. *J. Braz. Chem. Soc*., **29**, 482-489, 2018.

90. MENDOZA, F.; LU, R.; CEN, H. Comparison and fusion of four nondestructive sensors for predicting apple fruit firmness and soluble solids content. *Postharvest Biology and Technology*, **73**, 89–98, 2012.

91. MERDRIGNAC, I.; ESPINAT, D. Physicochemical characterization of petroleum fractions: The state of the art. *Oil & Gas Science and Technology*, **62**, 7-32, 2007.

92. MILLER, J. N.; MILLER, J. C. Statistics and chemometrics for analytical chemistry. 6ª Ed. New York: Prentice Hall, 2010. 278 p.

93. MISPELAAR, V. G.; SMILDE, A. K.; NOORD, O. E.; BLOMBERG, J.; SCHOENMAKERS, P. J. Classification of highly similar crude oils using data sets from comprehensive two-dimensional gas chromatography and multivariate techniques. *J Chromatogr. A.*, **1096**, 156-164, 2005.

94. MOLINA, D.; URIBE, U. N.; MURGICH, J. Partial least-squares (PLS) correlation between refined product yields and physicochemical properties with the $^{1}$H nuclear magnetic resonance (NMR) spectra of colombian crude oils. *Energ. Fuel*, **21**, 1674-80, 2007.

95. MORO, M. K.; NETO, A. C.; LACERDA, J. R. V.; ROMÃO, W.; CHINELATTO, JR. L. S.; CASTRO, E. V. R.; FILGUEIRAS, P. R. FTIR, $^{1}$H and $^{13}$C NMR data fusion to predict crude oils properties. *Fuel*, **263**, 2019.

96. MOHAMMADI, M.; KHORRAMI, M. K.; VATANI, A.; GHASEMZADEH, H.; VATANPARAST, H.; BAHRAMIAN, A.; FALLAH, A. Rapid determination and classification of crude oils by ATR-FTIR spectroscopy and chemometric methods. *Spectrochim. Acta, Part A*, **232**, 2020.

97. MUHAMMAD, A.; AZEREDO, R. B. V. $^{1}$H NMR spectroscopy and low-field relaxometry for predicting viscosity and API gravity of Brazilian crude oils – A comparative study. *Fuel*, **130**, 126-34, 2014

98. MÜLLER, A. L. H.; PICOLOTO, R. S.; MELLO, P. A.; FERRÃO, M. F.; SANTOS, M. F. P.; GUIMARÃES, R. C. L.; MÜLLER, E. I.; FLORES, E. M. M. Total sulfur determination in residues of crude oil distillation using FT-IR/ATR and variable selection methods. *Spectrochim. Acta, Part A*, **89**, 82-87, 2012.

99. MULLINS, O. C.; SHEU, E. Y.; HAMMAMI, A.; MARSHALL, A. G. Asphaltenes, heavy oils, and petroleomics. 1ª Ed. Nova Iorque: Springer-verlag, 2007.

100. NESPECA, M. G.; MUNHOZ, J. F. V. L.; FLUMIGNAN, D. L.; OLIVEIRA, J. E. Rapid and sensitive method for detecting adulterants in gasoline using ultra-fast gas chromatography and partial least square discriminant analysis. *Fuel*, **215**, 204-211, 2018.

101. OBISESAN, K. A.; JIMÉNEZ-CARVELO, A. M.; CUADROS-RODRIGUEZ, L.; RUISÁNCHEZ, I.; CALLAO, M. P. HPLC-UV and HPLC-CAD chromatographic data fusion for the authentication of the geographical origin of palm oil. *Talanta*, **170**, 413-418, 2017.

102. OZIGIS, M. S.; KADUK, J. D.; JARVIS, C. H. (2018). Mapping terrestrial oil spill impact using machine learning random forest and Landsat 8 OLI imagery: a case site within the Niger Delta region of Nigeria. *Environmental Science and Pollution Research*, **26**, 3621-3635, 2018.

103. OYGARD, K.; GRAHL-NIELSEN, O.; ULVOEN, S. Oil/oil correlation by aid of chemometrics. *Org. Geochem*., **6**, 561-7, 1984.

104. PABÓN, R. E. C.; SOUZA FILHO, C R. Crude oil spectral signatures and empirical models to derive API gravity. *Fuel*, 237, 1119–1131, 2019.

105. PAULO, E. H.; FOLLI, G. S.; NASCIMENTO, M. H. C.; MORO, M. K.; CUNHA, P. H. P.; CASTRO, E. V. R.; NETO, A. C.; FILGUEIRAS, P. R. Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties. *Fuel*, **279**, 2020.

106. PASQUINI, C. Near infrared spectroscopy: A mature analytical technique with new perspectives - A review. *Anal. Chim. Acta*, **1026**, 8-36, 2018.

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

108. PAVIA, D. L.; LAMPMAN, G. M.; KRIZ, G. S.; VYVYAN, J. R. Introduction to Spectroscopy. 4ª Ed. Cengage Learning, 2008. 752 p.

109. PEINDER, P.; VISSER, T.; PETRAUSKAS, D. D.; SALVATORI, F.; SOULIMANI, F.; WECKHUYSEN, B. M. Partial least squares modeling of combined infrared, $^{1}$H NMR and $^{13}$C NMR spectra to predict long residue properties of crude oils. *Vib. Spectrosc*., **51**, 205-212, 2009.

110. POJIĆ M, MASTILOVIĆ J, PALIĆ D, PESTORIĆ M. The development of near-infrared spectroscopy (NIRS) calibration for prediction of ash content in legumes on the basis of two different reference methods. *Food Chem.*, **123**, 800-805, 2010.

111. POVEDA, J. C.; MOLINA, D. R. Average molecular parameters of heavy crude oils and their fractions using NMR spectroscopy. *J. Petrol. Sci. Eng.*, **84-85**, 1-7, 2012.

112. PRADO, G. H. C.; RAO, Y.; KLERK, A. Nitrogen removal from oil: a review. *Energy Fuels*, **31**, 14-36, 2017.

113. RAINHA, K. P.; ROCHA, J. T. C.; RODRIGUES, R. R. T.; LOVATTI, B. P. O.; CASTRO, E. V. R.; FILGUEIRAS, P. R. 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.

114. RAMBO, M. K.; FERREIRA, M. M. C.; MELO, P. M.; JUNIOR, C. C. S.; BERTULOL, D. A.; RAMBO, M. C. D. Prediction of quality parameters of food residues using NIR spectroscopy and PLS models based on proximate analysis. *LWT - Food Sci. Technol.*, **40**, 444-450, 2020.

115. RAMOS, P. F. O.; TOLEDO, I. B.; NOGUEIRA, C. M.; NOVOTNY, E. H.; VIEIRA, A. J. M.; AZEREDO, R. B. V. Low field $^{1}$H NMR relaxometry and multivariate data analysis in crude oil viscosity prediction. *Chemom. Intell. Lab. Syst.*, **99**, 121-126, 2009.

116. RIAZI, M. R. Characterization and properties of petroleum fractions. 1ª Ed. ASTM Stock Number: MNL50, 2005. 407 p.

117. RODRIGUES, R. R. T.; ROCHA, J. T. C.; OLIVEIRA, M. S. L.; DIAS, J. C. M.; MULLER, E. I.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Evaluation of calibration transfer methods using the ATR-FTIR technique to predict density of crude oil. *Chemometr. Intell. Lab. Syst.*, **166**, 7-13, 2017.

118. RODRIGUES, E. 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.

119. ROUSEEL, S.; BELLON-MAUREL, V.; ROGER, J. M.; GRENIER, P. Autenticating white grape must variety with classification models based on aroma sensors, FT-IR and UV spectrometry. *J. Food. Eng.*, **60**, 407-19, 2003.

120. RUDSZUCK, T.; FÖRSTER, E.; NIRSCHL, H.; GUTHAUSEN, G. Low field NMR for quality control on oils. *Magn. Reson. Chem*., **57**, 777-793, 2019.

121. SANTOS, F. D.; SANTOS, L. P.; CUNHA, P. H. P.; BORGHI, F. T.; ROMÃO, W.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Discrimination of oils and fuels using a portable NIR spectrometer. *Fuel*, **283**, 2021.

122. SAVORANI, F.; TOMASI, G.; ENGELSEN, S. B. Alignment of 1D NMR data using the iCoshift tool: A tutorial. In: Magnetic Resonance in Food Science: From Food to Thought Edition: Special Publication No. 343, Publisher: The Royal Society of Chemistry, 14-24, 2013.

123. SHI, T.; CHEN, Y.; LIU, Y.; WU, G. Visible and near-infrared reflectance spectroscopy - An alternative for monitoring soil contamination by heavy metals. *J. Hazard. Mater*., **265**, 166-176, 2014.

124. SMOLINSKA, A.; BLANCHET, L.; COULIER, L.; AMPT, K. A. M.; LUIDER, T.; HINTZEN, R. Q.; WIJMENGA, S. S.; BUYDENS, L. M. Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis. *Plos One*, **7**, 2012.

125. 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 Fusion Methodology and Applications, *Data Handling in Science and Technology*, **31**, 51–79, 2019.

126. SPEIGHT, J. G. Petroleum asphaltenes part 1 - Asphaltenes, resins and the structure of petroleum. *Oil Gas Sci. Technol*., **59**, 467-477, 2004.

127. SPEIGHT, J. G. Handbook of petroleum product analysis. 2ª Ed. Nova Jérsia: Jonh Wiley & Sons Inc. Hoboken, 2015.

128. STEINMETZ, V.; SÉVILA, F.; BELLON-MAURELL, V. A methodology for sensor fusion design: application to fruit quality assessment. *J. Agr. Eng. Res*., **74**, 21-31, 1999.

129. SUN, D. W. Infrared spectroscopy for food quality analysis and control. Nova Iorque: Academic Press, 2009.

130. TAN, J.; LI, R.; JIANG, Z. Chemometric classification of Chinese lager beers according to manufacturer based on data fusion of fluorescence, UV and visible spectroscopies. *Food Chem.*, **184**, 30-36, 2015.

131. TIANGUANG, F.; JIANXIN, W.; BUCKLEY, J. Evaluating crude oils by SARA analysis. In: Improved Oil Recovery Symposium, 2002, Tulsa – Oklahoma.

132. THOMAS, J. E. Fundamentos *de* Engenharia de Petróleo. 2ª Ed. Rio de janeiro: Interciência, 2001. 271 p.

133. TELNAES, N.; BJORSETH, A.; CHRISTY, A. A.; KVALHEIM, O. M. Interpretation of multivariate data: Relationship between phenanthrenes in crude oils. *Chemometr. Intell. Lab*., **2**, 149-153, 1987.

134. TERRA, L. A.; FILGUEIRAS, P. R.; TOSE, L. V.; ROMÃO, W.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; VAZ, B. G.; POPPI, R. J. Laser desorption ionization FT-ICR mass spectrometry and CARSPLS for predicting basic nitrogen and aromatics contents in crude oils. *Fuel*, **160**, 274-281, 2015.

135. TOMASI, G.; SAVORANI, F.; ENGELSEN, S. B. Icoshift: an effective tool for the alignment of chromatographic data. *J. Chromatogr. A.*, **1218**, 7832-7840, 2012.

136. TRYGG, J.; WOLD, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.*, **16**, 119-128, 2002.

137. UOP Method 269-10, Nitrogen bases in hydrocarbons by potentiometric titration. West Conshohocken, PA, USA: American Society for Testing and Materials, 2010.

138. URDAL, K.; VOGT, N. B.; SPORSTOL, S. P.; LICHTENTHALER, R. G.; MOSTAD, H.; KOLSET, K.; NORDENSON, S.; ESBENSEN, K. Classification of weathered crude oils using multimethod chemical analysis, statistical methods and SIMCA pattern recognition. *Mar. Pollut. Bull.*, **17**, 366-373, 1986.

139. VEMPATAPU, B. P.; KANAUJIA, P. K. Monitoring petroleum fuel adulteration: a review of analytical methods. *Trends Analyt. Chem*., **92**, 1-11, 2017.

140. VENTURA, G. T.; HALL, G. J.; NELSON, R. K.; FRYSINGER, G. S.; RAGHURAMAN, B.; POMERANTZ, A. E.; MULLINS, O. C.; REDDY, C. M. Analysis of petroleum compositional similarity using multiway principal components analysis (MPCA) with comprehensive two-dimensional gas chromatographic data. *J. Chromatogr. A.*, **1218**, 2584-2592, 2011.

141. VERA, L.; ACEÑA, L.; GUASCH, J.; BOQUÉ, R.; MESTRES, M.; BUSTO, O. Discrimination and sensory description of beers through data fusion. *Talanta*, 87, 136–142, 2011.

142. VIEIRA, A. P.; PORTELA, N. A.; NETO, A. C.; LACERDA, V.; ROMÃO, W.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Determination of physicochemical properties of petroleum using $^{1}$H NMR spectroscopy combined with multivariate calibration. *Fuel*, **253**, 320-326, 2019.

143. VOET, H. V. D. Comparing the predictive accuracy of models using a simple randomization test. *Chemom. Intell. Lab. Syst*., **25**, 313-23, 1994.

144. WAPLES, D. W. Geochemistry in petroleum exploration. 1ª Ed. Boston: International Human Resources Development Corporation, 1985. 217 p.

145. WILT, B. K.; WELCH, W. Determination of asphaltenes in petroleum crude oils by fourier transform infrared spectroscopy. *Energy Fuels*, **12**, 1008-1012, 1998.

146. WOLD, H. Soft modeling by latent variables; the nonlinear iterative partial least squares approach, In: Perspectives in probability and statistics, 1975, J. Gani, Academic Press.

147. WOLD, S.; SJOSTROM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.*, **58**, 109–13, 2001.

148. WOODS, J.; KUNG, J.; KINGSTON, D.; KOTLYAR, L.; SPARKS, D.; MCCRACKEN, T. Canadian Crudes: A comparative study of SARA fractions from a modified HPLC separation technique. *Oil Gas Sci. Technol*., **1**, 151-63, 2008.

149. XIAOBO, Z.; JIEWEN, Z. Apple quality assessment by fusion three sensors. In: Sensors, 2005, Irvine – USA. IEE. p. 389.

150. YANG, B.; GAO, Y.; LI, H.; YE, S.; HE, H.; XIE, S. Rapid prediction of yellow tea free amino acids with hyperspectral images. *Plos One*, **14**, e0210084, 2019.

151. ZAFAR, F.; MANDAL, P. C.; SHAARI, K. Z. K.; MONIRUZZAMAN, M. Total acid number reduction of naphthenic acid using subcritical methanol and 1-butyl-3-methylimidazolium octylsulfate. *Procedia Engineering*, **148**, 1074-1080, 2016.

152. ZUDE, M.; HEROLD, B.; ROGER, J. M.; MAUREL, V. B.; LANDAHL, S. Non-destructive tests on the prediction of apple fruit flesh firmness and soluble solids content on tree and in shelf life. *J. Food Eng.*, **77**, 254-260, 2006.

## Apêndice A – Propriedades físico-químicas das amostras utilizadas neste trabalho

**Tabela A 1:** Propriedades físico-químicas das amostras utilizadas neste trabalho.

| Amostra / Propriedade | °API | S (%m/m) | NT (%m/m) | NB (%m/m) | NAT ($mgKOHg^{-1}$) | SAT (%m/m) | ARO (%m/m) | POL (%m/m) |
|---|---|---|---|---|---|---|---|---|
| AP-QMT-0001 | 54,0 | 0,084 | 0,0012 | 0,0009 | 0,060 | 88,40 | 11,60 | 0,00 |
| AP-QMT-0002 | 41,9 | 0,171 | 0,0716 | 0,0208 | 0,170 | 71,90 | 16,20 | 11,90 |
| AP-QMT-0003 | 17,0 | 0,562 | 0,3892 | 0,1360 | 3,090 | 41,60 | 29,80 | 28,60 |
| AP-QMT-0004 | 32,5 | 0,069 | 0,2800 | 0,0980 | 0,180 | 59,70 | 10,60 | 29,70 |
| AP-QMT-0005 | 25,5 | 0,351 | 0,2717 | 0,0908 | 1,180 | 58,60 | 23,40 | 18,00 |
| AP-QMT-0006 | 23,6 | 0,234 | 0,1300 | 0,0432 | 2,640 | 62,20 | 19,40 | 18,46 |
| AP-QMT-0007 | 29,9 | 0,216 | 0,2800 | 0,0859 | 0,070 | 52,70 | 33,60 | 13,57 |
| AP-QMT-0008 | 53,3 | 0,078 | 0,0014 | 0,0010 | 0,080 | 89,20 | 10,80 | 0,00 |
| AP-QMT-0009 | 11,4 | 0,520 | 0,3400 | 0,1300 | 3,400 | 37,00 | 28,70 | 33,00 |
| AP-QMT-0010 | 26,8 | 0,412 | 0,3529 | 0,1260 | 0,290 | 53,80 | 24,10 | 22,10 |
| AP-QMT-0011 | 26,7 | 0,500 | 0,2960 | 0,1130 | 0,190 | 56,00 | 26,90 | 17,10 |
| AP-QMT-0012 | 39,5 | 0,109 | 0,0565 | 0,0162 | 0,060 | 84,00 | 8,40 | 7,60 |
| AP-QMT-0013 | 29,0 | 0,310 | 0,3173 | 0,1100 | 0,100 | 52,10 | 27,20 | 20,70 |
| AP-QMT-0014 | 28,4 | 0,308 | 0,3355 | 0,1070 | 0,170 | 51,80 | 21,90 | 26,35 |
| AP-QMT-0015 | 29,4 | 0,360 | 0,2535 | 0,1020 | 0,350 | 58,30 | 21,90 | 19,80 |
| AP-QMT-0016 | 24,4 | 0,510 | 0,3477 | 0,1330 | 0,580 | 44,80 | 22,80 | 32,40 |
| AP-QMT-0017 | 29,6 | 0,361 | 0,2527 | 0,1030 | 0,240 | 60,30 | 22,70 | 17,00 |
| AP-QMT-0018 | 22,0 | 0,555 | 0,3747 | 0,1400 | 1,730 | 44,20 | 26,50 | 29,30 |
| AP-QMT-0019 | 30,8 | 0,323 | 0,2200 | 0,0813 | 0,050 | 57,70 | 24,20 | 18,01 |
| AP-QMT-0020 | 30,6 | 0,090 | 0,2400 | 0,0891 | 0,278 | 61,50 | 23,10 | 15,35 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AP-QMT-0021 | 23,4 | 0,962 | 0,4700 | 0,4700 | 0,350 | 48,00 | 28,60 | 23,20 |
| AP-QMT-0022 | 28,7 | 0,385 | 0,2836 | 0,1130 | 0,510 | 58,40 | 19,40 | 22,20 |
| AP-QMT-0023 | 29,2 | 0,352 | 0,2800 | 0,1040 | 0,270 | 54,00 | 24,00 | 22,00 |
| AP-QMT-0024 | 29,8 | 0,250 | 0,2259 | 0,0824 | 0,040 | 38,00 | 27,30 | 34,70 |
| AP-QMT-0025 | 32,1 | 0,266 | 0,1300 | 0,0428 | 0,450 | 56,60 | 24,40 | 19,00 |
| AP-QMT-0026 | 25,5 | 0,423 | 0,3200 | 0,1180 | 0,700 | 49,70 | 24,90 | 25,30 |
| AP-QMT-0028 | 30,2 | 0,353 | 0,1094 | 0,0978 | 0,320 | 63,10 | 18,20 | 18,70 |
| AP-QMT-0029 | 41,0 | 0,084 | 0,2401 | 0,0174 | 0,030 | 80,80 | 12,80 | 6,40 |
| AP-QMT-0030 | 25,9 | 0,190 | 0,0579 | 0,1280 | 0,815 | 48,60 | 22,20 | 29,20 |
| AP-QMT-0031 | 35,9 | 0,040 | 0,3674 | 0,0446 | 0,500 | 46,00 | 35,10 | 18,90 |
| AP-QMT-0032 | 28,0 | 0,535 | 0,3100 | 0,1080 | 0,140 | 53,50 | 29,40 | 17,10 |
| AP-QMT-0033 | 27,5 | 0,700 | 0,3200 | 0,1140 | 0,100 | 56,20 | 30,70 | 13,10 |
| AP-QMT-0034 | 29,5 | 0,229 | 0,2900 | 0,0996 | 0,060 | 61,50 | 22,60 | 15,49 |
| AP-QMT-0035 | 40,5 | 0,181 | 0,0757 | 0,0250 | 0,140 | 74,70 | 14,80 | 10,60 |
| AP-QMT-0036 | 21,0 | 0,496 | 0,5200 | 0,1560 | 0,890 | 38,20 | 27,40 | 4,70 |
| AP-QMT-0037 | 23,6 | 0,556 | 0,4009 | 0,1480 | 0,420 | 40,40 | 16,20 | 43,40 |
| AP-QMT-0038 | 52,0 | 0,072 | 0,0009 | 0,0007 | 0,040 | 90,60 | 9,40 | 0,00 |
| AP-QMT-0039 | 36,7 | 0,175 | 0,1060 | 0,0343 | 0,130 | 76,60 | 13,50 | 9,90 |
| AP-QMT-0040 | 32,9 | 0,120 | 0,0897 | 0,0357 | 0,170 | 70,80 | 14,70 | 14,73 |
| AP-QMT-0041 | 30,8 | 0,300 | 0,2200 | 0,0679 | 0,100 | 57,40 | 28,00 | 14,85 |
| AP-QMT-0042 | 34,0 | 0,220 | 0,1200 | 0,0385 | 0,310 | 62,40 | 27,50 | 10,10 |
| AP-QMT-0044 | 22,7 | 0,680 | 0,3900 | 0,1680 | 1,160 | 40,20 | 37,20 | 22,74 |
| AP-QMT-0045 | 28,4 | 0,376 | 0,3500 | 0,1070 | 0,360 | 52,60 | 27,20 | 20,20 |
| AP-QMT-0046 | 29,8 | 0,250 | 0,2259 | 0,0824 | 0,040 | 38,00 | 27,30 | 34,70 |
| AP-QMT-0047 | 27,8 | 0,360 | 0,3500 | 0,1130 | 0,300 | 55,10 | 27,30 | 17,60 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AP-QMT-0048 | 44,6 | 0,070 | 0,0130 | 0,0046 | 0,030 | 83,00 | 12,20 | 4,80 |
| AP-QMT-0049 | 33,0 | 0,224 | 0,1051 | 0,0362 | 0,330 | 62,50 | 26,60 | 10,85 |
| AP-QMT-0050 | 30,6 | 0,278 | 0,2400 | 0,0891 | 0,110 | 61,50 | 23,10 | 15,40 |
| AP-QMT-0053 | 29,6 | 0,352 | 0,2800 | 0,1040 | 0,270 | 54,00 | 24,00 | 22,00 |
| AP-QMT-0054 | 25,9 | 0,815 | 0,3674 | 0,1280 | 0,190 | 48,60 | 22,20 | 29,20 |
| AP-QMT-0055 | 25,6 | 0,351 | 0,2717 | 0,0908 | 1,180 | 58,60 | 23,40 | 18,00 |
| AP-QMT-0056 | 30,1 | 0,250 | 0,2800 | 0,0859 | 0,100 | 52,70 | 33,60 | 13,70 |
| AP-QMT-0057 | 29,6 | 0,360 | 0,2527 | 0,1030 | 0,240 | 60,30 | 22,70 | 17,00 |
| AP-QMT-0058 | 29,4 | 0,360 | 0,2535 | 0,1020 | 0,350 | 58,30 | 21,90 | 19,80 |
| AP-QMT-0059 | 29,0 | 0,310 | 0,3173 | 0,1100 | 0,100 | 52,10 | 27,20 | 20,70 |
| AP-QMT-0060 | 28,7 | 0,300 | 0,3166 | 0,1120 | 0,160 | 57,00 | 25,50 | 17,50 |
| AP-QMT-0061 | 31,7 | 0,155 | 0,2300 | 0,0782 | 0,030 | 57,10 | 26,20 | 16,71 |
| AP-QMT-0062 | 29,3 | 0,376 | 0,3100 | 0,1060 | 0,290 | 53,10 | 25,60 | 21,30 |
| AP-QMT-0064 | 23,6 | 0,556 | 0,3529 | 0,1480 | 0,420 | 40,40 | 16,20 | 43,45 |
| AP-QMT-0065 | 28,4 | 0,308 | 0,4009 | 0,1070 | 0,170 | 51,80 | 21,90 | 26,35 |
| AP-QMT-0066 | 41,0 | 0,084 | 0,3355 | 0,0174 | 0,030 | 80,80 | 12,80 | 6,40 |
| AP-QMT-0069 | 30,8 | 0,286 | 0,0579 | 0,0883 | 0,090 | 62,60 | 18,70 | 18,80 |
| AP-QMT-0073 | 28,1 | 0,370 | 0,4817 | 0,1140 | 0,240 | 59,70 | 23,40 | 16,80 |
| AP-QMT-0075 | 17,7 | 0,633 | 0,4200 | 0,1510 | 2,780 | 41,30 | 34,50 | 24,20 |
| AP-QMT-0076 | 26,9 | 0,610 | 0,3186 | 0,1160 | 0,110 | 53,20 | 27,30 | 19,50 |
| AP-QMT-0077 | 47,6 | 0,035 | 0,0126 | 0,0064 | 0,050 | 82,70 | 17,30 | 0,00 |
| AP-QMT-0078 | 22,1 | 0,663 | 0,3316 | 0,1200 | 1,320 | 41,30 | 31,30 | 27,40 |
| AP-QMT-0079 | 30,4 | 0,324 | 0,2358 | 0,0959 | 0,050 | 54,50 | 26,30 | 19,20 |
| AP-QMT-0080 | 28,1 | 0,350 | 0,3200 | 0,1130 | 0,200 | 59,40 | 21,00 | 19,60 |
| AP-QMT-0081 | 21,9 | 0,743 | 0,3674 | 0,1330 | 1,310 | 45,20 | 33,50 | 20,70 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AP-QMT-0082 | 28,8 | 0,368 | 0,3077 | 0,1390 | 0,230 | 55,30 | 27,40 | 17,30 |
| AP-QMT-0083 | 29,7 | 0,307 | 0,2401 | 0,0782 | 0,120 | 57,70 | 26,10 | 16,20 |
| AP-QMT-0084 | 36,1 | 0,060 | 0,1212 | 0,0485 | 0,090 | 77,50 | 7,00 | 15,50 |
| AP-QMT-0085 | 30,2 | 0,290 | 0,2352 | 0,0773 | 0,110 | 57,50 | 24,10 | 18,40 |
| AP-QMT-0086 | 32,3 | 0,308 | 0,2315 | 0,0821 | 0,230 | 64,90 | 21,90 | 13,20 |
| AP-QMT-0087 | 22,2 | 0,686 | 0,3500 | 0,1290 | 1,250 | 49,80 | 29,20 | 21,00 |
| AP-QMT-0088 | 31,4 | 0,336 | 0,2192 | 0,0887 | 0,350 | 66,20 | 20,20 | 13,60 |
| AP-QMT-0089 | 30,5 | 0,351 | 0,2366 | 0,1050 | 0,260 | 57,70 | 27,30 | 15,00 |
| AP-QMT-0091 | 28,1 | 0,312 | 0,3021 | 0,1050 | 0,140 | 59,20 | 23,50 | 17,35 |
| AP-QMT-0092 | 22,6 | 0,701 | 0,3500 | 0,1250 | 1,160 | 48,10 | 33,30 | 18,45 |
| AP-QMT-0094 | 17,7 | 0,609 | 0,4200 | 0,1440 | 3,070 | 36,70 | 27,60 | 35,60 |
| AP-QMT-0096 | 29,1 | 0,303 | 0,2765 | 0,1080 | 0,160 | 53,90 | 29,20 | 16,90 |
| AP-QMT-0098 | 22,2 | 0,686 | 0,3400 | 0,1250 | 0,650 | 45,10 | 32,60 | 22,00 |
| AP-QMT-0099 | 13,1 | 0,683 | 0,5671 | 0,2120 | 4,960 | 36,60 | 26,30 | 37,05 |
| AP-QMT-0100 | 37,8 | 0,169 | 0,1094 | 0,0449 | 0,730 | 74,10 | 13,90 | 11,95 |
| AP-QMT-0101 | 26,8 | 0,412 | 0,3529 | 0,1260 | 0,290 | 53,80 | 24,10 | 22,10 |
| AP-QMT-0102 | 30,2 | 0,353 | 0,2401 | 0,0978 | 0,320 | 63,10 | 18,20 | 18,70 |
| AP-QMT-0103 | 27,9 | 0,394 | 0,3147 | 0,1110 | 0,210 | 56,30 | 25,20 | 18,50 |
| AP-QMT-0104 | 49,4 | 0,004 | 0,0019 | 0,0006 | 0,360 | 87,00 | 13,00 | 0,00 |
| AP-QMT-0105 | 49,6 | 0,005 | 0,0014 | 0,0005 | 0,090 | 86,40 | 13,40 | 0,20 |
| AP-QMT-0106 | 52,0 | 0,072 | 0,0009 | 0,0007 | 0,040 | 90,60 | 9,40 | 0,00 |
| AP-QMT-0107 | 53,3 | 0,078 | 0,0014 | 0,0010 | 0,080 | 89,20 | 10,80 | 0,00 |
| AP-QMT-0108 | 54,0 | 0,084 | 0,0012 | 0,0009 | 0,060 | 88,40 | 11,60 | 0,00 |
| AP-QMT-0110 | 28,0 | 0,328 | 0,2813 | 0,0984 | 0,310 | 59,80 | 24,20 | 16,00 |
| AP-QMT-0112 | 19,4 | 0,767 | 0,4800 | 0,1710 | 1,250 | 40,20 | 33,30 | 26,50 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AP-QMT-0117 | 29,8 | 0,250 | 0,2259 | 0,0824 | 0,040 | 38,00 | 27,30 | 34,70 |
| AP-QMT-0121 | 28,5 | 0,440 | 0,3700 | 0,1100 | 0,440 | 51,30 | 29,10 | 19,70 |
| AP-QMT-0122 | 29,0 | 0,336 | 0,2912 | 0,1030 | 0,370 | 52,10 | 30,10 | 17,80 |
| AP-QMT-0124 | 26,7 | 0,500 | 0,2960 | 0,1130 | 0,190 | 56,00 | 26,90 | 17,10 |
| AP-QMT-0125 | 28,4 | 0,308 | 0,3355 | 0,1070 | 0,170 | 51,80 | 21,90 | 26,35 |
| AP-QMT-0126 | 29,7 | 0,300 | 0,2401 | 0,0782 | 0,120 | 57,70 | 26,10 | 16,20 |
| AP-QMT-0127 | 17,1 | 0,562 | 0,3423 | 0,1370 | 0,342 | n.d. | n.d. | n.d. |
| AP-QMT-0131 | 23,6 | 0,550 | 0,4004 | 0,1590 | 0,300 | 49,20 | 24,50 | 26,30 |
| AP-QMT-0132 | 28,7 | 0,367 | 0,2153 | 0,1060 | 0,280 | 62,30 | 20,60 | 17,10 |
| AP-QMT-0135 | 29,5 | 0,327 | 0,2195 | 0,0764 | 0,100 | 64,40 | 19,80 | 15,80 |
| AP-QMT-0136 | 19,0 | 0,452 | 0,3224 | 0,1110 | 1,840 | 65,90 | 33,60 | 0,50 |
| AP-QMT-0137 | 28,3 | 0,310 | 0,2937 | 0,1030 | 0,150 | 51,50 | 21,70 | 26,85 |
| AP-QMT-0139 | 28,8 | 0,368 | 0,3279 | 0,1090 | 0,270 | 50,80 | 22,90 | 26,30 |
| AP-QMT-0142 | 29,8 | 0,327 | 0,2630 | 0,1070 | 0,130 | 54,70 | 21,40 | 24,00 |
| AP-QMT-0143 | 28,5 | 0,383 | 0,3162 | 0,1130 | 0,300 | 53,40 | 24,70 | 21,90 |
| AP-QMT-0144 | 28,9 | 0,373 | 0,3208 | 0,1120 | 0,290 | 52,00 | 20,50 | 27,50 |
| AP-QMT-0145 | 18,3 | 0,797 | 0,4302 | 0,1500 | 1,430 | 39,40 | 31,50 | 29,00 |
| AP-QMT-0146 | 20,6 | 0,718 | 0,3758 | 0,1380 | 1,090 | n.d. | n.d. | n.d. |
| AP-QMT-0149 | 26,9 | 0,753 | 0,3288 | 0,1190 | 0,120 | 48,20 | 26,20 | 25,50 |
| AP-QMT-0152 | 29,3 | 0,740 | 0,1600 | 0,0469 | 0,290 | 56,60 | 22,70 | 20,60 |
| AP-QMT-0153 | 28,4 | 0,324 | 0,3141 | 0,1070 | 0,280 | 51,00 | 21,40 | 27,60 |
| AP-QMT-0154 | 29,0 | 0,394 | 0,3340 | 0,1250 | 0,280 | 48,60 | 24,60 | 26,80 |
| AP-QMT-0155 | 27,6 | 0,348 | 0,3459 | 0,1170 | 0,270 | 50,50 | 25,80 | 23,67 |
| AP-QMT-0156 | 18,4 | 0,549 | 0,3694 | 0,1360 | 2,110 | 44,90 | 32,20 | 22,90 |
| AP-QMT-0158 | 29,5 | 0,291 | 0,2408 | 0,0828 | 0,190 | 57,10 | 22,60 | 20,30 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AP-QMT-0159 | 27,6 | 0,397 | 0,3078 | 0,1200 | 0,310 | 48,50 | 23,40 | 28,10 |
| AP-QMT-0160 | 25,9 | 0,486 | 0,3631 | 0,1470 | 0,150 | 46,80 | 24,00 | 29,20 |
| AP-QMT-0161 | 27,7 | 0,394 | 0,3040 | 0,1190 | 0,300 | 52,30 | 24,50 | 23,20 |
| AP-QMT-0162 | 28,5 | 0,389 | 0,2919 | 0,1190 | 0,350 | 50,70 | 24,90 | 24,40 |
| AP-QMT-0164 | 15,2 | 0,729 | n.d. | n.d. | 2,020 | n.d. | n.d. | n.d. |