UNIVERSIDADE FEDERAL DE JUIZ DE FORA
Insituto de Ciências Exatas
PÓS-GRADUAÇÃO EM QUÍMICA

Julia Tristão do Carmo Rocha

Aplicação de espectroscopia no infravermelho próximo (NIR) e médio (MIR) associada a métodos quimiométricos, para avaliação de parâmetros físico-químicos em frações de petróleo

Juiz de Fora
2016

Julia Tristão do Carmo Rocha

Aplicação de espectroscopia no infravermelho próximo (NIR) e médio (MIR) associada a métodos quimiométricos, para avaliação de parâmetros físico-químicos em frações de petróleo

Tese apresentada ao Programa de Pós-graduação em Química, área de concentração: Química, da Universidade Federal de Juiz de Fora como requisito parcial para a obtenção do grau de Doutor.

Orientador: Prof. Dr. Marcone Augusto Leal de Oliveira

Co-orientador: Prof. Dr. Eustaquio Vinícius Ribeiro de Castro

Juiz de Fora

2016

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Rocha, Julia Tristão do Carmo.
Aplicação de espectroscopia no infravermelho próximo (NIR) e médio (MIR) associada a métodos quimiométricos, para avaliação de parâmetros físico-químicos em frações de petróleo / Julia Tristão do Carmo Rocha. -- 2016.
108 f. : il.

Orientador: Marcone Augusto Leal de Oliveira
Coorientador: Eustáquio Vinícius Ribeiro de Castro
Tese (doutorado) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Química, 2016.

1. Quimiometria. 2. MIR. 3. NIR. 4. Seleção de variáveis. 5. Derivados de Petróleo. I. Oliveira, Marcone Augusto Leal de, orient. II. Castro, Eustáquio Vinícius Ribeiro de, coorient. III. Título.

Às minhas filhas, Alice e Letícia,

o combustível da minha vida.

AGRADECIMENTOS

À Letícia e Alice, por existirem.

Ao Léo, pelo incentivo, apoio e paciência.

Aos meus pais, Rita e Paulo, pela educação, carinho e apoio.

Aos meus irmãos, Carmem e Lucas, por toda torcida e apoio.

Ao meu orientador, professor Marcone, pela paciência em me conduzir na execução deste trabalho, pelas valiosas críticas e sugestões, pelos ensinamentos, pelo exemplo de profissionalismo, pelas oportunidades de crescimento profissional e pessoal, pela amizade e pelo apoio constante.

Ao meu co-orientador, professor Eustáquio Vinícius Ribeiro de Castro, pelas sugestões, pela amizade e apoio durante todo o período de execução deste trabalho.

À Patrícia, por todo apoio, amizade e principalmente pela assistência durante esses anos de viagens a Juiz de Fora.

Aos amigos do LabPetro/UFES pela amizade e em especial a Karla, Betina e Rayza pela companhia, colaboração e amizade e ao Paulo, pela grande ajuda na análise estatística.

Aos amigos do CENPES/PETROBRAS, especialmente a Mirela e Júlio, pelo apoio no desenvolvimento deste trabalho.

Aos amigos do doutorado e do GQAQ, pela agradável receptividade.

Ao Labpetro/UFES, por todos os recursos que permitiram a elaboração deste trabalho.

Em especial, também, ao CENPES/Petrobras pelo apoio financeiro e suporte prestado na execução deste trabalho.

Ao PPGQUI/UFES, em especial ao professor e coordenador Valdemar, por ter disponibilizado a minha participação em disciplinas como aluna especial.

À FEST, pelo suporte financeiro.

À banca de qualificação, os professores Antônio Carlos, Denise e Richard, pelas correções, contribuições e sugestões.

À banca de defesa, os professores André Marcelo, Denise, Gustavo e Marcelo, por terem aceitado participar da mesma, com suas contribuições e sugestões.

Agradeço a todos aqueles que, de alguma forma, contribuíram para a conclusão deste trabalho.

"Mesmo quando tudo parece desabar, cabe a mim decidir entre rir ou chorar, ir ou ficar, desistir ou lutar; porque descobri, no caminho incerto da vida, que o mais importante é o decidir."

Cora Coralina

RESUMO

Os produtos petrolíferos em geral são altamente complexos e é exigido um esforço considerável para a caracterização de suas propriedades químicas e físicas. Às vezes tem-se urgência no resultado de determinadas análises e isto fica prejudicado pela forma como as análises são feitas. Assim, a quimiometria, associada à espectroscopia molecular (NIR e MIR em particular) vem gerando métodos alternativos para a caracterização e avaliação de propriedades físicas e químicas de petróleos e seus derivados com elevada exatidão, confiabilidade e rapidez. Para melhorar o desempenho previsor têm sido utilizados procedimentos apropriados para a seleção das regiões espectrais associadas com a propriedade de interesse. Desta forma, face às suas aplicabilidades, foi proposto neste trabalho a utilização das ferramentas quimiométricas com seleção de variáveis (método dos mínimos quadrados parciais por intervalos, iPLS, e por sinergismo de intervalos, siPLS; método de eliminação de variáveis não informativas por mínimos quadrados parciais, UVE; e algoritmo genético, GA), associada ao MIR e ao NIR, para a determinação das seguintes propriedades em frações de petróleo: Grau API, Índice de cetano, Índice de refração (a 20°C), Teor de Enxofre (%m/m), Ponto de fuligem (mm), Ponto de anilina (°C), Ponto de congelamento (°C), Ponto de entupimento (°C), Ponto de névoa (°C) e Ponto de fluidez (°C), avaliando, assim, a performance dos modelos obtidos, bem como as técnicas utilizadas na seleção de variáveis. Essa avaliação se deu pela determinação e análise do coeficiente de determinação (R^2), de diversos erros calculados para os conjuntos de calibração e previsão. Os modelos foram, ainda, submetidos a testes estatísticos (α=0,05), e tiveram suas figuras de mérito calculadas. Os melhores modelos para a previsão do Grau API e do ponto de névoa foram criados aplicando-se iPLS a dados de MIR, enquanto que para a previsão do teor de enxofre e pontos de refração, de fuligem e de anilina foram criados aplicando-se siPLS também ao MIR. Já para a previsão do índice de cetano e do teor de enxofre e do ponto de entupimento, os melhores modelos foram criados aplicando-se iPLS a dados de NIR. Nesse contexto, o melhor modelo para a predição do ponto de fluidez foi o GA. Finalmente, para a previsão do ponto de congelamento, nenhum método de seleção de variáveis melhorou a capacidade preditiva, quando comparados ao modelo criado aplicando-se PLS a dados de MIR. Dessa forma, conclui-se que houve um melhor desempenho dos modelos criados a partir de dados de MIR. Quanto aos métodos de seleção de variáveis, iPLS e siPLS obtiveram o melhor desempenho.

PALAVRAS-CHAVE: Quimiometria. MIR. NIR. Seleção de variáveis. Derivados de petróleo.

ABSTRACT

Petroleum products are, in general, highly complex and a considerable effort is needed to characterize their chemical and physical properties, though sometimes the results of several analyses are urgent and this is compromised by the way the analyses are carried out. Thus, chemometrics associated with molecular spectroscopy (particularly NIR and MIR) has good potential as a tool in analytical chemistry, creating alternative methods to characterize and evaluate physical and chemical properties of petroleum and its derivates with high precision, reliability and rapidity. To improve the predictor performance, appropriate procedures are being used to select spectral regions associated with the property of interest. In face of their applicabilities, this work proposes the use of chemometric tools, with variable selection (Interval Partial Least Square, iPLS and Sinergism Interval Partial Least Square, siPLS; Elimination of Uninformative Variables, UVE and Genetic Algorithm, GA), associated with mid infrared (MIR) and near infrared (NIR) spectroscopies to determine the following properties in petroleum fractions: API gravity, Cetane index, Refractive index (at 20°C), Sulfur content (%m/m), Smoke point (mm), Aniline point (°C), Freezing point (°C), Plugging point (°C), Cloud point (°C) and Pour point (°C), enabling the evaluation of performance of the obtained models, as well as the techniques used in variable selection. This evaluation was performed by determination and analyses of the following requirements: coefficient of determination (R^2), several calculated errors for the calibration and prediction set. The models were also subjected to statistical tests (α=0,05), and the figures of merit were calculated. The best models to predict API gravity and cloud point were created by applying iPLS to the MIR data, whereas for prediction of sulfur content, refractive index, and smoke and aniline points the models were created by applying iPLS to NIR data. In this context, the best model to predict the pour point was the GA. Finally, to predict freezing point, none of the variable selection methods improved the predictive capability when comparing to the model created using only PLS in MIR data. Thus, the conclusion is that a better performance was obtained for the models created from MIR data. Regarding efficiency of variable selection methods, the iPLS and siPLS methods resulted in a best performance.

KEYWORDS: Chemometrics. MIR. NIR. Variable selection. Petroleum derivatives.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

airPLS - Correção de linha de base utilizando um método iterativamente adaptativo por mínimos quadrados ponderados e penalizados, do inglês *baseline correction using Adaptative Iteratively Reweighted penalized least squares*

API – do inglês, *American Petroleum Institute*

CM – Centrado na média

CV – Validação cruzada, do inglês, cross-validation

GA – algoritmo genético, do inglês *Genetic Algorithm*

GA-PLS – Algoritmo genético por mínimos quadrados parciais, do inglês *Genetic Algorithm in Partial Least Square*

iPLS – Mínimos Quadrados Parciais por Intervalos, do inglês, *Interval Partial Least Square*

MIR – Infravermelho Médio, do inglês *Mid Infrared*

MSC – correção do espalhamento multiplicativo, do inglês *Multiplicative Scatter Correction*

MLR - Regressão Linear Múltipla, do inglês, *Multilinear Regression*

NIR – Infravermelho Próximo, do inglês, *Near Infrared*

PCA – Análise por Componentes Principais, do inglês *Principal Components Analysis*

PLS – Mínimos Quadrados Parciais, do inglês *Partial Least Squares*

R^2 – Coeficiente de determinação

R – Reprodutibilidade

RMSE – Raiz Quadrada do Erro Quadrático Médio, do inglês *Root Mean Square Error*

RMSEC - Raiz Quadrada do Erro Quadrático Médio de Calibração, do inglês *Root Mean Square Error of Calibration*

RMSECV – Raiz Quadrada do Erro Quadrático Médio de Validação Cruzada, do inglês *Root Mean Square Error of Cross-Validation.*

RMSEP - Raiz Quadrada do Erro Quadrático Médio de Previsão, do inglês *Root Mean Square Error of Prediction*

siPLS – Mínimos Quadrados Parciais por Sinergismo de Intervalos, do inglês *Sinergism Interval Partial Least Square*

SNV – Variação normal padrão, do inglês *Standard Normal Variate*

SVD – Desvio padrão dos erros de validação, do inglês *Standard Deviation of the Validation Errors*

UVE– método de eliminação de variáveis não informativas, do inglês *Uninformative Variables Elimination*

UVE-PLS – método de eliminação de variáveis não informativas por mínimos quadrados parciais, do inglês *Uninformative Variables Elimination in Partial Least Square*

VL – Variável Latente

SUMÁRIO

1. INTRODUÇÃO

A química analítica quantitativa sofreu um grande impacto causado pelo desenvolvimento dos métodos de calibração multivariada. Nesta, é possível estimar uma propriedade de interesse a partir de outras medições, que normalmente são espectros obtidos por procedimentos analíticos simples, rápidos, pouco onerosos e que dependem de pequena quantidade de amostra (FERREIRA et al., 1999). Os métodos de calibração multivariada permitem o tratamento de dados complexos do ponto de vista matemático e estatístico, correlacionando medidas instrumentais e valores para uma propriedade de interesse correspondente (BRERETON, 2003; SEKULIC et al., 1993).

Nesse aspecto, a quimiometria associada à espectroscopia molecular (infravermelho próximo e médio em particular) vem apresentando potencialidades como ferramenta para a química analítica, gerando métodos alternativos para a caracterização e avaliação de propriedades físicas e químicas de petróleos e seus derivados com elevada precisão, confiabilidade e rapidez (KHANMOHAMMADI et al., 2012; SOYEMI et al, 2000; PASQUINI e BUENO, 2007; FALLA et al., 2006; KALLEVIK, HVALHEIM e SJÖBLOM, 2000; KUPTSOV e ARBUZOVA, 2011; HANNISDAL, HEMMINGSEN e SJÖBLOM, 2005).

Cada vez mais tem se lançado mão da aplicação desses métodos em dados de petróleo e derivados, visto os produtos petrolíferos em geral serem altamente complexos e ser exigido um esforço considerável para a caracterização de suas propriedades químicas e físicas. Nesse contexto, às vezes tem-se urgência no resultado de determinadas análises para uma tomada de decisão e isto fica prejudicado pela forma como as análises são feitas (SPEIGHT, 2002; RIAZI, 2005; SIMANZHENKOV, 2003; LYONS e PLISGA, 2005). Diante disto, surge a necessidade de se desenvolver métodos mais rápidos, simples, econômicos, confiáveis e que provoquem menos impacto ambiental para a determinação dessas propriedades, tais como os métodos de calibração multivariada.

Nesse contexto, o método dos mínimos quadrados parciais (PLS – do inglês, *Partial Least Squares*) é o método de regressão mais utilizado para a construção de modelos de calibração multivariada a partir de dados de primeira ordem. Este método não requer um conhecimento exato de todos os componentes presentes nas amostras, podendo realizar a previsão de

amostras mesmo na presença de interferentes, desde que estes também estejam presentes por ocasião da construção do modelo (OLIVEIRA et al., 2004; SOARES et al., 2008).

A obtenção de modelos por calibração multivariada pode ser realizada utilizando a informação de toda faixa espectral de trabalho para construir um modelo de regressão correlacionando com a propriedade de interesse. No entanto, considerando o grande número de variáveis fornecidas por toda a faixa espectral, algumas delas podem interferir na modelagem, além de tornar o tratamento dos dados mais lento. Portanto, para melhorar o desempenho de técnicas de calibração multivariada, têm sido utilizados procedimentos apropriados para a seleção das regiões espectrais associadas com a propriedade de interesse (SOARES et al., 2008; NØRGAARD et al, 2000).

Já foram descritos na literatura alguns métodos para implementar a seleção de região espectral, os quais geralmente melhoram significativamente o desempenho dos métodos de calibração quando comparados com os métodos que utilizam os espectros totais. Esses métodos são chamados métodos de seleção de variáveis, que escolhem regiões específicas do espectro (um comprimento de onda ou um conjunto de comprimentos de onda) em que a colinearidade não é tão importante, enquanto gera modelos mais estáveis robustos e mais simples de interpretar. Na prática, a filosofia está baseada na identificação de um subconjunto de variáveis que podem produzir erros de previsão mais baixos (OLIVEIRA et al., 2004; SOARES et al., 2008).

Usualmente, esses métodos de seleção de variáveis são executados em análises PLS, o que implica que um número limitado de sinais é utilizado para construir o modelo multivariado, descartando os restantes. O propósito principal desta seleção é a construção de modelos a partir de dados espectrais, os quais carregam informações mais ricas sobre a substância ou propriedade de interesse, porém menos dados de sobreposição espectral com potenciais interferências. A teoria e a prática mostram que um melhor desempenho analítico do método PLS é alcançado após a aplicação de seleção de variáveis, o que explica o grande interesse nessas técnicas quimiométricas ao longo dos anos. Outras técnicas de análise multivariada podem também se beneficiar de seleção de variáveis, tanto para fins quantitativos ou de classificação (SOROL et al., 2010).

Os métodos de seleção de variáveis existentes se diferem com relação ao procedimento realizado para a seleção da região espectral, que pode ser contínuo ou discreto. Dentre os

métodos utilizados atualmente pode se destacar o método de mínimos quadrados parciais por intervalos (iPLS – do inglês, *Interval Partial Least Square*) (OLIVEIRA et al., 2004; NØRGAARD et al., 2000) e por sinergismo de intervalos (siPLS - do inglês, *Synergy Interval Partial Least Square*) (NØRGAARD et al., 2000), os quais são classificados como métodos de seleção de variáveis contínuos; e o método de eliminação de variáveis não informativas por mínimos quadrados parciais (UVE-PLS – do inglês, *Uninformative Variables Elimination in Partial Least Square*) (CENTNER e MASSART, 1996) e o algoritmo genético (GA – do inglês, *Genetic Algorithm*) (COSTA FILHO e POPPI, 1999), os quais são classificados como métodos de seleção de variáveis discretos. As técnicas de seleção de variáveis permitem a eliminação de informações não relevantes como, por exemplo, bandas que não contenham nenhuma informação das espécies ou propriedades a serem analisadas e linhas de base (OLIVEIRA et al., 2004; SOARES et al., 2008).

Desta forma, face às suas aplicabilidades, foi proposto neste trabalho a utilização das ferramentas quimiométricas, com seleção de variáveis (iPLS, siPLS, UVE e GA), associadas à espectroscopia no infravermelho médio (MIR) e próximo (NIR), para a determinação das seguintes propriedades em frações de petróleo: Grau API (°API); Índice de cetano; Índice de refração (a 20°C); Teor de Enxofre (%m/m); Ponto de fuligem (mm); Ponto de anilina (°C); Ponto de congelamento (°C); Ponto de entupimento (°C); Ponto de névoa (°C); Ponto de fluidez (°C), na qual foi realizada a avaliação dos modelos obtidos, bem como das técnicas utilizadas na seleção de variáveis.

1.1 PESQUISAS DESENVOLVIDAS – ANTECEDENTES

Antes da apresentação do trabalho propriamente dita, deve-se destacar que já existem estudos anteriores que indicam a aplicação da espectroscopia NIR ou MIR para a previsão de algumas dessas propriedades em alguns tipos de derivados de petróleo (alguns com desempenhos semelhantes aos apresentados nesse trabalho). Entretanto, nenhum deles envolve a aplicação, e comparação, de ambas as técnicas espectroscópicas em uma faixa de destilação tão ampla, como a utilizada neste trabalho (destilados entre 15 e 500ºC) para prever qualquer uma das propriedades apresentadas.

Entre esses trabalhos pode-se citar, por exemplo, o apresentado por Chung et al (2000), no qual a espectroscopia NIR foi aplicada com sucesso para a determinação da Grau API em

amostras de resíduo atmosférico. Morris et al.(2009) também apresentaram resultados eficazes para a determinação do índice de cetano em combustíveis de aviação e diesel, apenas por NIR. Ainda nesse contexto, Satya et al. (2007) aplicaram análise multivariada, combinada com espectroscopia no infravermelho próximo apenas, para prever algumas propriedades como a densidade e resíduo de carbono, entre outras, porém apenas em frações residuais de petróleo bruto. Soyemi et al. (2000) também descreveram uma série de métodos quimiométricos, aplicados exclusivamente em dados de NIR, e demonstraram a sua utilização na previsão de seis propriedades de combustível diesel, como o índice de cetano, densidade e ponto de congelamento. Então, como pode ser visto, nenhum trabalho testou as duas técnicas espectroscópicas para se prever qualquer uma destas propriedades, em uma faixa de amostras como a utilizada nesse trabalho.

Quanto à aplicação dos métodos de seleção de variáveis, Breitkreitz et al.(2003) propuseram um método para a determinação do enxofre total em amostras de diesel empregando espectroscopia NIR, seleção de variáveis e calibração multivariada. Entre as técnicas trabalhadas, eles usaram o PLS e o método de seleção de variáveis GA. Através da comparação dos valores de RMSE encontrados, pode notar-se que eles são semelhantes aos valores obtidos neste trabalho, o qual, como já mencionado, inclui uma gama maior de amostras de óleo destilado (entre 15 e 500 ° C). Neste contexto, Soares et al. (2008) apresentaram uma metodologia para a determinação do teor de enxofre, também apenas em amostras de diesel, utilizando dados espectroscópicos FTIR associados com calibração multivariada por PLS. Os modelos de calibração foram construídos utilizando todo o espectro e também foi aplicado o método de seleção de variáveis *stepwise* e, da mesma forma, os erros encontrados são semelhantes aos obtidos neste trabalho.

Um estudo já foi descrito para comparar a eficiência da aplicação das técnicas de seleção de variáveis utilizadas nessa pesquisa (iPLS, siPLS, UVE e GA), associadas à espectroscopia no infravermelho próximo, aplicadas para a análise de comprimidos (ABRAHAMSSON, 2003). Neste trabalho, os autores mostraram que os melhores resultados foram obtidos com a aplicação da técnica GA (em dados NIR). Entretanto, em sua conclusão o autor cita que os resultados que foram encontrados são válidos apenas para o tipo de matriz e o conjunto de dados trabalhados e que outras medições e investigações se faziam necessárias antes que qualquer conclusão geral pudesse ser tirada sobre esses métodos de seleção de variáveis.

2. FUNDAMENTOS TEÓRICOS

2.1 O PETRÓLEO

Petróleo é uma palavra derivada do Latim *petra* e *oleum*, que significa literalmente "óleo de pedra" e refere-se a líquidos ricos em hidrocarbonetos que acumularam-se em reservatórios subterrâneos. O petróleo (igualmente chamado de óleo cru) varia drasticamente nas propriedades da cor, do odor e da densidade e viscosidade as quais refletem a diversidade de sua origem (SPEIGHT, 2002).

Do ponto de vista químico o petróleo é uma mistura complexa composta em sua maioria de hidrocarbonetos, podendo possuir, em menor parte, compostos de oxigênio, nitrogênio e enxofre, combinados de forma variável, conferindo características diferenciadas aos diversos tipos de óleos crus encontrados na natureza (SPEIGHT, 2002; RIAZI, 2005).

Em estado bruto o petróleo não tem aplicabilidade prática, mas quando refinado ele fornece combustíveis líquidos valiosos, solventes, lubrificantes, e muitos outros produtos. Os combustíveis derivados do petróleo contribuem com aproximadamente um terço a um meio do suprimento total de energia no mundo (SPEIGHT, 2002).

O processo físico básico para a separação dos derivados do petróleo em uma refinaria é a destilação. O petróleo contém milhares de compostos diferentes que variam em massa molar de 16 $g \cdot mol^{-1}$ (metano, CH_4) a mais de 2000 $g \cdot mol^{-1}$ (SPEIGHT, 2002; RIAZI, 2005). Esta larga escala nas massas molares conduz aos pontos de ebulição que variam de -160º C (ponto de ebulição do metano) a mais do que 600º C, que é o ponto de ebulição de compostos pesados no óleo cru. Um grupo de hidrocarbonetos, portanto, pode ser separado com a destilação de acordo com o ponto de ebulição dos compostos mais leves e mais pesados nas misturas.

De fato, durante a destilação um petróleo bruto é convertido em uma série de frações do petróleo, na qual cada uma é uma mistura de um número limitado de hidrocarbonetos com uma escala específica do ponto de ebulição. As frações com uma escala mais larga de pontos de ebulição contêm o maior número de hidrocarbonetos. Todas as frações de uma coluna de destilação têm uma escala de ebulição conhecida, exceto o resíduo, para o qual o ponto de

ebulição superior geralmente não é conhecido (SPEIGHT, 2002). Porém, a natureza infinitamente variável de fatores composicionais faz com que todos os óleos crus e produtos de petróleo processados numa refinaria sejam diferentes entre si. Essa variabilidade representa uma característica química do tipo *fingerprint* ou impressão digital para cada petróleo e fornece uma base para caracterizá-lo (SPEIGHT, 2002; RIAZI, 2005). Na Tabela 1 são apresentados os principais derivados do petróleo com seus respectivos pontos de ebulição e hidrocarbonetos contidos.

Tabela 1 – Frações de petróleo e suas faixas de temperatura

Fração	**Hidrocarbonetos Contidos**	**Faixa de Temperatura (ºC)**
Gás	C_2-C_4	-90 – 1
Gasolina	C_4-C_{10}	-1 – 200
Naftas	C_4-C_{11}	-1 – 205
Combustível de Aviação	C_9-C_{14}	150 – 255
Querosene	C_{11}-C_{14}	205 – 255
Diesel	C_{11}-C_{16}	205 – 290
Óleo Combustível Leve	C_{14}-C_{18}	255 – 315
Óleo Combustível Pesado	C_{18}-C_{28}	315 – 425
Graxa	C_{18}-C_{36}	315 – 500
Óleo Lubrificante	>C_{25}	> 400
Óleo Combustível de Vácuo	C_{28}-C_{55}	425 – 600
Resíduo	>C_{55}	> 600

Fonte: RIAZI, 2005.

2.1.1 Avaliação de petróleo e derivados

O Petróleo apresenta grandes variações em sua composição e propriedades, e estes ocorrem não só em petróleos de diferentes áreas, mas também naqueles retirados de diferentes profundidades no mesmo poço de produção. Comercialmente sua composição, bem como de seus derivados, variam muito devido a suas diferentes origens e aos diferentes processos de refino pelos quais pode ser submetido. Além disso, sua qualidade tem mudado constantemente desde sua introdução no mercado como combustível (SPEIGHT, 2002; ANDRADE, 2009; PARISOTTO, 2007).

Dessa forma, os produtos petrolíferos em geral são altamente complexos, e é exigido um esforço considerável para a caracterização de suas propriedades químicas e físicas. Certamente, a análise desses produtos é necessária para se determinar as propriedades que podem ajudar a resolver um problema do processo assim como as propriedades que indicam a função e o desempenho do produto em questão. Portanto, de acordo com as propriedades de cada derivado, é definida sua habilidade em servir a específica finalidade (SPEIGHT, 2002).

Nesse contexto, o valor do petróleo depende da sua qualidade para o refino e de sua capacidade de ajuste a fim de gerar um determinado produto para atender a demanda do mercado. A partir de informações da qualidade do petróleo e derivados e do mercado a ser atendido, o refino moderno utiliza uma sofisticada combinação de calor, catalisador, e hidrogênio para reorganizar as moléculas de petróleo, de forma a aumentar o rendimento da formação de produtos ambientalmente saudáveis e com alto valor agregado. Processos de conversão incluem coque, hidrocraqueamento, e craqueamento catalítico para quebrar moléculas grandes em frações menores; hidrotratamento para reduzir teores de heteroátomos e aromáticos, criando produtos ambientalmente aceitáveis; e isomerização e rearranjo para reorganizar moléculas e formar compostos com alto valor agregado como, por exemplo, a gasolina com um elevado índice de octano e diesel com elevado nível de cetano. Além disso, o conhecimento da composição molecular de petróleo permite o ambientalista considerar o impacto biológico da exposição ambiental (SPEIGHT, 2002; RIAZI, 2005).

O valor do petróleo depende, portanto, da sua qualidade para o refino e de sua capacidade de ajuste a fim de gerar um determinado produto para atender a demanda do mercado. Assim, unidades de processamento de uma refinaria exigem métodos de testes analíticos que possam avaliar adequadamente matérias-primas e monitorar a qualidade do produto. Uma vez que as propriedades exigidas são determinadas, elas são controladas por testes e por análises apropriadas e auxiliam na elaboração de estratégias de transporte e refino, além de informarem sobre potenciais derivados esperados ((SPEIGHT, 2002; RIAZI, 2005; FILGUEIRAS, 2011).

Métodos de caracterização são necessários para o planejamento do processo, avaliação de petróleo bruto, e controle operacional[11]. Atualmente esses métodos, em sua forma tradicional, envolvem cerca de 700 ensaios físico-químicos, consumindo de 10 a 70 litros de petróleo, em não menos de 4 meses, ao custo estimado de mais de 80 mil dólares (ANP,2013; MAGALHÃES, 2005).

Dentre todos os ensaios utilizados para a avaliação de derivados de petróleo pode-se destacar, por exemplo, a grau API (do inglês, *American Petroleum Institute*), um tipo de medida de densidade muito utilizada na indústria petrolífera, calculada a partir da medida da densidade do petróleo e de seus derivados, determinada a 60 ºC, conforme a Equação 1, de forma que o Grau API e a densidade são medidas inversamente proporcionais. Portanto, quanto maior a densidade, mais pesado será o petróleo e menor o seu grau API e vice-versa. Essa medida é de grande importância a ser determinada, pois além de ser um indicador da qualidade do óleo, atualmente o petróleo, e especialmente seus derivados, são usualmente comprados e vendidos nessa base ou, com base em seu volume, e posteriormente convertidos para massa via densidade (SPEIGHT, 2002; RIAZI, 2005; ISO 12185:1996). O Índice de cetano também possui grande importância, pois mede a qualidade de ignição de um combustível diesel e tem influência direta na partida do motor e no seu funcionamento sob carga. Quanto menor o índice maior será o retardo da ignição e consequentemente, maior será a quantidade de combustível que permanecerá na câmara sem queimar no tempo certo. Isso leva a um mau funcionamento do motor, pois, quando a queima acontecer, gerará uma quantidade de energia superior àquela necessária. Em geral, Combustíveis com alto teor de parafinas apresentam alto número de cetano, enquanto produtos ricos em hidrocarbonetos aromáticos apresentam baixo número de cetano (SPEIGHT, 2002; RIAZI, 2005; ASTM D4737 – 10).

$$^{\circ}API = \frac{141{,}5}{d(60/60)} - 131{,}5 \qquad \textbf{(1)}$$

O índice de refração é também uma propriedade física fundamental, pois pode ser utilizado para a determinação da composição bruta do óleo de combustível residual e, muitas vezes, requer a sua medida em elevadas temperaturas. Além disso, o índice de refração, juntamente com a densidade, podem ser relacionados com a sua composição química e podem ser utilizados para tirar conclusões sobre a estrutura molecular (SPEIGHT, 2002; RIAZI, 2005; ASTM D1218 – 12).

O teor de enxofre também é uma propriedade importante, pois atualmente, para fins tributários, a legislação brasileira determina uma valoração do petróleo cujo cálculo é baseado

em apenas três propriedades: densidade, curva de destilação e teor de enxofre (ANP,2013). Ainda, os compostos contendo este elemento estão entre os componentes mais indesejáveis do petróleo. Isso ocorre devido à ação corrosiva que podem causar e à formação de gases tóxicos como o SO_2 (dióxido de enxofre) e o SO_3 (trióxido de enxofre), que ocorre durante a combustão do produto e são responsáveis por problemas como desativação dos catalisadores, degradação da coloração, corrosão e odores desagradáveis em seus derivados. Devido a isso, a legislação brasileira vem gradativamente reduzindo o teor máximo de enxofre permitido em seus combustíveis (SPEIGHT, 2002; RIAZI, 2005; ASTM D5453 – 12; ASTM D4294 – 10).

Como medida de controle, o ponto de fuligem também é bastante utilizado e indica a qualidade da queima do combustível. Sua medida é feita determinando-se a altura máxima de chama em milimetros à qual o óleo queima sem fuligem, quando testados sob condições específicas padronizadas. Através dessa determinação, é possível se correlacionar essa propriedade com o desempenho da combustão dos cortes na faixa dos combustíveis, pois a formação de carbono tende a aumentar com o ponto de ebulição. O ponto de fuligem também é de particular interesse para as indústrias de petróleo cujos processos expõem ou utilizam óleo mineral em temperaturas extremamente altas (SPEIGHT, 2002; RIAZI, 2005; ASTM D1322 – 15).

O Ponto de anilina é determinado pela menor temperatura na qual volumes iguais de anilina e óleo são completamente miscíveis e indica o grau de aromaticidade de uma fração do petróleo (compostos aromáticos apresentam os pontos de anilina mais baixos e os parafínicos, os mais altos. Cicloparafinas e olefinas exibem valores entre esses dois extremos) (SPEIGHT, 2002; RIAZI, 2005; ASTM D611 – 12). Em qualquer série homóloga de hidrocarbonetos o ponto de anilina aumenta com o aumento do peso molecular.

O ponto de congelamento é a temperatura na qual o líquido de hidrocarbonetos se solidifica em pressão atmosférica, e essa é uma das especificações importantes de propriedade para combustíveis de aviação, em virtude das temperaturas muito baixas encontradas em altitudes elevadas em aviões a jato. Essa propriedade é um índice da temperatura mais baixa em que ele pode ser utilizado para as aplicações previstas. Esse tipo de combustível deve apresentar valores aceitáveis do ponto de congelamento de modo que o fluxo de combustível para o motor seja adequado e mantido a alta altitude, evitando seu congelamento nas turbinas (SPEIGHT, 2002; RIAZI, 2005; ASTM D5972 – 05).

. Essa medida não deve ser confundida com o ponto de fluidez, uma propriedade também de extrema importância para avaliação de derivados de petróleo, o qual refere-se ao índice da menor temperatura em que o combustível irá fluir sob condições específicas e, portanto, é muito utilizada para se ajustar o processo de bombeio do mesmo(SPEIGHT, 2002; RIAZI, 2005; ASTM D5950 – 14).

Por um outro lado, uma propriedade análoga ao ponto de congelamento, o ponto de entupimento, é apropriada para estimar a menor temperatura na qual o combustível diesel possa fluir, sem dificuldades, em certos sistemas de combustível, de modo a evitar o entupimento do motor (SPEIGHT, 2002; RIAZI, 2005; ASTM D6371 – 05).

Nesse mesmo contexto, sob condições de baixa temperatura, constituintes parafínicos de combustível diesel podem ser precipitados como uma cera, bloqueando o sistema de linhas e filtros de combustível e causando mau funcionamento ou bloqueio do motor. A temperatura em que a precipitação ocorre depende da origem, do tipo, e do intervalo de ebulição do combustível. Quanto mais parafínico for o combustível, maior a sua temperatura de precipitação e menos adequado ele será para operações realizadas em baixa temperatura. A temperatura inicial na qual essa cera é precipitada é medida pelo ponto de névoa. Portanto, essa medida é um guia indicativo da temperatura na qual pode ocorrer entupimento dos sistemas de filtragem e restringir o fluxo do combustível. O ponto de névoa está se tornando cada vez mais importante para os combustíveis utilizados em motores a diesel de alta velocidade, especialmente devido à tendência para equipar estes motores com mais filtros finos (SPEIGHT, 2002; RIAZI, 2005; ASTM D5773 – 15).

Além das propriedades citadas acima, existem outras dezenas de ensaios que são realizados para que se haja uma caracterização e avaliação completa dos derivados de petróleo (SOYEMI, BUSCH e BUSCH, 2000; SATYA et al., 2007). Às vezes, tem-se urgência no resultado de determinadas análises para uma tomada de decisão e isto fica prejudicado pela forma como as análises são feitas. Diante disto, surge a necessidade de se desenvolver técnicas mais rápidas, simples, econômicas, confiáveis e que provoquem menos impacto ambiental (KHANMOHAMMADI et al, 2012; SOYEMI, BUSCH e BUSCH, 2000; MAGALHÃES, 2005). Por isso, existe atualmente grande interesse da comunidade científica em tentar otimizar esses procedimentos nos quesitos tempo, dinheiro e quantidade de amostra necessários para sua realização.

2.2 TÉCNICAS ESPECTROSCÓPICAS

Os métodos espectroscópicos podem ser uma alternativa eficaz na caracterização de petróleos, como meio de monitoramento e determinação de parâmetros físico-químicos devido à sua potencialidade, praticidade e rapidez analítica. As técnicas espectroscópicas são capazes de fornecer informações sobre o óleo em nível molecular. Entretanto, dificilmente o sinal analítico instrumental fornecerá a informação quantitativa de interesse diretamente. Nos casos onde não há separação do analito que possibilite uma análise multivariada, faz-se necessário a utilização da quimiometria para converter os sinais analíticos do instrumento à informação de interesse quantitativo (FILGUEIRAS, 2011; PEINDER, 2009).

Entre as várias técnicas espectroscópicas existentes, a espectroscopia no infravermelho é certamente uma das ferramentas analíticas mais importantes para o químico moderno, por ter uma ampla área de aplicação. A sua energia corresponde à região do espectro eletromagnético entre 12800 e 200 cm^{-1} (SKOOG e LEARY, 1992).

Nesse aspecto, a quimiometria, associada à espectroscopia (infravermelho próximo em particular) vem apresentando potencialidades como ferramenta para a química analítica gerando métodos confiáveis para a caracterização e avaliação de propriedades físico-químicas de petróleos e seus derivados (KHANMOHAMMADI et al., 2012; SOYEMI et al, 2000; PASQUINI e BUENO, 2007; FALLA et al., 2006; KALLEVIK, HVALHEIM e SJÖBLOM, 2000; KUPTSOV e ARBUZOVA, 2011; HANNISDAL, HEMMINGSEN e SJÖBLOM, 2005). Esses estudos têm avançado nos últimos anos, pois na atualidade já existem alguns modelos desenvolvidos tendo gerado patentes e estarem sendo usados na prática pela indústria petroquímica.

2.2.1 Infravermelho próximo

A espectroscopia no infravermelho próximo - NIR (do inglês, *Near Infrared*), com intervalo de número de onda compreendido entre 12800 – 4000 cm^{-1}, tem recebido muita atenção recentemente, particularmente na análise quantitativa de amostras com alta complexidade. Nesta faixa espectral é possível observar transições harmônicas (sobretons) e

combinações das transições fundamentais. As principais informações espectrais são correspondentes às vibrações das ligações C-H, N-H, S-H e O-H, ou seja, ligações covalentes que apresentam átomo de hidrogênio (SKOOG e LEARY, 1992; PAVIA et al., 2008; SILVERSTEIN, BASSLER e MORRILL, 1982; PASQUINI e BUENO, 2007).

As bandas de absortividade normalmente são largas e pouco seletivas sendo, assim, difícil de se atribuir os sinais analíticos a compostos específicos. Esta dificuldade pode ser contornada quando se utiliza o NIR vinculado às técnicas quimiométricas. O vidro pode ser usado como material de célula, tornando a análise mais simples e fácil (SKOOG e LEARY, 1992; ANDRADE, 2009). A espectroscopia NIR possui uma grande vantagem por permitir o uso de fibras ópticas, facilitando análise remota e permitindo a análise *in situ*, minimizando a necessidade de armazenagem adequada das amostras (SKOOG e LEARY, 1992; SILVERSTEIN, BASSLER e MORRILL, 1981).

2.2.2. Infravermelho médio

A espectroscopia na região do infravermelho médio - MIR (do inglês, *Mid Infrared*) abrange a região de números de onda 4000 – 200 cm^{-1} e geralmente compreende bandas intensas e bem definidas, que têm elevadas absorções, favorecendo a interpretação dos espectros, geralmente ricos em informações. Sendo assim, ela é uma técnica de caracterização molecular bastante ampla e bem aceita. Essa região é muito útil para a identificação de compostos, uma vez que pequenas diferenças na estrutura e composição das moléculas produzem alterações significativas no perfil e na distribuição das bandas de absorção.

A intensidade de uma banda de absorção no infravermelho depende da mudança no momento dipolar durante a vibração: uma mudança grande no momento dipolar dará um aumento forte na absorção enquanto que uma pequena alteração gera uma absorção de menor intensidade. Portanto, a espectroscopia no infravermelho médio é mais útil na determinação de grupos funcionais polares, tais como ligações C=O, N-H e O-H. Sendo que grupos menos polares como olefinas alifáticas, ligações C-H em aromáticos e vibrações C-C também exibem bandas características, tornando esta técnica muito valiosa para a análise do óleo bruto (PEINDER, 2009; SKOOG e LEARY, 1992; PAVIA et al., 2008; SILVERSTEIN, BASSLER e MORRILL, 1981).

2.3 MÉTODOS QUIMIOMÉTRICOS

A química analítica quantitativa teve grande impacto com o desenvolvimento das técnicas de calibração multivariada. Nesta, é possível estimar uma propriedade de interesse a partir de outras medições, que normalmente são espectros obtidos por procedimentos analíticos simples, rápidos, pouco onerosos e que dependem de pequena quantidade de amostra (FERREIRA et al., 1999).

A moderna instrumentação de análises química é capaz de gerar uma quantidade considerável de dados para uma única amostra, em um curto espaço de tempo. Um espectrômetro pode registrar sinais provenientes de mais de mil comprimentos de onda ou, um único cromatograma pode apresentar mais de cem picos (SKOOG e LEARY, 1992). Assim, para que uma informação útil seja obtida deste grande volume de dados, muitas vezes é necessária a utilização de técnicas matemáticas adequadas, sendo a quimiometria um dos campos de estudo da química que fornece tais ferramentas (FERREIRA et al., 1999).

A quimiometria pode ser definida como uma área da química na qual métodos matemáticos, estatísticos e computacionais são aplicados a dados de origens distintas para a obtenção de uma informação química desejada. Ela consiste em um conjunto de técnicas de cálculo com o objetivo de promover a obtenção de informação útil de um conjunto complexo de dados, englobando conceitos de planejamento experimental, pré-processamento de dados e análise estatística multivariada (OLIVEIRA, 2007; ZENI, 2005). A quimiometria é, portanto, muitas vezes utilizada como uma ferramenta de automação laboratorial e está relacionada com a análise multivariada. De um modo geral, a análise multivariada refere-se aos métodos estatísticos e matemáticos que analisam, simultaneamente, múltiplas medidas de um objeto sob investigação, seja ele de caráter químico ou não (HAIR et al., 1998).

Os métodos quimiométricos podem ser divididos em três grandes áreas: planejamento e otimização de experimentos, métodos qualitativos e de reconhecimento de padrões e calibração multivariada (NETO, SCARMÍNIO e BRUNS; 2006). A aplicação de um ou outro método, ou até mesmo a combinação deles depende da natureza do problema que se deseja resolver, ou do tipo de informação que se deseja obter (ARAUJO, 2007). Para a execução desse trabalho, métodos de calibração multivariada foram utilizados.

Os métodos de calibração multivariada relacionam-se ao desenvolvimento de modelos matemáticos que permitem estimar alguma propriedade de interesse, ou a concentração de algum(s) analito(s), com os sinais instrumentais. O desenvolvimento do método de calibração consiste em duas etapas: a calibração e a validação. A etapa de construção do modelo de calibração começa com a seleção de um conjunto de amostras, cuidadosamente escolhidas, para que sejam representativas de toda a região a ser modelada. Estas amostras (conjunto de calibração) serão utilizadas na construção de um modelo apropriado para relacionar as respostas instrumentais com a informação desejada. Durante essa etapa dois fatores são considerados cruciais: o número de componentes principais ou de variáveis latentes, e a detecção de amostras anômalas (*outliers*). Após, deve ser realizada a etapa de validação, verificando a capacidade preditiva do modelo. A validação consiste em testar o modelo com amostras externas, das quais se tem conhecimento prévio das propriedades (ou concentrações) que se deseja medir (FERREIRA et al., 1999; BRERETON, 2003).

Dessa forma, em calibração multivariada os dados são organizados algebricamente em vetores e matrizes. Considere o conjunto de dados $\{(x_1,y_1),...,(x_n,y_n)\}$, com $x_i \in R^m$ sendo o vetor contendo o espectro e $y_i \in R$ o valor de referência para a amostra *i*. Em problemas quantitativos y_i representa o valor da propriedade de interesse e em problemas qualitativos, uma variável categórica que representa a classe à qual a amostra pertence.

Os espectros são organizados na matriz de dados **X** de forma que cada amostra represente um vetor linha (Figura 1). Cada variável espectral (número de onda, deslocamento químico, etc.) é emparelhada na mesma coluna, assim todos os espectros devem ser medidos com mesma resolução espectral. Para *n* amostras com *m* variáveis espectrais, a matriz de dados terá dimensão $\mathbf{X}_{(n,m)}$ (*n* linhas por *m* colunas).

Figura 1 – Exemplo da representação de uma matriz dados espectrais.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

Fonte: O Autor.

Assim, um modelo de calibração é, na verdade, uma função matemática (f), que relaciona dois grupos de variáveis, uma delas denominada dependente **Y** e a outra denominada independente **X** (ARAUJO, 2007):

$$\mathbf{Y} = f(X) = \mathbf{X} * \mathbf{b} \qquad (2)$$

,onde **b** corresponde à matriz dos coeficientes de regressão do modelo, que são determinados matematicamente a partir de dados experimentais (BRERETON, 2000).

Existem, na literatura, diversos métodos de calibração multivariada. Tendo em vista que o método PLS foi utilizado no presente trabalho, este será abordado de forma preferencial.

2.3.1 Mínimos Quadrados Parciais (PLS)

Atualmente, a regressão PLS é o método de calibração multivariada mais utilizado em química analítica (BRERETON, 2000; ANDERSSON, 2009; COSTA FILHO e POPPI, 1999). Este método pode ser utilizado para a construção de modelos que permitem a previsão de propriedades físico-químicas em petróleos e derivados de uma forma mais econômica, mais rápida e com menor impacto ambiental, e os modelos construídos podem ser facilmente adaptados às rotinas de análise em laboratórios.

O PLS utiliza as respostas analíticas, bem como as informações de interesse, para capturar a variância dos dados da matriz **X** e do vetor com a propriedade de interesse **Y**, através de suas decomposições sucessivas e simultâneas, correlacionando-as (BRERETON, 2003; ANDERSSON, 2009). Nesse contexto, para o PLS, a primeira componente calculada é chamada de variável latente (VL) e descreve a direção de máxima variância que também se correlaciona com **X** e **Y**.

O modelo PLS é, portanto, obtido através de um processo iterativo, no qual se otimiza simultaneamente a projeção das amostras sobre o(s) peso(s) (**W**), para a determinação dos *scores*, e o ajuste por uma função linear dos *scores* da matriz **X** aos *scores* da matriz **Y** de modo a minimizar os desvios (VANDEGINSTE et al, 1998; BEEBE e KOWALSKI, 1987; CENTNER e MASSART, 1996). Vale lembrar que os pesos ou *loadings* (**W**) representam a

influência que cada variável possui na combinação linear das variáveis originais, ou seja, representam o peso que cada variável possui em cada variável latente formada. Já os escores correspondem à projeção de cada amostra no novo sistema de eixos formada e, portanto, cada amostra terá então um valor de escore para cada um dos novos eixos.

Portanto, a construção de um modelo PLS consiste de uma regressão entre os *scores* das matrizes **X** e **Y** em uma soma de "*h*" VL's. O modelo PLS pode ser definido através de relações externas, que correlacionam, individualmente, as matrizes **X** e **Y** (conforme as equações 3 e 4), enquanto as internas correlacionam ambas as matrizes (GELADI e KOWALSKI, 1986).

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\prime} + \boldsymbol{E}_X = \sum t_h p_h^{\prime} + \boldsymbol{E}_X \quad (\mathbf{3})$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\prime} + \boldsymbol{E}_Y = \sum u_h q_h^{\prime} + \boldsymbol{E}_Y \quad (\mathbf{4})$$

onde **X** é a matriz de dados (medida instrumental), **Y** é a matriz de resposta (concentração, por exemplo), **T** e **U** são os escores de **X** e **Y** respectivamente; os elementos **P** e **Q** são os "pesos" e E_X e E_y são os resíduos. A matriz de scores **T** é estimada pela combinação linear de **X** com coeficientes ponderados por **W** (pesos):

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W} \quad \mathbf{(5)}$$

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

$$b_{PLS} = \boldsymbol{W}(\boldsymbol{P}^{\prime}\boldsymbol{W})^{-1}q^{\prime} \quad \mathbf{(6)}$$

Finalmente, a partir do cálculo de $\boldsymbol{b}$, o modelo linear PLS pode ser representado por:

$$Y = \boldsymbol{X} * \boldsymbol{b}_{PLS} \quad \mathbf{(7)}$$

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

2.3.2 Método de seleção de variáveis

A construção de modelos por calibração multivariada pode ser realizada utilizando a informação de toda faixa espectral de trabalho para construir um modelo de regressão correlacionando-o com a propriedade de interesse. No entanto, considerando o grande número de variáveis fornecidas por toda a faixa espectral, algumas destas variáveis podem interferir na modelagem, além de tornar o tratamento dos dados mais lento. Portanto, para melhorar o desempenho de técnicas de calibração multivariada, têm sido utilizados procedimentos apropriados para a seleção das regiões espectrais associadas às propriedade de interesse (SOARES et al., 2008; NØRGAARD et al., 2000).

Alguns métodos têm sido descritos recentemente na literatura para implementar seleção de região espectral para melhorar significativamente o desempenho dos métodos de calibração de espectros totais. Esses métodos são chamados métodos de seleção de variáveis, que escolhem regiões específicas do espectro (um comprimento de onda ou um conjunto de comprimentos de onda) em que a colinearidade não é tão importante, enquanto geram modelos mais estáveis robustos e mais simples de interpretar. Na prática, a filosofia está baseada na identificação de um subconjunto de variáveis que produzirão erros de previsão mais baixos (OLIVEIRA et al., 2004; SOARES et al., 2008; NETO, 2005).

Os métodos de seleção de variáveis existentes se diferem com relação ao procedimento realizado para a seleção da região espectral, o qual pode ser realizado de forma contínua ou discreta (SOROL, 2010). Dentre os métodos contínuos utilizados atualmente pode-se destacar o iPLS e o siPLS. Dentro os métodos de seleção de variáveis discretos, estão os métodos UVE e o GA. Todos esses métodos foram testados neste trabalho. A técnica de seleção de variáveis permite a eliminação de informações não relevantes, como por exemplo, bandas que não contenham nenhuma informação das espécies ou propriedades a serem analisadas e permitem melhorar a razão sinal/ruído (OLVIEIRA et al., 2004; NETO, 2005).

2.3.2.1 Mínimos quadrados parciais por intervalos (iPLS) e por sinergismo de intervalos (siPLS)

O método iPLS é uma extensão do PLS, que desenvolve modelos locais PLS em subintervalos equidistantes de toda a região do espectro. Seu principal objetivo é prever

informação relevante nas diferentes subdivisões do espectro global, de forma a remover as regiões espectrais cujas variáveis se apresentem como supostamente de menor relevância e/ou contendo sinais apenas de interferentes. A partir deste ponto, um novo modelo PLS é construído a partir das variáveis selecionadas (NØRGAARD et al., 2000).

No iPLS são, portanto, realizadas regressões por mínimos quadrados parciais em subintervalos, de igual peso, de todo o espectro. O espectro é dividido em tantas partes quanto se desejar, até que, através de tentativa e erro, chega-se a uma divisão ótima, ou seja, obtêm-se regiões do espectro com menores erros de previsão e maiores valores de coeficiente de determinação (R^2). Amostras e/ou medidas anômalas detectadas pelo PLS devem ser geralmente removidas antes da aplicação do iPLS.

Os novos modelos construídos são avaliados, igualmente, como em um modelo PLS convencional. A diferença consiste, apenas, na divisão do conjunto de dados em intervalos iguais. O método é planejado para dar uma visão geral dos dados e pode ser útil para selecionar as variáveis mais representativas na construção de um modelo de calibração adequado. Porém, o método iPLS indica a região na qual está contida a informação, sendo uma aproximação univariada, em relação às regiões testadas, pois não fornece sinergismo das regiões espectrais envolvidas (LEARDI e NØRGAARD, 2004)

Para selecionar os subintervalos, a fim de obter melhores habilidades preditivas, pode ser utilizado, também, o algoritmo dos mínimos quadrados parciais por sinergismo de intervalos (siPLS), uma extensão do iPLS. Este possibilita selecionar a melhor combinação de intervalos, combinando 2 a 2, 3 a 3 e até 4 a 4 (ou mais) sub-regiões do espectro, fornecendo geralmente melhores coeficientes de determinação e os menores erros de previsão que o iPLS (NØRGAARD et al., 2000). A avaliação dos modelos criados por siPLS é realizada exatamente como para os modelos criados por iPLS, com a diferença apenas de que nesse primeiro há combinação dos subintervalos criados. Vale ressaltar que a aplicação de ambos os métodos, iPLS ou siPLS, deve sempre ser combinada com o conhecimento espectral do sistema.

2.3.2.2 Eliminação de variáveis não informativas por mínimos quadrados parciais (UVE-PLS) (CENTNER e MASSART, 1996; NETO, 2005; LEARDI e NØRGAARD, 2004)

No método UVE-PLS, o algoritmo descobre e elimina de um modelo PLS as variáveis não informativas. O critério usado para distinguir as variáveis informativas e não informativas é a confiança (estabilidade) dos coeficientes de regressão *b*, obtidos por validação cruzada do tipo *leave-one-out* (deixe um fora por vez). Nesse processo um coeficiente de regressão b é deixado de fora na estimativa da média dos coeficientes de regressão b_j (j = 1, ... , n) e do desvio padrão do vetor de n coeficientes b_{ij}, $s(b_j)$ e a seguir o coeficiente deixado de fora é previsto pelo modelo de regressão.

O critério de confiabilidade t_j (para cada variável *j*), abaixo do qual esses coeficientes são considerados muitos pequenos indicando que a variável correspondente pode ser removida, é calculado com base na Equação 8.

$$t_j = \frac{b_j}{s(b_j)}, para\ j = 1, \dots, p \qquad \textbf{(8)}$$

Onde b_j e $s(b_j)$ correspondem à media e desvio padrão dos coeficientes de regressão respectivamente e p é o número de variáveis das respostas instrumentais. Porém, os desvios padrão s(bj) para os coeficientes de regressão, contidos em b, para o modelo PLS não podem ser estimados diretamente apenas a partir das variáveis espectrais, sendo necessário usar uma matriz de variáveis aleatórias (que simula o ruído dos dados) anexada artificialmente aos dados experimentais.

Sendo assim, para a execução propriamente dita do método UVE-PLS, primeiramente o algoritmo cria uma matriz de números aleatórios [0 , 1] com mesma dimensão da matriz de dados **X**. Os números são multiplicados por uma constante pequena (por exemplo, 10^{-10}) dando-lhes pelo menos uma ordem de magnitude semelhante ao ruído instrumental. Essa multiplicação retém a variação das variáveis, mas torna a influência delas no modelo desprezível.

A nova matriz é acrescentada à matriz de dados **X** original formando uma matriz estendida com o dobro do número de variáveis da matriz original. Posteriormente, modelos

PLS são construídos para cada amostra usando validação cruzada do tipo *leave-one-out*, em que cada modelo contém todas amostras menos uma. Isto conduz a uma matriz de coeficientes de regressão b com n linhas (amostras) e uma coluna para cada variável, original e randômica. Os valores de t são calculados conforme a Equação 8 como sendo a média dos coeficientes b de cada coluna (variáveis) dividida pelo desvio padrão da referida coluna. Por fim, um limite de corte é fixado por uma faixa com os maiores valores positivos e negativos de t calculados para as variáveis randômicas. Todas variáveis com valor de t igual ou mais baixo, ou seja, dentro desta faixa, são eliminadas do modelo final. Isto significa que todas as variáveis, aleatórias e originais, assumindo conter nada mais que ruído são eliminadas.

Para estimar um corte, ou seja, o limite para as variáveis a serem incluídas, um critério de variáveis informativas ou não informativas é obtido a partir da Equação 9:

$$cutoff = k * \max\left(abs(c_{ruído})\right) \quad \textbf{(9)}$$

Onde k é um valor arbitrário (normalmente 2), $\boldsymbol{c_{ruído}}$ são os valores referentes às variáveis artificiais, e max $(abs(c_{ruído}))$ é o máximo valor absoluto do critério de confiabilidade.

Um novo modelo PLS é, portanto, construído a partir das variáveis selecionadas pelo UVE-PLS e sua eficiência comparada ao modelo PLS global.

2.3.2.3 Algoritmo genético (GA)

O GA é uma técnica de seleção de variáveis amplamente aplicada em quimiometria (NIAZI e LARDI, 2012). Um diferencial do GA é a possibilidade de otimização de um subconjunto de variáveis espectrais simultaneamente aos parâmetros de otimização do modelo de calibração.

Ela é uma técnica probabilística de busca e otimização matemática inspirada no princípio Darwiniano da evolução das espécies. Seu processo de otimização baseia-se no princípio de sobrevivência dos indivíduos mais aptos, reprodução e mutação dos mesmos (GOLDBERG, 1989). De acordo com a teoria de Darwin, o princípio de seleção privilegia os indivíduos mais

aptos com maior probabilidade de reprodução, passando seu código genético às próximas gerações.

Estes princípios biológicos foram motivadores para o desenvolvimento de algoritmos matemáticos de busca e otimização. Eles podem ser usados para encontrar solução numérica em problemas com grande número de variáveis, sendo esta uma de suas principais vantagens. Em aplicações químicas, normalmente o algoritmo genético é utilizado para selecionar variáveis em espectroscopia (COSTA FILHO e POPPI, 1999; NIAZI e LEARDI, 2012; LEARDI, 2000; XIN et al., 2012; FEI et al., 2009; LUCASIUS, BECKERS e KATEMAN, 1994).

O ponto de partida para a utilização do GA é a representação matemática do problema. Após a codificação, o algoritmo é inicializado e busca de forma iterativa pontos ótimos dentro do domínio amostral.

No caso da espectroscopia vibracional, cada espectro é tratado como sendo composto por um conjunto de genes (números de onda, por exemplo) que são dispostos em código binário. Cada variável ou número de onda pode receber a codificação binária “1” ou “0” (selecionada ou não selecionada) (COSTA FILHO e POPPI, 1999; ABRAHAMSSON et al., 2003).

O cromossomo original é perturbado randomicamente criando vários cromossomos que formam a população inicial. Para cada cromossomo, é avaliada a resposta associada com as condições experimentais correspondentes. Isto é feito construindo um modelo PLS para cada cromossomo, levando em consideração apenas as variáveis que foram codificadas com “1”. O modelo é então avaliado através de validação cruzada para adquirir um valor de aptidão que descreve a qualidade do modelo (ABRAHAMSSON et al., 2003; ZUPAN e GASTEIGER, 1993). A partir da população inicial, uma nova população a qual pode ser considerada como próxima geração, é obtida pelo cruzamento randômico entre material genético de cromossomos diferentes. No cruzamento, dois cromossomos pai são divididos geralmente em duas ou três partes, cada uma escolhida randomicamente, que são cruzadas e combinadas para formar dois cromossomos filhos que substituirão os cromossomos pai dentro de uma nova geração. Uma nova avaliação é realizada, e os cromossomos com valores de aptidão maiores têm uma probabilidade de reprodução maior que os cromossomos com menores aptidões, tudo para melhorar a aptidão global da população (COSTA FILHO e POPPI, 1999; ABRAHAMSSON et al., 2003; ZUPAN e GASTEIGER, 1993).

Mutações podem ser incorporadas ao modelo e são, às vezes, necessárias para superar alguns problemas na população, sendo utilizadas para dar nova informação genética à população, ou seja, uma variável não selecionada em quaisquer dos cromossomos originais, nunca seria selecionada na próxima geração se mutações não tivessem presentes. São utilizadas, também, para prevenir que a população se sature com cromossomos semelhantes (convergência prematura). Uma mutação nada mais é que a inversão de um gene no cromossomo. A taxa de mutação é usualmente definida e fixada de 0,001 - 0,01.

O algoritmo é repetido até que a condição de término é cumprida. A condição de término é baseada no critério de convergência, em que o algoritmo é encerrado quando uma certa porcentagem dos cromossomos for idêntica ou quando um determinado número de gerações é atingido (ABRAHAMSSON et al., 2003; ZUPAN e GASTEIGER, 1993). As variáveis do cromossomo mais apto são selecionadas para a construção do modelo de calibração, por PLS (GA-PLS), e sua eficiência comparada ao modelo PLS global.

2.3.3 Pré-processamento dos dados

Anteriormente à aplicação das ferramentas quimiométricas ao conjunto de dados a ser investigado, podem ser necessárias transformações dos dados espectrais originais, pois estes podem não ter uma distribuição adequada para a análise, dificultando a extração de informações úteis e interpretação dos mesmos. Basicamente, essas transformações consistem de pré-processamentos dos dados espectrais.

O pré-processamento dos dados tem por objetivo remover variações sistemáticas não desejadas dos espectros, como mudanças na linha de base, efeitos de espalhamento e fatores externos, não controláveis. A seguir são apresentados exemplos de alguns métodos que podem ser utilizados (FEARN et al., 2009; BARNES, DHANOA e LISTER, 1989; SAVITZKY e GOLAY, 1964; ZHANG, Z-M, CHEN e LIANG, 2010):

- Correção do espalhamento multiplicativo (MSC, do inglês *multiplicative scatter correction*): tem por finalidade remover a variabilidade que pode ser causada por efeitos de espalhamento multiplicativo. Através desse pré-processamento são eliminados desvios de linha de base não lineares (*drift*) que ocorrem em medidas de

reflectância em sólidos ou suspensões, devido ao espalhamento multiplicativo da luz. Este último se deve à falta de homogeneidade no tamanho das partículas.

- Variação normal padrão (SNV, do inglês *standard normal variate*): trata-se de outro método para correção do espalhamento multiplicativo da luz. Em geral as equações para a aplicação de SNV e MSC têm a mesma forma e, em muitos casos as duas abordagens produzem resultados semelhantes, de modo que eles são geralmente considerados como permutáveis. A diferença no cálculo ocorre basicamente no fato de que a SNV utiliza a média e o desvio padrão do espectro individual para sua execução, enquanto que o MSC depende de uma relação entre o espectro individual e um espectro de referência (coeficientes linear e angular) para a sua aplicação.
- Autoescalonamento: aplicado quando se deseja comparar variáveis com diferentes dimensões. Consiste, portanto, em se utilizar os dados adquiridos em uma mesma faixa de amplitude de sinal. O processo de autoescalonamento consiste em centrar os dados na média e dividi-los pelo respectivo desvio padrão, sendo um para cada comprimento de onda.
- Primeira e segunda derivadas: a aplicação das derivadas remove uma parte da deformação da linha de base, e melhor a definição de bandas que se encontram sobrepostas em uma mesma região espectral, revelando picos de pequena absorbância. Em geral a derivada de um ponto é calculada a partir de um polinômio aplicado em uma janela de pontos. A primeira derivada elimina desvios lineares da linha de base (termos constantes, enquanto que a segunda derivada elimina os desvios não lineares. Vale ressaltar que as derivadas ampliam ruído e, por isso, devem ser usadas em conjunto com métodos de alisamento como, por exemplo a suavização de Savitzky-Golay
- Correção de linha de base utilizando um método iterativamente adaptativo por mínimos quadrados ponderados e penalizados (airPLS): corresponde a um algoritmo de ajuste da linha de base rápido e flexível. Um procedimento é executado de forma iterativa e reponderada para aproximar gradualmente uma linha de base complexa. Os pesos de iteração são obtidos adaptativamente usando a soma dos erros quadrados entre uma linha de base previamente montada e os sinais originais. A fim de controlar a suavização da linha de base construída, uma penalidade é introduzida com base na soma dos quadrados derivados da mesma. Esse algoritmo é intuitivo e eficaz. Ao contrário da derivada, que também é um método de correção de linha de base, após a

execução do airPLS, o gráfico final possui forma semelhante ao espectro inicial, mantendo inclusive a mesma escala da absorbância.

Como recurso de pré-processamento todos os dados devem, ainda, ser centrados na média com o intuito de permitir que o sistema de eixo das variáveis latentes passe pela média, que consequentemente será a origem dos eixos (PARISOTTO, 2007).

2.3.4 Otimização dos modelos de calibração multivariada

A utilização dos modelos de calibração requer a otimização de alguns parâmetros. Quanto maior o número de parâmetros otimizados, maior o custo computacional e a possibilidade de superajuste aos dados. Para evitar este último problema, dois métodos de validação costumam ser aplicados (BRERETON, 2003):

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

1. *leave-one-out*: neste procedimento uma amostra é removida do conjunto de calibração para validação enquanto constrói-se o modelo com as n–1 amostras restantes. Neste caso, n modelos são construídos até que todas as amostras de calibração sejam utilizadas para validação;

2. *k-fold*: este procedimento consiste em dividir as *n* amostras de calibração em *k* subconjuntos mutuamente exclusivos de mesmo tamanho. Um subconjunto é utilizado para validação enquanto os *k-1* restantes são utilizados para construção do modelo. O procedimento é repetido *k* vezes até que todos os subgrupos tenham sido utilizados como validação. Se os subconjuntos são escolhidos em blocos, o método é denominado *contiguos block*; caso sejam escolhidos aleatoriamente, *random block*; ou sejam retiradas de forma ordenada *venetian blinds*.

No procedimento de validação cruzada as amostras do conjunto de calibração são previstas por diferentes modelos construídos com as próprias amostras de calibração, mas nunca uma mesma amostra participando da calibração e validação simultaneamente. A partir dos valores previstos das amostras de calibração, a raiz quadrada do erro quadrático médio de validação cruzada (RMSECV) é calculada (BRERETON, 2000):

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{N}(y_{prev,i} - y_{ref,i})^2}{n}} \quad \textbf{(10)}$$

onde y_{ref} e y_{prev} são os valores de referência de calibração e estimado pelo procedimento de validação cruzada para n amostras do conjunto de calibração.

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

Após otimização do modelo, este está pronto para ser aplicado em um conjunto de dados que não fez parte de sua construção ou otimização, conjunto este denominado de previsão ou teste. A qualidade dos modelos de calibração multivariada gerados normalmente é avaliada pela raiz quadrada do erro quadrático médio de previsão, RMSEP (VALDERRAMA, BRAGA e POPPI, 2007; VALDERRAMA, BRAGA e POPPI, 2009):

$$RMSEP = \sqrt{\frac{1}{nprev} * \sum_{i=1}^{nprev}(y_{prev,i} - y_{ref,i})^2} \quad \textbf{(11)}$$

onde *nprev* é o número de amostras de previsão (ou teste), y_{ref} e y_{prev} são respectivamente, os valores de referência e previsto pelo modelo para as amostras do conjunto de previsão.

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

$$R^2 = 1 - \frac{\sum_i (y_{ref,i} - y_{prev,i})^2}{\sum_i (y_{ref,i} - \bar{y}_{médio})^2} \quad \textbf{(12)}$$

Onde $\bar{y}_{médio}$ e $y_{prev,i}$ são os valores médio e previsto pelo modelo, respectivamente, podendo ser aplicado tanto às amostras do conjunto de calibração ou previsão. O valor de R^2 varia de 0 a 1, sendo que quanto mais próximo de 1 melhor o ajuste do modelo.

O *leverage*, que representa a distância de cada amostra ao centro dos dados, de cada amostra deve ser avaliado no procedimento de construção do modelo e comparado com os respectivos resíduos, pois amostras contendo alto *leverage* e alto resíduo podem corresponder a *outliers*. Dessa forma é possível se identificar as amostras anômalas presentes, para eliminar qualquer influência negativa nos modelos de calibração (VALDERRAMA, BRAGA e POPPI, 2007).

2.3.5 Avaliação dos modelos

Para avaliação dos modelos PLS, além dos cálculos dos erros já discutidos no tópico anterior, também devem estimadas suas figuras de mérito, como o intervalo de confiança, sensibilidade, seletividade, limite de detecção e limite de quantificação, cujos métodos já foram descritos em alguns trabalhos científicos (VALDERRAMA, 2009; VALDERRAMA, BRAGA e POPPI, 2007; OLIVIERI et al., 2006; ROCHA et al., 2012; ASTM E1655, 2012).

Uma importante etapa de avaliação dos modelos está na análise dos resíduos (e_i) do modelo, calculados por:

$$e_i = y_{ref,i} - y_{prev,i} \quad \textbf{(13)}$$

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

2.3.5.1 Avaliação dos erros sistemáticos

Os erros sistemáticos afetam a estimativa sempre no mesmo sentido, gerando resultados abaixo ou acima do valor esperado (FILGUEIRAS et al., 2014). A presença desses erros é calculada através do teste para viés, que representam o desvio médio de n medições (ASTM 1655, 2012):

$$viés = \frac{\sum(y_i - \hat{y}_i)}{n} \quad \textbf{(14)}$$

Esse teste é executado através de um teste-t bicaudal no qual as hipóteses testadas são:

H_0: viés = 0;

H_1: viés ≠ 0;

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

$$SVD = \sqrt{\frac{\sum_{i=1}^{n}\{[(y_{ref,i} - y_{est}) - viés]^2\}}{n-1}} \quad \textbf{(15)}$$

$$t_{calc} = \frac{|viés|\sqrt{n}}{SVD} \quad \textbf{(16)}$$

A partir daí, o valor de t_{calc} é comparado com o valor de t tabelado ($t_{tab)}$ da distribuição t-*student* com n-1 graus de liberdade e nível de significância α. Se $t_{calc} < t_{tab}$, aceita-se a hipótese nula e considera-se que não há evidências da presença de erros sistemáticos nos resíduos. Caso contrário, a hipótese nula é rejeitada e admite-se a presença de erros sistemáticos.

2.3.5.2 Avaliação dos erros de tendência

A presença de tendência não está diretamente associada à distribuição dos resíduos. Sendo assim, o procedimento normalmente adotado para avaliar tendência em resíduos de calibração multivariada é subjetivo, através da observação visual do gráfico dos resíduos em função dos valores de referência. Entretanto, neste procedimento diferentes analistas podem tomar decisões conflitantes quanto à existência de tendência ou não. Para evitar conclusões subjetivas foi implementado um teste de permutação não paramétrico para avaliar erros de tendência em resíduos de calibração multivariada (FILGUEIRAS et al., 2014). Esse teste baseia-se na probabilidade da tendência observada nos dados ser devida ao acaso portanto, em sua execução são testadas duas hipóteses:

H_0: os resíduos e_i são independentes de y_i;

H_1: os resíduos e_i não são independentes de y_i conforme a Equação:

$$e_i = g(y_i) + \varepsilon_i \quad \textbf{(17)}$$

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

Na hipótese alternativa, a dependência dos resíduos com os valores de referência é proposta, e presume-se que todo o efeito aleatoriamente presente em y_i é devido somente a variável ε_i. Haverá evidências de tendência nos resíduos se o coeficiente polinomial de maior ordem (b_n) da equação polinomial $g(y_i)$ for estatisticamente significativo, ao nível de significância α adotado. Assim, o grau do polinômio deve ser definido previamente à aplicação do teste. O teste pode ser resumido nos seguintes passos (FILGUEIRAS et al, 2014):

I. Calcular o coeficiente polinomial b_n ajustado para os dados originais dos resíduos em função dos valores de referência. Assim, o coeficiente será denominado b_n^*;
II. Permutar randomicamente somente o vetor **y** contendo a propriedade de interesse;
III. Calcular o coeficiente b_n^i para o *i*-ésimo vetor **y** permutado;
IV. Comparar b_n^* com b_n^i
V. Repetir as etapas (ii) a (iv) k-vezes.

Como os coeficientes b_n^i dos ajustes permutados não tem sentido físico, por serem aleatórios, a distribuição destes coeficientes constituirá a distribuição para o teste. Sendo assim, o p-valor do teste é determinado pela proporção do número de vezes em que $b_n^* > b_n^i$. Se o p-valor do teste for menor que o nível se significância α adotado, rejeita-se H_0, e a tendência observada é significativa. Caso contrário, aceita-se H_0 e a hipótese dos resíduos serem aleatórios.

3. OBJETIVOS

O objetivo principal deste trabalho é desenvolver uma metodologia analítica capaz de determinar os seguintes parâmetros físico-químicos em frações de destilação de petróleos:

1. Grau API (°API);
2. Índice de cetano
3. Índice de refração (a 20°C);
4. Enxofre (%m/m);
5. Ponto de fluidez (°C);
6. Ponto de fuligem (mm);
7. Ponto de anilina (°C);
8. Ponto de congelamento (°C);
9. Ponto de entupimento (°C) e
10. Ponto de névoa (°C);

Com este trabalho pretende-se, também:

1- Realizar medidas espectroscópicas NIR em derivados de petróleos com diferentes faixas temperaturas de destilação.

2- Realizar medidas espectroscópicas MIR em derivados de petróleos com diferentes faixas temperaturas de destilação.

3- Desenvolver e otimizar modelos por regressão multivariada para NIR para a determinação de propriedades físico-químicas dos derivados de petróleos.

4- Desenvolver e otimizar modelos por regressão multivariada para MIR para a determinação de propriedades físico-químicas dos derivados de petróleos.

5- Testar e comparar o desempenho dos métodos de seleção de variáveis iPLS, siPLS, GA e UVE, quando aplicados a espectros NIR, avaliando-se o desempenho dos respectivos modelos criados.

6- Testar e comparar o desempenho dos métodos de seleção de variáveis iPLS, siPLS, GA e UVE, quando aplicados a espectros MIR, avaliando-se o desempenho dos respectivos modelos criados.

7- Comparar a eficiência dos modelos de calibração multivariada otimizados para NIR e MIR para determinação de propriedades físico-químicas dos derivados de petróleos.

4. METODOLOGIA

Neste trabalho foram utilizados 104 frações de petróleo obtidas por destilação de nove diferentes petróleos, com grau API variando de 12,3 a 57,7, ou seja, foram incluídos amostras derivadas de diferentes tipos de petróleos, desde extra-leves até pesados. Os petróleos utilizados foram destilados na faixa de 15 °C a 500 °C conforme ASTM D 2892 (2011) e ASTM D 5236 (2007). Os cortes obtidos (amostras) foram cedidos pelo CENPES/Petrobras, armazenados em frascos apropriados e em temperatura adequada para posteriores análises. O processo para a destilação de um petróleo e obtenção de suas frações para análise custa, atualmente, em torno de R$15.000,00.

4.1 DETERMINAÇÃO DOS PARÂMETROS FÍSICO-QUÍMICOS

As propriedades físico-químicas foram determinadas conforme os respectivos métodos padrão descritos na Tabela 2. O teor de enxofre possui dois métodos relacionados, o ASTM D 5453 – 12 (2012), referente à análise dos cortes leves, e o ASTM D 4294 – 10 (2010), referente à análise de cortes médios e pesados.

Na Tabela 2 são reportados os valores de reprodutibilidade (R) dos respectivos métodos laboratoriais de cada propriedade trabalhada. Esses valores referentes ao teor de enxofre, ao ponto de entupimento e ao índice de acidez são heterocedásticos e, portanto, o valor apresentado na Tabela (2) refere-se à amostra, contida no conjunto de dados trabalhado, contendo o maior valor encontrado para cada uma dessas propriedades analisadas. Ainda nesse contexto, o valor de reprodutibilidade apresentado para o índice de cetano refere-se ao valor aceito pela indústria petroquímica, visto que o cálculo da reprodutibilidade para tal método é dependente da precisão das determinações da densidade e da temperatura de recuperação, que entram no cálculo e, por isso, esse valor também varia de amostra para amostra.

Tabela 2 – Métodos ASTM referentes às propriedades e suas reprodutibilidades

Propriedade	Método	Reprodutibilidade
Grau API	ISO 12185:1996 (2008)	1,1
Índice de Cetano	ASTM D 4737 – 10 (2010	10%
Índice de Refração	ASTM D 1218 – 12 (2012)	0,0005
Teor de Enxofre	ASTM D 5453 – 12 (2012) e ASTM D 4294 – 10 (2010)	até 0,1191 (a)
Ponto de fuligem	ASTM D 1322 – 15 (2015)	3 (b)
Ponto de Anilina	ASTM D 611 – 12 (2012)	1 (c)
Ponto de Congelamento	ASTM D 5972 – 05 (2010)	0,8 (c)
Ponto de Entupimento	ASTM D 6371 – 05 (2010)	até 7 (c)
Ponto de Névoa	ASTM D 5773 – 15 (2010)	2,5 (c)
Ponto de Fluidez	ASTM D 5950 – 14 (2012)	4,5 (c)

a = %m/m, **b** = mm, **c** = °C.

Fonte: O Autor.

A determinação dos parâmetros físico-químicos das amostras foi realizada no Laboratório CENPES/Petrobras.

4.2 INSTRUMENTAÇÃO

4.2.1 Infravermelho próximo (NIR)

Os espectros de infravermelho próximo (NIR) foram obtidos em um espectrômetro modelo Nicolet 380, do fabricante Thermo Fisher, equipado com detector de DTGS KBr e beamsplitter XT-KBr, usando como fonte luz branca. Como acessório para obtenção dos espectros, foi utilizada uma cubeta de vidro. O espectro registrado foi obtido como uma média de 64 varreduras consecutivas, com resolução de 8 cm^{-1} na faixa de trabalho de 3500 a 9200 cm^{-1}. Ao todo obtiveram-se 2956 comprimentos de onda (variáveis). Anteriormente a cada análise foi medido o branco. Para essa análise é necessário em torno de 1 mL de amostra.

Os dados de NIR, utilizados para a construção dos modelos quimiométricos foram obtidos no Laboratório CENPES/Petrobras.

4.2.2 Infravermelho médio (MIR)

O espectrômetro utilizado foi o ABB Bomen modelo MB 102, equipado com um detector de sulfato de triglicerina deuterado (DTGS). Os espectros foram obtidos em absorbância, na região do MIR, na faixa de 4000 a 600 cm^{-1} com uma varredura de 32 scans e uma resolução de 4 cm^{-1} em um cristal de seleneto de zinco com ângulo de incidência de 45ºC, de 80 mm de comprimento, 10 mm de largura, 4 mm de espessura e 10 reflexoes, do fabricante *Pike Technologies*. Ao todo obteve-se 1764 comprimentos de onda (variáveis). Anteriormente à obtenção do espectro de cada amostra foi feito o *background* utilizando-se o ar como referência para a correção da linha de base. Para aquisição dos espectros foi utilizado o programa GRAMS/AI 7.00. Para essa análise também é necessário em torno de 1 mL de amostra.

Os dados de MIR, utilizados para a construção dos modelos quimiométricos foram obtidos no Laboratório Labpetro/UFES.

4.3 QUIMIOMETRIA

O método quimiométrico principal aplicado aos dados foi o PLS, para o qual foram testados diferentes tipos de seleção de variáveis: iPLS, siPLS, UVE e GA. A construção dos modelos foi realizada em 3 etapas principais:

1 – Pré-processamento dos dados;

2 – Construção dos modelos propriamente dita, com e sem seleção de variáveis;

3 – Avaliação e validação dos modelos.

Todos os cálculos foram executados no software Matlab versão 7.

4.3.1 Tratamento e pré-processamento dos dados

Anteriormente à aplicação das ferramentas quimiométricas PLS simples, iPLS e siPLS ao conjunto de dados a ser investigado, foram aplicados e testados como recurso de tratamento dos dados a SNV e a 1ª. derivada com suavização de Savitzky-Golay (SAVITZKY e GOLAY, 1964) utilizando um polinômio de segunda ordem e uma janela de 7 pontos. Todos os dados foram, ainda, centrados na média antes da construção dos modelos.

Da mesma forma, anteriormente à aplicação das técnicas de seleção de variáveis UVE e GA , foi aplicado o método de correção de linha de base airPLS e, posteriormente, todos os dados também foram centrados na média antes da construção dos modelos.

4.3.2. Construção dos modelos

Anteriormente à construção dos modelos propriamente dita, foi feita a divisão das amostras, selecionando-se aquelas que constituíram os conjuntos de calibração e previsão. Para a seleção, aproximadamente 70% das amostras foram previamente selecionadas para o conjunto de calibração, enquanto que os 30% restantes foram para o conjunto de previsão, utilizando o procedimento de ordenar os valores de referência em ordem crescente e escolher uma amostra a cada três, método também conhecido como persianas.

Vale ressaltar que, embora tenham sido analisadas 104 amostras no total, alguns parâmetros não foram avaliados em todas elas, pois são específicos de determinados tipos de cortes. A decisão sobre a necessidade da realização de determinadas análises em determinados tipos de cortes depende tanto da rotina de avaliação e caracterização do petróleo, como do tipo de amostra a ser avaliada como, por exemplo, o ponto de congelamento, um índice de avaliação de combustíveis de aviação que, portanto, é medido apenas nos cortes com faixas de destilação específicas, próximas à correspondente a esse derivado especificamente (150 – 250 ºC). Entretanto, os conjuntos de calibração e validação, selecionados por parâmetro, eram exatamente os mesmos para ambos os modelos MIR e NIR

Então, feita essa separação dos conjuntos de calibração e validação, foram construídos os modelos globais utilizando PLS, para cada parâmetro analisado, e para cada tipo de pré-processamento e tratamento testados, utilizando toda a faixa espectral, por MIR e por NIR. Esse procedimento foi realizado utilizando o *software* PLS Toolbox versão 4.0 (WISE et al.,

2006). Nesse momento também foi realizada uma análise do *leverage* versus resíduos gerados por amostra, de forma a identificar possíveis *outliers*. Assim, amostras que apresentaram alto resíduo e alto *leverage* foram retiradas do conjunto de amostras a serem trabalhados (VALDERRAMA, BRAGA e POPPI, 2007).

O número de variáveis latentes para cada modelo de calibração foi otimizado pelo procedimento de validação cruzada utilizando o método *venetian blind 5-fold*. Dessa forma, o número de VL's a ser utilizado na construção do modelo foi determinado pelo gráfico de RMSECV em função do número de variáveis latentes, ou seja, foi selecionado aquele com o número de variáveis latentes correspondente ao menor valor de RMSECV e R^2 mais próximo de 1, não ultrapassando o número de 10 VL's.

4.3.3. Seleção de variáveis

Para a criação de modelos com seleção de variáveis foram empregados inicialmente os métodos de seleção de variáveis iPLS e siPLS utilizando o *software* iToolbox, desenvolvido por Lars Nørgaard disponível gratuitamente através do site {http://www.models.life.ku.dk/ipls} (WAGNER, 2000).

As 1764 absorbâncias dos espectros MIR originais (599,8 a 4000 cm^{-1}) foram divididas em 6, 12, 18, 24, 30 e 36 intervalos correspondendo a aproximadamente 294, 147, 98, 74, 59 e 49 variáveis em cada intervalo, respectivamente. Para se verificar o sinergismo entre os intervalos foi aplicado o método siPLS, utilizando-se combinações de até 3 intervalos.

Da mesma forma, as 2956 absorbâncias dos espectros NIR originais (3500,3 a 9199,2 $cm^{-1)}$ foram divididas em 10, 20, 30, 40, 50 e 60 intervalos (correspondendo a 296, 148, 95, 74, 59 e 49 variáveis em cada intervalo, respectivamente), de forma que os intervalos formados por MIR e NIR tivessem aproximadamente o mesmo número de variáveis originais utilizadas, facilitando, posteriormente, a comparação entre os modelos selecionados por esses dois métodos. Também na construção dos modelos a partir de espectros NIR foram feitas combinações de até 3 intervalos.

Os intervalos (e suas associações) com melhor desempenho foram selecionados para a construção dos diferentes modelos sendo, posteriormente, comparados com a eficiência do modelo global (modelo PLS utilizando espectro completo, com todos os números de onda) e com o modelo iPLS previamente selecionado. A partir dos valores obtidos foram selecionados

os intervalos mais correlacionados, e estes foram utilizados para a previsão de amostras externas.

Aos dados também foram aplicados os métodos de seleção de variáveis UVE, utilizando o pacote disponibilizado pelo autor (CENTNER e MASSART, 1996), e o GA utilizando o *software* PLS toolbox versão 4.0 (WISE et al., 2006).

O método UVE foi aplicado considerando-se um nível de confiança de 99% e os coeficientes de regressão foram obtidos por validação cruzada do tipo *leave-one-out* 5-fold.

A seleção de variáveis para a aplicação do GA foi realizada nas seguintes condições:

- Tamanho da população: 400 cromossomos para os espectros MIR e 600 para os espectros NIR;
- Número máximo de gerações: 200;
- Convergência: 80%;
- Taxa de mutação: 0,005 (0,5%);
- Número de variáveis por gene: 5;
- Número de indivíduos na população inicial: 10%;
- Número de partes em que os cromossomos são divididos: 2;
- Algoritmo: PLS;
- Número máximo de VLs para o PLS: 10;
- Tipo de Validação Cruzada: *venetian blind 5-fold*;
- Número de blocos para a Validação Cruzada: 1;

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

Ao final, todos os métodos de seleção de variáveis aplicados foram comparados quanto ao seu desempenho para previsão de cada propriedade avaliada.

4.3.4. Avaliação dos modelos

A avaliação dos modelos se deu pela determinação e análise dos seguintes requisitos: coeficiente de determinação (R^2), curva obtida entre os valores previstos versus medidos, erros de validação cruzada e previsão.

Foi estabelecido, para a execução desse trabalho, que o valor numérico de R^2 (Equação 12) deveria ser superior a 0,8, pois assume-se que valores inferiores a este indicam baixa qualidade preditiva do modelo. Portanto, modelos com valores de R^2 inferiores a 0,8, seja no conjunto de calibração ou de previsão, foram rejeitados.

Os pontos amostrais deveriam estar próximos à curva dos valores reais pelos valores de referência. Pontos distantes da curva demonstram inexatidão do resultado.

A eficiência dos modelos de calibração multivariada foi avaliada pelo cálculo de três tipos de raiz quadrada do erro (RMSE – do inglês, *root mean square error*): a raiz quadrado do erro quadrático médio de calibração (RMSEC), definido como na equação 18, onde n corresponde ao número de amostras e VL corresponde ao número de variáveis latentes selecionadas:

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{N}(y_{prev,i} - y_{ref,i})^2}{n - VL - 1}} \quad \textbf{(18)}$$

Onde y_{ref} e y_{prev} são respectivamente, os valores de referência e previsto pelo modelo para as amostras de calibração.

Já o RMSE, determinado no procedimento de validação cruzada (RMSECV, do inglês *root mean square error of cross validation*) e de previsão (RMSEP, do inglês *root mean square error of prediction*), foi calculado conforme a Equação 19 (BRERETON, 2000).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_{prev} - y_{ref}\right)^2}{N}} \quad \textbf{(19)}$$

Onde y_{ref} e y_{prev} são respectivamente, os valores de referência e previsto pelo modelo para as amostras de calibração ou validação e N corresponde ao número de amostras em cada conjunto (calibração ou previsão).

Esses valores de RMSE indicam a grandeza dos erros associados aos resultados obtidos e, como condição de aceitação dos modelos criados, não devem ser estatisticamente diferentes, ou seja, apresentar uma diferença grande em questão de unidades demonstrando, dessa forma, que os conjuntos de calibração e previsão são semelhantes. Para a execução desse trabalho foi estabelecido uma relação RMSEP/RMSEC de no máximo 2,5 como critério de aceitação do modelo, de forma a evitar sobreajuste. Como esses valores de RMSE são determinados na dimensão das propriedades que está se tentando determinar, sua avaliação depende da amplitude e da dimensão dos dados trabalhados e, nesse sentido, o cálculo do erro percentual de previsão facilita essa avaliação. Nesse contexto, também foi determinado o erro percentual de previsão a partir da Equação 20 (FILGUEIRAS et al., 2014):

$$RMSE\% = 100\frac{RMSEP}{\bar{y}_{previsão}} \quad \textbf{(20)}$$

onde $\bar{y}_{previsão}$ é a media dos valores de previsão.

Os modelos escolhidos foram submetidos a testes estatísticos, e tiveram suas figuras de mérito calculadas, como o intervalo de confiança e a sensibilidade analítica (VALDERRAMA, BRAGA e POPPI, 2007; OLIVIERI et al., 2006; FILGUEIRAS et al., 2014). Vale ressaltar que como o PLS trata-se de um método de calibração inversa, nos modelos deve ser avaliado o inverso da sensibilidade analítica, que representa a menor diferença que o modelo é capaz de distinguir.

O teste de viés foi aplicado para verificação da presença de erros sistemáticos e, para a avaliação de tendência nos resíduos do modelo, foi aplicado o teste de permutação (FILGUEIRAS et al., 2014). Quando algum destes testes indicava erro significativo, era realizada uma nova avaliação com a modificação do número VL's, de forma a tentar eliminar esses erros. Todos esses testes foram calculados, considerando o intervalo de 95% de confiança (neste trabalho sempre foi adotado $\alpha = 0{,}05$), a partir de algoritmos criados baseados na ASTM E1655 (2012).

Ainda, na avaliação dos resultados, os valores de repetibilidade e reprodutibilidade (R) são utilizados como indicadores do desempenho do modelo a ser alcançada. Isto é, considera-se que o RMSEP deve ser comparável à ASTM e a incerteza final de previsão é expressa pela Equação 21 (LAXALDE et al, 2011):

$$\sigma_{previsão} = \sqrt{RMSEP^2 + \sigma_{ASTM}^2} \quad \textbf{(21)}$$

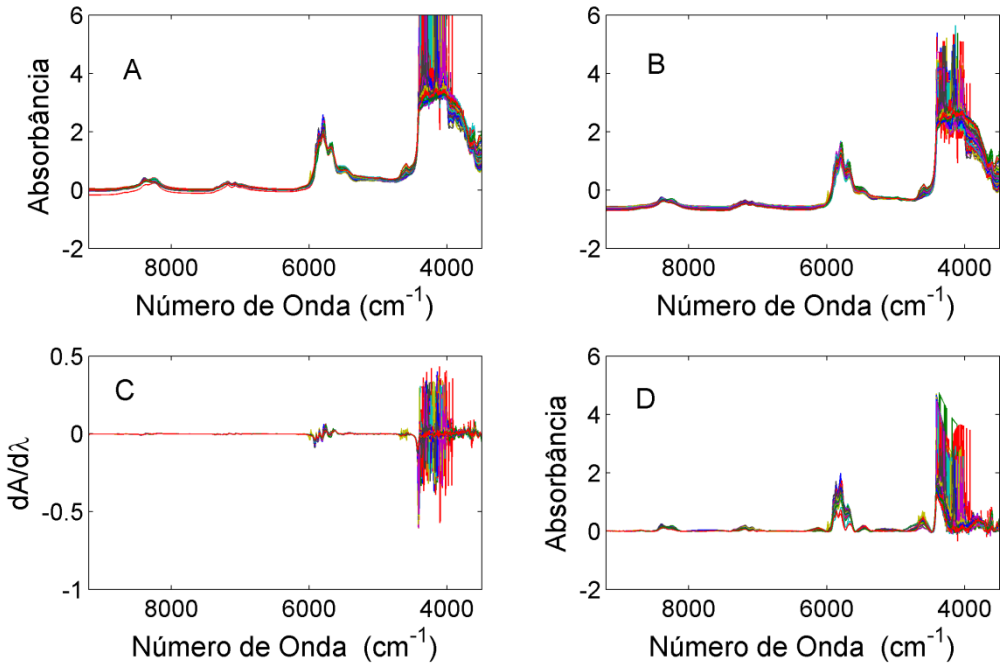
,onde $\sigma_{previsão}$ é o erro total de medição do modelo de previsão, RMSEP é a média da raiz quadrada do erro de previsão e σ_{ASTM} é o valor de variabilidade em medidas feitas em forma de repetibilidade ou reprodutibilidade. Neste trabalho esse cálculo foi feito apenas a partir dos valores de reprodubilidade, já apresentados na Tabela 2.

5. RESULTADOS E DISCUSSÕES

5.1 ESPECTROS NIR E MIR

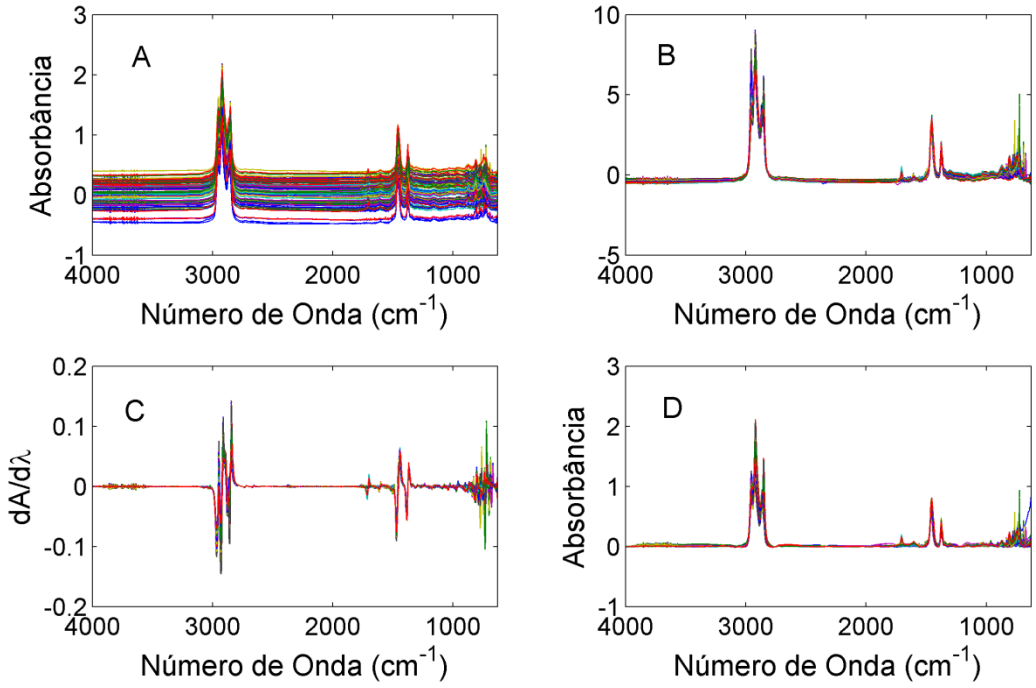
As Figuras 2 e 3 apresentam os respectivos espectros NIR e MIR obtidos para as amostras analisadas, bem como os efeitos dos tratamentos aplicados (SNV, derivada e airPLS) aos dados espectrais.

Figura 2 – Espectro NIR bruto (A), com aplicação dos pré-tratamentos: SNV (B), derivada (C) e airPLS (D) das 104 amostras.



Fonte: O Autor.

Figura 3 – Espectro MIR bruto (A), com aplicação dos pré-tratamentos: SNV (B), derivada (C) e airPLS (D) das 104 amostras.



Fonte: O Autor.

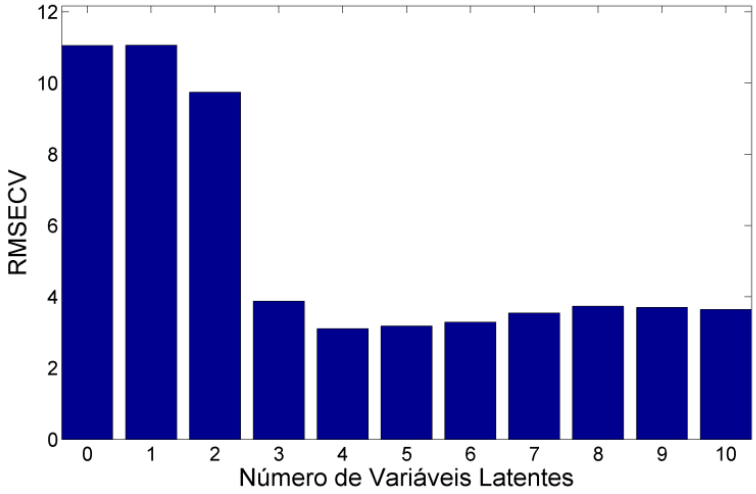
A partir das Figuras 2 e 3, pode se observar que os espectros MIR possuem maior ruído e deslocamento de linha de base quando comparados com os espectros NIR, os quais apresentam um ruído relevante apenas na região de 3500 a 4400 cm^{-1} aproximadamente, a qual não foi possível ser eliminada com a aplicação dos métodos de pré-tratamento aplicados. Consequentemente a isso, houve um maior ganho quanto à aplicação dos métodos de tratamento e pré-processamento aos espectros MIR, quando comparados ao NIR. Também, o desempenho preditivo dos espectros NIR, em geral melhoraram quando essa região (3500 a 4400 cm^{-1} foi excluída, conforme será apresentado a seguir. Vale destacar que esse ruído pode ter ocorrido pelo fato de essa região estar no limite de leitura do equipamento.

5.2 MÉTODO DOS MÍNIMOS QUADRADOS PARCIAIS SIMPLES, POR INTERVALOS E POR SINERGISMO DE INTERVALOS

Como já citado anteriormente, a construção dos modelos por PLS, iPLS e siPLS é feita de forma bastante semelhante e, neste trabalho, foram testados inclusive os mesmos tipos de pré-processamento dos dados.

Exemplificando, para cada uma dessas técnicas avaliadas, a escolha do melhor número de VL's dos modelos construídos foi baseada na avaliação do RMSECV. Nos métodos de calibração multivariada, o número apropriado de VL's é aquele que melhor descreve a variabilidade do conjunto de dados, tendo por objetivo avaliar a dimensionalidade dos modelos desenvolvidos, sem que ocorra subajuste ou sobreajuste. Como exemplo, na Figura 4 é mostrado o gráfico do número de RMSECV *vs* variáveis latentes..

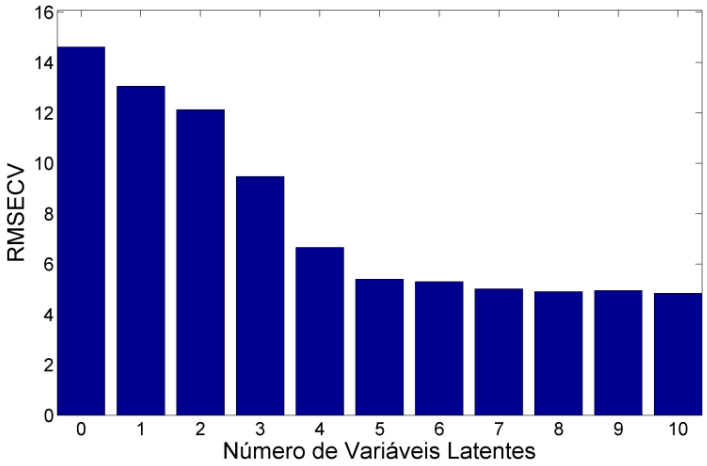
Figura 4 – Gráfico utilizado para a seleção das VL's em função do RMSECV para a previsão do Índice de cetano por MIR sem pré-tratamento.



Fonte: O Autor.

Com base no gráfico da Figura 4, o algoritmo determina 4 VL's, pois não há mais redução do RMSECV em relação ao valor referente à variável latente posterior. Porém, o número de VL's pode apresentar uma evolução monotônica a partir de determinadas VL's, como pode ser observado na Figura 5. Nesse caso, ao invés de utilizar 10 VL's, pode-se utilizar 5 VL's, pois não há uma diferença significativa no valor de RMSECV entre esses dois modelos não justificando, portanto, o uso de 5 VL's a mais.

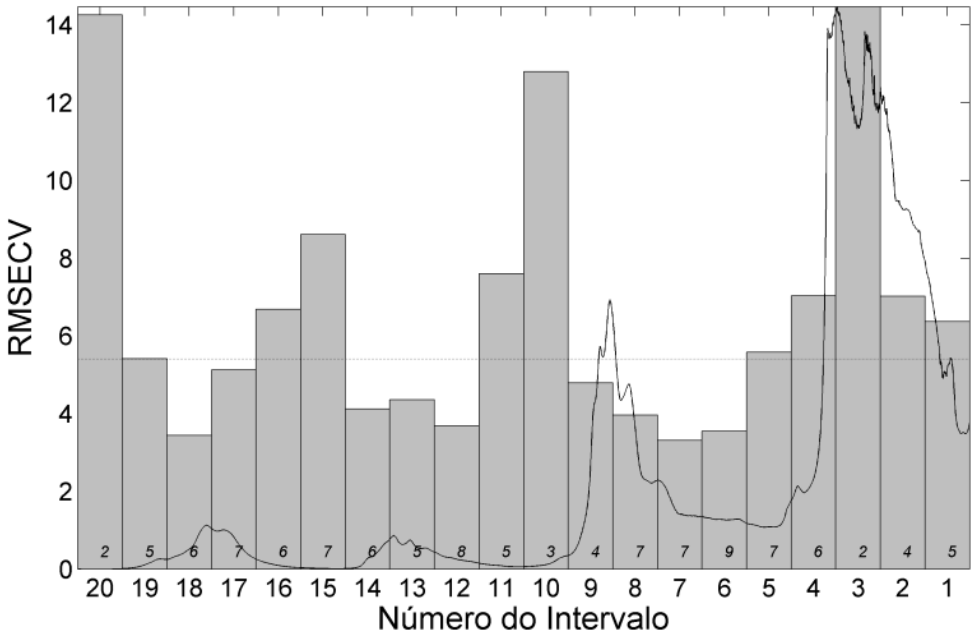
Figura 5 – Gráfico utilizado para a seleção das VL's em função do RMSECV para a previsão do ponto de anilina por MIR sem pré-tratamento.



Fonte: O Autor.

Os espectros das amostras NIR e MIR foram divididos pelos algoritmos iPLS e siPLS em vários subintervalos, conforme o exemplo ilustrado na Figura 6.

Figura 6 – Representação da divisão das variáveis (números de onda) NIR em 20 intervalos.



Fonte: O Autor.

Neste exemplo, o espectro NIR foi dividido em 20 intervalos. A linha tracejada corresponde ao valor do RMSECV para o modelo global, enquanto que as barras correspondem ao RMSECV dos intervalos e o número na base da barra corresponde às variáveis latentes utilizadas para construir cada modelo. Assim, na Figura 6 o subintervalo 7 possui o menor RMSECV e indica a utilização de 7 variáveis latentes. Deve-se verificar a faixa de número de onda que está compreendida nesse intervalo e, neste caso, o intervalo 7 compreende a faixa espectral de 5496,4 a 5212,9 cm^{-1}.

Desta forma, todo esse processo foi repetido para a divisão dos espectros MIR em 6, 12, 18, 24, 30 e 36 intervalos e NIR em 10, 20, 30, 40, 50 e 60 intervalos para cada parâmetro físico-químico predito e em cada divisão foi aplicado, também, o método siPLS com associação de até 3 intervalos.

5.2.2 Modelos selecionados por PLS

Inicialmente foram criados modelos para a previsão das propriedades físico-químicas por PLS a partir de dados de NIR e MIR, utilizando-se todas as variáveis espectrais (PLS global). Sendo assim, os modelos selecionados por essa técnica quimiométrica, para a previsão de cada propriedade, estão descritos na Tabela 3.

A partir da Tabela 3 pode se verificar que para a previsão por NIR, maior parte dos modelos foi selecionada sem a aplicação de algum tratamento prévio dos dados (exceto pela centragem na média), entretanto, para algumas porpriedades (pontos de anilina, de entupimento, de névoa e de fluidez) o melhor modelo foi encontrado aplicando-se a SNV. Por outro lado, para a previsão das propriedades a partir de dados de MIR foi necessária a aplicação de um ou outro método de tratamento testado (SNV e derivada), exceto apenas para a previsão do teor de enxofre, para o qual o melhor modelo foi encontrado apenas centrando os dados na média.

Vale destacar que alguns modelos selecionados por NIR geraram resultados fora do estipulado primordialmente para a aceitação dos modelos. São esses os modelos criados para a previsão do ponto de fuligem e ponto de névoa, que resultaram em altos valores de RMSEP e baixos valores de R^2 (menores que 0,8, valor previamente estabelecido para aceitação dos modelos) e aqueles criados para a previsão dos pontos de congelamento e de entupimento que,

embora tenham gerado valores aceitáveis de R^2 para as amostras de calibração, o mesmo sucesso não foi obtido para a previsão de amostras externas, gerando valores de R^2 bastante baixos. Essas peculiaridades podem também ser facilmente visualizadas na Figura 7, através da observação dessa relação entre os valores medidos e previstos. Observando-se, nessa Figura 7, a relação entre valores medidos e previstos observa-se que o modelos selecionados para a previsão dos pontos de entupimento e de névoa tiveram desempenho bastante ruim, como todos os valores previstos praticamente constantes, para todo o conjunto testado e com intervalos de confiança bastante extensos. Por outro lado, de uma forma geral, os modelos apresentados na Tabela 3 para a previsão das propriedades a partir de dados de MIR, geraram resultados satisfatórios e suas respectivas relações entre valores medidos e previstos são mostradas na Figura 8.

Tabela 3 – Valores calculados para a validação dos modelos selecionados para cada propriedade físico-química, por PLS.

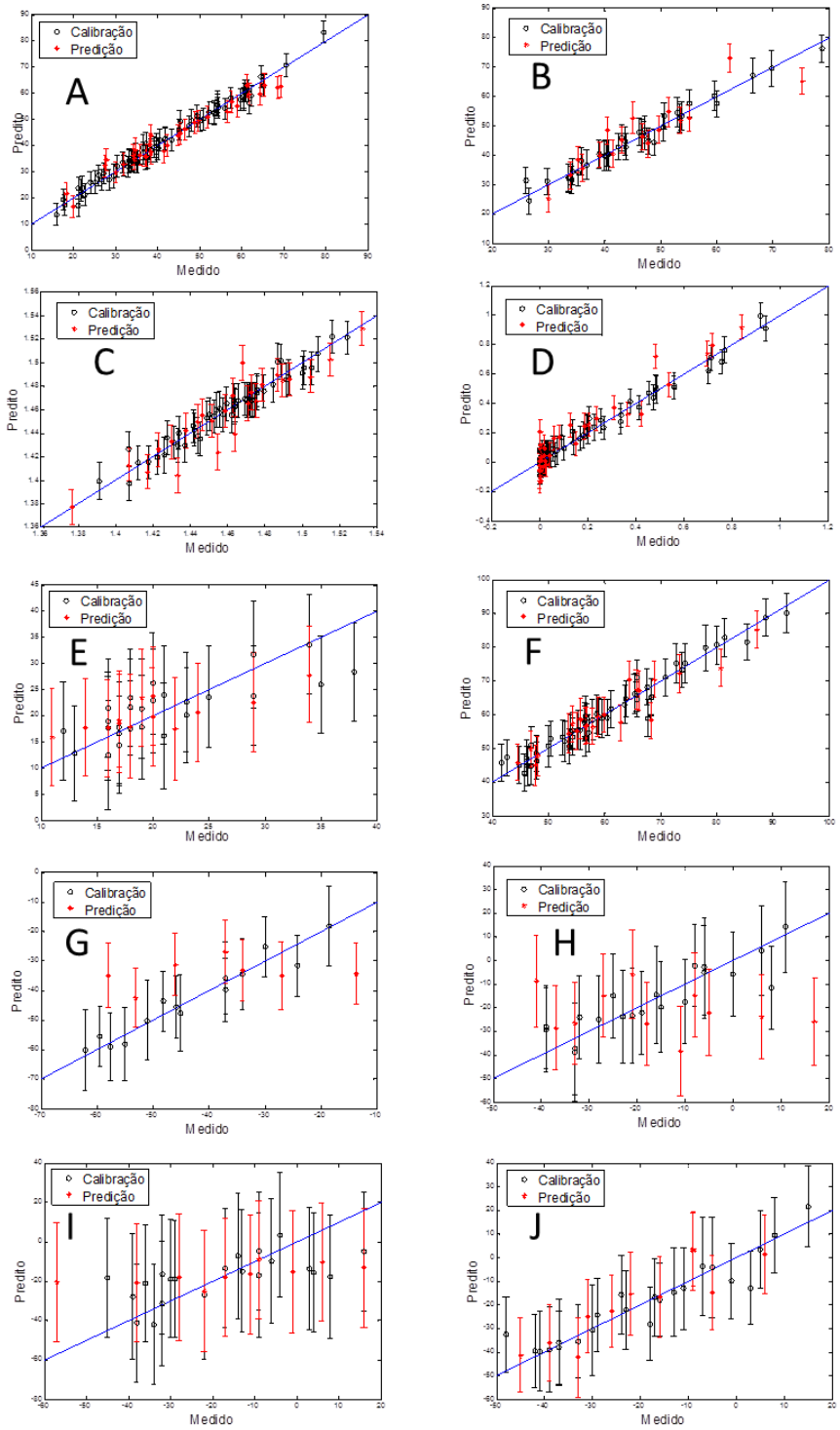
IR	Parâmetros \ Propriedades	º API	Índice de Cetano	Índice de Refração	Teor de Enxofre	Ponto de fuligem	Ponto de Anilina	Ponto de Congelamento	Ponto de Entupimento	Ponto de Névoa	Ponto de Fluidez
NIR	Tratamento	CM	CM	CM	CM	CM	SNV + CM	CM	SNV + CM	SNV + CM	SNV + CM
	VL's	7	7	5	10	3	7	6	4	2	6
	RMSEC	2,0	2,2	0,0068	0,0400(a)	4,4(b)	2,59(c)	4,6(c)	8(c)	14(c)	7(c)
	RMSEP	3,2	4,9	0,0132	0,0819(a)	3,9(b)	3,42(c)	14,4(c)	22(c)	18(c)	7(c)
	$R^2_{calibração}$	0,9833	0,9732	0,9545	0,9788	0,6184	0,9602	0,9383	0,7779	0,4129	0,8787
	$R^2_{previsão}$	0,9600	0,8143	0,8663	0,9214	0,6357	0,8951	0,0664	0,0695	0,4441	0,8002
	Sens. Anal. $^{-1}$	0,6	0,8	0,0012	0,0147(a)	0,3(b)	0,62(c)	0,4(c)	0,9(c)	0,6(c)	1,1(c)
MIR	Tratamento	SNV + CM	SNV + CM	D + CM	CM	D + CM	CM	D + CM	SNV + CM	D + CM	D + CM
	VL's	3	10	6	7	6	8	7	5	3	8
	RMSEC	1,6	1,0	0,0023	0,0359(a)	2,0(b)	1,21(c)	0,9(c)	2(c)	4(c)	2(c)
	RMSEP	3,2	3,9	0,0023	0,0582(a)	2,5(b)	2,68(c)	1,8(c)	9(c)	6(c)	6(c)
	$R^2_{calibração}$	0,9889	0,9947	0,9957	0,9814	0,9351	0,9909	0,9983	0,9906	0,9464	0,9887
	$R^2_{previsão}$	0,9509	0,9302	0,9940	0,9458	0,8457	0,9464	0,9778	0,8603	0,9506	0,8910
	Sens. Anal. $^{-1}$	0,3	0,4	0,0005	0,0069(a)	0,3(b)	0,41(c)	0,2(c)	0,5(c)	0,5(c)	0,4(c)

a = %m/m, **b** = mm, **c** =(°C); SNV = standard normal variate; CM = centrado na media; D = suavização de savitzky-golay.

Sens. Anal.$^{-1}$ = Inverso da Sensibilidade analítica.

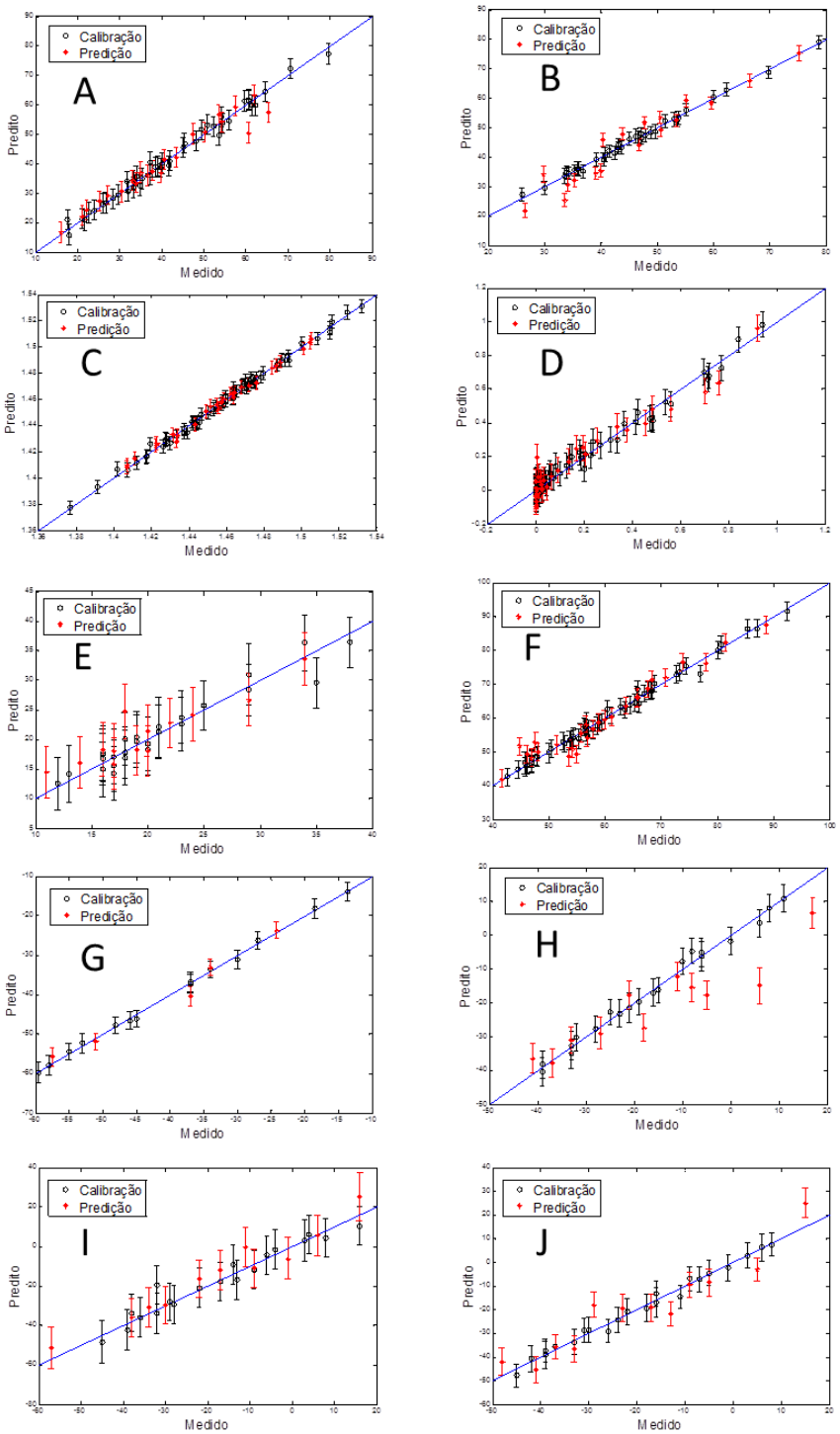
Fonte: O Autor.

Figura 7 – Relação entre valores medidos e previstos referentes aos modelos criados por PLS a partir de dados de NIR (A – Grau API, B – Índice de Cetano, C – Índice de Refração, D – Teor de Enxofre (%m/m), E – Ponto de Fuligem (mm), F – Ponto de Anilina (°C), G – Ponto de Congelamento (°C), H – Ponto de Entupimento (°C), I – Ponto de Névoa (°C), J – Ponto de Fluidez (°C)).



Fonte: O Autor.

Figura 8 – Relação entre valores medidos e previstos referentes aos modelos criados por PLS a partir de dados de MIR (A – Grau API, B – Índice de Cetano, C – Índice de Refração, D – Teor de Enxofre (%m/m), E – Ponto de Fuligem (mm), F – Ponto de Anilina (°C), G – Ponto de Congelamento (°C), H – Ponto de Entupimento (°C), I – Ponto de Névoa (°C), J – Ponto de Fluidez (°C)).



Fonte: O Autor.

Entretanto, nenhum modelo apresentado na Tabela 3 apresentou erros sistemáticos ou de tendência nos resíduos de calibração ou de previsão, visto que todos os valores de *p* encontrados foram maiores do que o nível de significância adotado de 0,05. Os respectivos valores de *p* calculados, em ambos os testes, para as amostras de previsão são mostrados na Tabela 4. Vale destacar, ainda, que não houve presença de erros de tendência ou sistemático para o conjunto de calibração em nenhum modelo selecionado.

Tabela 4 – Resultados da avaliação de erros sistemáticos e de tendência nos resíduos de previsão dos modelos criados por PLS

Propriedades	NIR		MIR	
	Erro Sistemático	Erro de Tendência	Erro Sistemático	Erro de Tendência
Grau API	0,37	0,05	1,00	0,28
Índice de Cetano	0,65	0,49	0,57	0,05
Índice de Refração	0,33	0,42	0,42	0,29
Teor de Enxofre	0,13	0,05	0,85	0,15
Ponto de fuligem	0,91	0,41	0,21	0,41
Ponto de Anilina	0,65	0,32	0,59	0,31
Ponto de Congelamento	0,46	0,25	0,83	0,11
Ponto de Entupimento	0,46	0,29	0,06	0,26
Ponto de Névoa	0,93	0,46	0,05	0,09
Ponto de Fluidez	0,38	0,37	0,95	0,37

Fonte: O Autor.

A partir do exposto, nas Tabelas e Figuras acima, verifica-se que a previsão das propriedades, por PLS, a partir de NIR não foi tão eficiente, mesmo para aqueles modelos que estão estatisticamente aprovados, apresentando muitas amostras com intervalos de confiança distante da linha de calibração, o que demonstra uma inexatidão do modelo. O mesmo ocorreu para a previsão a partir de MIR, porém em uma dimensão menor.

Dessa forma, comparando-se o desempenho previsor de cada propriedade (Tabela 3), há um desempenho melhor dos modelos criados a partir dos dados de MIR do que de NIR, inclusive para aqueles modelos aprovados estatisticamente por um ou outro método. Esse comportamento pode ser melhor visualizado nas Figuras 7 e 8, que mostram a relação entre os valores medidos e previstos dos modelos selecionados por NIR e MIR, respectivamente.

Comparando-se essas Figuras 7 e 8 observa-se que os modelos criados a partir de dados de MIR estão bem mais ajustados e com menores valores de intervalo de confiança.

5.2.3 Modelos selecionados por iPLS ou siPLS aplicados aos espectros NIR

Conforme citado anteriormente, o siPLS é criado de forma bastante semelhante ao iPLS, podendo ser considerado até mesmo como uma extensão do método. Dessa forma, seu uso só é justificado quando há melhora no desempenho do modelo em comparação ao criado por iPLS. Sendo assim, nesse trabalho serão apresentados apenas os modelos selecionados por um ou outro método, tendo o siPLS sido utilizado apenas quando houve ganho significativo no poder previsor do modelo em relação àquele selecionado por iPLS, para cada propriedade avaliada. Inicialmente serão apresentados os modelos criados a partir de dados de NIR e, posteriormente, serão discutidos aqueles gerados a partir do MIR.

Dessa forma, para a construção dos modelos por iPLS e siPLS também foram testados os pré-processamentos: SNV e derivada antes de centrar na média. O resultado final de seu uso nos modelos selecionados, a partir de dados de NIR, pode ser visualizado na Tabela 5. Pode-se observar nessa Tabela (5) que, contrariamente aos modelos criados por PLS a partir de dados de NIR, para todas as propriedades o melhor modelo preditivo foi obtido aplicando-se algum dos dois tipos de pré-processamento testados, exceto para o índice de refração, para o qual o melhor modelo foi obtido apenas centrando-se os dados na média.

Tabela 5 – Região espectral utilizada para a construção dos modelos por NIR e pré-processamento aplicados aos dados.

Propriedade	Região Espectral (cm^{-1})	Nº de Variáveis	Processamento
Grau API	5212,9 - 5498,3	149	SNV+CM
Índice de Cetano	5600,5 - 5791,4	100	SNV+CM
Índice de Refração	6925,4 - 7492,4	295	CM
Teor de Enxofre	5212,9 - 5787,8	296	D+CM
Ponto de fuligem	6354,6 - 6923,5	296	SNV+CM
Ponto de Anilina	4927,4 - 5496,4	296	SNV+CM
Ponto de Congelamento	8063,3 - 8630,3	295	D + CM
Ponto de Entupimento	4642,0 a 4925,5, 6925,4 a 7208,9, 7210,8 a 7494,3	444	SNV + CM
Ponto de Névoa	5498,3 a 5781,8	148	D + CM
Ponto de Fluidez	5212,9 a 6352,6	592	SNV + CM

SNV = *standard normal variate*; CM = centrado na média; D = suavização de Saviszky-Golay
Fonte: O Autor

Na aplicação dos métodos para seleção da região espectral, foi observado que já com a aplicação do método iPLS houve uma grande melhora no desempenho preditivo dos modelos em relação à aplicação do método PLS com espectro completo, gerando modelos bem ajustados e com resultados satisfatórios. Ao aplicar o método siPLS não houve melhora significativa nos resultados que justificasse a junção de duas regiões distintas, exceto para a previsão do ponto de entupimento, pois a pequena diferença encontrada na previsão das outras propriedades, com a aplicação dessa técnica, poderia ainda comprometer a robustez do modelo.

Dentre os modelos selecionados houve uma grande redução do número de variáveis originais utilizadas na construção de cada um em relação ao modelo com espectro completo (2956 variáveis). A região selecionada para a previsão de cada parâmetro variou bastante não havendo, portanto, uma região específica que se pudesse determinar todas as características previstas das amostras, conforme mostrado na Tabela 5. Entretanto, verifica-se que a região espectral de 3500 a 4400 cm^{-1}, que apresentou maior ruído (Figura 2), não foi utilizada na construção de nenhum modelo.

A partir da Tabela 5 observa-se que a região de previsão do Grau API é caracterizada por ligações entre carbono e oxigênio e oxigênio e hidrogênio o que sugere que essa propriedade possa estar relacionada com a quantidade de grupo ácido presente no meio, tradicionalmente

hidrofílico, e que conferem incremento da polaridade ao meio. Já é de conhecimento prévio que o ponto de fuligem se trata de uma propriedade totalmente relacionada com os heteroátomos o que é confirmado, também, pelas absorções características da região onde o modelo foi criado para a previsão do mesmo. A previsão do teor de enxofre, como já era de se esperar, obteve melhor desempenho em região de ligações entre enxofre e hidrogênio, carbono e hidrogênio e oxigênio e hidrogênio. As predições das propriedades restantes ocorrem em regiões onde prevalecem as ligações entre carbono e hidrogênio, o que vai ao encontro do esperado, visto serem propriedades totalmente relacionadas com a quantidade de parafinas e isoparafinas presentes.

Posteriormente à criação dos modelos de calibração, partiu-se para a previsão de amostras externas, cujos resultados são mostrados na Tabela 6. Observa-se nessa Tabela 6 que os modelos também resultaram em baixos valores de erros de previsão e erros percentuais de previsão representados pelos valores de RMSEP e RMSEP% mostrando, dessa forma, que há boa exatidão para a previsão dos parâmetros em amostras externas, não incluídas na calibração. Foram encontradas, também, respostas aceitáveis para sensibilidade e seletividade de acordo com o teste F aplicado. Os valores de referência e previstos não apresentaram diferença significativa para os modelos analisados. A relação entre valores medidos e previstos para os modelos selecionados é mostrada na Figura 9.

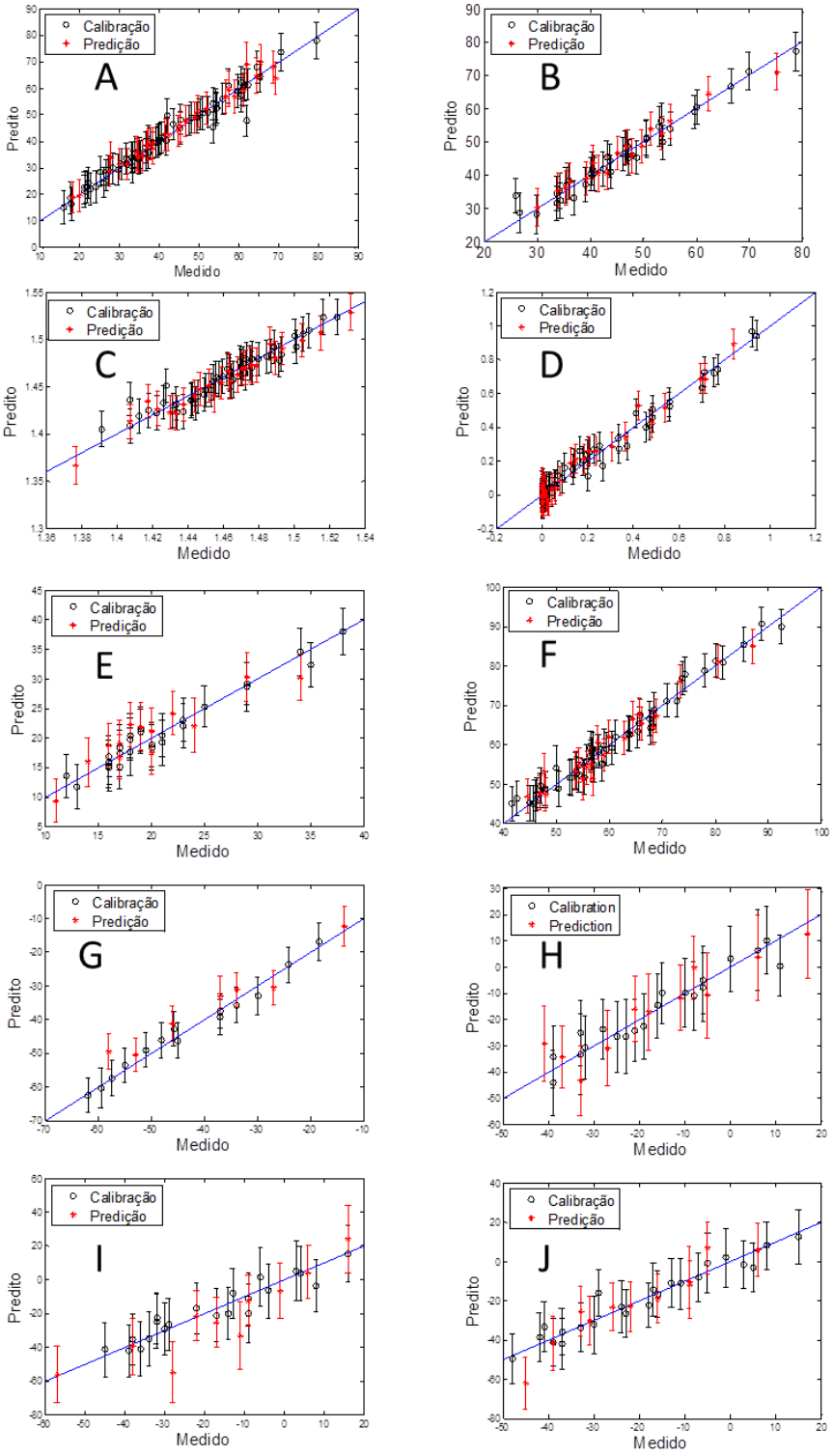
Tabela 6 – Valores calculados para a validação dos modelos selecionados para cada propriedade físico-química, por iPLS.

IR	Parâmetros \ Propriedades	º API	Índice de Cetano	Índice de Refração	Teor de Enxofre	Ponto de fuligem	Ponto de Anilina	Ponto de Congelamento	Ponto de Entupimento	Ponto de Névoa	Ponto de Fluidez
NIR	VL's	6	5	3	9	7	7	4	7	6	7
	RMSEC	2,9	2,4	0,0086	0,0418(a)	1,6(b)	2,04(c)	2,1(c)	5(c)	7(c)	5(c)
	RMSEP	2,3	1,9	0,0072	0,0419 (a)	2,5(b)	2,53(c)	4,5(c)	6(c)	5(c)	7(c)
	RMSEP%	5,2	4,2	0,4941	23,9126(a)	11,9(b)	4,19(c)	-12,6(c)	-38(c)	-30(c)	-32(c)
	R^2 calibração	0,9653	0,9637	0,9238	0,9764	0,9582	0,9753	0,9824	0,9264	0,8766	0,9377
	R^2 previsão	0,9736	0,9703	0,9622	0,9711	0,8391	0,9423	0,9591	0,8743	0,9551	0,8899
	Bias calibração	$-7{,}00E^{-16}$	$1{,}00E^{-14}$	$-3{,}24E^{-17}$	$-2{,}02E^{-17}$(a)	$4{,}0E^{-15}$(b)	$-1{,}79E^{-14}$(c)	$7{,}4E^{-15}$(c)	$-7E^{-15}$(c)	0(c)	-1,00E-15
	Bias previsão	$-6{,}00E^{-01}$	-0,2608	0,0027	-0,007(a)	-0,7(b)	0,17(c)	-3,0(c)	-0,1(c)	0(c)	0(c)
	Sens. Anal.	1,7	2,3	1199,875	49,4888(a^{-1})	3,1(b^{-1})	1,23(c^{-1})	3,3(c^{-1})	2(c^{-1})	1(c^{-1})	1(c^{-1})
	Sens. Anal. $^{-1}$	0,6	0,4	0,0008	0,0202(a)	0,3(b)	0,81(c)	0,3(c)	1(c)	1(c)	0,8(c)
	σ previsão	2,6	10,2	0,0072	0,1263(a)	3,9(b)	2,72(c)	4,5(c)	9(c)	5(c)	8(c)
MIR	VL's	7	3	8	7	6	8	7	3	3	4
	RMSEC	1,2	2,4	0,0018	0,0347(a)	1,2(b)	1,94(c)	1,6 (c)	3(c)	5(c)	4(c)
	RMSEP	2,1	3,3	0,0016	0,0517(a)	1,7(b)	1,71(c)	2,8(c)	4(c)	5(c)	6(c)
	RMSEP%	5,2	7,3	0,1098	30,0377(a)	8,8(b)	2,85(c)	-6,5(c)	-22(c)	-33(c)	-34(c)
	R^2 calibração	0,995	0,956	0,9976	0,9825	0,9744	0,9766	0,9940	0,97	0,9206	0,9505
	R^2 previsão	0,978	0,9346	0,9968	0,9593	0,9128	0,9787	0,9152	0,9631	0,9463	0,8750
	Bias calibração	$4{,}20E^{-15}$	$1{,}02E^{-14}$	$-3{,}08E^{-16}$	$-4{,}81E^{-17}$(a)	$-3{,}14E^{-15}$(b)	$7{,}64E^{-15}$(c)	$-1{,}5E^{-15}$(c)	$-3{,}11E^{-15}$(c)	$4{,}71E^{-15}$(c)	$-6{,}13E^{-14}$(c)
	Bias previsão	-0,1637	0,81	-1,85E-04	0,0033(a)	0,1797(b)	-0,0137(c)	-1,7(c)	2(c)	-2(c)	-0,2188(c)
	Sens. Anal.	2,8	4	889,4622	63,6853(a^{-1})	1,6(b^{-1})	0,94(c^{-1})	1,0(c^{-1})	2(c^{-1})	3(c^{-1})	1(c^{-1})
	Sens. Anal. $^{-1}$	0,4	0,3	0,0011	0,0157(a)	0,6(b)	1,06(c)	0,9(c)	0,5(c)	0,3(c)	0,7(c)
	σ previsão	2,3	10,5	0,0017	0,1298(a)	3,5(b)	1,98(c)	2,9 (c)	8,1(c)	5,6(c)	8(c)

Sens. Anal. = Sensibilidade Analítica; **a** = %m/m, **b** = mm, **c** = (°C).

Fonte: O Autor.

Figura 9 – Relação entre valores medidos e previstos referentes aos modelos criados por iPLS/siPLS a partir de dados de NIR (A – Grau API, B – Índice de Cetano, C – Índice de Refração, D – Teor de Enxofre (%m/m), E – Ponto de Fuligem (mm), F – Ponto de Anilina (°C), G – Ponto de Congelamento (°C), H – Ponto de Entupimento (°C), I – Ponto de Névoa (°C), J – Ponto de Fluidez (°C)).



Fonte: O Autor.

Nenhum modelo, também, apresentou erros sistemáticos ou de tendência nos resíduos de calibração ou de previsão, visto que todos os valores de *p* encontrados foram maiores do que o nível de significância adotado de 0,05. Esse valores, calculados para as amostras de previsão, são mostrados na Tabela 7. Da mesma forma, não houve presença de erros de tendência ou sistemático para o conjunto de calibração em nenhum modelo selecionado.

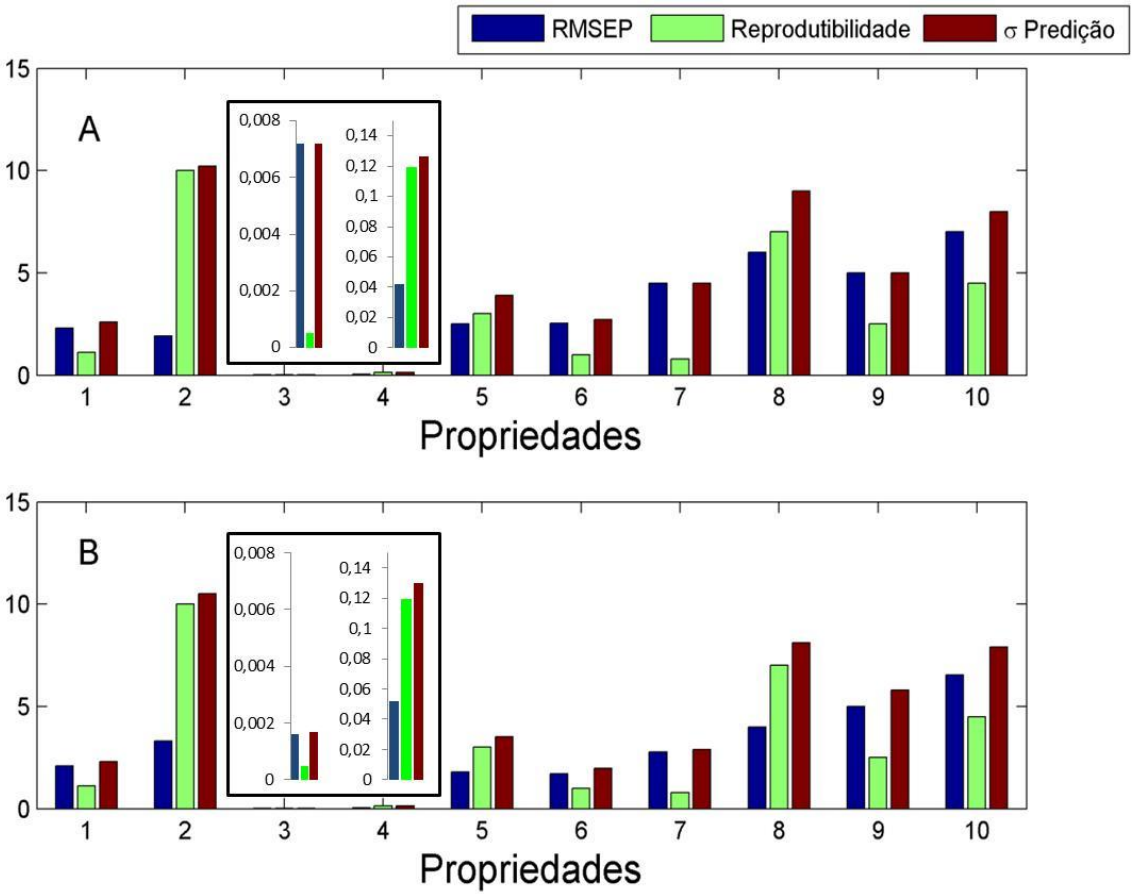
Tabela 7 – Resultados da avaliação de erros sistemáticos e de tendência nos resíduos de previsão dos modelos criados por NIR

Propriedades	Erro Sistemático p-valor	Erro de Tendência p-valor
Grau API	0,15	0,21
Índice de Cetano	0,59	0,13
Índice de Refração	0,06	0,23
Teor de Enxofre	0,35	0,14
Ponto de fuligem	0,35	0,06
Ponto de Anilina	0,76	0,38
Ponto de Congelamento	0,07	0,29
Ponto de Entupimento	0,95	0,07
Ponto de Névoa	0,93	0,08
Ponto de Fluidez	0,83	0,20

Fonte: O Autor.

Comparando-se, ainda, os valores de reprodutibilidade dos respectivos métodos laboratoriais (Tabela 2) para determinação das propriedades estudadas e os valores de RMSEP e desvio padrão de previsão calculados (Tabela 6), observa-se que, para a previsão do índice de cetano e teor de enxofre, principalmente, o erro adicionado à previsão devido ao modelo estatístico é baixo quando comparado à reprodutibilidade do método padrão. Já para a previsão do ponto de fuligem e do ponto de entupimento há uma contribuição proporcional do modelo estatístico e do método laboratorial ao erro de previsão calculado. Entretanto, embora tenha se encontrado modelos estatisticamente aceitáveis para a previsão da grau API, índice de refração, ponto de anilina, ponto de congelamento e ponto de fluidez, os valores de desvio padrão de previsão calculados indicam que a maior parte do erro encontrado é devido ao modelo estatístico. Essa comparação entre os valores de RMSEP, reprodutibilidade e desvio padrão de previsão também pode ser visualizada na Figura 10.

Figura 10 – Relação entre os erros experimentais e calculados para as propriedades (1 – Grau API, 2 – Índice de Cetano, 3 – Índice de Refração, 4 – Teor de Enxofre (%m/m), 5 – Ponto de Fuligem (mm), 6 – Ponto de Anilina (°C), 7 – Ponto de Congelamento (°C), 8 – Ponto de Entupimento (°C), 9 – Ponto de Névoa (°C), 10 – Ponto de Fluidez (°C)) por (A) NIR e (B) MIR.



Fonte: O Autor

Finalmente, comparando-se esses modelos com aqueles gerados por PLS a partir de todo o espectro NIR, verifica-se que há um ganho no poder preditivo para todas as propriedades avaliadas, pois houve uma clara redução dos intervalos de confiança calculados e os mesmos se encontram melhor ajustados à linha de calibração (ainda com alguns desvios, porém em menor número e bem menores do que aqueles apresentados na Figura 7) sendo, portanto, vantajosa a aplicação dos métodos de seleção de variáveis iPLS e/ou siPLS a esse conjunto de dados.

5.2.4 Modelos selecionados por iPLS ou siPLS aplicados aos espectros MIR

Os modelos criados aplicando-se iPLS e siPLS aos espectros MIR foram construídos, também, conforme aqueles gerados a partir dos dados de NIR. Portanto, da mesma forma também serão apresentados apenas os resultados de um ou outro método. Nesse contexto, a Tabela 8 apresenta a região onde foram encontrados os melhores modelos para a previsão de cada propriedade.

Nessa Tabela 8, observa-se que o método siPLS obteve um melhor desempenho para a previsão de 3 propriedades (índice de refração, ponto de fuligem e ponto de anilina). Verifica-se também que, assim como para a previsão por NIR, para todas as propriedades o melhor modelo preditivo foi obtido aplicando-se algum dos dois tipos de pré-processamento testados, exceto para o índice de refração, para o qual o melhor modelo foi obtido apenas centrando-se os dados na média.

A região selecionada para a previsão de cada parâmetro variou bastante não havendo, portanto, uma região específica com a qual se possa determinar todas as características previstas das frações conforme pode ser visto na Tabela 8. Entretanto, verifica-se uma certa prevalência das regiões por volta de 1250 a 1500 e 2800 a 3200 cm^{-1}, nas quais ocorre maior intensidade nas absorbâncias espectrais (Figura 3).

Tabela 8 – Região espectral utilizada para a construção dos modelos por MIR e pré-processamento aplicados aos dados.

Propriedade	Região Espectral (cm^{-1})	Número de Variáveis	Processamento
Grau API	2867,9 a 3149,5	147	SNV + CM
Índice de Cetano	1355,8 a 1542,9	98	SNV + CM
Índice de Refração	1396,3 a 1508,2 e 2875,6 a 2987,5	118	CM
Teor de Enxofre	941,2 a 1280,6	177	D + CM
Ponto de fuligem	1355,8 a 1542,9 e 2678,9 a 2866,0	196	SNV + CM
Ponto de Anilina	713,6 a 825,5 e 1282,5 a 1394,4	118	D + CM
Ponto de Congelamento	1396,3 a 1508,2 e 2875,6 a 2987,5	118	D + CM
Ponto de Entupimento	2867,9 a 3149,5	147	SNV + CM
Ponto de Névoa	2867,9 a 3433,0	294	SNV + CM
Ponto de Fluidez	2875,6 a 3014,5 e 3157,2 a 3296,1	146	D + CM

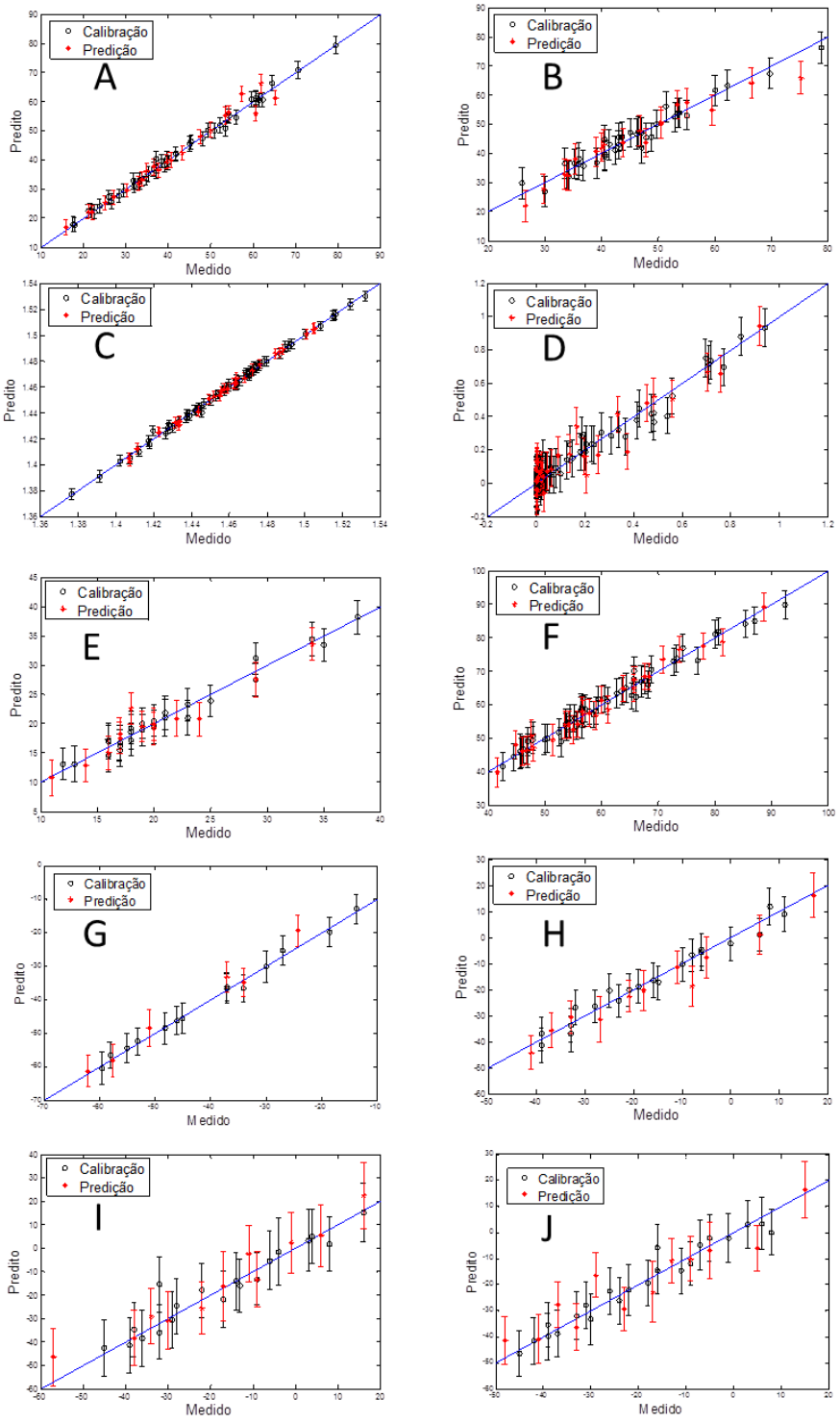
SNV = *standard normal variate*; CM = centrado na média ; D = suavização de Savitzky-Golay
Fonte: O Autor.

Verificando mais minuciosamente a região em que cada modelo foi construído, constata-se que a previsão do ponto de fuligem por MIR, assim como aconteceu em sua previsão por NIR, ocorreu em região indicativa de presença de heteroátomos. A previsão do ponto de anilina, como já era de se esperar, ocorreu na região caracterizado por aromáticos, visto essa propriedade ser indicativa do grau de aromaticidade do composto, como já discutido previamente. A previsão do teor de enxofre ocorreu em região de absorção de ligações entre esse elemento e o oxigênio. A previsão dos pontos de cetano, entupimento, névoa e índice de refração ocorreram em regiões caracterizadas por diferentes ligações entre carbono e hidrogênio, assim como ocorreu em suas predições por NIR. Ainda, o modelo criado para a previsão do grau API foi encontrado em região de ligações entre carbono e hidrogênio de alifáticos, aromáticos e alcenos, o que também pode estar relacionado com a hidrossolubilidade do meio. Finalmente, conforme será discutido mais adiante, embora tenham sido apresentados modelos para a previsão dos pontos de congelamento e de fluidez, por siPLS a partir de espectros MIR, não houve uma região que se destacasse com desempenho previsor melhor do que o modelo criado aplicando-se todo o espectro para esse

fim; sendo assim, não há uma região espectral no MIR que melhor se correlacione com essas propriedades.

A Tabela 6 também apresenta os resultados encontrados para a previsão por MIR onde observa-se que que os modelos também resultaram em baixos valores de erro de previsão e erro percentual de previsão mostrando, também, que há boa exatidão para a previsão dos parâmetros em amostras desconhecidas. Foram encontradas, também, respostas aceitáveis para sensibilidade e seletividade de acordo com o teste F aplicado. Os valores de referência e previstos não apresentaram diferença significativa para os modelos analisados. A relação entre valores medidos e previstos para os modelos selecionados é mostrada na Figura 11.

Figura 11 – Relação entre valores medidos e previstos referentes aos modelos criados por iPLS/siPLS a partir de dados de MIR (A – Grau API, B – Índice de Cetano, C – Índice de Refração, D – Teor de Enxofre (%m/m), E – Ponto de Fuligem (mm), F – Ponto de Anilina (°C), G – Ponto de Congelamento (°C), H – Ponto de Entupimento (°C), I – Ponto de Névoa (°C), J – Ponto de Fluidez (°C)).



Fonte: O Autor.

Quanto ao teste para verificação de tendência nos resíduos, verifica-se que a mesma ocorreu para a previsão do índice de cetano, conforme pode ser visto na Tabela 9, em que apresentou o valor de *p* menor do que 0,05 e mesmo com a variação do número de VL's, não pôde ser eliminado. Nesta Tabela 9 observa-se também que há a presença de erros sistemáticos para a previsão do ponto de entupimento, os quais também não puderam ser eliminados mesmo com a variação do número de VL's. Por um outro lado, não houve presença de erros de tendência ou sistemático para o conjunto de calibração em nenhum modelo selecionado.

Tabela 9 – Resultados da avaliação de erros sistemáticos e de tendência nos resíduos de previsão dos modelos criados por MIR

Propriedades	Erro Sistemático p valor	Erro de Tendência p valor
Grau API	0,73	0,06
Índice de Cetano	0,39	0,003
Índice de Refração	0,56	0,09
Teor de Enxofre	0,60	0,10
Ponto de fuligem	0,73	0,31
Ponto de Anilina	0,96	0,31
Ponto de Congelamento	0,13	0,33
Ponto de Entupimento	0,04	0,24
Ponto de Névoa	0,16	0,05
Ponto de Fluidez	0,91	0,38

Fonte: O Autor.

Através da comparação feita entre os valores de R (Tabela 2), RMSEP e desvio padrão dos valores de previsão calculados (Tabela 6), apresentada na Figura 10, o mesmo comportamento é comparável àquele já discutido para os modelos de previsão a partir de dados de NIR.

Apesar de que todos os modelos, selecionados a partir da aplicação dos métodos iPLS ou siPLS a dados de MIR, tenham apresentado altas capacidades previsoras, com elevados valores de R^2, não houve melhora no desempenho previsor do modelo para a previsão dos pontos de congelamento e fluidez, quando comparados àqueles selecionados por PLS global, o que é facilmente visualizado comparando-se os resultados apresentados nas Tabelas 3 e 6.

Comparando-se, ainda, os resultados apresentados nas Figuras 8 e 11, observa-se que há uma redução no intervalo de confiança calculado, após a aplicação dos métodos iPLS e siPLS, bem como há um maior ajuste à reta dos resultados de previsão apresentados havendo, portanto, um ganho no poder preditivo dessas propriedades após a aplicação desses métodos de seleção de variáveis, exceto para a previsão dos pontos de congelamento e fluidez, conforme já discutido anteriormente.

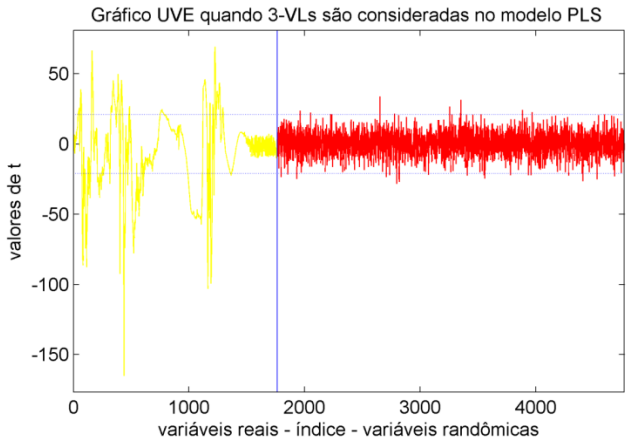
Por um outro lado, avaliando-se ainda os resultados referentes aos modelos selecionados após a aplicação dos métodos iPLS e siPLS a dados de NIR e MIR (Tabela 6), observa-se que das 10 propriedades avaliadas, apenas o índice de cetano, o teor de enxofre e o ponto de entupimento obtiveram um melhor modelo por NIR, quando comparado com o MIR. Vale lembrar que, embora o modelo selecionado para a previsão do ponto de entupimento a partir de dados de MIR, com a aplicação de iPLS ou siPLS, tenha apresentado resultados melhores do que aqueles apresentados por NIR, ocorreu a presença de erro sistemático no mesmo e, portanto, também considera-se que o melhor modelo para a previsão dessa propriedade tenha sido aquele selecionado por NIR. Entretanto, pode se considerar que os modelos criados para a previsão do grau API e do ponto de fluidez, a partir da aplicação dos métodos iPLS e siPLS a dados de NIR e MIR obtiveram desempenhos equivalentes. Nesse contexto, o restante das propriedades com melhor resultados previsores foram obtidos por MIR, quando há a aplicação de iPLS ou siPLS.

Ainda nesse contexto pode se verificar que, embora os modelos criados por iPLS e siPLS selecionados para a previsão dos pontos de congelamento e de fluidez por MIR, tenham gerados erros maiores do que aqueles criados pelo PLS global, esses resultados ainda foram melhores do que aqueles criados por PLS ou iPLS ou siPLS a partir dos dados de NIR.

5.3 MODELOS SELECIONADOS POR UVE-PLS APLICADO A ESPECTROS NIR E MIR

Foram, então, criados os modelos aplicando-se o método de seleção de variáveis UVE e o método de calibração multivariada PLS. Inicialmente, então, foi feita a seleção das variáveis por UVE, separadamente para cada uma das propriedades, por NIR e por MIR. A Figura 12 exemplifica, para o modelo de calibração multivariada, para previsão do grau API, o gráfico do UVE referente à seleção das variáveis por esse método.

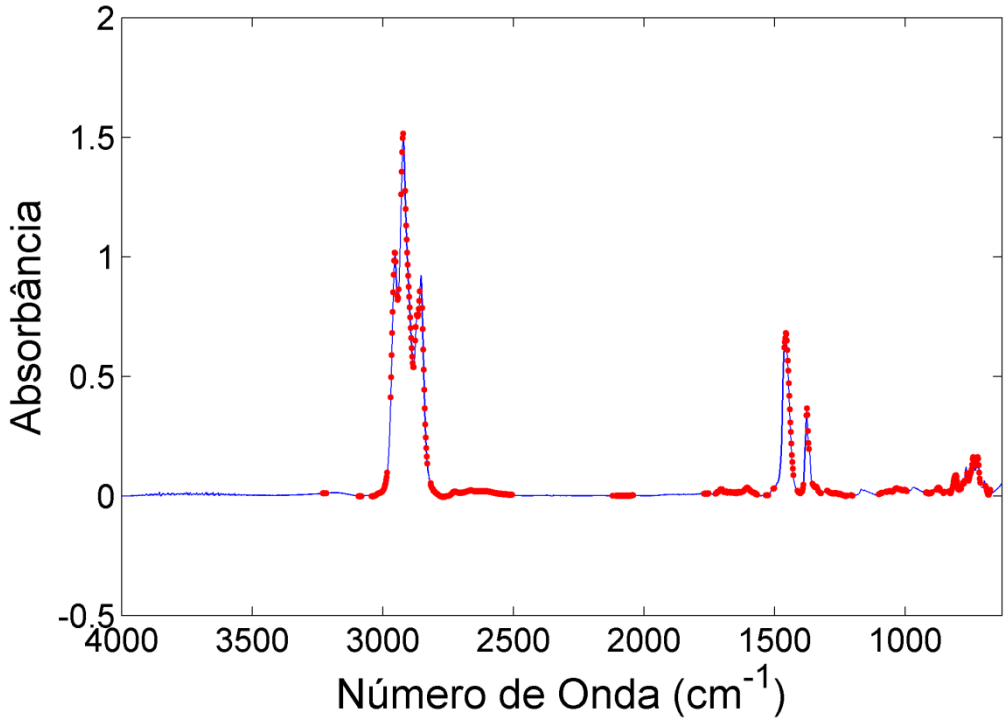
Figura 12 – Gráfico do modelo UVE: valores de *t* para as variáveis experimentais (1-1.764, em amarelo) e randômicas (1.765-3.528, em vermelho). O limite de corte é indicado pela linha azul tracejada.



Fonte: O Autor.

Um novo modelo PLS (UVE-PLS) foi, então, construído a partir das variáveis selecionadas pelo UVE e sua exatidão comparada ao modelo PLS global. Para exemplificar essa fase, a Figura 13 indica as variáveis selecionadas pelo modelo UVE-PLS para a previsão do grau API por MIR.

Figura 13 – Gráfico das variáveis selecionadas pelo modelo UVE-PLS, em vermelho.



Fonte: O Autor.

A Tabela 10 apresenta os modelos criados por UVE-PLS, a partir de dados de NIR e MIR. Vale ressaltar que, conforme apresentado, os modelos criados para a pedição dos pontos de fuligem, congelamento, névoa e fluidez, a partir dos dados de NIR apresentaram valores de R^2 menores do que 0,8, valor previamente estabelecido para aceitação dos modelos e, por isso, os mesmos encontram-se estatisticamente reprovados. Já os modelos criados a partir de dados de MIR estão todos aprovados conforme requisitos previamente estabelecidos.

A partir dessa Tabela 10 observa-se, ainda, que de uma forma geral para todos os modelos, houve um desempenho melhor dos modelos selecionados por MIR, com menores valores de RMSEP e maiores valores de R^2, quando comparados com os respectivos modelos criados a partir de dados de NIR. Isso mostra, portanto, que a técnica de seleção de variáveis UVE obteve um melhor desempenho quando aplicada aos dados de infravermelho médio. A relação entre valores medidos e previstos para as amostras de previsão é mostrada nas Figuras 14 e 15, nas quais pode se observar, também, que os valores de intervalo de confiança referentes aos modelos NIR foram

maiores do que aqueles dos modelos MIR para todas as propriedades avaliadas, especialmente naqueles modelos que foram reprovados estatisticamente. Não havendo, por exemplo, distinção nos valores para a previsão do ponto de entupimento, por NIR, para o qual todos os valores calculados aproximam-se de uma constante.

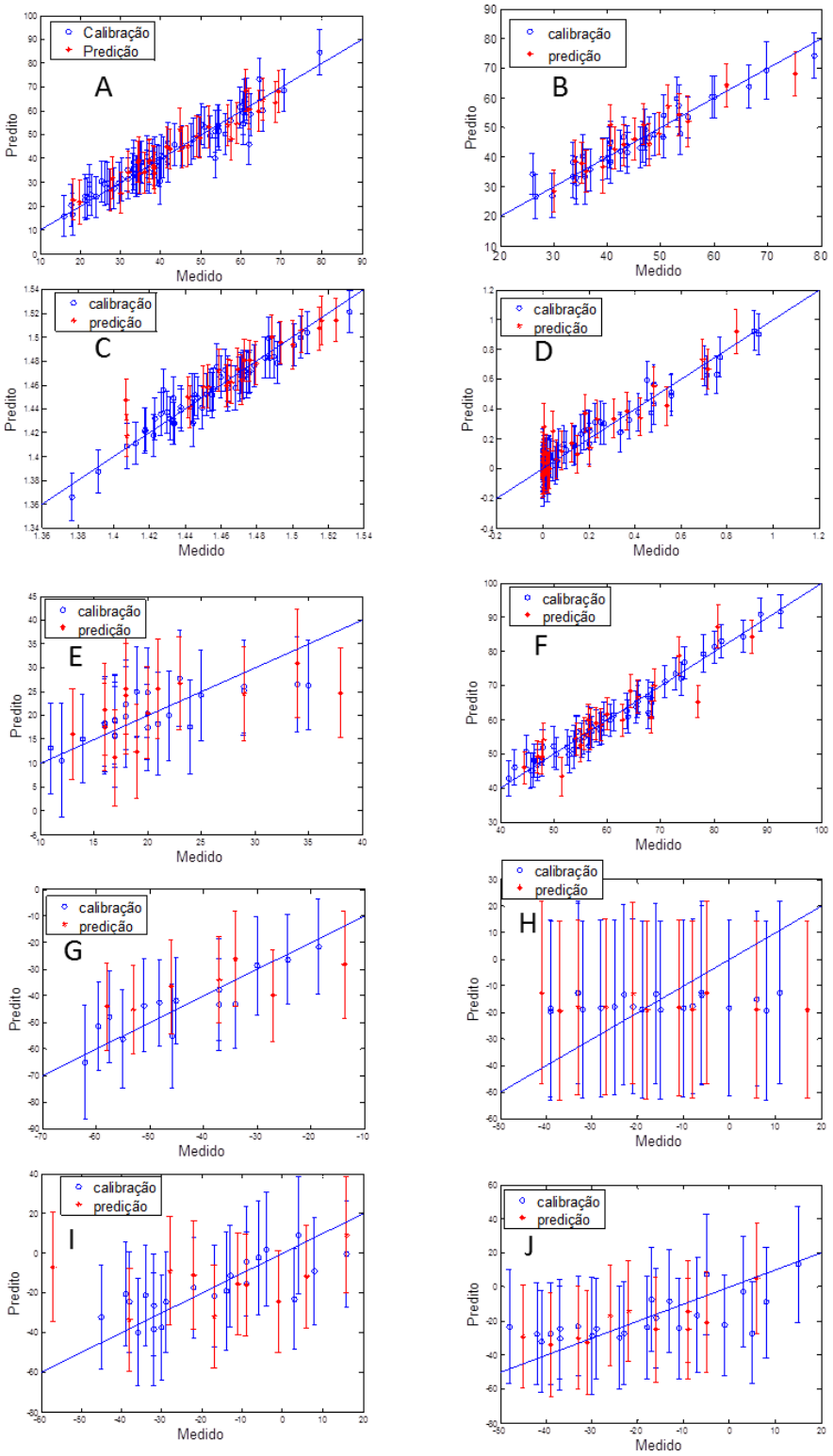
Tabela 10 – Valores calculados para a validação dos modelos selecionados, por UVE, para cada propriedade físico-química.

IR	Parâmetros \ Propriedades	º API	Índice de Cetano	Índice de Refração	Teor de Enxofre	Ponto de fuligem	Ponto de Anilina	Ponto de Congelamento	Ponto de Entupimento	Ponto de Névoa	Ponto de Fluidez
NIR	VL's	4	3	5	8	5	7	4	1	2	3
	RMSEC	4,2	3,4	0,0086	0,0633(a)	4,4(b)	2,26(c)	7,3(c)	16(c)	12(c)	14(c)
	RMSEP	3,7	4	0,0110	0,0872(a)	5,9(b)	4,46(c)	10,6(c)	18(c)	22(c)	10(c)
	RMSEP%	8,3	9	0,7459	45,8889(a)	27,3(b)	7,31(c)	-29,4(c)	-107(c)	-147(c)	22(c)
	R^2 calibração	0,9212	0,9227	0,9306	0,9452	0,577	0,9698	0,796	0,0327	0,575	0,4646
	R^2 previsão	0,9293	0,8684	0,905	0,8837	0,3577	0,8467	0,5238	0,0654	0,0512	0,5845
	Bias calibração	$-7,00E^{-16}$	$9,18E^{-15}$	$4,54E^{-18}$	$2,46E^{-16}$(a)	$-5,60E^{-15}$(b)	$-1,83E^{-14}$(c)	$-3,81E^{-15}$(c)	$-1,78E^{-15}$(c)	$-1,22E^{-14}$(c)	$4,02E^{-14}$(c)
	Bias previsão	-0,02	-0,8152	-0.0036	-0,0218(a)	0,0515(b)	-0,0043(c)	-2,15(c)	0,83(c)	-1,47(c)	0,7040(c)
	Sens. Anal.	1,3	3,3	551,614	69(a^{-1})	1,6(b^{-1})	2,37(c^{-1})	1,2(c^{-1})	4(c^{-1})	0,6(c^{-1})	0,9(c^{-1})
	Sens. Anal. $^{-1}$	0,8	0,3	0,0018	0,01(a)	0,6(b)	0,42(c)	0,9(c)	0,2(c)	1,8(c)	1,2(c)
	σ previsão	3,8	11	0,0110	0,1476(a)	6,6(b)	4,57(c)	10,7(c)	19(c)	22(c)	11(c)
MIR	VL's	3	4	5	4	3	3	3	3	3	3
	RMSEC	1,7	1,7	0,002	0,0407(a)	2,9(b)	2,09(c)	2,3(c)	3(c)	4(c)	5(c)
	RMSEP	2,5	2,4	0,0024	0,0521(a)	2,1(b)	2,35(c)	5,9(c)	8(c)	8(c)	7(c)
	RMSEP%	6,1	5,1	0,1647	28,4197(a)	10,5(b)	3,88(c)	-14,3	-35(c)	-55(c)	-36(c)
	R^2 calibração	0,9887	0,9786	0,9967	0,9747	0,8402	0.9698	0,9803	0,969	0,9553	0,9111
	R^2 previsão	0,9681	0,9686	0,9927	0,9583	0,9044	0.9608	0,8771	0,9279	0,8971	0,8693
	Bias calibração	$7,91E^{-15}$	$7,21E^{-15}$	$-3,19E^{-16}$	$-5,63E^{-17}$(a)	$-2,12E^{-15}$(b)	$-3,35E^{-15}$(c)	$7,36E^{-15}$(c)	$2,84E^{-15}$(c)	0(c)	$4,68E^{-15}$(c)
	Bias previsão	-0,0023	-0,3605	-0,0005	-0.0112(a)	-0.4097(b)	-0.5876(c)	-3,1038(c)	5,7651(c)	-3,9741(c)	-1,0084(c)
	Sens. Anal.	2,3	1,7	$1,06E^{+03}$	56,7684(a^{-1})	0,8(b^{-1})	1,35(c^{-1})	1,5(c^{-1})	1,4(c^{-1})	1,02(c^{-1})	1,4(c^{-1})
	Sens. Anal. $^{-1}$	0,4	0,6	$9,40E^{-04}$	0,0176(a)	1,2(b)	0,74(c)	0,7(c)	0,7(c)	0,98(c)	0,7(c)
	σ previsão	2,7	10,3	0,0025	0,1300	3,7(b)	2,56(c)	6,0(c)	10(c)	8(c)	8(c)

Sens. Anal. = Sensibilidade Analítica; **a** = %m/m, **b** = mm, **c** = (°C).

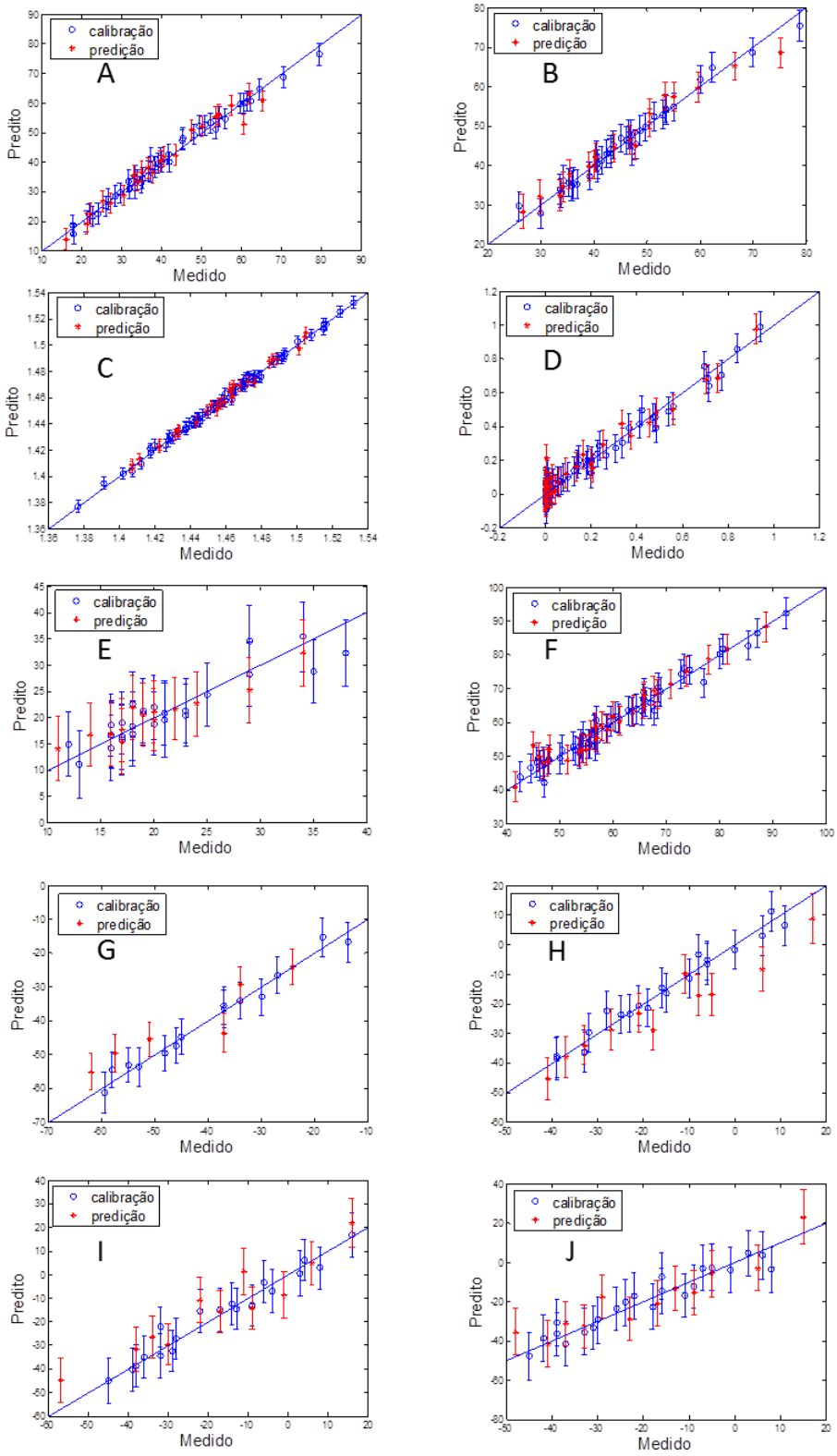
Fonte: O Autor.

Figura 14 – Relação entre valores medidos e previstos referentes aos modelos criados por UVE-PLS a partir de dados de NIR (A – Grau API, B – Índice de Cetano, C – Índice de Refração, D – Teor de Enxofre (%m/m), E – Ponto de Fuligem (mm), F – Ponto de Anilina (°C), G – Ponto de Congelamento (°C), H – Ponto de Entupimento (°C), I – Ponto de Névoa (°C), J – Ponto de Fluidez (°C)).



Fonte: O Autor.

Figura 15 – Relação entre valores medidos e previstos referentes aos modelos criados por UVE-PLS a partir de dados de MIR (A – Grau API, B – Índice de Cetano, C – Índice de Refração, D – Teor de Enxofre (%m/m), E – Ponto de Fuligem (mm), F – Ponto de Anilina (°C), G – Ponto de Congelamento (°C), H – Ponto de Entupimento (°C), I – Ponto de Névoa (°C), J – Ponto de Fluidez (°C)).



Fonte: O Autor.

Apenas para a predição do ponto de entupimento, por MIR, foi encontrado erro sistemático (Tabela 11), o qual não pôde ser eliminado mesmo com a variação do número de VL's. Em nenhum outro modelo selecionado pela aplicação de UVE-PLS para o conjunto de previsão (Tabela 11) ou de calibração.

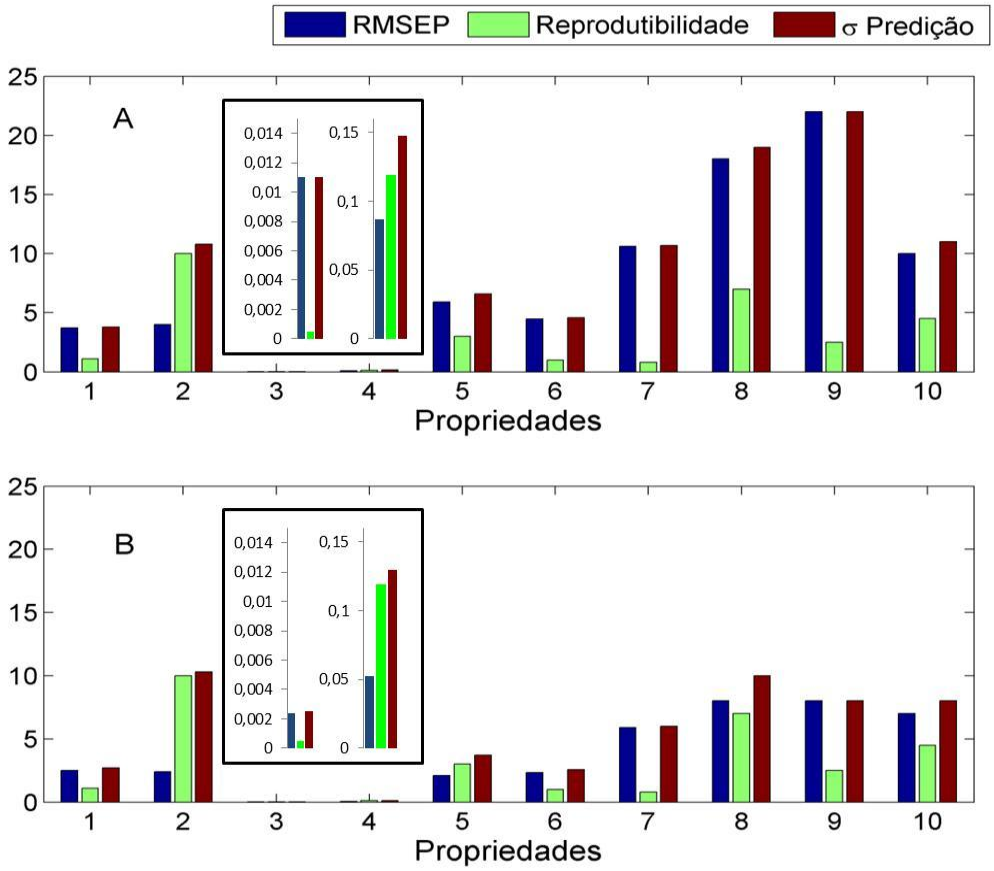
Tabela 11 – Resultados da avaliação de erros sistemáticos e de tendência nos resíduos de previsão dos modelos criados por UVE

Propriedades	NIR		MIR	
	Erro Sistemático	Erro de Tendência	Erro Sistemático	Erro de Tendência
Grau API	0,98	0,32	1,00	0,14
Índice de Cetano	0,41	0,05	0,54	0,08
Índice de Refração	0,13	0,49	0,25	0,13
Teor de Enxofre	0,16	0,05	0,21	0,28
Ponto de fuligem	0,98	0,11	0,51	0,49
Ponto de Anilina	1,00	0,40	0,20	0,27
Ponto de Congelamento	0,63	0,25	0,23	0,26
Ponto de Entupimento	0,89	0,33	0,00	0,38
Ponto de Névoa	0,84	0,37	0,08	0,11
Ponto de Fluidez	0,82	0,29	0,62	0,10

Fonte: O Autor.

Uma comparação entre os valores de reprodutibilidade dos respectivos métodos laboratoriais (Tabela 2) e os valores de RMSEP e desvio padrão de previsão calculados (Tabela 10) é mostrada na Figura 16. Nessa, é possível verificar-se que apenas para a previsão do índice de cetano e do teor de enxofre, o erro adicionado à previsão devido ao modelo estatístico é baixo quando comparado à reprodutibilidade do método padrão, tanto por NIR quanto por MIR. Para a previsão do índice de refração, pelas duas técnicas espectroscópicas, e dos pontos de fuligem e ponto de entupimento, apenas por MIR, verifica-se que existe uma contribuição proporcional do modelo estatístico e do método laboratorial ao erro de previsão calculado. Por um outro lado, para o restante dos modelos observa-se que o erro adicionado à previsão devido ao modelo estatístico é alto quando comparado à reprodutibilidade do método padrão, tanto por NIR quanto por MIR.

Figura 16 – Relação entre os erros experimentais e calculados para as propriedades (1 – Grau API, 2 – Índice de Cetano, 3 – Índice de Refração, 4 – Teor de Enxofre (%m/m), 5 – Ponto de Fuligem (mm), 6 – Ponto de Anilina (°C), 7 – Ponto de Congelamento (°C), 8 – Ponto de Entupimento (°C), 9 – Ponto de Névoa (°C), 10 – Ponto de Fluidez (°C)) por (A) NIR e (B) MIR, a partir da aplicação de UVE-PLS.



Fonte: O Autor.

Vale destacar, ainda, que não houve melhora com relação à aplicação dessa técnica de seleção de variáveis (UVE), quando comparado com os métodos aplicados anteriormente (iPLS e siPLS), em ambas as técnicas espectroscópicas avaliadas , quando se compara os valores de RMSEP, intervalos de confiança e R^2 gerados pelos modelos. Entretanto, em geral os resultados são equivalentes, principalmente para a previsão das seguintes propriedades: grau API, índice de cetano, índice de refração, teor de enxofre, ponto de fuligem e ponto de anilina.

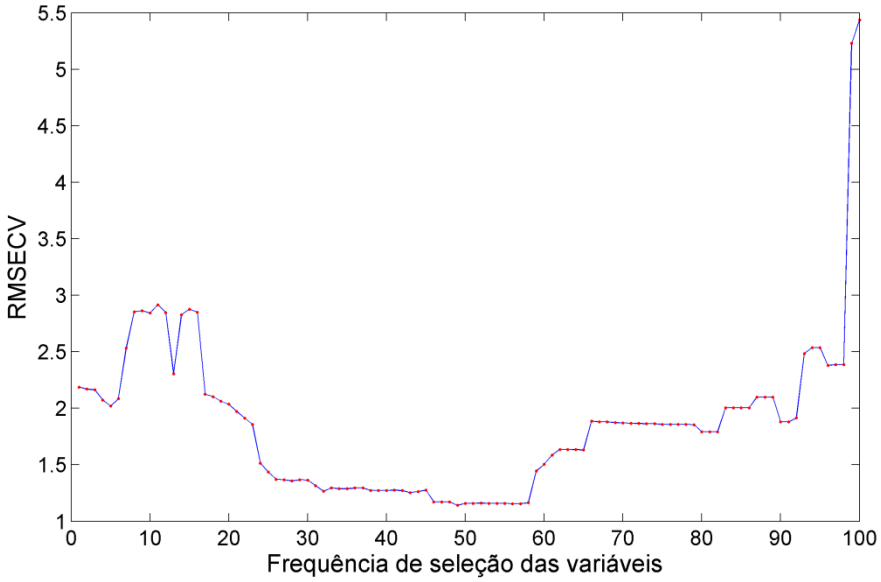
5.4 MODELOS SELECIONADOS POR GA-PLS APLICADO A ESPECTROS NIR E MIR

Conforme discutido anteriormente, na avaliação da resposta, ou seja aptidão, do modelo criado por GA deve-se encontrar o valor associado à eficiência de cada cromossomo relacionado ao sistema de interesse, sendo o resultado mais importante no procedimento do algoritmo genético. A aptidão é uma característica intrínseca ao indivíduo, que representa sua habilidade de produzir a melhor resposta. O objetivo é encontrar o menor erro possível, e este será o responsável direto pela vida ou morte dos indivíduos. Após as gerações, o número de vezes em que as variáveis selecionadas aparecem nos indivíduos mais aptos é representado pela frequência das variáveis incluídas nos modelos.

Sendo assim, a Figura 17 exemplifica para o modelo de calibração multivariada, para previsão do grau API por MIR, o gráfico do GA correlacionando o valor de RMSECV *versus* a frequência do número de variáveis utilizadas para a construção do modelo. Nessa Figura (17) verifica-se que há um valor mínimo com a utilização das variáveis selecionadas em uma frequência de 49% equivalendo, nesse caso a 210 variáveis utilizadas, porém com um patamar entre 46 e 58%.

.

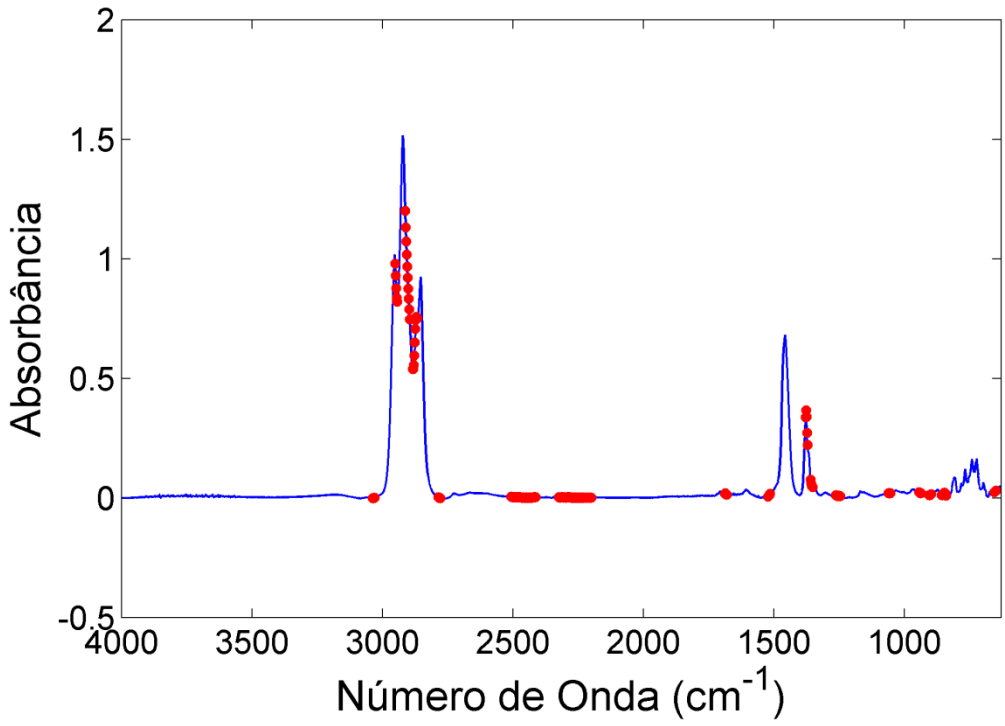
Figura 17 – Gráfico RMSECV versus número de variáveis selecionadas por GA.



Fonte: O Autor.

Um novo modelo PLS (GA-PLS) deve, então, ser construído a partir das variáveis selecionadas pelo GA e sua eficiência comparada ao modelo PLS global. A Figura 18 indica as variáveis selecionadas pelo modelo GA para a previsão do grau API por MIR.

Figura 18 – Gráfico das variáveis selecionadas pelo método GA, em vermelho.



Fonte: O Autor.

Esse procedimento foi realizado na construção dos modelos de todas as propriedades avaliadas. Nesse contexto, a Tabela 12 apresenta os resultados gerados a partir dos modelos selecionados por GA-PLS, a partir de dados de NIR e MIR. Observa-se nessa Tabela que os modelos criados para a previsão dos pontos de congelamento, entupimento, névoa e fluidez, a partir de dados de NIR, estão estatisticamente reprovados, pois apresentaram valores de R^2 menor do que aquele previamente estabelecido (0,8).

Ainda, a partir dessa Tabela (12) verifica-se que houve um desempenho melhor para a previsão do índice de refração e dos pontos de anilina, congelamento, entupimento,

névoa e fluidez por MIR em relação aos resultados encontrados por NIR. Enquanto que para a previsão do grau API, índice de cetano, teor de enxofre e ponto de fuligem houve um desempenho equivalente não havendo, portanto, nenhuma propriedade em que o modelo criado por NIR tenha obtido um melhor desempenho. Vale destacar, ainda, que todos os modelos MIR resultaram em baixos valores de erro de previsão e erro percentual de previsão mostrando, também, que há boa exatidão para a previsão dos parâmetros em amostras desconhecidas.

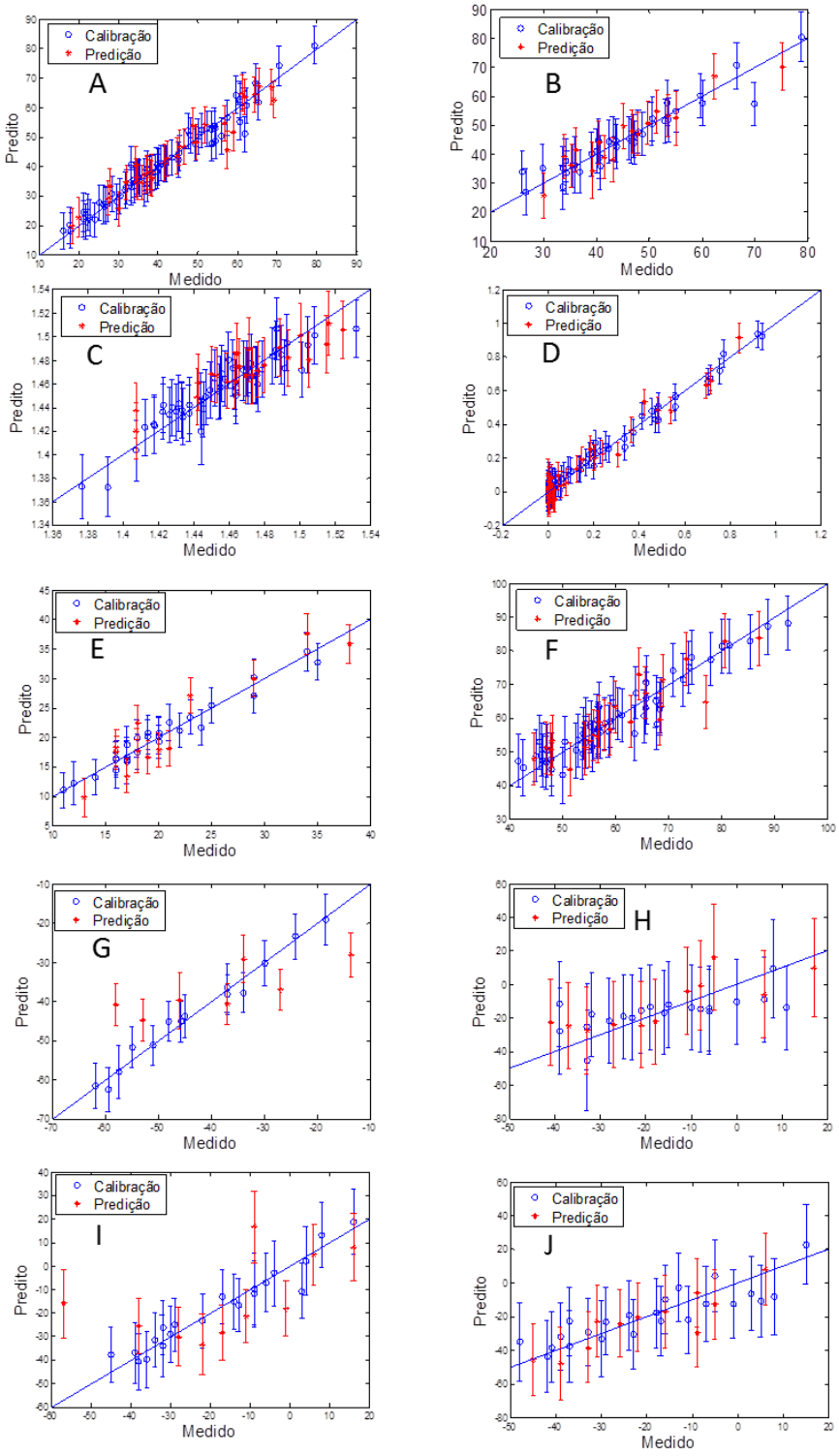
A relação entre valores medidos e previstos para as amostras de previsão dos modelos selecionados por GA-PLS a partir de dados de NIR e MIR é mostrada nas Figuras 19 e 20, respectivamente, nas quais pode se verificar que os intervalos de confiança calculados para os modelos gerados por NIR são, em geral, maiores do que aqueles referentes aos modelos criados a partir de dados de MIR. Entretanto, houve um desvio significativo dos valores calculados em relação à linha de calibração em ambas as técnicas espectroscópicas avaliadas, especialmente para as amostras pertencentes ao conjunto de previsão.

Tabela 12 – Valores calculados para a validação dos modelos selecionados, por GA, para cada propriedade físico-química.

IR	Parâmetros \ Propriedades	º API	Índice de Cetano	Índice de Refração	Teor de Enxofre	Ponto de fuligem	Ponto de Anilina	Ponto de Congelamento	Ponto de Entupimento	Ponto de Névoa	Ponto de Fluidez
NIR	VL's	5	5	3	8	5	3	4	1	6	3
	RMSEC	2,4	1,9	0,0117	0,037(a)	1,4(b)	3,80(c)	2,3(c)	12(c)	5(c)	10(c)
	RMSEP	3,8	2,5	0,0144	0,0527(a)	2,8(b)	4,79(c)	10,3(c)	11(c)	18(c)	8(c)
	RMSEP%	8,5	5,5	0,9769	31,6377(a)	12,9(b)	7,86(c)	-27,7(c)	-93(c)	-126(c)	-34(c)
	R^2 calibração	0,9749	0,9778	0,8657	0,9813	0,959	0,9069	0,9795	0,4231	0,9344	0,7511
	R^2 previsão	0,9262	0,9561	0,841	0,9533	0,8834	0,8181	0,6299	0,6738	0,3053	0,7883
	Bias calibração	$1,10E^{-15}$	$1,90E^{-14}$	$2,61E^{-17}$	$1,80E^{-16}$(a)	$1,3E^{-14}$(b)	$-2,65E^{-14}$(c)	$2,9E^{-14}$(c)	$-1,07E^{-15}$(c)	$4,60E^{-10}$(c)	$-1,60E^{-14}$(c)
	Bias previsão	0,3	0,5	$-3,00E^{-03}$	0,0017(a)	-0,04(b)	0,15(c)	-1,2(c)	-4(c)	18(c)	2(c)
	Sens. Anal.	1,2	0,9	667,1226	82,2708(a^{-1})	3,1(b^{-1})	2,22(c^{-1})	1,8(c^{-1})	0,5(c^{-1})	1,2(c^{-1})	0,4(c^{-1})
	Sens. Anal. $^{-1}$	0,8	1,1	0,0015	0,0122(a)	0,3(b)	0,45(c)	0,6(c)	1,8(c)	0,8(c)	2,3(c)
	σ previsão	3,9	10,3	0,0144	0,1302(a)	4,1(b)	4,89(c)	10,3(c)	13(c)	18(c)	9(c)
MIR	VL's	3	7	5	5	8	5	3	5	4	6
	RMSEC	1,5	0,7	0,002	0,0375(a)	0,7(b)	1,14(c)	1,7(c)	1(c)	2(c)	1(c)
	RMSEP	3,8	2,9	0,0025	0,0587(a)	2,2(b)	2,76(c)	4,7(c)	4(c)	9(c)	5(c)
	RMSEP%	9,3	6,3	0,1716	31,6976(a)	10,8(b)	4,61(c)	-11,1(c)	-22(c)	-62(c)	-30(c)
	R^2 calibração	0,9908	0,9967	0,9967	0,9789	0,9931	0,9914	0,9894	0,9942	0,9854	0,9954
	R^2 previsão	0,9271	0,9452	0,993	0,9492	0,8709	0,9435	0,9263	0,9772	0,8142	0,9264
	Bias calibração	$6,30E^{-15}$	$1,32E^{-15}$	$-3,51E^{-16}$	$-2,03E^{-16}$(a)	$-6,70E^{-15}$(b)	$5,50E^{-15}$(c)	$1,90E^{-14}$(c)	$-7,02E^{-15}$(c)	$-4,33E^{-14}$(c)	$2,9068E^{-15}$(c)
	Bias previsão	-0,4669	-0,1355	-0,0007	-0.0131(a)	-0,0964(b)	0,2318(c)	-2,0081(c)	3,7597(c)	-3,0703(c)	-1,6956(c)
	Sens. Anal.	1,3	1,8	690,5379	64,1958(a^{-1})	2,2(b^{-1})	1,13(c^{-1})	1,3(c^{-1})	1,2(c^{-1})	0,6(c^{-1})	1,02(c^{-1})
	Sens. Anal. $^{-1}$	0,8	0,5	0,0014	0,0156(a)	0,5(b)	0,89(c)	0,8(c)	0,9(c)	1,7(c)	0,98(c)
	σ previsão	4,0	10,4	0,0025	0,1328(a)	3,7(b)	2,93(c)	4,8(c)	8(c)	10(c)	7(c)

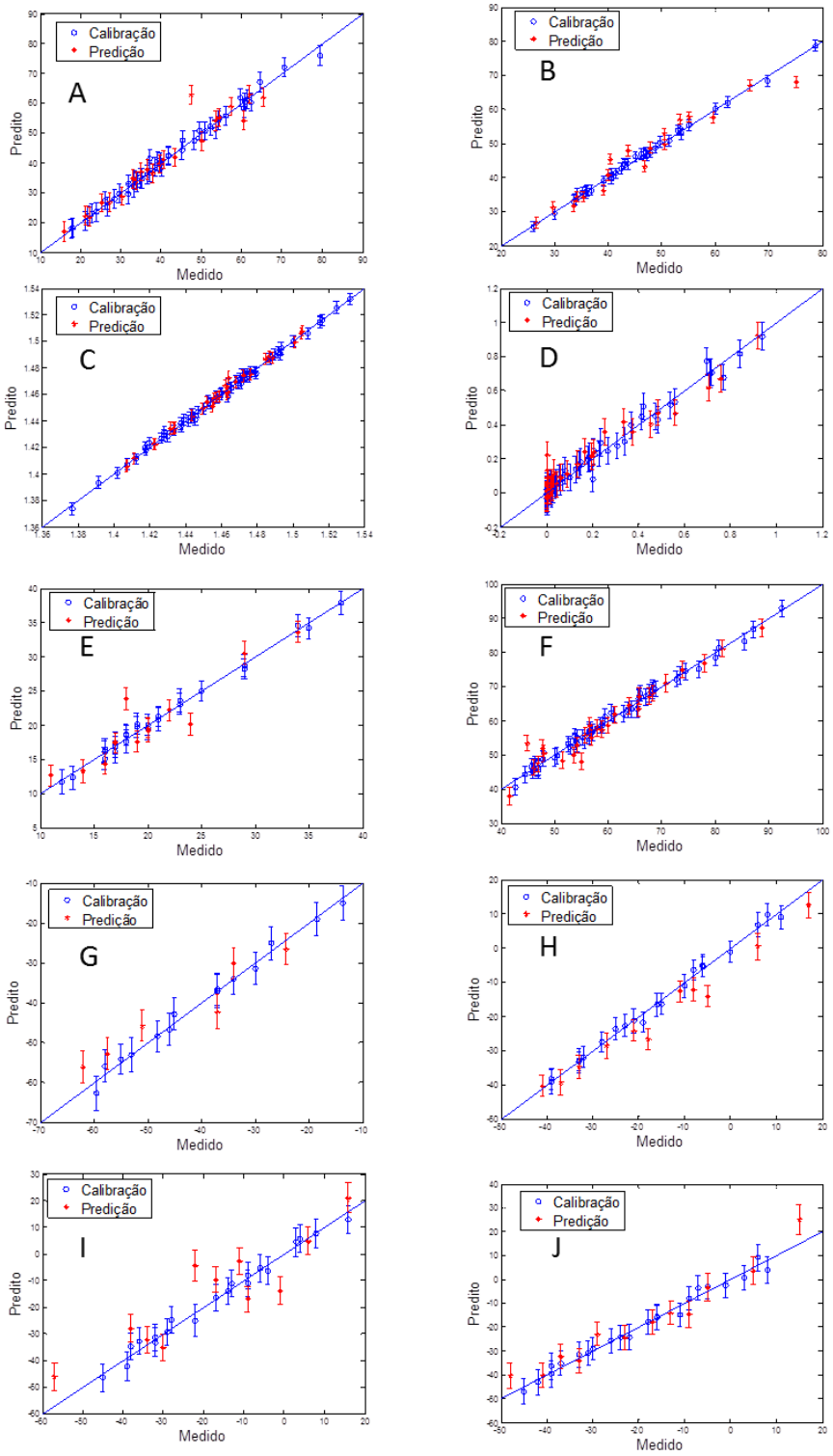
a = %m/m, **b** = mm, **c** = (°C);
Sens. Anal. = Sensibilidade Analítica.
Fonte: O Autor.

Figura 19 – Relação entre valores medidos e previstos referentes aos modelos criados por GA-PLS a partir de dados de NIR (A – Grau API, B – Índice de Cetano, C – Índice de Refração, D – Teor de Enxofre (%m/m), E – Ponto de Fuligem (mm), F – Ponto de Anilina (°C), G – Ponto de Congelamento (°C), H – Ponto de Entupimento (°C), I – Ponto de Névoa (°C), J – Ponto de Fluidez (°C)).



Fonte: O Autor.

Figura 20 – Relação entre valores medidos e previstos referentes aos modelos criados por GA-PLS a partir de dados de MIR (A – Grau API, B – Índice de Cetano, C – Índice de Refração, D – Teor de Enxofre (%m/m), E – Ponto de Fuligem (mm), F – Ponto de Anilina (°C), G – Ponto de Congelamento (°C), H – Ponto de Entupimento (°C), I – Ponto de Névoa (°C), J – Ponto de Fluidez (°C)).



Fonte: O Autor.

Quanto à presença de erros de tendência ou sistemático, os mesmos foram encontrados nos modelos criados, por NIR, para a previsão do índice de cetano e entupimento, respectivamente, no conjunto de previsão (Tabela 13); estando os outros modelos aprovados estatisticamente de acordo com os requisitos previamente estabelecidos. Não houve presença de erros de tendência ou sistemático para o conjunto de calibração em nenhum modelo selecionado.

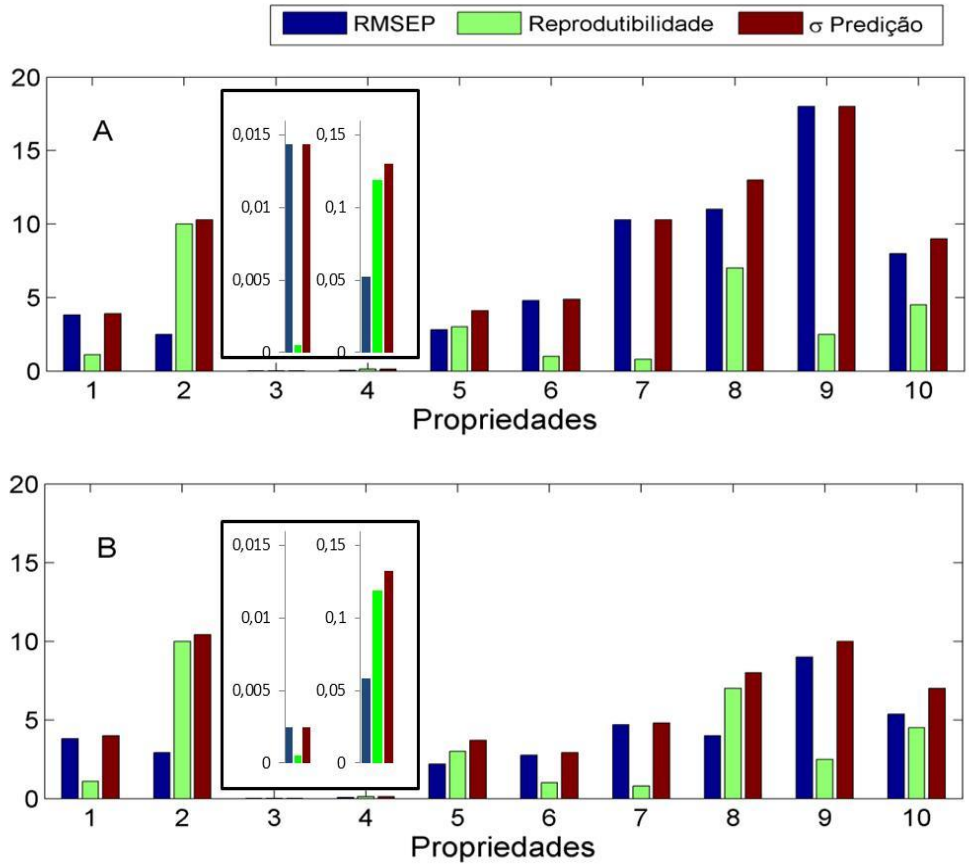
Tabela 13 – Resultados da avaliação de erros sistemáticos e de tendência nos resíduos de previsão dos modelos criados por GA

Propriedades	NIR		MIR	
	Erro Sistemático	Erro de Tendência	Erro Sistemático	Erro de Tendência
Grau API	0,64	0,33	0,58	0,10
Índice de Cetano	0,45	0,01	0,62	0,05
Índice de Refração	0,35	0,48	0,85	0,41
Teor de Enxofre	0,86	0,19	0,20	0,05
Ponto de fuligem	0,96	0,07	0,82	0,45
Ponto de Anilina	0,89	0,47	0,67	0,06
Ponto de Congelamento	0,79	0,42	0,34	0,29
Ponto de Entupimento	0,18	0,09	0,00	0,13
Ponto de Névoa	0,77	0,46	0,29	0,28
Ponto de Fluidez	0,33	0,53	0,29	0,05

Fonte: O Autor.

A Figura 21 apresenta a comparação realizada entre os valores de reprodutibilidade dos respectivos métodos laboratoriais (Tabela 2) e os valores de RMSEP e desvio padrão de previsão calculados (Tabela 12). Através dessa comparação, foi possível observar que, assim como para os outros métodos de seleção de variáveis já avaliados, apenas para a previsão do índice de cetano e do teor de enxofre, o erro adicionado à previsão devido ao modelo estatístico é baixo quando comparado à reprodutibilidade do método padrão, tanto por NIR quanto por MIR. Para a previsão do índice de refração, ponto de fuligem, por ambas técnicas espectroscópicas, e do ponto fluidez, apenas por MIR, verifica-se que existe uma contribuição proporcional do modelo estatístico e do método laboratorial ao erro de previsão calculado. Já para o restante dos modelos observa-se que o erro adicionado à previsão devido ao modelo estatístico é alto quando comparado à reprodutibilidade do método padrão, tanto por NIR quanto por MIR.

Figura 21 – Relação entre os erros experimentais e calculados para as propriedades (1 – Grau API, 2 – Índice de Cetano, 3 – Índice de Refração, 4 – Teor de Enxofre (%m/m), 5 – Ponto de Fuligem (mm), 6 – Ponto de Anilina (°C), 7 – Ponto de Congelamento (°C), 8 – Ponto de Entupimento (°C), 9 – Ponto de Névoa (°C), 10 – Ponto de Fluidez (°C)) por (A) NIR e (B) MIR, a partir da aplicação de GA-PLS.



Fonte: O Autor.

Ainda, quando é realizada a comparação dos melhores modelos criados por GA-PLS, para cada propriedade, em relação aos melhores criados por UVE-PLS, verifica-se que há um desempenho comparável em ambos, tendo apenas a previsão do grau API por UVE-PLS se sobressaído em relação ao GA, enquanto a previsão dos pontos de entupimento, congelamento, névoa e fluidez apresentaram maior exatidão por GA do que por UVE-PLS.

Por outro lado, quando os modelos criados por GA-PLS são comparados com aqueles apresentados por PLS global, iPLS, siPLS, verifica-se que apenas os modelos criados para a previsão do grau API e pontos de congelamento e névoa obtiveram um resultado melhor por uma dessas últimas técnicas. Sendo todos os outros modelos com

resultados equiparáveis. Entretanto, houve um maior desvio dos intervalos de confiança calculados, em relação ao centro dos dados, dos modelos criados a partir do GA-PLS.

5.5 COMPARAÇÃO ENTRE OS MODELOS SELECIONADOS PELOS DIFERENTES MÉTODOS DE SELEÇÃO DE VARIÁVEIS APLICADOS

Por fim, após a apresentação e discussão dos modelos selecionados pelos diferentes métodos de seleção de variáveis aplicados é possível se apresentar uma comparação de seus desempenhos previsores (Figura 22).

A partir dessa comparação observa-se que, conforme já citado anteriormente, há um desempenho melhor dos modelos criados a partir dos dados de MIR do que de NIR quando não há aplicação de nenhum método de seleção de variáveis.

Quando se aplica os diversos métodos de seleção de variáveis aos espectros NIR, há uma diminuição no valor de RMSEP (e aumento de R^2) em todos os modelos selecionados por iPLS ou siPLS. Entretanto, embora para a previsão de muitas propriedades por UVE-PLS ou GA-PLS também ocorra essa redução, esse valor aumenta quando o método UVE-PLS é aplicado para a previsão do grau API, do índice de refração e dos pontos de fuligem, anilina, névoa e fluidez. Da mesma forma, esse mesmo comportamento é observado quando o método GA-PLS é aplicado para a previsão do grau API, teor de enxofre e pontos de anilina, névoa e fluidez. Sendo assim, ao se comparar a eficiência dos métodos de seleção de variáveis aplicados a dados de NIR, nota-se uma melhor performance dos métodos iPLS e siPLS para a previsão das propriedades estudadas.

Já quando se aplica as técnicas de seleção de variáveis aos espectros MIR, observa-se uma redução no valor de RMSEP (e aumento de R^2) quando os métodos iPLS e siPLS são utilizados para a previsão de quase todas as propriedades, exceto na previsão dos pontos de congelamento e de fluidez. Esse aumento, na previsão dessas duas propriedades, também ocorreu quando foi aplicado o método UVE-PLS na predição de ambas as propriedades. Por um outro lado, embora o método GA-PLS também não tenha gerado redução no valor de RMSEP para a predição do ponto de congelamento,

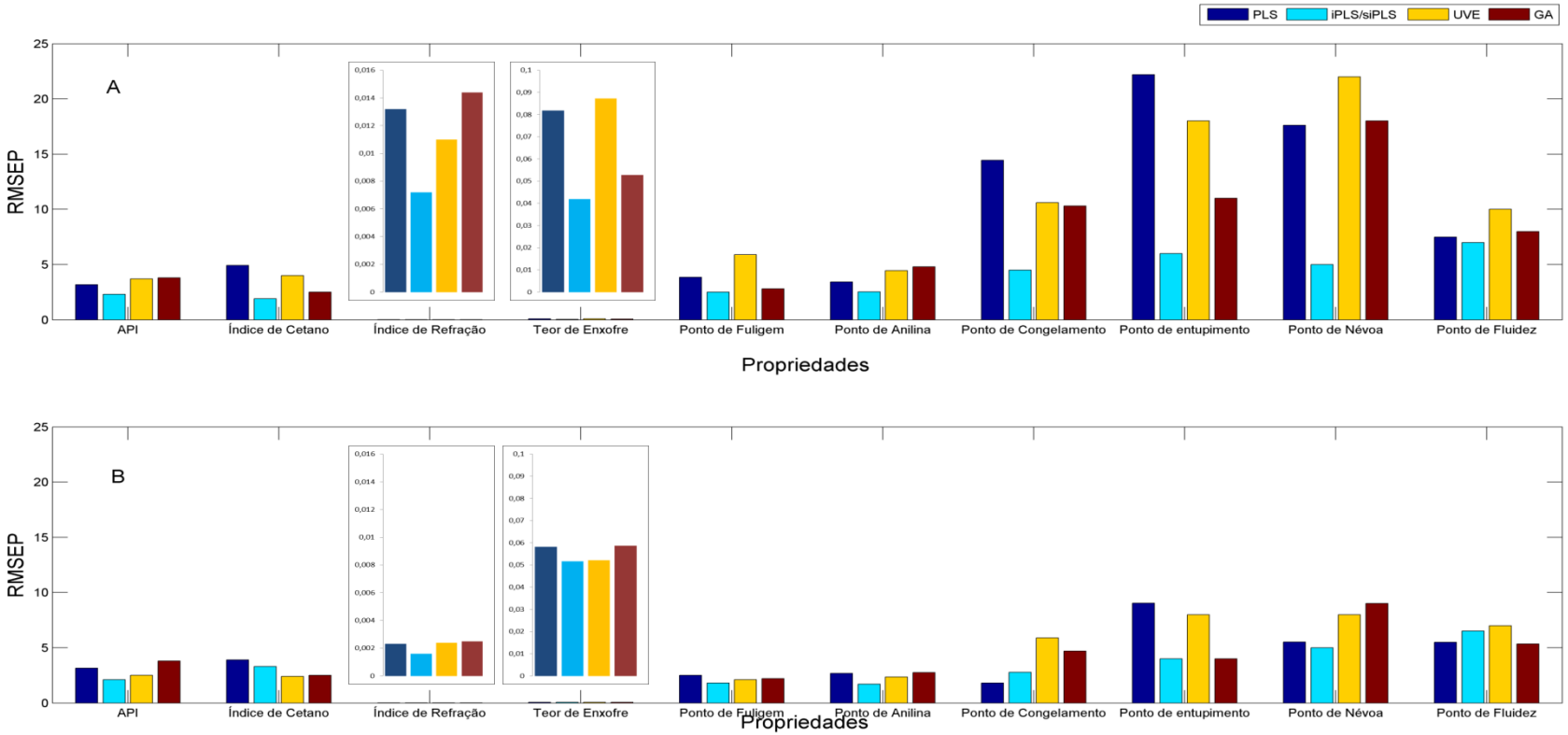
por MIR ou NIR, essa técnica teve um desempenho melhor quando aplicada para a predição do ponto de fluidez. Nesse contexto, o único método que melhorou a predição do ponto de fluidez, comparado àquele criado sem seleção de variáveis, foi o GA aplicado ao MIR. Ainda nesse contexto, houve também uma piora no poder preditivo do índice de refração e do ponto de névoa quando os métodos UVE-PLS e GA-PLS foram aplicados a dados de MIR. Para o restante das propriedades, também ocorreu melhora na performance dos modelos selecionados.

Dessa forma, ao se comparar a eficiência dos métodos de seleção de variáveis aplicados a dados de MIR nota-se, igualmente como para a previsão a partir de dados de NIR, uma melhor performance dos métodos iPLS e siPLS para a maior parte das propriedades estudadas, exceto para a previsão dos pontos de congelamento e ponto de fluidez para os quais, conforme discutido anteriormente, não houve melhora preditiva do modelo quando esses métodos foram aplicados, comparados com os modelos PLS global. Entretanto, pode se dizer que, para esse conjunto de dados, o desempenho dos 4 métodos (iPLS, siPLS, UVE e GA) foi semelhante em muitos casos, com algumas particularidades, pois o fato de um método ser escolhido para a previsão de determinada propriedade não implica que outros também não estejam adequados.

Vale destacar, ainda, que assim como o desempenho dos modelos criados a partir de dados de MIR foi melhor do que aqueles criados a partir do NIR, pela aplicação do PLS global, isso também ocorreu quando os modelos criados a partir de todos os métodos de seleção de variáveis aplicados a NIR e MIR são comparados, exceto quanto à determinação do índice de cetano, teor de enxofre e ponto de entupimento que, conforme mencionado anteriormente, obtiveram melhores resultados a partir da aplicação dos métodos iPLS e siPLS aos dados de NIR.

Quando se compara o tempo necessário para o computador executar os algoritmos testados, os métodos GA e siPLS são os mais demorados, quando comparados ao tempo necessário para a aplicação dos métodos iPLS e UVE.

Figura 22 – Comparação entre os valores de RMSEP dos modelos selecionados pelos diferentes métodos de seleção de variáveis por (A) NIR e (B) MIR



Fonte: O Autor.

5.6 ESCOLHA DO MELHOR MODELO PARA CADA PROPRIEDADE

Após toda a discussão e análise feita dos modelos apresentados, finalmente é possível selecionar a melhor técnica quimiométrica e espectroscópica para a determinação de cada propriedade físico-química no conjunto de amostras avaliadas. Entretanto, embora tenha se escolhido uma técnica para a determinação de cada propriedade, existem outros modelos apresentados, com resultados semelhantes, que também podem ser utilizados sem que haja algum prejuízo significante no desempenho de previsão.

Nesse contexto, na Tabela 14 são apresentadas as técnicas utilizadas nos modelos que geraram os melhores resultados para a previsão de cada propriedade estudada. Essa Tabela também mostra os resultados medidos e previstos para 10 amostras diferentes, assim como a faixa de destilação de cada uma delas. Vale destacar que como os conjuntos de calibração e previsão eram diferentes para cada propriedade, essas amostras foram selecionadas de acordo com aquelas que mais frequentemente se encontravam no conjunto de previsão dos modelos criados para cada propriedade. Foram escolhidas, também, amostras com diferentes temperaturas de corte, de forma a englobar quase toda a faixa de destilação trabalhada (15 a 500 ºC).

Quanto aos tipos de pré-processamento aplicados aos modelos apresentados na Tabela 14, aqueles selecionados para a predição do grau API, do índice de cetano e dos pontos de fuligem, entupimento e névoa, foram criados aplicando-se a SNV aos dados. Por outro lado, modelos criados após a aplicação da 1ª. derivada como recurso de pré-processamento foram selecionados para a predição do teor de enxofre e dos pontos de anilina e congelamento. Vale ressaltar que, assim como esperado, o melhor modelo para a predição do índice de refração foi criado apenas centrando-se os dados na média. Isso ocorre justamente por essa propriedade se tratar de uma propriedade física, relacionada ao espalhamento da luz e, portanto, quando um método de pré-tratamento é aplicado, esse efeito é minimizado (ou eliminado) dificultando, assim, a predição dessa propriedade.

Tabela 14 – Modelos selecionados e previsão de algumas amostras externas.

Propriedade			Grau API	Índice de Cetano	Índice de Refração	Teor de Enxofre (a)	Ponto de fuligem (b)	Ponto de Anilina (c)	Ponto de Congelamento (c)	Ponto de Entupimento (c)	Ponto de Névoa (c)	Ponto de Fluidez (c)
Técnica espectroscópica			MIR	NIR	MIR	NIR	MIR	MIR	MIR	NIR	MIR	MIR
Técnica quimiométrica			iPLS	iPLS	siPLS	iPLS	siPLS	siPLS	PLS	siPLS	iPLS	GA
Amostra	Faixa de corte (°C)											
A	15 a 154	medido	65,3	-	-	0,00093	-	-	-	-	-	-
		predito	61,3	-	-	0,01330	-	-	-	-	-	-
B	77 a 98	medido	61,9	-	-	1,407	-	-	-	-	-	-
		predito	66,2	-	-	1,405	-	-	-	-	-	-
C	196 a 221	medido	-	33,9	1,4621	0,185	-	48,03	-	-	-	-
		predito	-	35,3	1,4629	0,214	-	47,16	-	-	-	-
D	204 a 230	medido	-	-	1,4557	-	22	-	-51	-	-	-
		predito	-	-	1,4566	-	21	-	-52	-	-	-
E	250 a 281	medido	40,6	-	-	-	24	73,82	-	-18	-17	-17
		predito	40,1	-	-	-	21	76,60	-	-17	-16	-17
F	257 a 271	medido	-	46,8	-	0,235	14	59,95	-24,2	-	-	-33
		predito	-	48,5	-	0,259	13	61,28	-23,6	-	-	-37
G	273 a 296	medido	33,1	49,7	1,4761	0,203	-	-	-	-21	-22	-23
		predito	32,1	48,7	1,4775	0,259	-	-	-	-16	-26	-23
H	331 a 349	medido	36,6	-	-	0,0416	-	88,71	-	17	16	15
		predito	37,6	-	-	0,0372	-	89,17	-	13	23	25
I	340 a 400	medido	25,2	-	1,501	-	-	68,38	-	-	-	5
		predito	25,3	-	1,501	-	-	68,37	-	-	-	3
J	400 a 437	medido	22,2	-	-	0,422	-	-	-	-	-	-
		predito	21,9	-	-	0,528	-	-	-	-	-	-

a = %m/m, **b** = mm, **c** = (°C)

Fonte: O Autor.

6. CONCLUSÃO

A maior parte dos modelos desenvolvidos obteve um resultado satisfatório, demonstrando que tanto os espectros NIR quanto MIR, combinados com métodos quimiométricos, podem ser usados para determinar propriedades em frações de petróleo. Vale destacar que as propriedades químicas com influência direta no perfil espectral, tais como grau API, índice de cetano, índice de refração, teor de enxofre, ponto de fuligem e ponto de anilina geraram modelos mais robustos e com melhor capacidade preditora, quando comparados aos modelos selecionados para a predição das propriedades restantes.

De modo geral, comparando-se as duas técnicas espectroscópicas testadas, houve uma melhor performance daqueles criados a partir de MIR, entretanto os modelos desenvolvidos por NIR não podem ser descartados, pois forneceram resultados aceitáveis e na prática laboratorial, principalmente devido à possibilidade do uso de sondas ópticas, podem também ser bastante úteis.

A maior parte dos modelos, tanto por NIR quanto por MIR, tiveram sua habilidade previsora melhorada quando os diversos métodos de seleção de variáveis foram aplicados, havendo uma superioridade preditiva quanto à aplicação dos métodos iPLS e siPLS, em relação aos métodos UVE e GA. Nesse contexto, conclui-se que os métodos de seleção de variáveis contínuos são mais adequados para o conjunto de dados trabalhado. Isso se deve principalmente ao tipo de espectroscopia utilizada, pois espectros NIR e MIR em geral, fornecem informações por faixas e não por pontos.

Vale destacar que, embora apenas para a previsão do índice de cetano e do teor de enxofre, o erro adicionado à previsão devido ao modelo estatístico é baixo quando comparado à reprodutibilidade do método padrão, tanto por NIR quanto por MIR em todos os métodos quimiométricos avaliados, o erro final dos melhores modelos selecionados para previsão para as propriedades restantes pode ser viável para aplicação prática. Isso ocorre pois métodos muito automatizados em geral possuem alta precisão e exatidão, mas para a indústria do petróleo uma rápida estimativa do parâmetro com menor exatidão, na maioria das vezes, é aceitável.

REFERÊNCIAS

ABRAHAMSSON, C.; JOHANSSON, J.; SPARÉN, A.; LINDGREN, F. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, **Chemom. Intell. Lab. Syst.**, v. 69, n. 1-2, 3-12, 2003.

ANDERSSON, M. A. comparison of nine PLS1 algorithms. **Journal of Chemometrics**, v. 23, n. 10, 518-522, 2009.

ANDRADE, G. H. **Estudo da espectroscopia na região do infravermelho médio e próximo para previsão das propriedades do petróleo e emulsão de petróleo do tipo água em óleo**. Dissertação de Mestrado. Departamento de Química, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2009.

ANP. **Agência Nacional de Petróleo, Gás Natural e Biocombustíveis**. Disponível em: <http://www.anp.gov.br> Acesso em: 13 nov. 2013.

ARAUJO, T. P. **Emprego de espectroscopia no infravermelho e métodos quimiométricos para a análise direta de tetracilinas em leite bovino**. Dissertação de Mestrado. Instituto de Química, Universidade Estadual de Campinas (Unicamp), Campinas, 2007.

ASTM International. **ASTM D1218 – 12. Standard test method for refractive index and refractive dispersion of hydrocarbon liquids**. West Conshohocken, Pennsylvania, USA, 2012.

ASTM International. **ASTM D1322 – 15. Standard test method for smoke point of kerosin and aviation turbine fuel**. West Conshohocken, Pennsylvania, USA,2015.

ASTM International. **ASTM D2892. Standard Test Method for Distillation of Crude Petroleum (15-Theoretical Plate Column)**. West Conshohocken, Pennsylvania, USA, 2011.

ASTM International. **ASTM D4294 – 10. Standard test method for sulfur in petroleum and petroleum products by energy dispersive X-ray fluorescence spectrometry**. West Conshohocken, Pennsylvania, USA,2010.

ASTM International. **ASTM D4737 – 10. Standard test method for calculated cetane index by four variable equation**. West Conshohocken, Pennsylvania, USA,2010.

ASTM International. **ASTM D5236**. **Standard Test Method for Distillation of Heavy Hydrocarbon Mixtures (Vacuum Potstill Method)**. West Conshohocken, Pennsylvania, USA, 2007.

ASTM International. **ASTM D5453 – 12. Standard test method for determination of total sulfur in light hydrocarbons, spark ignition engine fuel, diesel engine fuel, and engine oil by ultraviolet fluorescence**. West Conshohocken, Pennsylvania, USA, 2012.

ASTM International. **ASTM D5773 – 15. Standard test methods for cloud point of petroleum products**. West Conshohocken, Pennsylvania, USA,2010.

ASTM International. **ASTM D5950 – 14. Standard test methods for pour point of petroleum products**. West Conshohocken, Pennsylvania, USA, 2012.

ASTM International. **ASTM D5972 – 05. Standard test methods for freezing point of aviation fuels**. West Conshohocken, Pennsylvania, USA,2010.

ASTM International. **ASTM D611 – 12. Standard test methods for aniline point and mixed aniline point of petroleum products and hydrocarbon solvents**. West Conshohocken, Pennsylvania, USA,2012.

ASTM International. **ASTM D6371 – 05. Standard test methods for cold filter polugging point of diesel and heating fuels**. West Conshohocken, Pennsylvania, USA,2010.

ASTM International. **ASTM E1655-12. Standards, standards practices for infrared, multivariate, quantitative analysis**. vol.03.06, West Conshohocken, Pennsylvania, USA, 2012.

BARNES R. J.; DHANOA M. S.; LISTER S. J.; Standard Normal Variate transformation and De-trending of Near-infrared Diffuse Reflectance Spectra. **Appl. Spectrosc.**, v. 43, n. 5, 772-777, 1989.

BEEBE, K. R.; KOWALSKI, B. R.; An Introduction to Multivariate Calibration and Analysis. **Anal. Chem.**, v.59, n.17, 1007A-1017A, 1987.

BREITKREITZ, M. C.; RAIMUNDO, I. M.; ROHWEDDER, J. J. R.; PASQUINI, C.; FILHO, H. A. D.; JOSÉ, G. E.; ARAÚJO, M. C. U. Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration. **The Analyst,** v. 128, 1204-1207, 2003.

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

BRERETON, R. G.; **Chemometrics:** Data Analysis for the laboratory and Chemical Plant. Chichester: John Wiley & Sons Ltda, 2003, 489 p.

CENTNER, V.; MASSART, D. Elimination of uniformative variables for multivariate calibration, **Anal. Chem.**, v. 68, n. 21, 3851-3858, 1996.

Chung, H.; KU, M. Comparison of Near-Infrared, Infrared, and Raman Spectroscopy for the analysis of heavy petroleum products. **Applied Spectroscopy**, v. 54, n. 2, 239-245, 2000.

COSTA FILHO, P.A.; POPPI, R.J.; Algoritmo genético em química, **Quím. Nova**, v. 22, n. 3, 405-411, 1999.

FALLA, F.S.; LARINIA, C.; LE, ROUX GAC; QUINA, F.H.; MORO, L.F.L.; NASCIMENTO, C.A.O. Characterization of crude petroleum by NIR. **J. Pet. Sci. Eng.**, v. 51, 127-37, 2006.

FEARN, T.; RICCIOLI, C.; GARRIDO-VARO, A.; GUERRERO-GINEL, J. E. On the geometry of SNV and MSC. **Chemometrics and Intelligent Laboratory Systems**, v. 96, n. 1, 22–26, 2009.

FEI, Q.; LI, M.; WANG, B.; HUAN, Y.; FENG, G.; REN, Y.; Analysis of cefalexin with NIR spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection. **Chemometrics and Intelligent Laboratory Systems**, v. 97, n. 2, 127–131, 2009.

FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L.O. Quimiometria I: calibração multivariada, um tutorial. **Química. Nova**, v. 22, n. 5, 724-731, 1999.

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**, v. 116, 123-130, 2014.

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. **Chemome. Intel. Lab. Syst.**, v. 133, 33-41, 2014.

FILGUEIRAS, P. R. **Determinação da composição de blends de petróleo utilizando FTIR-ATR e calibração multivariada.** Dissertação de mestrado. Departamento de Química, Universidade Federal do Espírito Santo (UFES), Vitória, 2011.

GELADI P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Anal. Chim. Acta**, v. 185, 1-17, 1986.

GOLDBERG, D. E.; **Genetic Algorithms in search, optimization and machine learning**, Reading, Addison-Wesley; 1989.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Multivariate Data Analysis**, 5 ed., Londres: Prentice-Hall, 1998.

HANNISDAL, A.; HEMMINGSEN, P.V.; SJÖBLOM, J. Group-type analysis of heavy crude oils using vibration spectroscopy in combinations with multivariate analysis**. Ind. Eng. Chem. Res.,** v. 44, 1349-1357, 2005.

ISO 12185:1996, **Crude petroleum and petroleum products – determination of density – oscillating U-tube method, International Organization for standardization**. Conshohocken, PA, USA: American Society for testing and Materials; 2008.

KALLEVIK, H.; KVALHEIM, O.M.; SJÖBLOM, J. Quantitative determination of asphaltenes and resins in solution by means of near-infrared spectroscopy. Correlations to emulstion stability. **J. Colloid. Interface Sci.**, v. 225, 494-504, 2000.

KHANMOHAMMADI, M.; GARMARUDI, A.B.; GUARDIA, M. Characterization of petroleum-based products by infrared spectroscopy and chemometrics. **TrAC Trends Anal Chem,** v. 35, 135-49, 2012.

KUPTSOV, A.K.H.; ARBUZOVA, T.V. A study of heavy oil fractions by Fourier-transform near-infrared raman spectroscopy. **Petroleum Chemistry**, v. 51, 203-211, 2011.

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**, v. 705, n. 1-2, 227-234, 2011.

LEARDI, R.; Application of genetic algorithm–PLS for feature selection in spectral data sets. **Journal of Chemometrics**, v. 14, n. 5-6, 643-655, 2000.

LEARDI, R.; NØRGAARD, L.; Sequential application of backward inerval partial least squares and genetic algorithms for the selection of relevant spectral regions. **J. Chemom.**, v. 18, n. 11, 486-497, 2004.

LUCASIUS, C. B.; BECKERS, M. L. M.; KATEMAN, G.; Genetic algorithms in wavelength selection: a comparative study. **Analytica Chimica Acta**, v. 286, n. 2, 135-153, 1994.

LYONS, W. C.; PLISGA, G. J. **Standard Handbook of Petroleum & Natural Gas Engeneering**. 2 ed. Amsterdam: Elsevier; 2005, 1822 p.

MAGALHÃES, J. C. D. **Estudo exploratório das propriedades de caracterização de petróleos brasileiros.** Dissertação de Mestrado. Departamento de Química Analítica, Instituto de Química, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2005.

MORRIS, R. E.; HAMMOND, M. H.; CRAMER, J. A.; JOHNSON, K. J.; GIORDANO, B. C.; KRAMER, K. E..; ROSE-PEHRSSON, S. L. Rapid fuel quality surveillance through chemometric modeling of near-infrared spectra. **Energy & Fuels**, v. 23, 1610-1618, 2009.

MUNCK, L.; NIELSEN, J. P.; MØLLER, B.; JACOBSEN, S.; SØNDERGAARD, I.; ENGELSEN. S. B.; NØRGAARD, L.; BRO, R. Exploring the phenotipyc expression of a regulatory proteome-altering gene by sepectroscopy and chemometrics. **Anal Chim Acta**, v. 446, n. 1–2, 169–184, 2001.

NETO, B. B.; SCARMÍNIO, I. S.; BRUNS, R. E. 25 anos de Quimiometria no Brasil. **Química Nova**, v. 29, n. 6, 1401-1406, 2006.

NETO**,** W. B.; **Parâmetros de qualidade de lubrificantes e óleo de oliva através de espectroscopia vibracional, calibração multivariada e seleção de variáveis**. Tese de Doutorado. Instituto de Química, Universidade Estadual de Campinas (Unicamp), Campinas, 2005.

NIAZI, A.; LEARDI, R.; Genetic algorithms in chemometrics. **Journal of Chemometrics**, v. 26, n. 6, 345-351, 2012.

NØRGAARD, L.; SAUDLAND, A.; WAGNER, J.; NIELSEN, J. P.; MUNCK, L; ENGELSEN, S.B. Interval partial least-square regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. **Appl. Spectrosc.**, v.54, n.3, 413-419, 2000.

OLIVEIRA, F. C. C.; SOUZA, A. T. P. C.; DIAS, J. A.; DIAS, S. C. L.; RUBIM, J. C. A escolha da faixa espectral no uso combinado de métodos espectroscópicos e quimiométricos. **Química Nova**, v. 27, n. 2, 218-225, 2004.

OLIVEIRA, J. S. **Avaliação da qualidade de biodiesel por espectroscopia FTIR e FTNIR associadas à quimiometria**. Dissertação de Mestrado. Instituto de Química, Universidade de Brasília (UnB), Distrito Federal, 2007.

OLIVIERI A.C.; FABER N.K.M.; FERRÉ J.; BOQUÉ R.; KALIVAS J.H.; MARK H. Uncertainty estimation and figures of merit for multivariate calibration. **Pure Appl. Chem.**, v. 78, n. 3, 633-661, 2006.

PARISOTTO, G.; **Determinação do número de acidez total em resíduos de destilação atmosférica e de vácuo do petróleo empregando a espectroscopia no infravermelho (ATR-FTIR) e calibração multivariada.** Dissertação de Mestrado. Departamento de Química, Universidade Federal de Santa Maria (UFSM), Santa Maria, 2007.

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

PAVIA D. L.; LAMPMAN G. M.; KRIZ G. S.; VYVYAN J. R. **Introduction to spectroscopy**. 4 ed. Washington: Brooks/Cole, 2009. 656 p.

PEINDER, P. **Characterization and classification of crude oils using a combination of spectroscopy and chemometrics.** Tese de doutorado. Utrecht University, Holanda, 2009.

RIAZI, M. R. **Characterization and properties of petroleum fracions**. 1 ed. Baltimore: American Society for Testing and Materials (ASTM), 2005.

ROCHA, W. F. C.; NOGUEIRA, R.; VAZ, B. C. Validation of model of multivariate calibration: an application to the determination of biodiesel blend levels in diesel by near-infrared spectroscopy. **Journal of Chemometrics**, v. 26, 456-461, 2012.

SATYA, S.; ROEHNER, R. M.; DEO, M. D.; HANSON, F. V. Estimation of properties of crude oil residual fractions using chemometrics. **Energy & Fuels**, v. 21, n. 2, 998−1005, 2007.

SAVITZKY, A.; GOLAY M.J.E. Smoothing and differentiation of data by simplified least squares procedures. **Anal. Chem.**, v. 36, n. 8, 1627 – 1639, 1964.

SEKULIC, S.; SEASHOLTZ, M. B.; WANG, Z.; KOWALSKI, B. R. Nonlinear multivariate calibration methods in analytical chemistry. **Anal. Chem**., v. 65, n. 19, A835-A845, 1993.

SILVERSTEIN, R. M.; WEBSTER, F. X. **Spectrometric Identification of Organic Compounds**. 6 ed. New York: Wiley, 1998.

SIMANZHENKOV, V.; IDEM R. **Crude oil chemistry**. New York: Marcel Dekker, InC, 2003, 402 p.

SKOOG, D A.; LEARY, J. J. **Principles of Instrumental Analysis.** 4 ed., Orlando: Saunders College Publishing, 1992, 700p.

SOARES, I.P.; REZENDE, T.F.; SILVA, R.C.; CASTRO, E.V.R.; FORTES, I.C.P. Multivariate Calibration by Variable Selection for Blends of Raw Soybean

Oil/Biodiesel from Different Sources Using Fourier Transform Infrared Spectroscopy (FTIR) Spectra Data. **Energy & Fuels**, v. 22, 2079-2083, 2008.

SOROL, N.; ARANCIBIA, E.; BORTOLATO, S. A.; OLIVIERI, A. C. Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice A test field for variable selection methods. **Chemometrics and Intelligent Laboratory Systems**, v. 102, 100-109, 2010.

SOYEMI, O.O.; BUSCH, M.A.; BUSCH, K.W. Multivariate analysis of near-infrared spectra using the G-programming language. **J. Chem. Inf. Comput. Sci.**, v. 40, 1093-100, 2000.

SPEIGHT, J. G. **Handbook of Petroleum Product analysis**. New Jersey: Wiley-Interscience. 2002

TERRA, L. A.; FILGUEIRAS, P. R.; TOSE, L. V.; ROMÃO, W.;DE SOUZA, D. D., DE CASTRO, E. V. R.; DE OLIVEIRA, L. M. L.; DIAS, J. C. M.; POPPI, R. J; Petroleomics by electrospray ionization FT-ICR mass spectrometry coupled to partial least squares with variable selection methods: prediction of the total acid number of crude oils. **Analyst**, v. 139, n.19, 4908-4916, 2014.

VALDERRAMA P.; BRAGA J.W.; POPPI R.J. Variable selection, outlier detection, and figures of merit estimation in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy. **J. Agric. Food Chem.**, v. 55, n. 21, 8331-8338, 2007.

VALDERRAMA, P.; BRAGA, J. W. B.; POPPI, R. J.; Estado da arte de figuras de mérito em calibração multivariada. **Química Nova**, v. 32, 1278-1287, 2009.

VANDEGINSTE, B. G. M.; MASSART, D. L.; BUYDENS, L. M. C.; JING, S.; LEWI, P. J.; SMEYERS-VERBEKE, J. **Handbook of Chemometrics and Qualimetrics: Part B**. Amsterdan: Elsevier, 1998.

WAGNER, J. **The graphical iPLS toolbox for MATLAB**, v. 2.1, 2000. Disponível em : < http://www.models.life.ku.dk/ipls>.

WISE, B. M.; GALLAGHER, N.B.; BRO, R.; SHAVER, J. M.; WINDIG, W.; KOCH, R. S. **PLS Toolbox Version 4.0 for Use with Matlab**. Wenatchee: Eigenvector Research Inc., 2006.

XIN, N.; GU, X.; WU, H.; HU, Y.; YANG, Z. Application of Genetic Algorithm-Support Vector Regression (GA-SVR) for quantitative analysis of herbal medicines. **Journal of Chemometrics**, v. 26, n. 7, 353-360, 2012.

ZENI, D. **Determinação de cloridrato de propranolol em medicamentos por espectroscopia no infravermelho com calibração multivariada (PLS)**. Dissertação de Mestrado. Departamento de Química, Universidade Federal de Santa Maria (UFSM), Santa Maria, 2005.

ZHANG, Z-M; CHEN, S.; LIANG, Y-Z.; Baseline Correction using adaptative iteratively reweighted penalized least squares. **Analyst**, v.135,1138-1146, 2010.

ZUPAN, J.; GASTEIGER, J.**; Neural Networks for Chemistry: an introduction**. Weinheim: VCH, 1993.

APÊNDICE 1 – TRABALHO PUBLICADO

energy&fuels

Article

pubs.acs.org/EF

Sulfur Determination in Brazilian Petroleum Fractions by Mid-infrared and Near-infrared Spectroscopy and Partial Least Squares Associated with Variable Selection Methods

Julia T. C. Rocha,*[,†,‡] Lize M. S. L. Oliveira,[§] Julio C. M. Dias,[§] Ulysses B. Pinto,[§] Maria de Lourdes S. P. Marques,[§] Betina P. Oliveira,[‡] Paulo R. Filgueiras,[‡] Eustáquio V. R. Castro,[‡] and Marcone A. L. de Oliveira[†]

[†]Grupo de Química Analítica e Quimiometria, Department of Chemistry, Federal University of Juiz de Fora, 36036-900 Juiz de Fora, Minas Gerais, Brazil

[‡]Laboratory of Research and Development of Methodologies for the Analysis of Oils, Department of Chemistry, Federal University of Espírito Santo, Avenida Fernando Ferrari, 514, Goiabeiras, 29075-910 Vitória, Espírito Santo, Brazil

[§]Centro de Pesquisas Leopoldo Américo Miguez de Mello, Petrobras, Avenida Horacio Macedo 950, University City, Rio de Janeiro 21941-598, Brazil

APÊNDICE 2 – TRABALHOS PARALELOS

Modern Research in Catalysis, 2013, 2, 63-67
http://dx.doi.org/10.4236/mrc.2013.23010 Published Online July 2013 (http://www.scirp.org/journal/mrc)

Scientific Research

Effect of $Nb_2O_5 \cdot nH_2O$ Termal Treatment on the Esterification of a Fatty Acid

Deborah A. dos Santos, Valdemar Lacerda Jr., Júlia Tristão do C. Rocha, Reginaldo B. dos Santos, Sandro J. Greco, Alvaro C. Neto, Renzo C. Silva, Eustáquio V. R. de Castro
Chemistry Department, Espírito Santo Federal University, Vitória, Brazil
Email: vljuniorqui@gmail.com

Received February 7, 2013; revised March 20, 2013; accepted April 23, 2013

APÊNDICE 3 – PARTICIPAÇÃO EM EVENTOS

1. V Encontro Capixaba de Química (ENCAQUI), 2015.
    a. Apresentação oral / pôster: Aplicação de métodos quimiométricos, associados às espectroscopias MIR e NIR, para a previsão do teor de enxofre em frações de petróleo.
    b. Pôster: Determinação de enxofre total em petróleo bruto por espectroscopias no infravermelho médio e quimiometria.
    c. Pôster: Determinação de similaridade entre petróleos brasileiros por espectroscopia NIR e análise por componentes principais.

17º Encontro Nacional de Química Analítica (ENQA), 2013.

    d. Apresentação oral / pôster: Determinação de Parâmetros físico-químicos em petróleos a partir do estudo de correlação entre suas propriedades.
    e. Pôster: Estudo de propriedades de petróleo utilizando quimiometria associada a infravermelho próximo.

2. Escola de Inverno de Quimiometria, 2013.

3. 3º Encontro Petrobras e Universidades de Novas Tecnologias para Avaliação de Petróleos, 2013
    a. Pôster: Determinação de Parâmetros físico-químicos de petróleo a partir do estudo de correlações entre suas propriedades.
    b. Pôster: Estudo de algumas Propriedades de Petróleos Utilizando Quimiometria Associada a Infravermelho Próximo (NIR).

4. IV Encontro Capixaba de Química, 2013.
    a. Pôster: Determinação de nitrogênio básico e resíduo de carbono em petróleos a partir de estudo de correlações entre suas propriedades. 2013. (Encontro).
    b. Pôster: Teor de Nitrogênio Total e Número de Acidez Total de Petróleos Utilizando Quimiometria Associada a Infravermelho Próximo (NIR).