UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS GRADUAÇÃO EM QUÍMICA

DANÚBIO LEONARDO BERNARDINO DE OLIVEIRA

COMPARAÇÃO ENTRE ALGORITMOS DE SELEÇÃO DE VARIÁVEIS EM REGRESSÃO PLS PARA DETERMINAÇÃO DE PARÂMETROS DE QUALIDADE DE BIODIESEL POR ESPECTROMETRIA NIR

CAMPINA GRANDE – PB
2021

DANÚBIO LEONARDO BERNARDINO DE OLIVEIRA

COMPARAÇÃO ENTRE ALGORITMOS DE SELEÇÃO DE VARIÁVEIS EM REGRESSÃO PLS PARA DETERMINAÇÃO DE PARÂMETROS DE QUALIDADE DE BIODIESEL POR ESPECTROMETRIA NIR

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Estadual da Paraíba, como requisito obrigatório à obtenção de título de Mestre em Química.

Área de Concentração: Química Analítica

Orientador: Prof. Dr. JOSÉ GERMANO VÉRAS NETO

CAMPINA GRANDE – PB

2021

O48c Oliveira, Danúbio Leonardo Bernardino de.

Comparação entre algoritmos de seleção de variáveis em Regressão PLS para determinação de parâmetros de qualidade de biodiesel por espectrometria NIR [manuscrito] / Danubio Leonardo Bernardino de Oliveira. - 2021.

59 p. : il. colorido.

Digitado.

Dissertação (Mestrado em Química - Mestrado) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2021.

"Orientação : Prof. Dr. José Germano Véras Neto , Departamento de Química - CCT."

1. Algoritmo Firefly. 2. Seleção por intervalos. 3. Calibração multivariada. 4. Quimiometria. I. Título

21. ed. CDD 543

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

DANÚBIO LEONARDO BERNARDINO DE OLIVEIRA

COMPARAÇÃO ENTRE ALGORITMOS DE SELEÇÃO DE VARIÁVEIS EM REGRESSÃO PLS PARA DETERMINAÇÃO DE PARÂMETROS DE QUALIDADE DE BIODIESEL POR ESPECTROMETRIA NIR

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Estadual da Paraíba, como requisito obrigatório à obtenção de título de Mestre em Química.

Área de Concentração: Química Analítica

Aprovada em 15 de fevereiro de 2021.

BANCA EXAMINADORA

Prof. Dr. José Germano Véras Neto (Orientador)
Universidade Estadual da Paraíba (UEPB)

Profa. Dra. Simone da Silva Simões (Examinadora interna)
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Neirivaldo Cavalcante da Silva (Examinador Externo)
Universidade Federal do Pará (UFPA)

AGRADECIMENTOS

Ao término deste trabalho e conclusão deste tão esperado curso, vibro com uma força oriunda da alma, uma nova e importante conquista em minha vida. Por isso, não anseio em agradecer as pessoas que estiveram comigo, lado a lado, rumo a essa vitória:

A Deus, amado Pai Celestial, Ser e Poder supremo e sobrenatural, pelo dom da vida e a capacidade intelectual para poder concluir este trabalho.

A uma das grandes conquistas de minha vida, meus amigos, colegas de universidade e profissão, pessoas que tracei, desde o início do curso duras trilhas, como também pelos bons momentos que ficaram marcados em nossa história.

Aos Companheiros de trabalho, e amigos na vida, os quais incentivaram massivamente para o término desse trabalho.

E como não poderia faltar, os amigos de jornada, dos quais estar neste passo dependeu de incessáveis decisões tomadas quando jovens.

À Banca examinadora, desde a qualificação, Profa. Dra. Sara Regina e Prof. Dr. Rodrigo Oliveira, bem como os da dissertação, Simone Simões e Neirivaldo Cavalcante pelas correções, sugestões e apontamentos, mas não apenas isto: palavras de força e incentivo para o término do trabalho mesmo no período pandêmico da COVID-19.

Ao LQAQ pelo desenvolvimento da pesquisa, e oportunizar o engrandecimento como profissional, desde o campo profissional ao humano. Hilthon, Rômulo, Odilon, Mariana, Victor, Wedja e Lucas que passaram de pesquisadores a espectadores durante meu estágio docência.

Ao PPGQ-UEPB por tornar acessível aos alunos do interior da Paraíba ensino gratuito e de qualidade em nível de pós-graduação em Química, fazendo um papel de extrema importância na interiorização do ensino. Ao secretário do curso, David, que sempre estava disponível a resolver os trâmites administrativos.

Ao IFPB e a Pró-reitoria de pesquisa, inovação e pós-graduação pelo incentivo financeiro, através do Programa de Incentivo à Qualificação do Servidor (PIQIFPB), Edital 07/2020-PRPIPG-IFPB. “As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do(s) autor(es) e não necessariamente refletem a visão do IFPB”.

Em especial ao Orientador, Germano Véras, que permitiu o aprofundamento no campo de atuação que sigo hoje, desde a época de graduação, e que foi um mentor durante a jornada de mestrado. Quando eu pensei não ter mais saída, estava ele para dar uma segunda chance, de fundamental importância, e instigar o pesquisador em mim que estava em pleno de desenvolvimento.

A minha mamãe, Diva, pelo exemplo de força e esperança, pela dedicação e amor, por ter, até antes de me gerado, já traçado com desejo o caminho no qual eu deveria tomar para o adiante, me munindo com o escudo da educação.

Ao meu pai, Oliveira (*in memorian*), meu herói de infância no qual me espelhei por muito tempo, ele me ensinou do seu jeito único a ter caráter, atitude ou sentido diferencial na vida de uma pessoa. E que mesmo nos seus últimos momentos, mostrou-me o amor fraternal. Estarás sempre no meu pensamento e ao meu lado.

E por fim, mas de suprema importância, A minha amada, Camila Lima do Nascimento, companheira de todas as horas e ponto norte da minha direção.

RESUMO

O Algoritmo bioinspirado firefly (FFiPLS) é baseado na atratividade entre vaga-lumes como comportamento de enxame ao procurar por comida. A técnica de seleção de variáveis é capaz de reduzir o número de variáveis por intervalos a fim de obter melhoria das previsões do modelo quimiométrico. Neste sentido, o presente estudo tem como objetivo comparar o desempenho de modelos quimiométricos de regressão PLS utilizando algoritmos de seleção de variáveis por intervalos FFiPLS frente ao iSPA-PLS e iPLS para determinação de parâmetros de qualidade de biodiesel por espectrometria NIR. O desenvolvimento deste trabalho contou com a obtenção das amostras de biodiesel a partir de óleo de soja via rota metílica. Posteriormente, foi mensurada a densidade do biodiesel e por fim adicionado ao diesel em proporções 5, 10, 15, 20, 25, 30, 35, 40, 45 e 50 (%v/v), o índice de biodiesel em diesel, ambos realizados em triplicata. Construiu-se um banco de dados a partir das medidas dos espectros de absorção no infravermelho próximo com 100 amostras de biodiesel, na faixa dos comprimentos de onda (λ) de 410 a 2500 nm. A partir desta faixa foram escolhidas duas faixas espectrais, de 441-1551 nm e de 1100-1600 nm. As regiões foram escolhidas por estar nas regiões de primeira e segunda região de sobretom do C–H. O tratamento quimiométrico consistiu na seleção das amostras dos conjuntos de calibração e predição utilizando algoritmo SPXY, e para remoção de outliers o teste T^2 de Hotelling. Empregou-se como pré-processamento espectral a derivação Savitzky-Golay, polinômio de 2º grau e janela de 17 pontos. O software utilizado para os pré-tratamentos e para a remoção de outliers foi o The Unscrambler® versão 9.7. Para seleção de amostras e variáveis bem como para a construção dos modelos foi utilizado o software Matlab® versão R2016a. Foram comparados ao algoritmo FFiPLS outros dois algoritmos para seleção de variáveis por intervalos: o iSPA-PLS e o iPLS. Para o índice de biodiesel em diesel, os modelos com faixa de 441-1551 nm com pré-processamento espectral apresentaram baixos valores de variáveis latentes, com melhores valores para a etapa de predição (RMSEP, $R^2_{predição}$ e REP) para os três algoritmos de seleção de variáveis, iSPA-PLS, iPLS e FFiPLS. Já para a densidade, O modelo com faixa de 410-2500 nm com pré-processamento espectral, o algoritmo FFiPLS apresentou melhores valores na etapa de predição (RMSEP, $R^2_{predição}$ e REP).

PALAVRAS-CHAVE: Algoritmo Firefly. Seleção por intervalos. Calibração Multivariada. Quimiometria.

ABSTRACT

The Firefly Bioinspired Algorithm (FFiPLS) is based on the attractiveness of fireflies as swarming behavior when looking for food. The variable selection technique is able to reduce the number of variables by intervals in order to obtain better predictions from the chemometric model. In this sense, the present study aims to compare the performance of PLS regression chemometric models using FFiPLS interval selection algorithms against the iSPA-PLS and iPLS to determine biodiesel quality parameters by NIR spectrometry. The development of this work involved obtaining samples of biodiesel from soybean oil via methyl route. Subsequently, the density of biodiesel was measured and finally added to diesel in proportions 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 (% v / v), the biodiesel index in diesel, both carried out in triplicate. A database was built from the measurements of the absorption spectra in the near infrared with 100 samples of biodiesel, in the range of wavelengths (λ) from 410 to 2500 nm. From this range, two spectral bands were chosen, 441-1551 nm and 1100-1600 nm. The regions were chosen because they are in the first and second overtone regions of the C – H. The chemometric treatment consisted of selecting the samples from the calibration and prediction sets using the SPXY algorithm, and to remove outliers the Hotelling T^2 test. The Savitzky-Golay derivation, a 2nd degree polynomial and a 17-point window, was used as spectral pre-processing. The software used for pre-treatments and for removing outliers was The Unscrambler® version 9.7. Matlab® software version R2016a was used to select samples and variables as well as to build the models. Two other algorithms for selecting variables by intervals were compared to the FFiPLS algorithm: iSPA-PLS and iPLS. For the diesel biodiesel index, models with a range of 441-1551 nm with spectral pre-processing showed low values of latent variables, with better values for the prediction step (RMSEP, $R^2_{prediction}$ and REP) for the three selection algorithms of variables, iSPA-PLS, iPLS and FFiPLS. As for density, the model with a range of 410-2500 nm with spectral pre-processing, the FFiPLS algorithm presented better values in the prediction step (RMSEP, $R^2_{prediction}$ and REP).

KEYWORDS: Firefly Algorithm. Interval selection. Multivariate Calibration. Chemometrics.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

Abreviaturas e siglas	Português	Inglês
%v/v	Percentual volume/volume	–
ABC	Colônia de Abelhas Artificiais	*Artificial Bee Colony*
ACO	Otimização de Colônias de Formigas	*Ant Colony Optimization*
CNPE	Conselho Nacional de Política Energética	
CV	Validação Cruzada	*Cross-Validation*
EJCR	Região Elíptica De Confiança Conjunta	*Elliptical Joint Confidence Region*
FA	Algoritmo Firefly	*Firefly Algorithm*
FFiPLS	Algoritmo Vaga-lume para Seleção de Intervalos em PLS	*Firefly Algorithm for Interval Selection in PLS*
fpop	Número de firefly para a população	–
GA	Algoritmos Genéticos	*Genetic Algorithm*
I_max	Quantidade de Intervalos Máximos	–
iPLS	Seleção de Intervalos em PLS	*Interval Selection in PLS*
iSPA-PLS	Algoritmo das Projeções Sucessivas para seleção de intervalos em PLS	*Successive Projections Algorithm for Interval Selection in PLS*
KS	Kennard-Stone	Kennard-Stone
LOO	Deixar de fora	*leave-one-out*
LV	Variáveis Latentes	*latent variables*
Matlab	–	*MATrix LABoratory*
MLR	Regressão Linear Múltipla	*Multiple Linear Regression*
NIPALS	Iteração não linear por mínimos quadrados parciais	*Nonlinear Iterative Partial Least Squares*
NIRS	Espectroscopia no Infravermelho Próximo	*Near-infrared spectroscopy*
nm	Nanômetros	–
PCA	Análise de Componentes Principais	Principal Component Analysis
PCR	Regressão em Componentes Principais	*Principal Components Regression*
pH	Potencial Hidrogeniônico	–
PLSR	Regressão por Mínimos Quadrados Parciais	*Partial Least Squares Regression*

PSO	Otimização de Enxame de Partículas	*Particle Swarm Optimization*
REP	Erro relativo de predição	*Relative Error of Predictions*
RMSEC	Raiz quadrada do erro médio quadrático de calibração	*Root Mean Square Error of Calibration*
RMSECV	Raiz quadrada do erro médio quadrático de validação cruzada	*Root Mean Square Error of Cross-Validation*
RMSEP	Raiz quadrada do erro médio quadrático de predição	*Root Mean Squared Error of Prediction*
SPA-MLR	Algoritmo das Projeções Sucessivas em MLR	Successive Projections Algorithm in MLR
SPXy	Partição do conjunto de amostras baseado nas distâncias conjuntas	*Sample set Partitioning based on joint X-y distances*
SDV	Desvio Padrão dos Erros de Validação	*Standard Deviation of Validation*
SVD	Decomposição por Valores Singulares	*Singular Value Decomposition*
SW	Inteligência de Enxame	*Swarm Intelligence*

SUMÁRIO

1 INTRODUÇÃO

O surgimento da quimiometria, durante a década de 70, bem como o uso de computadores possibilitou a manipulação de grandes bancos de dados na área de Química, tanto para reconhecimento de padrões, quanto na calibração multivariada (BARROS NETO et al., 2006). Com isto, as técnicas analíticas que obtinham muitas informações sobre a amostra se tornaram viáveis para aplicações práticas, com possibilidade de aumentar a capacidade preditiva das técnicas analíticas mesmo na presença de interferentes e/ou em matrizes complexas com grande número de contamitantes.

Dentre as técnicas analíticas que se beneficiaram significativamente do uso da quimiometria com a possibilidade de tratamento de grande banco de dados estão as espectroscopias e, mais especificamente, a espectroscopia de absorção molecular na região do infravermelho próximo (NIR, do inglês *Near-infrared*). Isto se justifica pois o NIR gera bandas largas e sobrepostas, indicando sinais multicorrelacionados e com baixa intensidade (PASQUINI, 2018).

Apesar da quimiometria conter ferramentas que tornam possível que se processe alto número de variáveis obtidas em uma amostra, devido às medidas de absorbâncias em cada comprimento de onda, isto demanda considerável desempenho computacional e justifica o uso de algoritmos que permitam escolher comprimentos de onda ou intervalos destes que sejam mais correlacionados com a variável de interesse (PAULA et al., 2014). Esta etapa de seleção de varáveis identifica os comprimentos de onda mais informativos, reduz zonas ruidosas e regiões não informativas (SHI et al., 2016) sobretudo também para identificação de variáveis para construção de espectrômetros multiespectrais dedicados muito comuns em aplicações industriais e permitir que usuários não especialistas construam modelos confiáveis (XIABO, et al., 2010).

O uso de técnicas de seleção de variáveis é empregado em diversos algoritmos de calibração multivariada, dentro os quais Regressão Linear Múltipla (MLR, do inglês *Multiple Linear Regression*) e Regressão por Mínimos Quadrados Parciais (PLSR, do inglês *Partial Least Squares Regression*), sendo que para o MLR a seleção é uma alternativa necessária à construção do modelo devido a multicolinearidade, mas para o modelo de regressão PLS também pode ser implementado, já que as regiões espectrais podem conter ruídos heterocedásticos, bem como a inclusão de variáveis

redundantes que afetam a preditividade do modelo quimiométrico (BRERETON, 2003).

O desempenho dos modelos de regressão PLS pode ser otimizado pois a seleção de variáveis por intervalos elege regiões do espectro que podem correlacionar com os parâmetros de interesse. Apesar da regressão PLS deslocar as variáveis não informativas para variáveis latentes com menor variância, a seleção de variáveis pode influenciar o resultado, introduzindo erro aos modelos quimiométricos construídos.

Dentre os algoritmos de seleção de varáveis para otimização de modelos quimiométricos preditivos se destacam os meta-heurísticos que são inspirados por comportamentos biológicos de animais, insetos ou aves a fim de encontrar a solução ideal através de busca cega ou pesquisa informada mediante função heurística (DARWISH, 2018). Deste modo, algoritmos evolutivos realizam busca iterativa com o propósito de determinar a solução para o problema definido pela convergência de respostas ou pelo número de ciclos.

Algoritmo bioinspirado firefly ou vaga-lume baseado no comportamento luminoso de vaga-lumes foi descrito através de modelo de regressão PLS, utilizando seleção de variáveis, para predizer propriedades físicas, elementos terras-raras e tório em amostras de solo da região semiárida brasileira, superando modelos iPLS (do inglês *Interval Selection in PLS*), iSPA-PLS (do inglês *Successive Projections Algorithm for Interval Selection in PLS*) e *full* PLS. O algoritmo firefly (FFiPLS, do inglês *Firefly Algorithm for Interval Selection in PLS*) atinge níveis mais baixos para as figuras de mérito RMSEP, bias e REP (OLIVEIRA et al., 2021).

Poucos artigos foram publicados na área de química analítica com o algoritmo firefly, perfazendo uma grande lacuna no conhecimento em torno de pesquisas que consolidem a eficácia deste algoritmo para seleção de variáveis por intervalos.

Neste sentido, objetiva-se utilizar o algoritmo de seleção de variáveis (FFiPLS*)* em comparativo ao iSPA-PLS e ao iPLS para determinar parâmetros de qualidade do Biodiesel em proporções de 5, 10, 15, 20, 25, 30, 35, 40, 45 e 50 (%v/v), empregando espectroscopia NIR.

1.1 Objetivo geral

Comparar o desempenho de modelos quimiométricos de regressão PLS utilizando o algoritmo de seleção de variáveis FFiPLS frente ao iSPA-PLS e ao iPLS para determinação de densidade e de índice de biodiesel em diesel em biodiesel por espectrometria NIR.

1.1.1 Objetivos Específicos

- Comparar o desempenho dos algoritmos de seleção de variáveis FFiPLS, iSPA-PLS e iPLS empregando espectros brutos e espectros pré-processados;
- Avaliar a capacidade preditiva do modelo de regressão PLS a partir dos algoritmos seleção de variáveis FFiPLS, iSPA-PLS e iPLS em parâmetros de qualidade de biodiesel a partir de dados de espectros NIR.
- Aplicar ferramentas de diagnóstico para caracterizar a performance dos modelos construídos com base nos algoritmos FFiPLS, iSPA-PLS e iPLS.

2 REFERENCIAL TEÓRICO

2.1 Quimiometria

O estudo da quimiometria inicia formalmente na primeira metade da década de 70, onde se estruturou com o advento dos microcomputadores, realizando cálculos extensos e demorados em computadores de grande porte, eram encontrados apenas em grandes centros universitários. Em anos posteriores os computadores tornaram-se objetos comuns nos laboratórios químicos (BARROS NETO, et al., 2006).

Os químicos analíticos são os principais usuários de quimiometria, sendo, portanto, de fundamental importância a compreensão a fundo os métodos que usam a partir de conhecimentos da estatística e da matemática (BRERETON, 2003). Paralelamente, o usuário de quimiometria se depara com sequências de equações, onde, predominantemente, os cálculos envolvidos são complexos, requirindo o uso de softwares. No entanto, a interpretação dos resultados obtidos não demanda do analista a realização de equações matemáticas manualmente, já que alguns métodos são amplamente difundidos no meio científico através de algoritmos que executam rotinas pré-determinadas (BRERETON, 2018).

2.2 Calibração multivariada

A calibração multivariada é uma das áreas da Quimiometria bastante investigada dentro do campo das metodologias analíticas alternativas, bem como nas áreas de interesse de produtos alimentícios, farmacêuticos, polímeros e combustíveis (PASQUINI, 2018). A calibração consiste em estabelecer modelos para encontrar uma relação matemática entre a Matriz **X** (resposta instrumental, como dados espectrais, por exemplo) e a Matriz **Y** (propriedades ou parâmetros atribuídos a amostra). O modelo pode estimar propriedades de interesse, normalmente desconhecidas, através de regressão matemática inversa, partindo da matriz **X** e obtendo a propriedade indicada em **Y** (HONORATO, 2006). Como exemplo pode-se citar propriedades de interesse (variável dependente) concentração de uma determinada amostra, pH, densidade, condutividade, índice de acidez, dentre outras.

A capacidade preditiva do modelo pode ser avaliada utilizando métodos de validação interna ou externa. Para a validação externa, utiliza-se um conjunto de

amostras independentes, externas ao modelo de calibração. Neste sentido, os parâmetros de interesse deste subconjunto de validação externa devem ser conhecidos a fim de certificar a relação com as previstas no modelo (NUNES, 2008). Os métodos de validação interna são utilizados, geralmente, quando se tem um pequeno número de amostras. Um processo interno, bastante utilizado na calibração multivariada é a validação cruzada (do inglês *cross-validation*), sendo bastante utilizado nos modelos de regressão PLS desenvolvidos com espectroscopia NIR (LI et al., 2015; ATTIA et al., 2017; XU et al., 2018). A validação cruzada consiste em dividir o grupamento de calibração, anteriormente construído, em vários segmentos, e a partir daí, remover amostra(s) por vez e construir o modelo de calibração, testando o modelo com a(s) amostra(s) retirada(s) para validar o modelo. A abordagem mais comum é a validação cruzada *leave-one-out* (LOO), onde o procedimento de retirar uma amostra por vez é repetido até que todas as amostras tenham sido deixadas de fora (BRERETON, 2007).

2.2.1 Métodos de Calibração Multivariada

Diversas ferramentas quimiométricas são utilizadas para a calibração multivariada em bancos de dados que trazem informação química, por exemplo. Para isso, os métodos mais comuns são a Regressão Linear Múltipla (MLR, do inglês *Multiple Linear Regression*), a Regressão em Componentes Principais (PCR, do inglês *Principal Components Regression*) e a Regressão por Mínimos Quadrados Parciais (PLSR, do inglês *Partial Least Squares Regression*) (HONORATO, 2006).

2.2.1.1 Regressão Linear Múltipla – MLR

A técnica de MLR, introduzida por Sternberg et al. (1960), estabelece a aplicação de mínimos quadrados a partir da relação linear entre a matriz **X**, com a informação do sinal analítico obtido instrumentalmente, com dimensão (*m* x *k*), sendo *m* o número de amostras e *k* o número de variáveis e a matriz coluna *y*, com dimensão (*m* x 1) que contém a propriedade de interesse. Cada variável dependente de *y* é expressa como uma combinação linear das variáveis independentes da matriz **X**, em que o resíduo não modelado é representado por E, na Equação (1):

$$y = Xb_{MLR} + E \quad (1)$$

O vetor *b* contém os coeficientes da regressão e é calculado a partir da pseudo-inversa de **X**, na Equação (2):

$$b_{MLR} = (X^T X)^{-1} X^T y \quad (2)$$

Onde os índices sobrescritos -1 e **T** representam a inversão e a transposição da matriz, respectivamente.

Algebricamente a inversão da matriz ($\mathbf{X^TX}$) possui restrições: o número de amostras do conjunto de calibração *m* deve ser igual ou maior ao número de variáveis *k*, o que incidirá, do contrário, em um sistema de equações indeterminado; as variáveis devem ser linearmente independentes, senão incorrerá numa matriz singular (GOMES, 2012). A fim de corrigir estas limitações, pode-se ajustar o banco de dados a partir da técnica de seleção de variáveis, reduzindo as dimensões das colunas da matriz (variáveis), atendendo a premissa $m \geq k$.

2.2.1.2 Regressão em componentes principais – PCR

O PCR é útil na espectroscopia principalmente quando apenas alguns compostos podem ser identificados na mistura, diferentemente do MLR que apresentam melhores resultados com misturas quando todos os componentes são identificados (BRERETON, 2007).

A decomposição da matriz **X** em PCR é uma análise por componentes principais, onde é decomposta em duas outras matrizes, denominada de matriz de escores **(T)** e matriz de pesos **(P)**, mais a matriz de resíduos de **X (E)** conforme representada na Equação (3):

$$X = TP^T + E \quad (3)$$

Neste sentido, após executar a Análise de componentes principais - PCA (do inglês, *principal component analysis*) há que definir o número de componentes principais (PC). No campo da idealidade, o número de PCs deve ser igual ao número de compostos pertencentes à mistura analisada. No entanto, partindo da impossibilidade de determinar todos os componentes pertencentes à amostra em

questão, define-se o número de PCs devido a correlações entre as propriedades dos compostos, similaridades espectrais, reduzindo o número de componentes. Antagonicamente, há aumento no número de PCs em consequência a ruído instrumentais ou deslocamento de linha de base (BRERETON, 2007).

Após obtenção da matriz **T**, pode-se obter uma equação de regressão entre a propriedade a ser determinada *y* e a matriz de escores **T**. A relação existente entre essas duas propriedades pode ser descrita pela Equação (4):

$$y = T_{(m\ x\ k)} b_{PCR} + F \qquad (4)$$

Onde *k* equivale ao número de PCs utilizados na obtenção dos coeficientes de regressão e F aos resíduos.

A fim de determinar a matriz **T** através da PCA, utilizam-se algoritmos iterativos com o objetivo de atingir a convergência. Dentre os algoritmos bastante utilizados destacam-se o NIPALS e a decomposição por valores singulares (SVD, do inglês *Singular Value Decomposition*) (FERNANDES, 2013).

2.2.1.3 Regressão por mínimos quadrados parciais – PLSR

A PLSR é construída baseada no algoritmo de Iteração não linear por mínimos quadrados parciais (NIPALS, do inglês *Nonlinear Iterative Partial Least Squares*). (GELADI et al., 1986). O método utiliza matrizes de dados multivariados, a partir de dados instrumentais (matriz **X**), como as informações de concentrações (matriz **Y**), de tal forma que a propriedade seja função da resposta instrumental, através da mudança das variáveis originais por um sub-conjunto truncado das variáveis latentes dos dados originais (SIMÕES, 2008).

As premissas básicas desses métodos são decompor todas as variáveis independentes e dependentes, **X** e **Y**, respectivamente, em um produto com duas matrizes menores, mais uma matriz de erro, como descritas nas Equações (5) e (6) (BRERETON, 2003; SIMÕES, 2008):

$$X = TP^T + E \qquad (5)$$

$$Y = UQ^T + F \tag{6}$$

Para as Equações (5) e (6), **T** e **U** são as matrizes de *scores*, **P** e **Q** são as matrizes de *loadings*, e **E** e **F** são as matrizes de erro de **X** e **Y** respectivamente.

O procedimento a seguir, descritos nas Equações (7) e (8), será a relação linear entre **U** e **T** (BRERETON, 2003; SIMÕES, 2008):

$$U = BT + G \tag{7}$$

$$Y = BTQ^T + H \tag{8}$$

Os termos nas Equações (7) e (8) são o coeficiente ajustado, **B**, calculado geralmente usando o algoritmo NIPALS, **G** é a matriz de resíduos dos *scores*, e **H** a matriz de resíduos de concentração (BRERETON, 2003; SIMÕES, 2008).

Atualmente, o PLSR, bem como suas variantes, mantém sua forte abrangência dentre as técnicas de regressão multivariada empregada na maioria dos métodos analíticos descritos na literatura para espectroscopia NIR (PASQUINI, 2018). Porém, deve ser observado que ao desenvolver o modelo de calibração PLSR deve-se levar em conta o número de variáveis latentes, valor este crítico, já que um baixo número de LV incidirá em um modelo bastante simples e subajustado (do inglês, *underfitting*). Antagonicamente, ampliar o número de LV aumenta o risco de sobreajuste (do inglês, *overfitting*) e degrada a eficiência de generalização do modelo (XU et al., 2018).

Semelhante ao modelo PCR, O PLSR é eficiente em quantificar analitos mesmo na presença de interferentes, bastando incluí-los no conjunto de calibração. Conceitualmente, há distinção entre os componentes principais (modeladas em ordem decrescente da variância explicada) na PCR e os fatores estimados, de maneira mais ampla, denominado variáveis latentes no PLSR. As variáveis latentes explicam concomitantemente a variância espectral e da propriedade de interesse, havendo perda de ortogonalidade (SENA et al., 2018).

Para que os modelos de regressão sejam construídos é necessário que o conjunto de amostras seja dividido em três: conjunto de calibração, conjunto de validação e conjunto de predição. Os dois primeiros são usados para a construção propriamente dita do modelo e o terceiro envolve amostras que não compuseram os outros dois conjuntos. Para que estas amostras sejam divididas podem ser utilizadas

algumas estratégias, dentre as quais escolha randômica ou com apoio de algoritmos como o Kennard-Stone (KENNARD et al., 1969) e o SPXy (GALVÃO et al., 2005).

2.2.2 Técnicas de seleção de variáveis

A seleção de variáveis bem como a redução de dimensionalidade está sendo investigada na construção de modelos eficientes tanto para classificação quanto para regressão, auxiliando ao analista na tomada de decisão (FISTER et al., 2013; GOMES et al., 2013; GOODARZI et al., 2014; SHI et al., 2016; ATTIA et al., 2017; ZHANG et al., 2018). Banco de dados com grande número de variáveis, devem originar um bom subconjunto, pois contribuem para aprimorar o desempenho do modelo. No entanto, deve-se levar em consideração que muitas das variáveis possuem informação contraditória ou mesmo redundantes, resultando na degradação da performance. Neste sentido, a seleção de variáveis constitui maneira de aperfeiçoar e otimizar dados de forma parcimoniosa (ZHANG et al., 2018).

Independentemente dos métodos de calibração multivariada, como o PLSR e o PCR, possibilitarem a construção de modelos mesmo quando o número de variáveis é maior que o número de amostras, o uso de técnicas de seleção de variáveis é capaz de reduzir o número de variáveis a fim de obter melhoria das previsões do modelo quimiométrico, melhor interpretação e menor dispêndio de medição instrumental. Do mesmo modo, a remoção de variáveis irrelevantes, ruidosas ou não confiáveis usualmente melhoram a capacidade de previsão e podem reduzir a complexidade do modelo (ANDERSEN et al., 2010).

2.2.3 Algoritmos Determinísticos e Estocásticos

Um critério a ser ponderado durante a avaliação do método de seleção de variáveis, deve considerar qual maneira o algoritmo de seleção de variáveis é executado para um grande número de variáveis. Algoritmos Determinísticos, que expressam ao final do cálculo uma solução única apresentam bom desempenho para pequenos problemas, porém falham ao passo que o número de variáveis é aumentado. Em detrimento, Algoritmos Estocásticos exibem melhores resultados em dimensões mais altas, apesar de realçar não exclusivamente um resultado, mas a

tendência de encontrar a melhor solução global (RAPHAEL et al., 2003), devido ao uso de randomização em busca de um conjunto de soluções (FISTER et al., 2013).

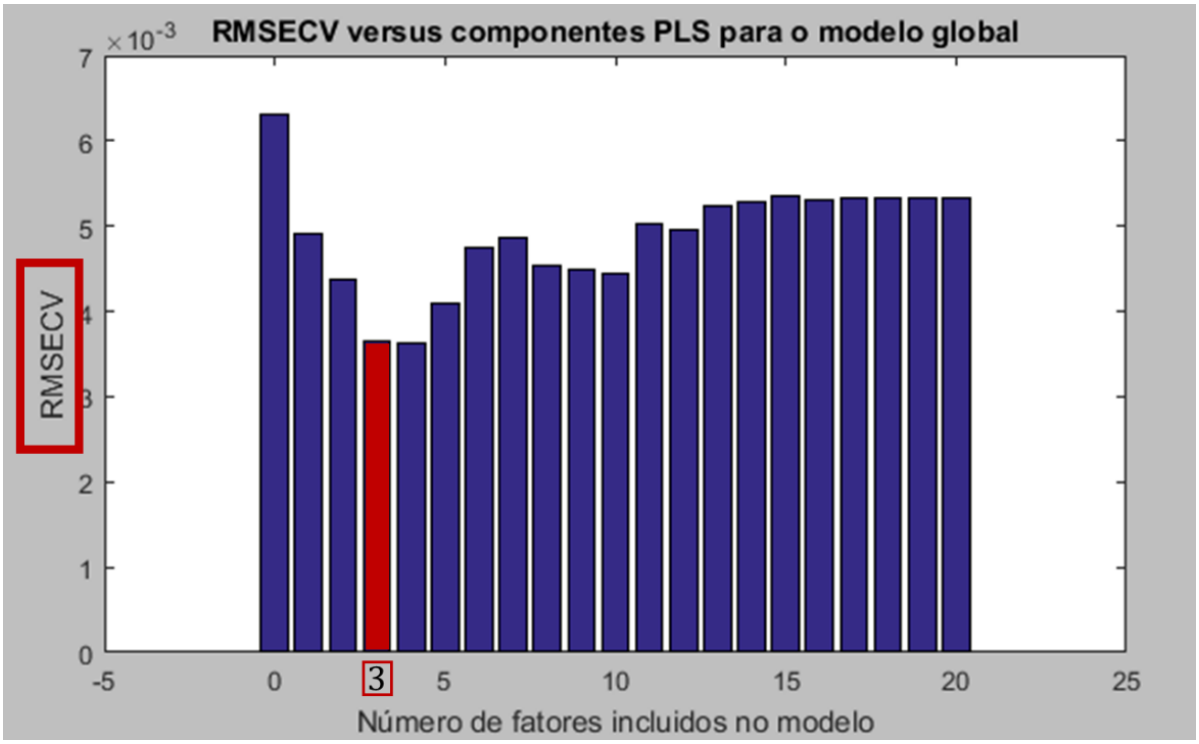
Para fins da construção deste trabalho foram utilizados algoritmos determinísticos, o iSPA-PLS (GOMES et al., 2013), o iPLS (NØRGAARD et al. 2000) e algoritmo estocástico FFiPLS (OLIVEIRA et al., 2021) a fim de correlacionar os parâmetros de interesse à resposta instrumental.

2.2.3.1 iSPA-PLS

O Algoritmo das Projeções Sucessivas para seleção de intervalos em PLS (iSPA-PLS, do inglês *Successive Projections Algorithm for Interval Selection in PLS*), proposto por Gomes et al. (2013), é uma extensão do SPA-MLR (Algoritmo das Projeções Sucessivas em MLR, do inglês *Successive Projections Algorithm in MLR*). Os modelos SPA-MLR apresentam vantagem com relação à simplicidade e facilidade de interpretação, em comparação aos modelos PLSR. Entretanto, a vantagem do iSPA-PLS frente ao SPA-MLR é ser menos sensível ao ruído instrumental, assim como apresentou modelos mais robustos (GOMES et al., 2013). Outra vantagem do iSPA-PLS obteve como resultado ser mais parcimonioso e com maior capacidade preditiva, quando confrontado a modelo *full* PLS, e iPLS em amostras de hamburguer de frango (KREPPER et al., 2018).

Para fins de seleção de variáveis por intervalos, o iSPA-PLS foi desenvolvido para ser processado em duas fases, onde preliminarmente calcula o número ótimo de fatores para o modelo *full* PLS, através de validação por série de teste ou validação cruzada, empregando uma quantidade de fatores determinado pelo usuário. Com isso, há saída gráfica apresentando os valores de RMSECV para o PLS global em função do número de fatores incluídos no modelo com objetivo de orientar a escolha do número de fatores (GOMES, 2012). A saída Gráfica é mostrada na Figura 1.

Figura 1: Gráfico de RMSECV x número de fatores para o PLS global.



Fonte: Própria, 2021.

A validação cruzada é empregada para determinar um número apropriado de variáveis latentes em cada modelo PLS. A melhor combinação de intervalos é então escolhida com base no menor RMSECV (GOMES, et al, 2013).

Na primeira fase, a matriz de repostas instrumentais é centrada na média das colunas, dividindo em *w* intervalos não sobrepostos. A condição estabelecida é que o número de variáveis *k* incluído no intervalo tem que ser maior que o número ótimo de fatores. Posteriormente, os intervalos são submetidos à etapa de projeção via SPA e logo após a matriz SEL é obtida, onde as colunas de SEL incluem os índices das cadeias dos intervalos. Na segunda fase, o conjunto de intervalos selecionados é utilizado na construção dos modelos PLS, empregando validação cruzada, o qual selecionará o que apresentar o menor RMSECV ou RMSE (FERNANDES, 2016).

2.2.3.2 Regressão por mínimos quadrados parciais por intervalos – iPLS

Implementado inicialmente para NIRS, a regressão por mínimos quadrados parciais por intervalos – iPLS, proposto por Nørgaard et al. (2000) busca um intervalo espectral que contém informação que pode ser atrelado ao parâmetro de interesse.

A fim de constituir uma sequência para o algoritmo iPLS, o espectro é dividido previamente em intervalos de tamanhos iguais e então modelos PLS locais são construídos para que cada intervalo seja avaliado, destacando regiões espectrais importantes e removendo interferência de outras regiões. Posteriormente, os modelos construídos com cada intervalos são comparados em termos de RMSECV, bem como R^2, *slope* (inclinação) como também *offset* (deslocamento) (NØRGAARD et al. 2000). Deve-se ponderar durante a escolha da quantidade de intervalos as quais os espetros serão divididos, uma vez que optar por um número pequeno de intervalos acarreta em faixas maiores, que podem apresentar informações desnecessárias, acarretando em sobreajuste, enquanto segmentar os espectros em intervalos muito estreitos pode fragmentar as informações úteis dos dados, gerando possível subajuste (FERNANDES, 2016).

2.2.3.3 FFiPLS

A inteligência de enxame (do inglês *Swarm Intelligence*) é inspirado no comportamento coletivo de enxames sociais de formigas, cupins, abelhas e vermes, que apesar de serem indivíduos relativamente pouco sofisticados, apresentam comportamento coordenado que direciona os enxames para os objetivos desejados. Algoritmos bioinspirados se apresentam como métodos de otimização, a exemplo da otimização de colônias de formigas (do inglês *ant colony optimization - ACO*), otimização de enxame de partículas (do inglês *particle swarm optimization* - PSO), colônia de abelhas artificiais (do inglês *artificial bee colony* - ABC), além do promissor algoritmo firefly ou algoritmo vagalume (do inglês *firefly algorithm* - FA) (FISTER et al., 2013).

A estes algoritmos bioinspirados podem ser atrelados técnicas analíticas e quimiométricas a fim de estabelecer, por exemplo, à calibração multivariada, a partir da técnica de Espectroscopia no Infravermelho Próximo, seleção de variáveis e Regressão por Mínimos Quadrados Parciais (do inglês *Partial Least Squares* - PLS) a fim de otimizar técnicas de seleção de variáveis existentes.

O algoritmo firefly foi desenvolvido por Yang (2008) é estocástico e difere dos algoritmos determinísticos, pois utiliza randomização na busca de um conjunto de soluções. Pode ser aplicado para solucionar problemas mais difíceis de otimização. Não há garantia que a solução ideal seja encontrada em um período de tempo

razoável. Em contraponto, a randomização permite que o processo de pesquisa evite que a solução fique presa em ótimos locais (FISTER et al., 2013).

Há de observar que poucos artigos foram publicados na área de química analítica com o algoritmo firefly, perfazendo uma grande lacuna no conhecimento em torno de pesquisas que consolidem a eficácia deste algoritmo para seleção de variáveis. No entanto, os algoritmos que envolvem o conceito firefly atrelado a calibração multivariada em PLS apenas selecionam variáveis individuais (LI et al., 2015; ATTIA et al., 2017; XU et al.,2018). Já o algoritmo firefly (FFiPLS), proposto por Oliveira et al., 2021, selecionam intervalos a partir do comportamento inspirado heuristicamente pelas características dos vaga-lumes, por meio do modelo de regressão PLS, que emprega o cálculo em todo o espectro e posteriormente aos intervalos selecionados. O algoritmo FFiPLS pode ser descrito como uma abordagem de inteligência de enxame (do inglês, *swarm intelligence*). A matriz inicial de calibração (Xcal) é particionada em *n* intervalos (I) não sobrepostos e definidos pelo usuário, onde são posteriormente otimizados através do método. (OLIVEIRA et al., 2021).

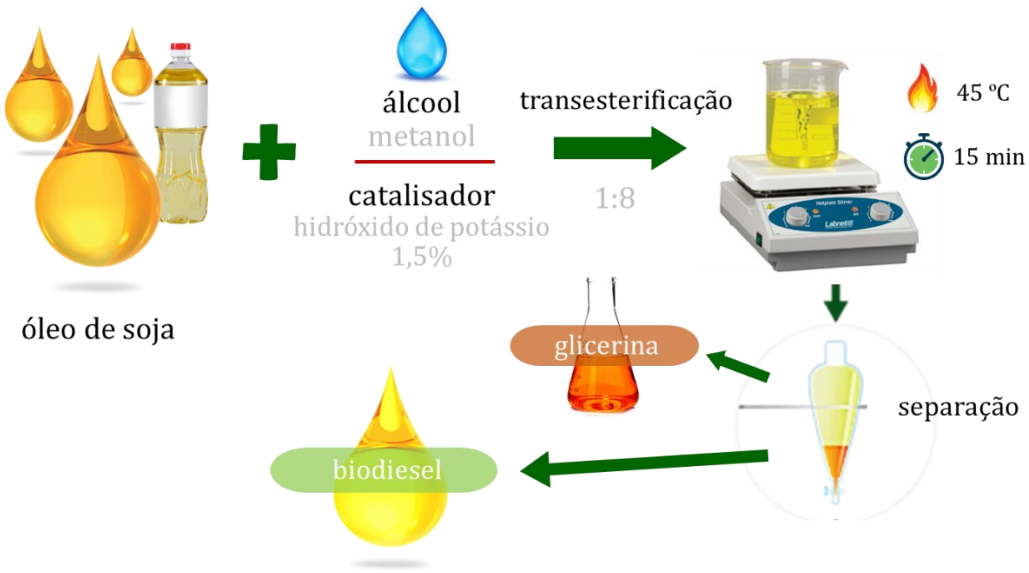
Os Parâmetros de entrada ffpop corresponde à quantidade de vaga-lumes na população inicial gerada aleatoriamente; Ciclos é número de iteração do processo de otimização; ω_0 é a atratividade em r (distância) = 0; α é a porcentagem aleatória do movimento do vaga-lume; e γ é o brilho. Cada vaga-lume artificial terá brilho proporcional à sua capacidade de produzir um bom condicionamento dado pelo RMSECV, onde posteriormente é apresentado graficamente. Em cada ciclo do otimizador, o comportamento do enxame será guiado por um componente aleatório α, e pelo brilho (γ). Todo vaga-lume recebe uma atratividade que diminui rapidamente com a distância. Inicialmente, todos os vaga-lumes têm atratividade unitária (i = 1). Ao final dos ciclos de otimização, o vaga-lume de maior brilho e atratividade corresponde à solução otimizada do problema de seleção de variáveis. Em seguida, o modelo PLSR final é construído levando em consideração apenas os intervalos armazenados no vaga-lume mais brilhante (OLIVEIRA et al., 2021).

3 MATERIAL E MÉTODOS

3.1 Amostras de biodiesel

A obtenção das amostras de biodiesel foi dada a partir do produto de reação de transesterificação de óleo de soja via rota metílica empregando hidróxido de potássio como catalisador, em proporção de 1:8 óleo de soja/álcool metílico, sendo adicionado 1,5% do catalisador em relação à massa do óleo. A mistura foi submetida a agitação magnética e aquecimento a 45ºC por 15 minutos. Aguardou-se a separação entre o biodiesel e a glicerina com posterior secagem e purificação do biodiesel (FERNANDES et al., 2011). O esquema da produção de biodiesel é apresentado na Figura 2.

Figura 2: Sequência esquemática da produção do biodiesel.



Fonte: Própria, 2021.

As amostras de diesel o qual foi produzido a mistura biodiesel/diesel foi fornecido pela Petrobrás Distribuidora, localizada no município de Cabedelo/PB (FERNANDES et al., 2011).

3.2 Medidas espectrais

Utilizou-se banco de dados de biodiesel de óleo de soja em diesel com 100 amostras e 2093 variáveis, que consistem nos comprimentos de onda (λ) da varredura do espectro de 410 a 2500 nm, compreendido na faixa do infravermelho próximo, com 1 nm de resolução e realizados em triplicata, conduzido através de espectrofotômetro NIR da PerkinElmer®, modelo Lambda 750, equipado com célula de quartzo com 1 cm de caminho ótico, com fonte de tungstênio e tubo fotomultiplicador R928 e sistema de detecção de PbS (FERNANDES et al., 2011).

3.3 Parâmetros de interesse

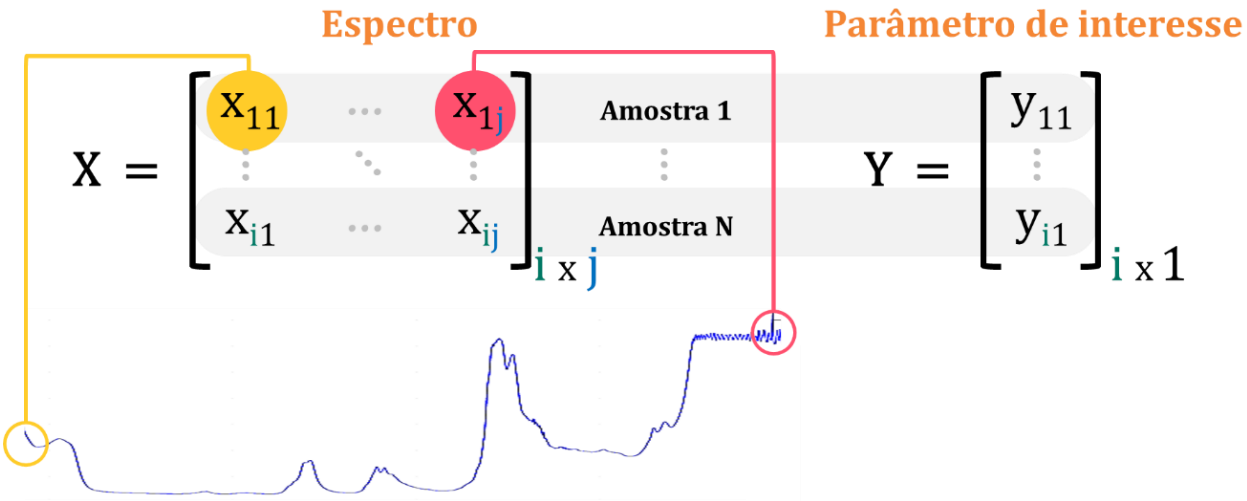
Os parâmetros de interesse nas amostras de biodiesel aqui estudados foram o índice de biodiesel em diesel, com um intervalo de 5 a 50 (%v/v) de biodiesel em diesel, nas seguintes proporções 5, 10, 15, 20, 25, 30, 35, 40, 45 e 50 (%v/v) e a densidade do biodiesel.

3.4 Banco de dados

Utilizou-se banco de dados de biodiesel de óleo de soja em diesel com 100 amostras e 2093 variáveis para os dados brutos. Em seguida, o espectro foi dividido em duas faixas espectrais: conjunto 1 - faixas de 441 a 1551 nm; conjunto 2 - faixas de 1100 a 1600 nm. O conjunto 1 foi escolhido devido a banda de absorção no visível com pico na faixa espectral de comprimento de onda em cerca de 530 nm, correspondente aos ésteres metílicos contendo ligações duplas conjugadas e triglicerídeos não transesterificados enquanto para as bandas da região NIR os picos entre 1400, correspondente ao primeiro sobretom das bandas de combinação de ligações C–H e os picos em torno de 1200 nm equivalente ao segundo sobretom dos modos de estiramento das ligações C–H (XIABO et al., 2010; FERNANDES et al., 2011). Já o conjunto 2 foi escolhido por conter apenas a região NIR. A matriz de dados é construída a partir dos dados instrumentais (espectrais), cada coluna é uma variável e está relacionada a um comprimento de onda do espectro NIR bem como cada linha refere-se as amostras na matriz X. Para a matriz **Y**, ou especificamente vetor **Y**, a

coluna é o parâmetro de interesse e cada amostra está nas linhas de **Y**. A Figura 3 ilustra a organização dos dados na matriz de dados.

Figura 3: Organização matricial dos dados instrumentais pelos parâmetros de interesse.



Fonte: Própria, 2021.

3.4.1 Métodos quimiométricos

Após a construção do banco de dados, com dados espectrais NIR e parâmetros de interesse, os dados brutos (*full* espectro) e os dados pré-processados foram construídos os conjuntos de calibração e predição através de algoritmo de seleção de amostras SPXy, com posterior seleção de variáveis e ao final a construção dos modelos de regressão PLS, conforme mostrado na Figura 4.

Figura 4: Sequência de métodos quimiométricos utilizados.

Fonte: Própria, 2021.

3.4.1.1 Seleção de amostras

As amostras dos conjuntos de calibração (60 amostras) e predição (40 amostras) foram selecionadas através do algoritmo SPXy a partir da interface *Data Hand Gui*, em Matlab® versão R2016a (GALVÃO et al., 2005).

3.4.1.2 Pré-tratamentos espectrais

Realizou-se pré-tratamentos espectrais no conjunto de dados. Inicialmente, avaliou-se o espectro dos dados brutos, removendo *a priori* regiões pobres em informação e/ou que contenham ruído de alta frequência além de buscar ferramentas quimiométricas com o intuito de suavizar sinais espúrios.

Com isso, utilizou-se os pré-processamentos espectrais para a faixa completa, para o conjunto 1 - faixas de 441 a 1551 nm; conjunto 2 - faixas de 1100 a 1600 nm. Empregou-se a derivação Savitzky-Golay, polinômio de 2º grau e janela de 17 pontos. O software utilizado para os pré-tratamentos foi o The Unscrambler® versão 9.7.

3.4.1.3 Seleção de variáveis

A fim de buscar o algoritmo que selecionasse as variáveis mais relacionadas ao problema (parâmetros de interesse) estudado, comparou-se ao algoritmo estocástico FFiPLS e outros dois algoritmos determinísticos, o iSPA-PLS e o iPLS.

Os modelos FFiPLS, por serem estocásticos, foram realizados em decuplicata. A fim de construir o modelo de regressão, foram utilizadas as matrizes de calibração e de predição ($X_{calibração}$, $Y_{calibração}$, $X_{predição}$, $Y_{predição}$). Posteriormente, foi estipulada a quantidade de intervalos na qual será dividido o espectro, bem como estipular a quantidade de variáveis latentes calculadas para o PLS modelo global, a fim de definir o menor RMSECV, para o iSPA-PLS, o iPLS e o FFiPLS. Consecutivamente, adotou-se o valor de 100 firefly para a população (ffpop) bem como para os ciclos (gerações) para convergência do modelo. Por fim, atribuiu os valores para w_0, gama (γ) e alfa (α), respectivamente, 0,97, 1,0 e 0,2. O software utilizado para a seleção de variáveis e posterior construção dos modelos de regressão foi o Matlab® versão R2016a.

4 RESULTADOS E DISCUSSÃO

4.1 Avaliação das propriedades de interesse

A Tabela 1 apresenta as medidas de tendência central e de dispersão das propriedades de interesse, a fim de avaliar, de maneira preliminar, o conjunto de dados que foram obtidos a partir de ensaios físico-químicos.

Tabela 1: Medidas de tendência central e dispersão das propriedades de interesse das amostras de biodiesel.

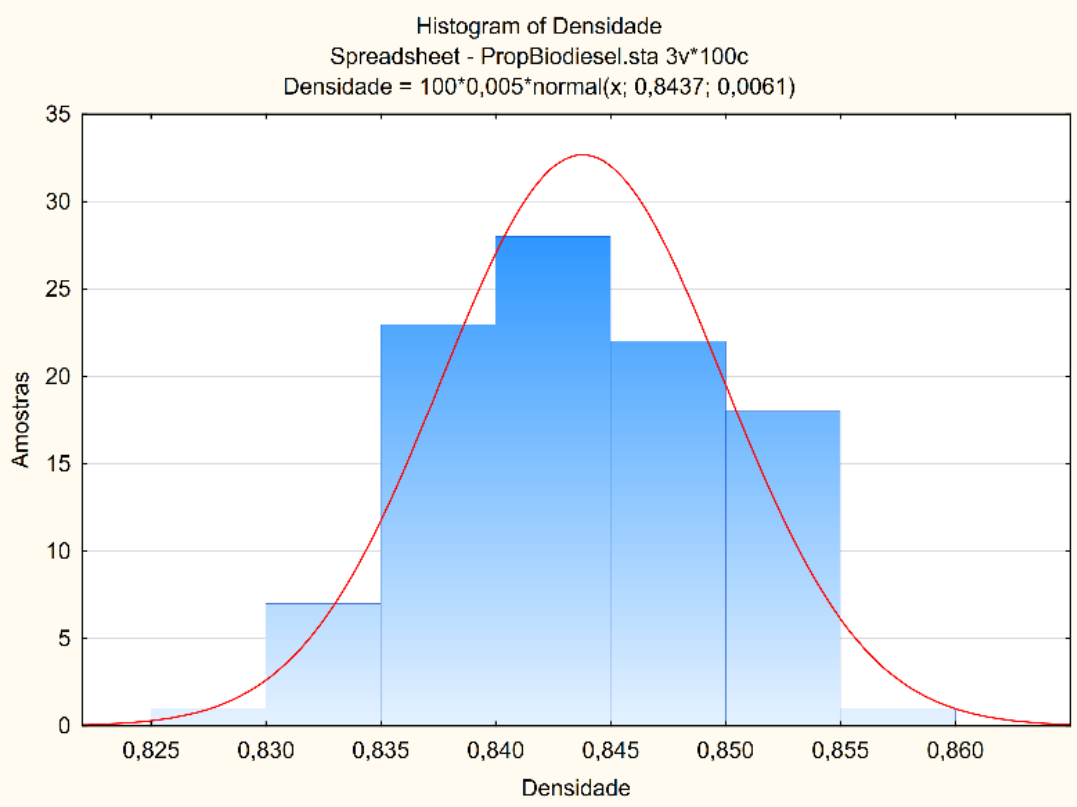
Propriedades de interesse	Medidas	Valores
Índice de biodiesel em diesel (%v/v)	Mínimo	5
	Máximo	50
	Mediana	27,5
	Média	27,5
	Desvio Padrão	14,4
	Amplitude	45
Densidade (g/cm³)	Mínimo	0,8270
	Máximo	0,8551
	Mediana	0,8438
	Média	0,8437
	Desvio Padrão	0,0061
	Amplitude	0,0280

Fonte: Própria, 2021.

Os valores compreendidos entre 5 a 50% foram considerados devido à evolução da adição de biodiesel ao diesel estipulada pelo Conselho Nacional de Política Energética - CNPE, onde fixa o percentual de até 15% de adição até o ano de 2023 (CNPE, 2018).

Constata-se que para a propriedade densidade, os valores apresentam uma amplitude na casa de centésimos e desvio padrão amostral na casa de milésimos, sendo medidas de dispersão baixas. As medidas de tendência central, aqui avaliadas em média e mediana são notadamente próximas. Na Figura 5 apresenta-se o histograma e a gaussiana do comportamento amostral da densidade das amostras de mistura de biodiesel/diesel.

Figura 5: Histograma dos valores de densidade das amostras da mistura de biodiesel/diesel.



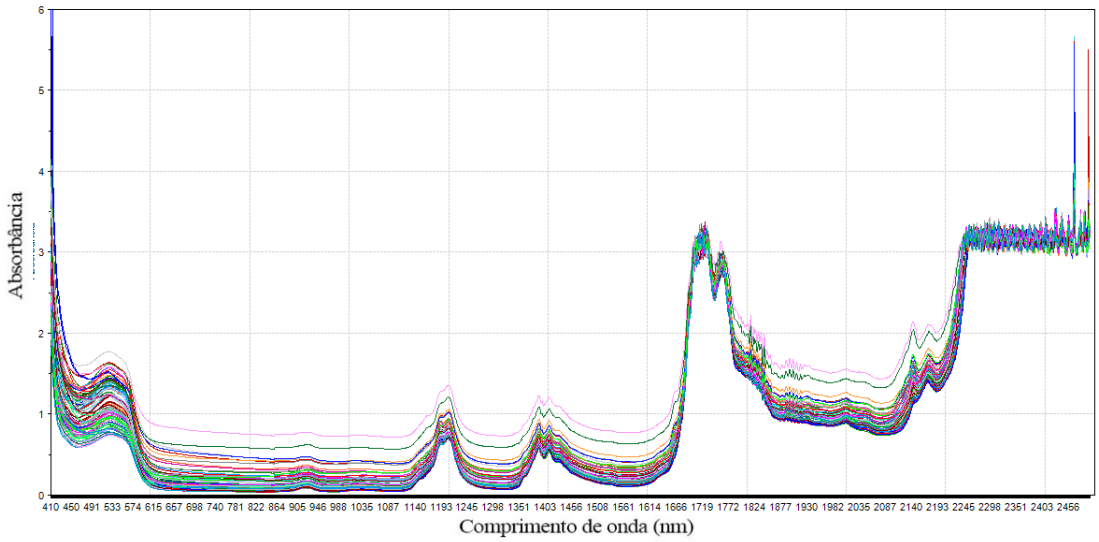
Fonte: Própria, 2021.

Nota-se a distribuição dos valores da densidade das amostras de biodiesel com média de 0,8437 g/cm³, e os dados tem distribuição normal.

4.2 Análise espectral

Na Figura 6 é apresentado o comportamento espectral dos dados brutos na região NIR para as 100 amostras das misturas diesel/biodiesel. Observa-se que na região de 2250 a 2500 nm, há a presença de ruído espectral de alta frequência.

Figura 6: Comportamento espectral na região NIR de 100 amostras de misturas diesel/biodiesel a partir dos dados brutos.



Fonte: Própria, 2021.

No entanto, a região de ruído de alta frequência, como também as regiões de alta absorbância (acima de 3,0, neste espectro), exibido tanto no início do espectro como no final, observada anteriormente na Figura 6, seria comum ao processo de remoção de variáveis *a priori* por incluir ao modelo regiões que contém informações ruidosas, a partir de informações redundantes e/ou não informativas, que podem comprometer o modelo. No entanto a região não será descartada para fins de seleção de amostras e de variáveis e será considerado todo espectro. Neste sentido, busca-se apresentar capacidade do algoritmo em não selecionar variáveis e intervalos de variáveis que compreendem regiões onde possivelmente não evidenciam correlação com o parâmetro de interesse.

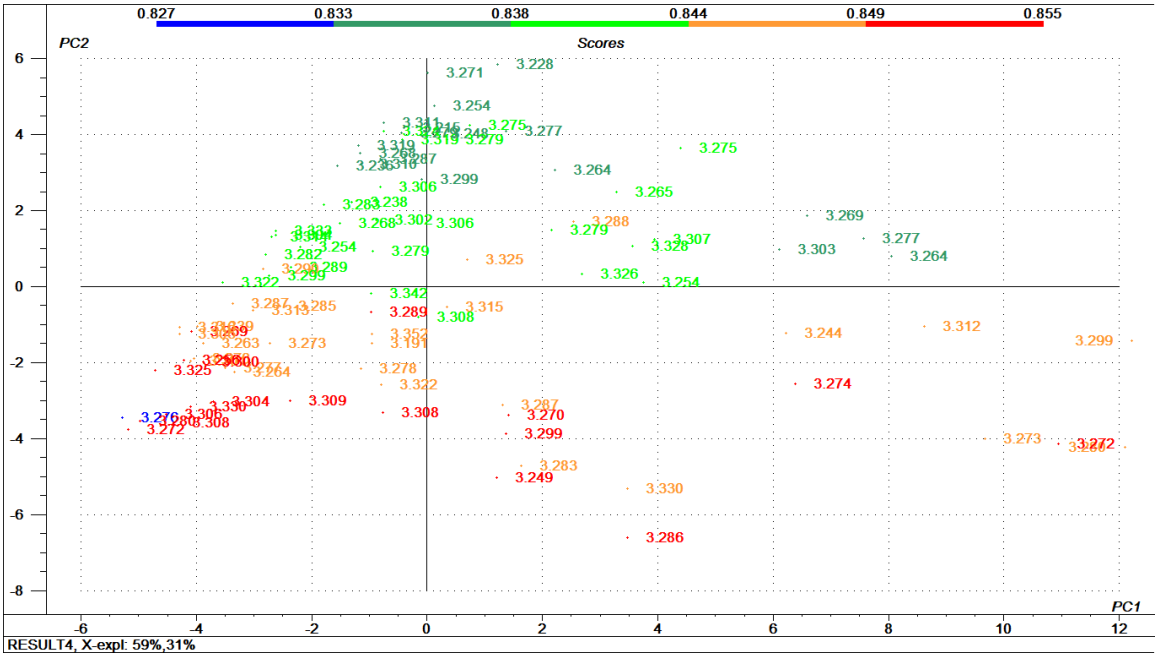
Posteriormente, realizou-se o procedimento de seleção de amostras através do algoritmo SPXy. A seleção de variáveis foi realizada através dos algoritmos iSPA-PLS, iPLS e FFiPLS que serão apresentados separadamente e, ao final, discutidos.

4.3 Análise de outliers

Para detectar amostras anômalas, foi realizada PCA (Análise de Componentes Principais, do inglês *Principal Component Analysis*) e posteriormente analisada a medida de T^2 de Hotelling. onde relaciona a distância da amostra até o centro dos dados. Para cada faixa espectral, 410-2500 nm, 441-1551 nm e 1100-1600 nm, com

e sem pré-processamento, foram retiradas amostras que estava fora da elipse da medida de T^2 de Hotelling. Para a propriedade densidade, além das amostras anteriormente retiradas, o gráfico de scores para os valores de densidade aponta que a amostra 10 está separada das demais, marcada em azul, conforme Figura 7.

Figura 7: Gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados brutos.



Fonte: Própria, 2021.

Neste sentido, foram dispostas a amostras consideradas outliers para as faixas espectrais, com e sem pré-processamento, para os dois parâmetros de interesse, o índice de biodiesel em diesel e a densidade, na Tabela 2.

Tabela 2: Medidas de tendência central e dispersão das propriedades de interesse das amostras de biodiesel.

Parâmetro de interesse	Faixa espectral	Sem Pré-processamento	Com Pré-processamento
Índice de biodiesel em diesel (%v/v)	410-2500	1, 80, 84, 90, 91	1, 80, 91
	441-1551	80, 84, 90	1, 91
	1100-1600	80, 90	23, 51, 61, 73
Densidade (g/cm³)	410-2500	1, 10, 80, 84, 90, 91	1, 10, 80, 91
	441-1551	10, 80, 84, 90	1, 10, 91
	1100-1600	10, 80, 90	10, 23, 51, 61, 73

Fonte: Própria, 2021.

4.4 Modelos PLS – índice de biodiesel em diesel

Serão apresentados os melhores modelos de regressão PLS, considerando o menor número de variáveis latentes, os maiores valores de $R^2_{calibração}$, $R^2_{validação}$, $R^2_{predição}$, e menores valores de RMSEC, RMSECV, RMSEP e menor REP - erro relativo de predição (do inglês, *Relative Error of Predictions*).

Dentre os modelos de regressão PLS construídos, baseados nos espectros com e sem pré-processamento, com faixas espectrais de 410-2500 nm, 441-1551 nm e 1100-1600 nm e com a seleção de variáveis iSPA-PLS, iPLS e FFiPLS. Com isso, retirou-se os modelos com o teste de bias significativo, onde $t_{calculado} > t_{crítico}$.

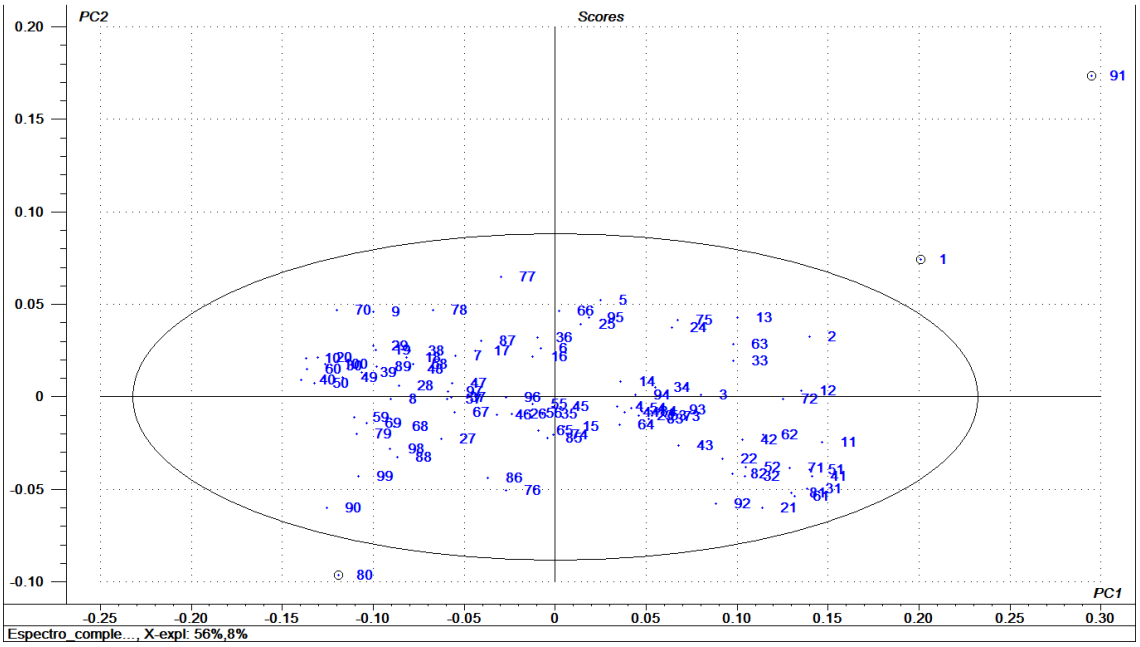
Para escolha do modelo, para o índice de biodiesel em diesel, os modelos pré-processados (1D2P17J), com teste de bias não significativo, apresentaram menores valores de variáveis latentes (entre 1 e 2), restringindo as faixas espectrais de 410-2500 nm (espectro completo) e de 441-1551 nm.

4.4.1 Faixa 410-2500 nm (1D2P17J) – índice de biodiesel em diesel

4.4.1.1 Análise Exploratória - Faixa 410-2500 nm (1D2P17J)

A partir do gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados pré-processados, aplicando a 1ª derivada Savitzky-Golay, polinômio de 2º grau e janela de 17 pontos (1D2P17J), tendo como faixa espectral o espectro completo. Observa-se, a partir da Figura 8, que três amostras (1, 80 e 91) estão fora da elipse da medida T^2 de Hotelling.

Figura 8: Gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados pré-processados (1D2P17J) com faixa de espectro completo.

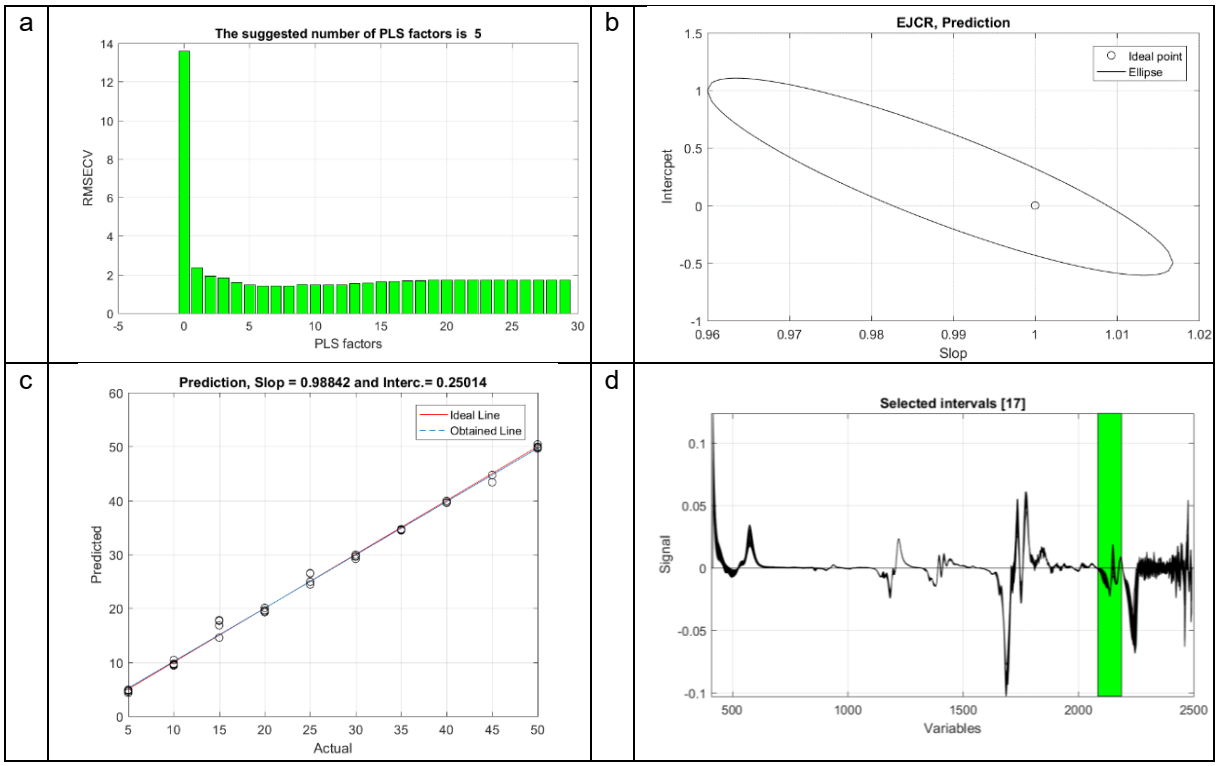


Fonte: Própria, 2021.

4.4.1.2 Modelo iSPA-PLS e iPLS - Faixa 410-2500 nm (1D2P17J)

Utilizando as 58 amostras do conjunto de calibração, os modelos iSPA-PLS e iPLS foram construídos, com toda a faixa espectral (410-2500 nm) e com pré-processamento espectral (1D2P17J), para o parâmetro índice de biodiesel em diesel. O espectro foi dividido em 20 intervalos para seleção de variáveis e foi escolhido o a partir do gráfico de estimativa que calcula o número estimado de variáveis latentes considerando o modelo *full* PLS, a partir do menor RMSECV (Figura 9 a). O modelo iSPA-PLS, para esta faixa espectral, foi construído com duas variáveis latentes. Foi utilizada a validação interna *full cross validation*. O gráfico EJCR (Região elíptica de confiança conjunta, do inglês *Elliptical joint confidence region*) será mostrado a fim de identificar se o ponto ideal estará na elipse de confiança para a predição (Figura 9 b). Posteriormente, o gráfico valor predito *versus* valor real, para predição (Figura 9 c). Os modelos iSPA-PLS e iPLS selecionaram o mesmo intervalo (intervalo 17 – faixa espectral de 2085-2188 nm) (Figura 9 d).

Figura 9: a. Gráfico de estimativa do número de fatores sugeridos pelo modelo iSPA-PLS e iPLS para a propriedade índice de biodiesel em diesel; b. EJCR; c. Predito x Real; d. Intervalo selecionado.

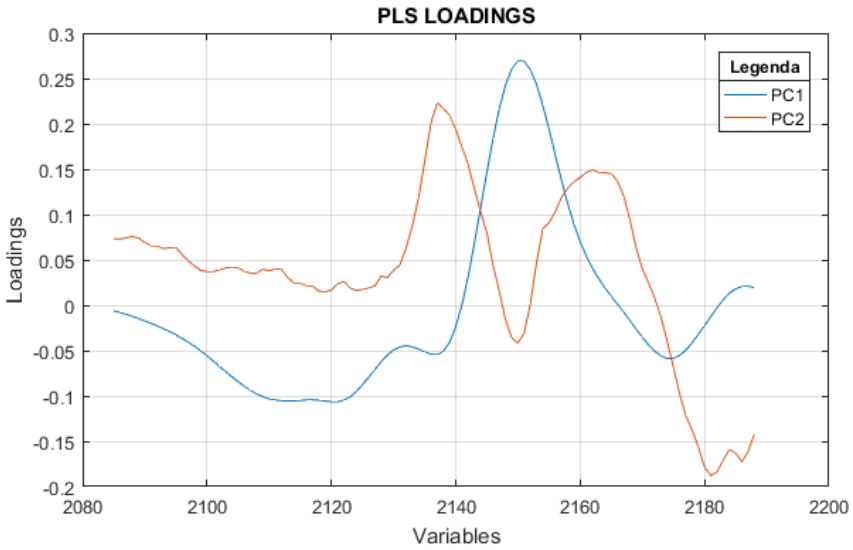


Fonte: Própria, 2021.

O gráfico EJCR mostra que caso o ponto (0,1), onde a inclinação (do inglês slop) é zero e o intercepto (do inglês intercept) é um, denominado ponto ideal, está dentro da elipse de confiança, o teste de bias está ausente (GONZÁLEZ et al., 1999), ou seja, o teste de bias não é significativo, o que coaduna com os valores de $t_{calculado}(0,383) < t_{crítico}(1,686)$. As amostras se apresentam próximas da reta de ajuste, com valores de $R^2_{predição}$ igual a 0,9807, RMSEP igual a 0,8996 e REP igual a 3,4062%.

O gráfico de loadings, para a faixa espectral de 2085-2188 nm, com duas PCs mostra que há variabilidade nos espectros, porém não há ruído (Figura 10).

Figura 10: a. Gráfico de loadings com duas PCs para os modelos iSPA-PLS e iPLS para a propriedade índice de biodiesel em diesel, com faixa de 2085-2188 nm, com pré-processamento.

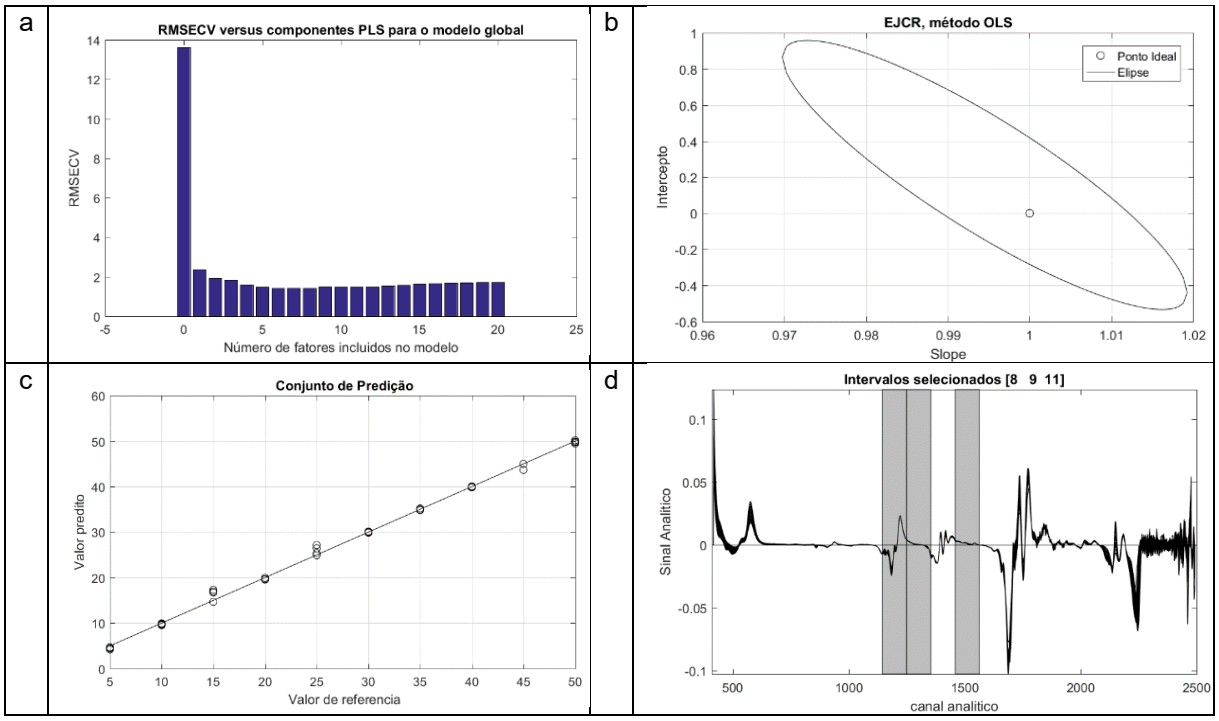


Fonte: Própria, 2021.

4.4.1.3 Modelo FFiPLS - Faixa 410-2500 nm (1D2P17J)

Utilizando as 58 amostras do conjunto de calibração, os modelos FFiPLS foram construídos, com toda a faixa espectral (410-2500 nm) e com pré-processamento espectral (1D2P17J), para o parâmetro índice de biodiesel em diesel. O espectro foi dividido em 20 intervalos para seleção de variáveis e foi escolhido o a partir do gráfico de estimativa que calcula o número estimado de variáveis latentes considerando o modelo *full* PLS, a partir do menor RMSECV (Figura 11 a). O modelo FFiPLS, para esta faixa espectral, foi construído com uma variável latente. Foi utilizada a validação interna *full cross validation*. O gráfico EJCR será mostrado a fim de identificar se o ponto ideal estará na elipse de confiança para a predição (Figura 11 b). Posteriormente, o gráfico valor predito *versus* valor real, para predição (Figura 11 c). O modelo FFiPLS selecionou os intervalos 8, 9 e 11 (faixa espectral de 1145-1249 nm, 1250-1354 nm e 1460-1564 nm, respectivamente) (Figura 11 d). O modelo considerado foi o FFiPLS repetição 03.

Figura 11: a. Gráfico de estimativa do número de fatores sugeridos pelo modelo FFiPLS para a propriedade índice de biodiesel em diesel; b. EJCR; c. Predito x Real; d. Intervalo selecionado.

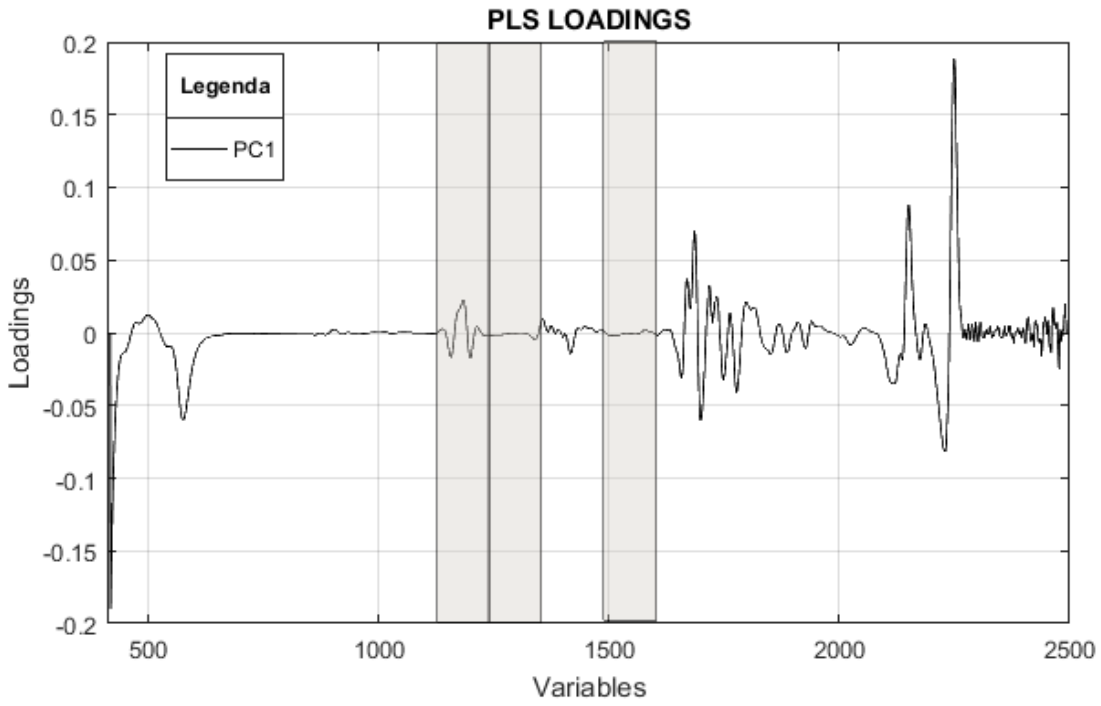


Fonte: Própria, 2021.

O gráfico EJCR mostra que o ponto ideal está dentro da elipse de confiança, e portanto o teste de bias não é significativo (GONZÁLEZ et al., 1999), o que coaduna com os valores de $t_{calculado}(0,532) < t_{crítico}(1,686)$. As amostras se apresentam próximas da reta de ajuste, com valores de $R^2_{predição}$ igual a 0,9972, RMSEP igual a 0,7753 e REP igual a 2,9356%.

O gráfico de loadings, para os intervalos 8, 9 e 11, faixa espectral de 1145-1249 nm, 1250-1354 nm e 1460-1564 nm, respectivamente, com uma PC mostra que a região ruidosa não está selecionada, apesar de grande variabilidade (Figura 12).

Figura 12: a. Gráfico de loadings com uma PC para o modelo FFiPLS para a propriedade índice de biodiesel em diesel, para os três intervalos selecionados, com pré-processamento.



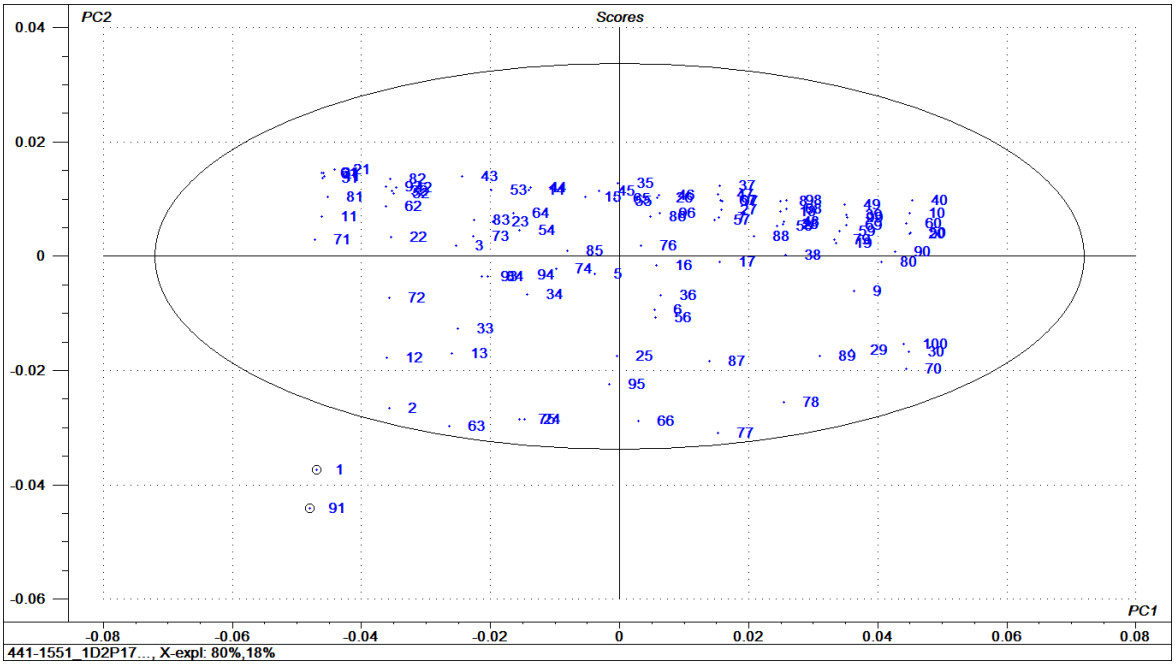
Fonte: Própria, 2021.

4.4.2 Faixa 441-1551 nm (1D2P17J) – índice de biodiesel em diesel

4.4.2.1 Análise Exploratória - Faixa 441-1551 nm (1D2P17J)

A partir do gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados pré-processados, aplicando a 1ª derivada Savitzky-Golay, polinômio de 2º grau e janela de 17 pontos (1D2P17J), tendo como faixa espectral de 441-1551 nm. Observa-se, a partir da Figura 13, que três amostras (1, 91) estão fora da elipse da medida T² de Hotelling, conforme.

Figura 13: Gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados pré-processados (1D2P17J) com faixa espectral de 441-1551 nm.

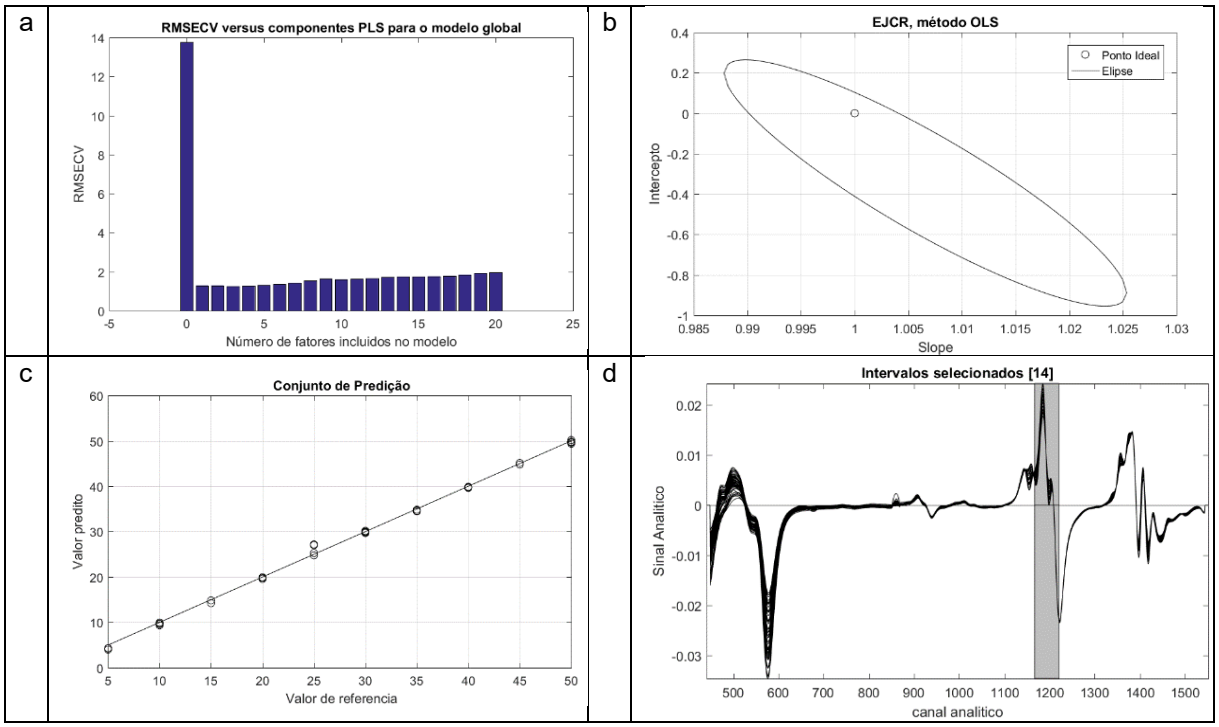


Fonte: Própria, 2021.

4.4.2.2 Modelo iSPA-PLS, iPLS e FFiPLS- Faixa 441-1551 nm (1D2P17J)

Utilizando as 59 amostras do conjunto de calibração, os modelos iSPA-PLS, Ipls e FFiPLS foram construídos, com a faixa espectral (441-1551 nm) e com pré--processamento espectral (1D2P17J), para o parâmetro índice de biodiesel em diesel. O espectro foi dividido em 20 intervalos para seleção de variáveis e foi escolhido o a partir do gráfico de estimativa que calcula o número estimado de variáveis latentes considerando o modelo *full* PLS, a partir do menor RMSECV (Figura 14 a). O modelo iSPA-PLS, iPLS e FFiPLS para esta faixa espectral, foi construído com uma variável latente. Foi utilizada a validação interna *full cross validation*. O gráfico EJCR será mostrado a fim de identificar se o ponto ideal estará na elipse de confiança para a predição (Figura 14 b). Posteriormente, o gráfico valor predito *versus* valor real, para predição (Figura 14 c). Os modelos iSPA-PLS, iPLS e FFiPLS selecionaram o mesmo intervalo (intervalo 14 – faixa espectral de 1167-1221 nm) (Figura 14 d). O modelo FFiPLS considerado foi o FFiPLS repetição 08.

Figura 14: a. Gráfico de estimativa do número de fatores sugeridos pelo modelo FFiPLS para a propriedade índice de biodiesel em diesel; b. EJCR; c. Predito x Real; d. Intervalo selecionado.

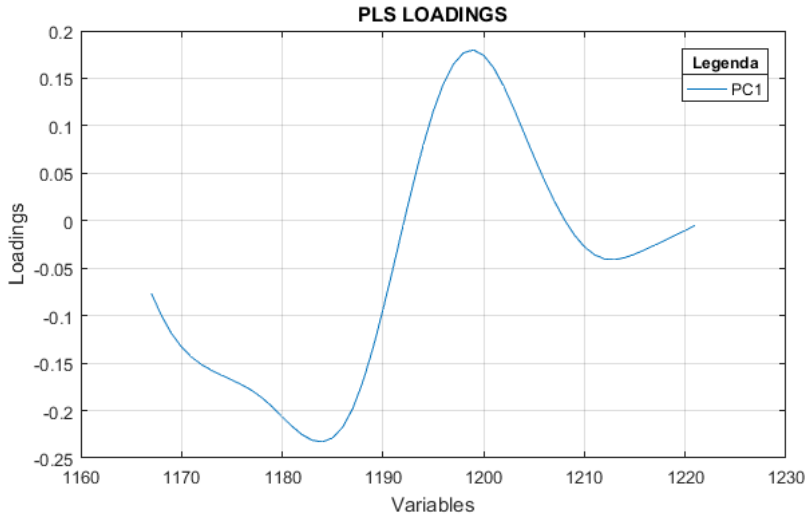


Fonte: Própria, 2021.

O gráfico EJCR mostra que o ponto ideal está dentro da elipse de confiança, portanto teste de bias não é significativo (GONZÁLEZ et al., 1999), o que coaduna com os valores de $t_{calculado}(1,5976) < t_{crítico}(1,686)$. As amostras se apresentam próximas da reta de ajuste, com valores de $R^2_{predição}$ igual a 0,9984, RMSEP igual a 0,6144 e REP igual a 2,1204%.

O gráfico de loadings, para o intervalo 14, faixa espectral de 1167-1221 nm, mostra que há variabilidade nos espectros, porém não há ruído (Figura 15).

Figura 15: a. Gráfico de loadings com uma PCs para os modelos iSPA-PLS, iPLS e FFiPLS para a propriedade índice de biodiesel em diesel, com faixa de 1167-1221 nm, com pré-processamento.



Fonte: Própria, 2021.

4.4.3 Comparação entre métricas - índice de biodiesel em diesel

A Tabela 3 compara os algoritmos de seleção de variáveis para a propriedade Índice de biodiesel em diesel através das figuras de mérito LV, RMSEC, $R^2_{calibração}$, RMSECV, $R^2_{validação}$, RMSEP e $R^2_{predição}$ e REP.

Tabela 3: Comparação de métricas obtidas com os espectros Vis-NIR na faixa de 410-2500 nm e 441-1551 nm com pré-processamento (1D2P17J) das amostras de biodiesel para a propriedade Índice de biodiesel em diesel.

Faixa Espectral	MODELOS	LV	RMSEC	$R^2_{calibração}$	RMSECV	$R^2_{validação}$	RMSEP	$R^2_{predição}$	REP
410-2500 nm	ISPA-PLS	2	1,1147	0,9936	1,1434	0,9910	0,8996	0,9807	3,4062%
	IPLS	2	1,1147	0,9936	1,1434	0,9910	0,8996	0,9807	3,4062%
	FFiPLS_R03	1	1,1146	0,9935	1,1226	0,9932	0,7753	0,9972	2,9356%
441-1551 nm	ISPA-PLS	1	1,2333	0,9922	1,2450	0,9918	0,6144	0,9984	2,1204%
	IPLS	1	1,2333	0,9922	1,2450	0,9918	0,6144	0,9984	2,1204%
	FFiPLS_R08	1	1,2333	0,9922	1,2450	0,9918	0,6144	0,9984	2,1204%

Fonte: Própria, 2021.

Como se pode observar, os modelos apresentaram baixos valores de variáveis latentes, com predominância dos modelos para a faixa espectral de 441-1551 nm com pré-processamento espectral. Apesar dos modelos da faixa 441-1551 nm apresentarem maior RMSEC e menor $R^2_{calibração}$ que os modelos da faixa 410-2500 nm, exibem melhores figuras de mérito de predição, com menor RMSEP e REP para

os algoritmos ISPA-PLS, IPLS e FFIPLS. Já para a faixa 410-2500 nm, o modelo FFiPLS apresentou menor variáveis latentes, menor RMSECV, RMSEP e REP e maiores $R^2_{validação}$ e $R^2_{predição}$ que os modelos iSPA-PLS e iPLS.

4.5 Modelos PLS – densidade

Serão apresentados os melhores modelos de regressão PLS, considerando o menor número de variáveis latentes, os maiores valores de $R^2_{calibração}$, $R^2_{validação}$, $R^2_{predição}$, e menores valores de RMSEC, RMSECV, RMSEP e menor REP.

Dentre os modelos de regressão PLS construídos, baseados nos espectros com e sem pré-processamento, com faixas espectrais de 410-2500 nm, 441-1551 nm e 1100-1600 nm e com a seleção de variáveis iSPA-PLS, iPLS e FFiPLS. Com isso, retirou-se os modelos com o teste de bias significativo, onde $t_{calculado} > t_{crítico}$.

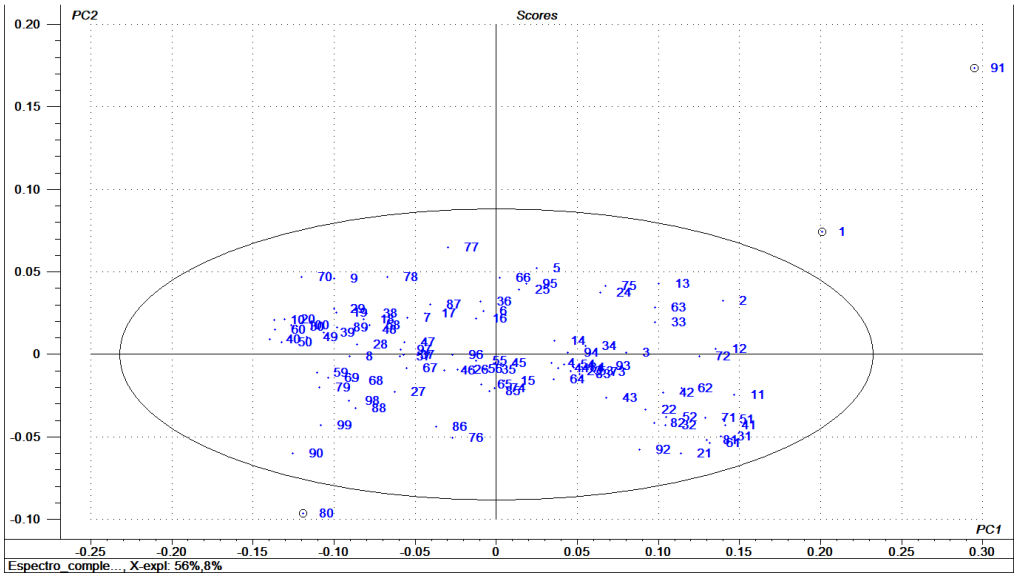
Para escolha do modelo, para o parâmetro de interesse densidade, os modelos pré-processados (1D2P17J) com faixa espectral de 410-2500 nm (espectro completo), e sem pré-processamento, com faixa espectral de 1100-1600 nm, com teste de bias não significativo, apresentaram menores valores de variáveis latentes (3).

4.5.1 Faixa 410-2500 nm (1D2P17J) – densidade

4.5.1.1 Análise Exploratória - Faixa 410-2500 nm (1D2P17J)

A partir do gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados pré-processados, aplicando a 1ª derivada Savitzky-Golay, polinômio de 2º grau e janela de 17 pontos (1D2P17J), tendo como faixa espectral o espectro completo. Observa-se, a partir da Figura 16, que três amostras (1, 80 e 91) estão fora da elipse da medida T^2 de Hotelling.

Figura 16: Gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados pré-processados (1D2P17J) com faixa de espectro completo.

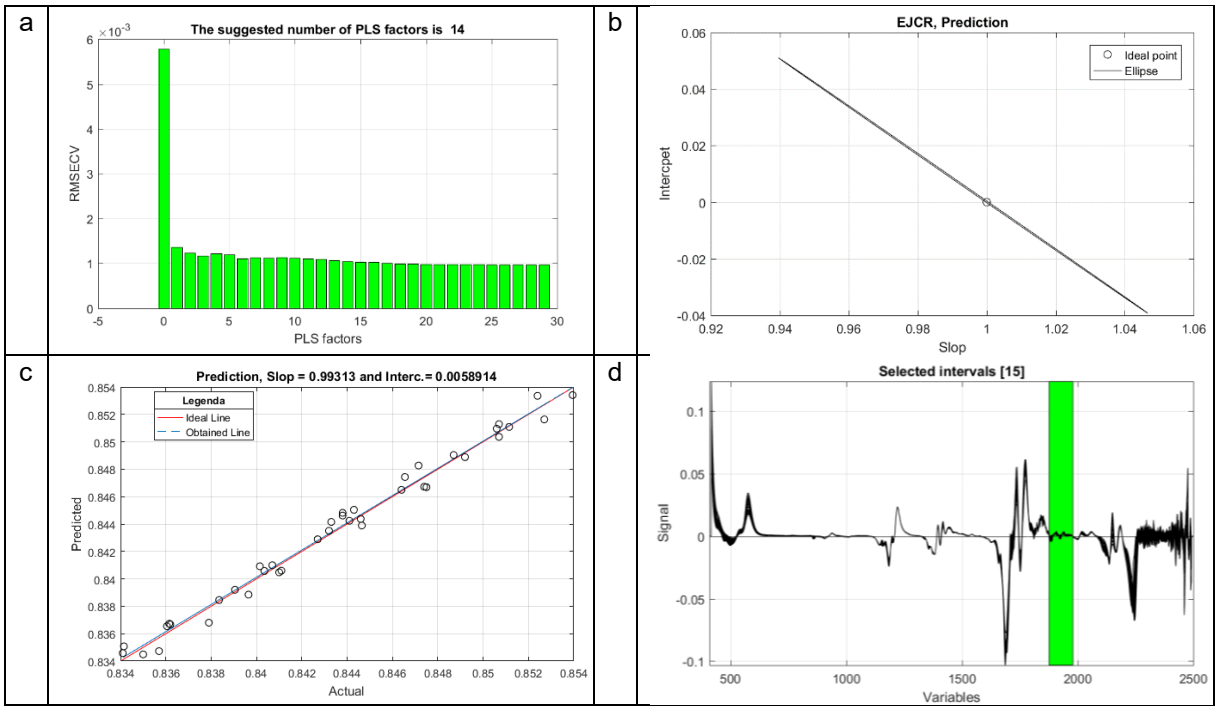


Fonte: Própria, 2021.

4.5.1.2 Modelo iSPA-PLS e iPLS - Faixa 410-2500 nm (1D2P17J)

Utilizando as 57 amostras do conjunto de calibração, os modelos iSPA-PLS e iPLS foram construídos, com toda a faixa espectral (410-2500 nm) e com pré-processamento espectral (1D2P17J), para o parâmetro densidade. O espectro foi dividido em 20 intervalos para seleção de variáveis e foi escolhido o a partir do gráfico de estimativa que calcula o número estimado de variáveis latentes considerando o modelo *full* PLS, a partir do menor RMSECV (Figura 17 a). O modelo iSPA-PLS e o iPLS, para esta faixa espectral, foi construído com três variáveis latentes. Foi utilizada a validação interna *full cross validation*. O gráfico EJCR será mostrado a fim de identificar se o ponto ideal estará na elipse de confiança para a predição (Figura 17 b). Posteriormente, o gráfico valor predito *versus* valor real, para predição (Figura 17 c). Os modelos iSPA-PLS e iPLS selecionaram o mesmo intervalo (intervalo 15 – faixa espectral de 1877-1980 nm) (Figura 17 d).

Figura 17: a. Gráfico de estimativa do número de fatores sugeridos pelo modelo iSPA-PLS e iPLS para a propriedade densidade; b. EJCR; c. Predito x Real; d. Intervalo selecionado.

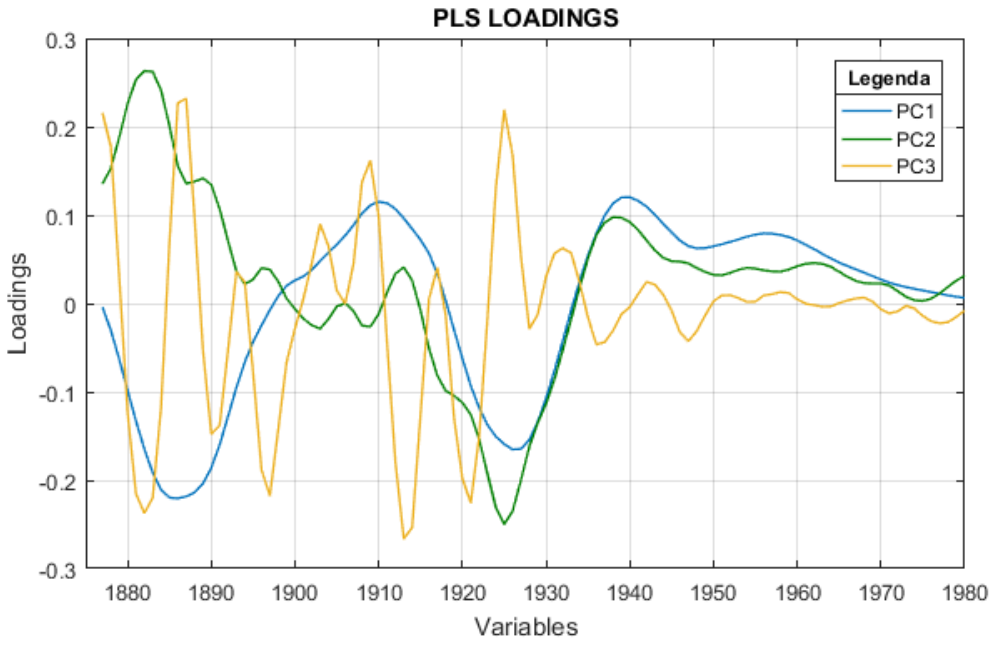


Fonte: Própria, 2021.

O gráfico EJCR mostra que o ponto ideal está dentro da elipse de confiança, portanto teste de bias não é significativo (GONZÁLEZ et al., 1999), o que coaduna com os valores de $t_{calculado}(0,8715) < t_{crítico}(1,686)$, ou seja, o teste de bias não é significativo. As amostras se apresentam próximas da reta de ajuste, com valores de $R^2_{predição}$ igual a 0,9931, RMSEP igual a 6.10^{-4} e REP igual a 0,0766%.

O gráfico de loadings, para a faixa espectral de 1877-1980 nm, com três PCs mostra que há variabilidade nos espectros, apesar da PC3 adicionar ruído aos loadings (Figura 18).

Figura 18: a. Gráfico de loadings com três PCs para os modelos iSPA-PLS e iPLS para a propriedade densidade, com faixa de 1877-1980 nm, com pré-processamento.

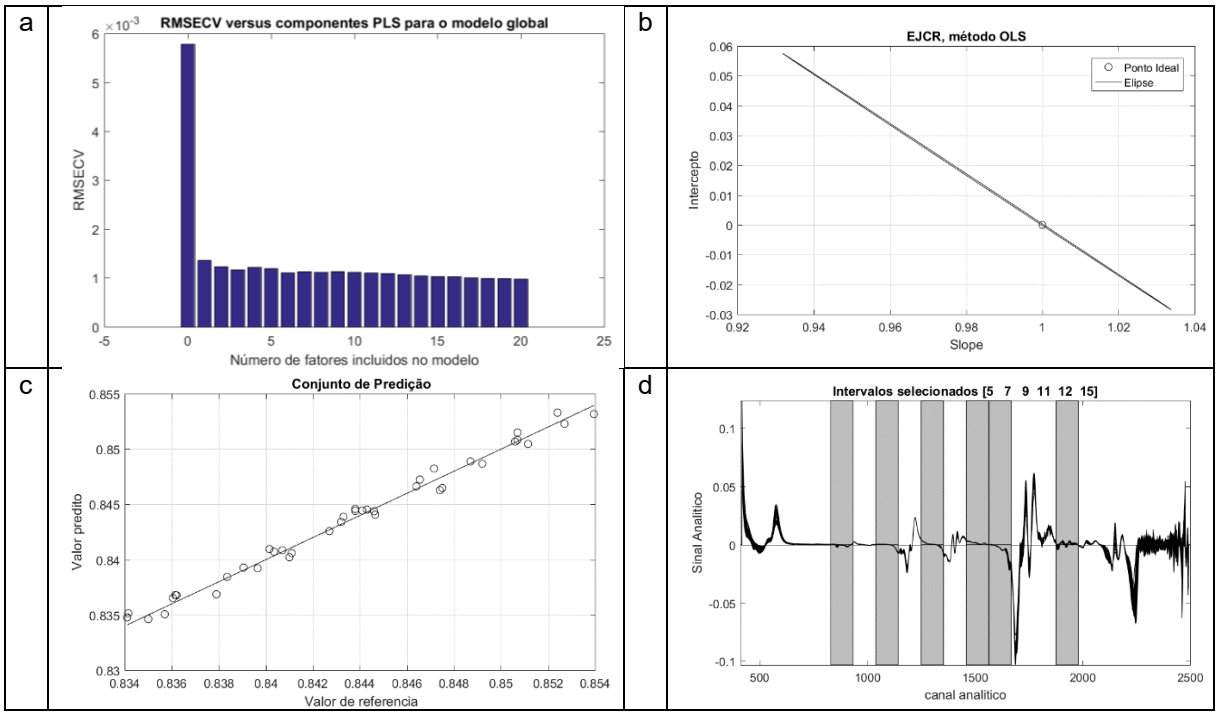


Fonte: Própria, 2021.

4.5.1.3 Modelo FFiPLS - Faixa 410-2500 nm (1D2P17J)

Utilizando as 57 amostras do conjunto de calibração, os modelos FFiPLS foram construídos, com toda a faixa espectral (410-2500 nm) e com pré-processamento espectral (1D2P17J), para a densidade. O espectro foi dividido em 20 intervalos para seleção de variáveis e foi escolhido o a partir do gráfico de estimativa que calcula o número estimado de variáveis latentes considerando o modelo *full* PLS, a partir do menor RMSECV (Figura 19 a). O modelo FFiPLS, para esta faixa espectral, foi construído com três variáveis latentes. Foi utilizada a validação interna *full cross validation*. O gráfico EJCR será mostrado a fim de identificar se o ponto ideal estará na elipse de confiança para a predição (Figura 19 b). Posteriormente, o gráfico valor predito *versus* valor real, para predição (Figura 19 c). O modelo FFiPLS selecionou os intervalos 5, 7, 9, 11, 12, e 15 (faixa espectral de 830-934 nm, 1040-1144 nm, 1250-1354 nm, 1460-1564 nm, 1565-1668 nm e 1877-1980 nm, respectivamente) (Figura 19 d). O modelo considerado foi o FFiPLS repetição 02.

Figura 19: a. Gráfico de estimativa do número de fatores sugeridos pelo modelo FFiPLS para a propriedade densidade; b. EJCR; c. Predito x Real; d. Intervalo selecionado.

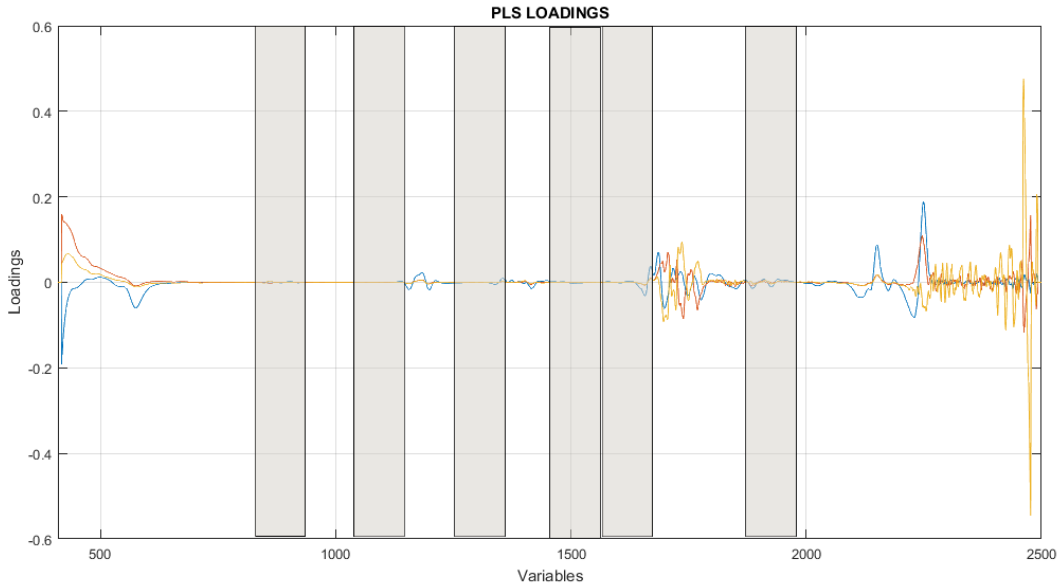


Fonte: Própria, 2021.

O gráfico EJCR mostra que o ponto ideal está dentro da elipse de confiança, e portanto o teste de bias não é significativo (GONZÁLEZ et al., 1999), o que coaduna com os valores de $t_{calculado}(0,6841) < t_{crítico}(1,686)$. As amostras se apresentam próximas da reta de ajuste, com valores de $R^2_{predição}$ igual a 0,9879, RMSEP igual a $6,2.10^{-4}$ e REP igual a 0,0735%.

O gráfico de loadings, para os intervalos 5, 7, 9, 11, 12, e 15, com três PCs mostra que a região ruidosa não está selecionada, apesar de grande variabilidade (Figura 20).

Figura 20: a. Gráfico de loadings com uma PC para o modelo FFiPLS para a propriedade densidade, para os seis intervalos selecionados, com pré-processamento.



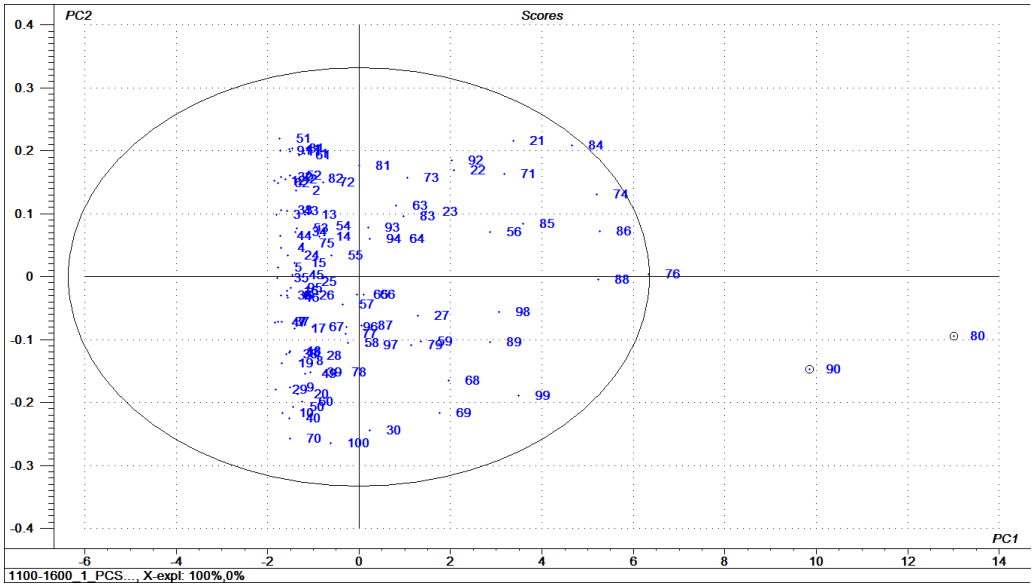
Fonte: Própria, 2021.

4.5.2 Faixa 1100-1600 nm – densidade

4.5.2.1 Análise Exploratória - Faixa 1100-1600 nm

A partir do gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados não pré-processados, aplicando a 1ª derivada Savitzky-Golay, polinômio de 2º grau e janela de 17 pontos (1D2P17J), tendo como faixa espectral de 1100-1600 nm. Observa-se, a partir da Figura 21, que duas amostras (80, 90) estão fora da elipse da medida T² de Hotelling, conforme Figura 21.

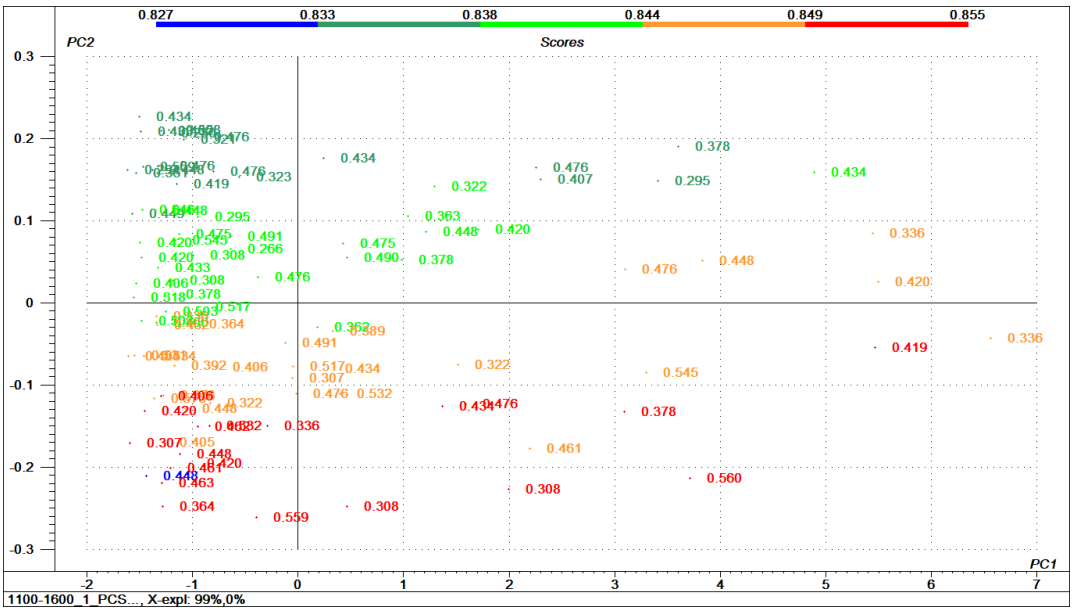
Figura 21: Gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir dos dados não pré-processados com faixa espectral de 1100-1600 nm.



Fonte: Própria, 2021.

E a partir do gráfico de scores, observa-se que para os valores de densidade, a amostra 10 está separada das demais, marcada em azul, sendo considerado um outlier, conforme Figura 22.

Figura 22: Gráfico de scores das 100 amostras de misturas diesel/biodiesel a partir da faixa espectral de 1100-1600 nm sem pré processamento espectral para a propriedade densidade.

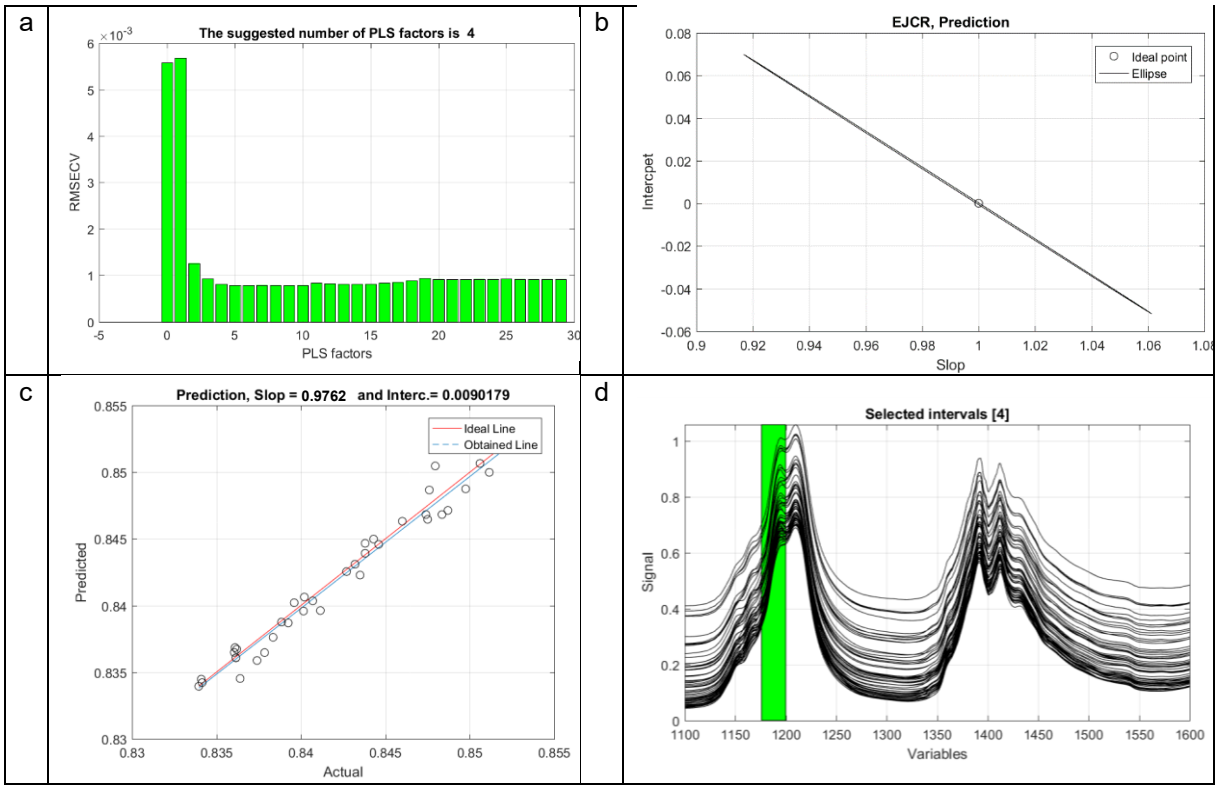


Fonte: Própria, 2021.

4.5.2.2 Modelos iPLS e FFiPLS - Faixa 1100-1600 nm

Utilizando as 58 amostras do conjunto de calibração, o modelo iPLS foi construído, com a faixa espectral (1100-1600 nm) e sem pré-processamento espectral, para o parâmetro densidade. O espectro foi dividido em 20 intervalos para seleção de variáveis e foi escolhido o a partir do gráfico de estimativa que calcula o número estimado de variáveis latentes considerando o modelo *full* PLS, a partir do menor RMSECV (Figura 23 a). Os modelos iPLS e FFiPLS para esta faixa espectral, foi construído com três variáveis latentes. Foi utilizada a validação interna *full cross validation*. O gráfico EJCR será mostrado a fim de identificar se o ponto ideal estará na elipse de confiança para a predição (Figura 23 b). Posteriormente, o gráfico valor predito *versus* valor real, para predição (Figura 23 c). Os modelos iPLS e FFiPLS selecionaram o intervalo 4 – faixa espectral de 1176-1200 nm) (Figura 23 d).

Figura 23: a. Gráfico de estimativa do número de fatores sugeridos pelo modelo iPLS e FFiPLS para a propriedade densidade; b. EJCR; c. Predito x Real; d. Intervalo selecionado.



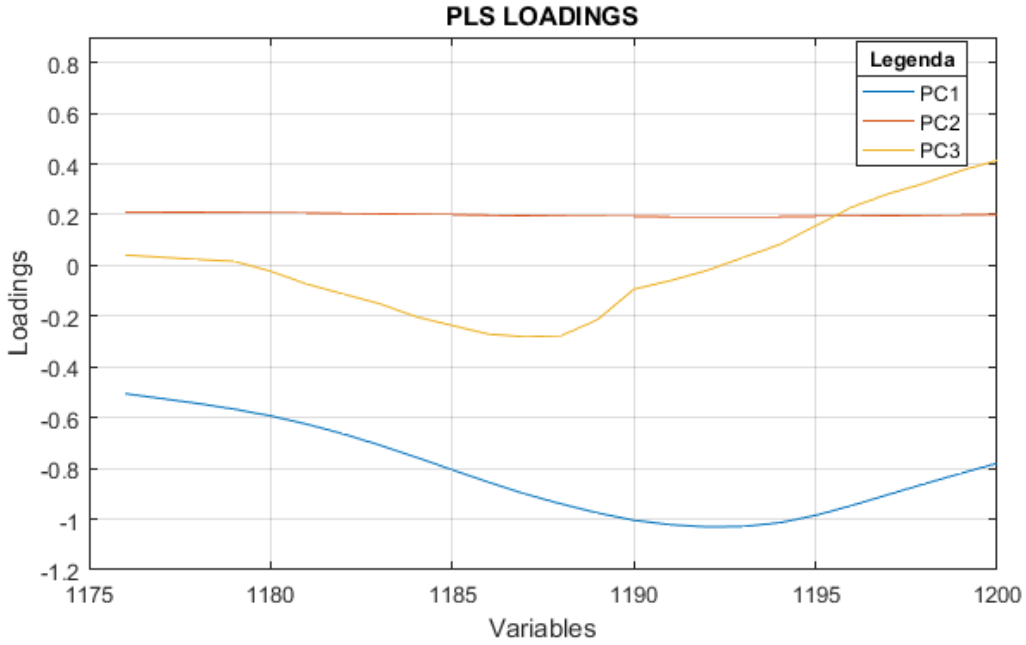
Fonte: Própria, 2021.

O gráfico EJCR mostra que o ponto ideal está dentro da elipse de confiança, portanto teste de bias não é significativo (GONZÁLEZ et al., 1999), o que coaduna

com os valores de $t_{calculado}(1,6789) < t_{crítico}(1,686)$. As amostras se apresentam próximas da reta de ajuste, com valores de $R^2_{predição}$ igual a 0,9762, RMSEP igual a $9,2.10^{-4}$ e REP igual a 0,1095%.

O gráfico de loadings, para o intervalo 4, faixa espectral de 1176-1200 nm, mostra que há variabilidade nos espectros, com três PCs, porém não há ruído (Figura 24).

Figura 24: a. Gráfico de loadings com uma PCs para o modelo iPLS para a propriedade densidade, com faixa de 1176-1200 nm, sem pré-processamento.



Fonte: Própria, 2021.

4.5.3 Comparação entre métricas - densidade

A Tabela 3 compara os algoritmos de seleção de variáveis para a propriedade densidade através das figuras de mérito LV, RMSEC, $R^2_{calibração}$, RMSECV, $R^2_{validação}$, RMSEP e $R^2_{predição}$ e REP.

Tabela 4: Comparação de métricas obtidas com os espectros Vis-NIR na faixa de 410-2500 nm com pré-processamento espectral (1D2P17J) e 1100-1600 nm sem pré-processamento espectral das amostras de biodiesel para a propriedade densidade.

Faixa Espectral	MODELOS	LV	RMSEC	$R^2_{calibração}$	RMSECV	$R^2_{validação}$	RMSEP	$R^2_{predição}$	REP
410-2500 nm	ISPA-PLS	3	$8{,}1.10^{-4}$	0,9820	$8{,}6.10^{-4}$	0,9890	$6{,}5.10^{-4}$	0,9871	0,0766%
	IPLS	3	$8{,}1.10^{-4}$	0,9820	$8{,}6.10^{-4}$	0,9890	$6{,}5.10^{-4}$	0,9871	0,0766%
	FFiPLS_R02	3	$8{,}3.10^{-4}$	0,9810	$8{,}8.10^{-4}$	0,9768	$6{,}2.10^{-4}$	0,9879	0,0735%
1100-1600 nm	ISPA-PLS	4	$6{,}8.10^{-4}$	0,9862	$7{,}3.10^{-4}$	0,9860	$8{,}6.10^{-4}$	0,9790	0,1018%
	IPLS	3	$7{,}6.10^{-4}$	0,9827	$7{,}9.10^{-4}$	0,9798	$9{,}2.10^{-4}$	0,9762	0,1095%
	FFiPLS_R01	3	$7{,}6.10^{-4}$	0,9827	$7{,}9.10^{-4}$	0,9798	$9{,}2.10^{-4}$	0,9762	0,1095%

Fonte: Própria, 2021.

Como se pode observar, os modelos apresentaram baixos valores de variáveis latentes, para a faixa espectral de 410-2500 nm com pré-processamento espectral e para 1100-1600 sem pré-processamento espectral. Apesar dos modelos da faixa 410-2500 nm, com pré-processamento espectral, apresentarem maior RMSEC e RMSECV que os modelos da faixa 1100-1600 nm, sem pré-processamento espectral, exibem melhores figuras de mérito de predição, com menor RMSEP e REP e maiores $R^2_{predição}$ para os algoritmos ISPA-PLS, IPLS e FFIPLS.

5 CONCLUSÕES

Os algoritmos iSPA-PLS, iPLS e FFiPLS, foram comparados e utilizados em banco de dados de biodiesel, com dados espectrais NIR e a partir de dois parâmetros de interesse: índice de biodiesel em diesel e densidade. Os modelos de regressão foram construídos com pré-processamento e sem pré-processamento espectral, na faixa 410-2500 nm, na faixa de 441-1551 nm e na faixa de 1100-1600 nm.

Para o índice de biodiesel em diesel, ambas as faixas com pré-processamento espectral apresentaram baixos valores de variáveis latentes (faixa de 410-2500 nm e faixa de 441-1551 nm), com melhores valores para a etapa de calibração e validação (RMSEC, RMSECV, $R^2_{calibrção}$ e $R^2_{validação}$) para o modelo FFiPLS (faixa de 410-2500 nm) que os modelos iSPA-PLS e iPLS na faixa 441-1551 nm, e com etapa de predição (RMSEP, $R^2_{predição}$ e REP) com melhores valores para os três algoritmos de seleção de variáveis, iSPA-PLS, iPLS e FFiPLS, na faixa 441-1551 nm.

Já para a propriedade densidade, a faixa de 410-2500 nm com pré-processamento espectral apresentou melhores valores na etapa de predição (RMSEP, $R^2_{predição}$ e REP) que na faixa 1100-1600 nm sem pré-processamento. Para a faixa 1100-1600 nm, os modelos apresentam valores aproximados, com exceção do modelo iSPA-PLS, que apresentou uma variável a mais que o iPLS e o FFiPLS.

As regiões NIR cos picos entre 1400, correspondente ao primeiro sobretom das bandas de combinação de ligações C–H e os picos em torno de 1200 nm equivalente ao segundo sobretom dos modos de estiramento das ligações C–H foram selecionadas pelos algoritmos iSPA-PLS, iPLS e FFiPLS nos modelos de regressão para os dois parâmetros de interesse, mostrando correlação entre os espectros NIR analisados. Para a faixa completa (410-2500 nm) os algoritmos seleção de variáveis, iSPA-PLS, iPLS e FFiPLS, não selecionaram as regiões de ruídos de alta frequência tampouco a região de alta absorbância, afirmando a capacidade do algoritmo em não selecionar variáveis e intervalos de variáveis que compreendem regiões onde possivelmente não evidenciam correlação com o parâmetro de interesse.

Com isso, o algoritmo FFiPLS apresenta-se como ferramenta interessante para seleção de variáveis para a construção de modelos de calibração multivariada baseados em PLS para amostras de biodiesel, comparadas aos modelos iSPA-PLS e iPLS, com principal vantagem na determinação a partir de espectro completo e pré-processado.

REFERÊNCIAS

ANDERSEN, C.M; BRO, R. Variable selection in regression—a tutorial. **Journal of Chemometrics**, v. 24, p. 728-737, 2010. doi: 10.1002/cem.1360

ATTIA, K. A. M.; NASSAR, M. W. I.; EL-ZEINY, M. B.; SERAG, A. Firefly algorithm versus genetic algorithm as powerful variable selection tools and their effect on different multivariate calibration models in spectroscopy: A comparative study. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 170, p. 117–123, 2017. doi:10.1016/j.saa.2016.07.016

BARROS NETO, Benício de; SCARMINIO, Ieda S; BRUNS, Roy E.; 25 anos de quimiometria no Brasil. **Quím. Nova** [online], v. 29, n. 6, p.1401-1406. 2006. doi:10.1590/S0100-40422006000600042

BRERETON, R. G. **Applied chemometrics for scientists**. John Wiley & Sons, 2007. 379 p.

BRERETON, R. G. **Chemometrics: Data Analysis for the Laboratory and Chemical Plant**. John Wiley & Sons, USA, 2003. 489 p.

BRERETON, R. G. **Chemometrics: Data Driven Extraction for Science**. John Wiley & Sons, USA, 2018. 439 p.

CNPE. Resolução CNPE nº 16, de 29 de outubro 2018. Dispõe sobre a evolução da adição obrigatória de biodiesel ao óleo diesel vendido ao consumidor final, em qualquer parte do território nacional. **Diário Oficial da União**, Brasília, DF, 30 dez. 2017. Disponível em: http://www.mme.gov.br/documents/10584/71068545/Resolucao _16_CNPE_29-10-18.pdf. Acesso em: 02 fev. 2021.

DARWISH, A. Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications. **Future Computing and Informatics Journal**, v. 3, n. 2, p. 231–246, 2018. doi:10.1016/j.fcij.2018.06.001

FERNANDES, D. D. S.; GOMES, A. A.; COSTA, G. B.; SILVA, G. W. B.; VERAS, G. Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection. **Talanta**, v. 87, p. 30–34, 2011. doi: 10.1016/j.talanta.2005.03.025

FERNANDES, D. D. S. **Espectroscopia uv-vis para avaliação de biodiesel e misturas biodiesel/diesel**. Dissertação (Mestrado em Ciências Agrárias) - Universidade Estadual da Paraíba, Campina Grande, 2013. 75p.

FERNANDES, D. D. S. **Novas estratégias para seleção de variáveis por intervalos em problemas de classificação**. Tese (Doutorado em Química) - Universidade Federal da Paraíba, João Pessoa, 2016. 137p.

FISTER, I.; FISTER, I.; YANG, X.-S.; BREST, J. A comprehensive review of firefly algorithms. **Swarm and Evolutionary Computation**, v. 13, p. 34–46, 2013. doi:10.1016/j.swevo.2013.06.001

GALVÃO, R. K. H.; ARAÚJO, M. C. U.; JOSÉ, G. E.; PONTES, M. J. C.; SILVA, E. C.; SALDANHA, T. C. B. A method for calibration and validation subset partitioning. **Talanta**, v. 67, p. 736-740, 2005. doi: 10.1016/j.talanta.2005.03.025

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Analytica Chimica Acta**, v. 185, p. 1–17, 1986. doi:10.1016/0003-2670(86)80028-9

GOMES, A. A. **Algoritmo das projeções sucessivas aplicado à seleção de variáveis em regressão PLS**. Dissertação (Mestrado em Química) - Universidade Federal da Paraíba, João Pessoa, 2012. 120p.

GOMES, A. A.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; VÉRAS, G.; SILVA, E. C. The successive projections algorithm for interval selection in PLS. **Microchemical Journal**, v. 110, p. 202–208, 2013. doi:10.1016/j.microc.2013.03.015

GONZÁLEZ, A.G. HERRADOR, M. A. ASUERO, A. G. Intra-laboratory testing of method accuracy from recovery assays. Talanta, v. 48, p. 729-736, 1999. doi: 10.1016/S0039-9140(98)00271-9

GOODARZI, M.; COELHO, L. S. Firefly as a novel swarm intelligence variable selection method in spectroscopy. **Analytica Chimica Acta**, v. 852, p. 20–27, 2014. doi:10.1016/j.aca.2014.09.045

HONORATO, F. A. **Previsão de propriedades das gasolinas do Nordeste empregando espectroscopia NIR/MIR e transferência de calibração**. Tese (Doutorado em Química) - Universidade Federal de Pernambuco, Recife, 2006. 91 p.

KENNARD, R. W., & STONE, L. A. Computer Aided Design of Experiments. **Technometrics**, v. 11, n. 1 p. 137-148, 1969. doi:10.2307/1266770

KREPPER, G.; ROMEO, F.; FERNANDES, D. D. DE S.; DINIZ, P. H. G. D.; ARAÚJO, M. C. U.; DI NEZIO, M. S.; CENTURIÓN, M. E. Determination of fat content in chicken hamburgers using NIR spectroscopy and the Successive Projections Algorithm for interval selection in PLS regression (*i*SPA-PLS). **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 189, p. 300–306, 2018. doi:10.1016/j.saa.2017.08.046

LI, W.; HAN, H.; ZHANG, L.; ZHANG, Y.; QU, H. A feasibility study on the non-invasive analysis of bottled Compound E Jiao oral liquid using near infrared spectroscopy. **Sensors and Actuators B: Chemical**, v. 211, p. 131-137, 2015. doi:10.1016/j.snb.2015.01.073

LINDEN, Ricardo. **Algoritmos genéticos**. 2 ed. Rio de Janeiro: Brasport, 2008. 402 p.

NØRGAARD, L.; SAUDLAND, A.; WAGNER, J.; NIELSEN, J. P.; MUNCK, L.; ENGELSEN, S. B. Interval Partial Least-Squares Regression (*i*PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. **Applied Spectroscopy**, v. 54, n. 3, p. 413–419, 2000. doi:10.1366/0003702001949500

NUNES, P. G. A. **Uma nova técnica para seleção de variáveis em calibração multivariada aplicada às espectrometrias UV-VIS e NIR**. Tese (Doutorado em Química) - Universidade Federal da Paraíba, João Pessoa, 2008. 106p.

OLIVEIRA, D. L. B.; PEREIRA, L. H. S.; SCHNEIDER, M. P.; SILVA, Y. J. A. B.; NASCIMENTO, C.W.A.; van STRAATEN, P.; SILVA, Y. J. A. B.; GOMES, A. A.; VÉRAS, G. Bio-inspired algorithm for variable selection in i-PLSR to determine physical properties, thorium and rare earth elements in soils from Brazilian semiarid region. **Microchemical Journal**, v. 160, Part A, p. 1-7, 2021. doi:10.1016/j.microc.2020.105640

PASQUINI, C. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. **Analytica Chimica Acta**, v. 1026, p. 8–36, 2018. doi:10.1016/j.aca.2018.04.004

PAULA, L. C. M.; SOARES, A. S.; LIMA, T. W.; DELBEM, A. C. B.; COELHO, C. J.; GALVÃO FILHO, A. R. A gpu-based implementation of the firefly algorithm for variable selection in multivariate calibration problems. **PloS one**, v. 9, n. 12, 2014. doi:10.1371/journal.pone.0114145

RAPHAEL, B.; SMITH, I. F. C. A direct stochastic algorithm for global search. **Applied Mathematics and computation**, v. 146, n. 2-3, p. 729-758, 2003. doi:10.1016/s0096-3003(02)00629-x

SENA, M. M.; ALMEIDA, M. R. Quimiometria aplicada aos dados espectrais no Infravermelho Próximo. TIBOLA, C. S. et al. (Org.). In: **Espectroscopia no Infravermelho Próximo para Avaliar Indicadores de Qualidade Tecnológica e Contaminantes em Grãos**. Brasília, DF: Embrapa, cap. 2. p. 31–50. 2018.

SHI, J.; HU, X.; ZOU, X.; ZHAO, J.; ZHANG, W.; HUANG, X.; ZHU, Y.; LI, Z.; XU, Y. A heuristic and parallel simulated annealing algorithm for variable selection in near-infrared spectroscopy analysis. **Journal of Chemometrics**, v. 30, n. 8, p. 442-450, 2016. doi:10.1002/cem.2812

SIMÕES, S. S. **Desenvolvimento de métodos validados para a determinação de captopril usando espectrometria NIRR e calibração multivariada**. Tese (Doutorado em Química) - Universidade Federal da Paraíba, João Pessoa, 2008. 83p.

STERNBERG, J. C.; STILLO, H. S.; SCHWENDEMAN, R. H. Spectrophotometric Analysis of Multicomponent Systems Using Least Squares Method in Matrix Form. Ergosterol Irradiation System. **Analytical Chemistry**, v. 32, n. 1, p. 84–90, 1960. doi:10.1021/ac60157a025

XIAOBO, Z.; JIEWEN, Z.; POVEY, M. J. W.; HOLMES, M.; HANPIN, M. Variables selection methods in near-infrared spectroscopy. **Analytica Chimica Acta**, v. 667, 14-32. 2010. doi:10.1016/j.aca.2010.03.048

XU, L.; FU, H.-Y.; GOODARZI, M.; CAI, C.-B.; YIN, Q.-B.; WU, Y.; TANG, B.-C.; SHE, Y.-B. Stochastic cross validation. **Chemometrics and Intelligent Laboratory Systems**, v. 175, p. 74–81, 2018. doi:10.1016/j.chemolab.2018.02.008

YANG, X. S. **Nature-Inspired Metaheuristic Algorithms**. 1 ed. 2008. 116 p.

ZHANG, L.; MISTRY, K.; LIM, C. P.; NEOH, S. C. Feature selection using firefly optimization for classification and regression models. **Decision Support Systems**, v. 106, p. 64–85, 2018. doi:10.1016/j.dss.2017.12.001