UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I – CAMPINA GRANDE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA
CURSO DE MESTRADO EM QUÍMICA

GIOVANNA DE FÁTIMA ABRANTES OLIVEIRA

SELEÇÃO DE VARIÁVEIS BASEADA EM ALGORITMO FIREFLY E ANÁLISE DISCRIMINANTE POR MÍNIMOS QUADRADOS PARCIAIS PARA PROBLEMAS DE CLASSIFICAÇÃO

CAMPINA GRANDE - PB
2023

GIOVANNA DE FÁTIMA ABRANTES OLIVEIRA

SELEÇÃO DE VARIÁVEIS BASEADA EM ALGORITMO FIREFLY E ANÁLISE DISCRIMINANTE POR MÍNIMOS QUADRADOS PARCIAIS PARA PROBLEMAS DE CLASSIFICAÇÃO

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

Área de Concentração: Química Analítica

Orientador: Prof. Dr. José Germano Véras Neto

Coorientador: Dr. David Douglas de Sousa Fernandes

CAMPINA GRANDE - PB
2023

O48s Oliveira, Giovanna de Fátima Abrantes.

Seleção de variáveis baseada em algoritmo *firefly* e análise discriminante por mínimos quadrados parciais para problemas de classificação [manuscrito] / Giovanna de Fátima Abrantes Oliveira. - 2023.

61 p. : il. colorido.

Digitado.

Dissertação (Mestrado em Química) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.

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

"Coorientação: Prof. Dr. David Douglas de Sousa Fernandes , Departamento de Química - CCT."

1. Seleção de variáveis. 2. Inteligência artificial. 3. Algoritmos bioinspirados. I. Título

21. ed. CDD 005.3

Elaborada por Talita M. A. Tavares - CRB - CRB 15/971 BC/UEPB

GIOVANNA DE FÁTIMA ABRANTES OLIVEIRA

SELEÇÃO DE VARIÁVEIS BASEADA EM ALGORITMO FIREFLY E ANÁLISE DISCRIMINANTE POR MÍNIMOS QUADRADOS PARCIAIS PARA PROBLEMAS DE CLASSIFICAÇÃO

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

Área de Concentração: Química Analítica

Aprovada em: 31 / 07 / 2023

BANCA EXAMINADORA

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

Profa. Dra. Simone da Silva Simões (Examinador)
Universidade Estadual da Paraíba

Prof. Dr. Marcelo Fabián Pistonesi (Examinador)
Universidad Nacional del Sur

AGRADECIMENTOS

À minha avó Adalgisa, minha mãe Fátima e ao meu pai Jurandi por todo amor, apoio e incentivo e se fazerem a base sólida que firma meus pés neste chão. Em especial a minha avó que me remete o maior significado de amor e dedicação, obrigado pelo afeto, por ser tão presente e por nunca me deixar hesitar em continuar buscando conhecimento.

À Valkyria por toda paz e felicidade que trouxe para minha vida.

A Emmanuel pela paciência infinita, atenção e lealdade, independente da fase que me encontrei durante essa etapa.

Ao meu orientador Dr Germano Véras pela valorosa orientação, pela compreensão na tomada de decisões nessa etapa e principalmente pela paciência e confiança depositada.

Ao meu coorientador Dr David Douglas pela disponibilidade e incentivo que foram fundamentais para prosseguir e concluir este estudo.

Aos meus colegas do LQAQ e do LABDEM pela parceria e companhia diária, foi um privilégio pertencer a um grupo tão formidável.

A irmã que a vida escolheu para mim, Dayane, por sempre acreditar no meu potencial, pela lealdade e amizade verdadeira, sincera e que carrega um valor imensurável.

Aos meus amigos que estiveram presentes e aceitaram minha nova versão.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. A CAPES pela bolsa concedida.

RESUMO

A seleção de variáveis consiste em uma ferramenta potencial que busca filtrar informações relevantes para resolução de problemas envolvendo matrizes complexas. Buscando melhorar a precisão dos modelos quimiométricos e a robustez atendendo ao princípio da parcimônia diferentes algoritmos têm sido desenvolvidos utilizando a seleção de variáveis. Com os avanços da inteligência artificial o uso de algoritmos bioinspirados para a otimização e resolução de problemas complexos tornou-se uma ferramenta interessante para diversas aplicações em classificação multivariada. Nesse contexto, o presente estudo propõe um novo algoritmo bionspirado no comportamento dos vagalumes denominado FA-PLS-DA para seleção de variáveis empregando a Análise Discriminante Linear buscando superar problemas que envolvem elevada multicolinearidade entre as variáveis. Para avaliar o desempenho do algoritmo proposto, foram utilizados três bancos de dados espectrométricos na região NIR de domínio público e dados com informação simulada, sendo os dados brutos e pré-processados. O primeiro banco de dados composto de espectros ATR-FTIR na faixa de 4000 a 650 cm^{-1} de 104 amostras de saliva para avaliar a presença ou ausência de SARS-CoV-2. O segundo banco de dados utilizados consiste também em espectros NIR de 192 amostras de leite de cabra para avaliar a adulteração pela adição de leite de vaca. O terceiro banco de dados é também composto por espectros NIR de 120 amostras de azeite de oliva extra-virgem provenientes de quatro países diferentes. Para estudo com informação simulada o banco de dados compreendeu as 90 observações com 600 variáveis usando quatro fatores para gerar três classes distintas, as amostras foram divididas em conjuntos de treinamento e teste usando o algoritmo *Kennard-Stone.* O desempenho do FA-PLS-DA foi comparado com os resultados da Análise Discriminante Linear por Mínimos Quadrados Parciais (PLS-DA) aplicando diferentes pré-processamentos aos dados. O tratamento dos dados foi realizado em ambiente Matlab. Foram selecionados os modelos com os dados pré-processados que apresentaram maior Taxa Correta de Classificação (TCC), o algoritmo FA-PLS-DA selecionou uma quantidade menor variáveis latentes (LVs) para todos os bancos de dados. Ademais, o algoritmo proposto apresentou TCC de 100% para o conjunto de treinamento do banco de dados de COVID, enquanto o PLS-DA apresentou TCC de 98,72% empregando um número maior de variábeis latentes. Para o banco de dados de leite de cabra o algoritmo proposto apresentou TCC de 95,92%, já o PLS-DA mostrou 100% de TCC, apesar de superar o algoritmo proposto em termos de TCC, o PLS-DA empregou um número elevado de LVs para construção dos modelos. O algoritmo proposto superou o PLS-DA na construção dos modelos do banco de dados de azeite de oliva, em que

obteve 100% de TCC para os conjuntos de treinamento e teste empregando o menor número de LVs. Para os dados com informação simulada o FA-PLS-DA apresentou 82,22% de TCC para o conjunto de teste, enquanto a TCC do PLS-DA foi de 77,78%. Em todos os bancos de dados o algoritmo FA-PLS-DA mostrou ser mais parcimonioso que o PLS-DA tendo sua performance comparada ao desempenho do PLS-DA, sendo robusto e capaz de classificar as amostras adequadamente a partir das variáveis selecionadas corroborando com sua viabilidade.

Palavras-chave: classificação; *firefly*; algoritmo bioinspirado.

ABSTRACT

Variable selection is a potential tool that seeks to filter relevant information to solve problems involving complex matrices. Seeking to improve the accuracy of chemometric models and robustness to the principle of parsimony different algorithms have been developed using variable selection. With the advances in artificial intelligence the use of bio-inspired algorithms for the optimization and resolution of complex problems has become an interesting tool for several applications in the context of multivariate calibration and classification. In this context, the present study proposes a new bio-inspired algorithm in the behavior of fireflies called FA-PLS-DA for variable selection employing Linear Discriminant Analysis seeking to overcome problems involving high multicollinearity among variables. To evaluate the performance of the proposed algorithm, we stipulated population conditions of 50 fireflies in 50 life cycles and used three case studies involving public domain NIR spectrometric databases and a database with simulated information. The first database consists of ATR-FTIR spectra in the 4000 to 650 cm-1 range of 104 saliva samples to assess the presence or absence of SARS-CoV-2. The second database used also consists of NIR spectra of 192 goat milk samples to assess adulteration by the addition of cow's milk. The third database also consists of NIR spectra of 120 samples of extra virgin olive oil from four different countries. For study with simulated information the database comprised 90 observations with 600 variables using four factors to generate three distinct classes, the samples were divided into training and test sets using the Kennard-Stone algorithm. The performance of FA-PLS-DA was compared with the results of Partial Least Squares Linear Discriminant Analysis (PLS-DA) applying different preprocessing to the data, and the data was treated in programs developed in Matlab environment. The preprocessed models that presented the highest TCC were selected, the FA-PLS-DA algorithm selected 3, 7, 9 and 7 latent variables (LVs) for the COVID, goat milk, extra-virgin olive oil and simulated databases respectively, while the PLS-DA selected 4, 16, 16 and 4 LVs. Furthermore, the proposed algorithm showed a correct classification rate (CCR) of 100% for the COVID database training set, while PLS-DA showed a rate of 98.72% by employing a larger number of latent variables. For the goat milk database, the proposed algorithm showed a CCR of 95.92%, while PLS-DA showed 100% CCR. Despite outperforming the proposed algorithm in terms of CCR, PLS-DA employed a high number of LVs for model building. The proposed algorithm outperformed PLS-DA in building the models for the olive oil database, where it obtained 100% CCR for the training and test sets by employing the smallest number of LVs. For the data with simulated information FA-PLS-DA showed 82, 22% CCR for the test

set, while PLS-DA showed 77.78%. For all the databases, the FA-PLS-DA algorithm proved to be more parsimonious than PLS-DA, and its performance compared to PLS-DA was robust and able to classify the samples properly from the selected variables, corroborating its viability.

Keywords: classification, firefly, bioinspired algorithm

LISTA DE ILUSTRAÇÕES

Página

LISTA DE TABELAS

Página

LISTA DE ABREVIATURAS E SIGLAS

CV – Validação cruzada, do inglês *Cross Validation*

FA – Algoritmo Vagalume, do inglês *Firefly Algorithm*

FF-iPLS – Algoritmo Vagalume para Seleção de Intervalos em PLS, do inglês *Firefly Algorithm for Interval Selection in PLS*

FFpop – Número de vagalumes para a população

GA – Algoritmo Genético, do inglês *Genetic Algorithm*

I_max – Quantidade de intervalos máximos

iSPA-PLS – Algoritmo das Projeções Sucessivas para seleção de intervalos em PLS, do inglês *Successive Projections Algorithm for Interval Selection in PLS*

KS – Algoritmo *Kennard-Stone*

LDA – Análise Discriminante Linear, do inglês *Linear Discriminant Analysis*

LOO – Deixar um de fora, do inglês *leave-one-out*

LV – Variáveis Latentes, do inglês *latent variables*

MSC – Correção de espalhamento multiplicativo, do inglês *Multiplicative Signal Correction*

NIPALS – Interação não linear por mínimos quadrados parciais, do inglês *Nonlinear Iterative Partial Least Squares*

NIR – Espectroscopia no Infravermelho próximo, do inglês *Near-infrared spectroscopy*

nm - Nanometros

PCA – Análise de Componentes Principais, do inglês *Principal Component Analysis*

PLSR – Regressão por mínimos quadrados parciais, do inglês *Partial Least Squares Regression*

PLS-DA – Análise discriminante por mínimos quadrados parciais, do inglês *Partial Least Squares – Discriminant Analysis*

PSO - Otimização por enxame de partículas, do inglês *Particle Swarm Optimization*

SPXY – Partição do conjunto de amostras, do inglês *Sample set Partitioning based on joint X-y distances*

SW – Inteligência de Enxame, do inglês *Swarm Intelligence*

SUMÁRIO

1 INTRODUÇÃO

As técnicas de seleção de variáveis são ferramentas para reduzir dimensionalidade em um grande conjunto de dados. Através destas ferramentas é possível minimizar redundância de informações, excluir variáveis ruidosas e/ou não informativas e extrair informações significativas em modelos quimiométricos desenvolvidos para diferentes aplicações. Em essência, seleção de variáveis consiste em encontrar um subconjunto de variáveis em uma matriz de dados que apresente uma melhor correlação com a variável de interesse (SSEGANE et al, 2012; GOMES et al, 2022). Essas técnicas podem auxiliar na melhoria do desempenho de algoritmos em termos de precisão para construção de modelos que expliquem tais relações presentes entre variáveis preditoras e a resposta. Diferentes modelos podem ser empregados como os modelos determinísticos e estocásticos (HEINZE; WALLISCH; DUNKLER, 2018; ANDERSEN; BRO, 2010).

A seleção de variáveis pode ser aplicada na quimiometria na construção de modelos de calibração, bem como para modelos de classificação. Nestes modelos uma relação é feita entre uma matriz multivariada independente a um vetor de respostas que admite valores pertencentes a diferentes classes. Dentre os algoritmos multivariados de classificação, os modelos discriminantes são focados no grau de similaridade existente entre as amostras de um mesmo grupo e dissimilaridade entre amostras em grupos distintos (BEEBE, 1998).

Particular atenção em relação às técnicas discriminantes se dá na análise discriminante por mínimos quadrados parciais (PLS-DA, do inglês *Partial Least Square in Discriminant Analysis*) visto que a definição de classes se dá por regressão linear entre uma matriz de variáveis independentes **X** e uma matriz de variáveis dependentes indicativas de diferentes classes as quais as amostras pertencem (RUIZ-PÉREZ et al, 2020). Entretanto, esta técnica possui como inconveniente o aumento de erros de classificação devido a grande quantidade de variáveis visto que podem ser não informativas ou apresentar sinais espúrios. Nesse sentido, diferentes algoritmos voltados à seleção de variáveis têm sido desenvolvidos para resolução de tais problemas como é o caso do algoritmo Firefly (FA, do inglês *Firefly Algorithm*), que consiste em um algoritmo estocástico, classificado como bioinspirado e que apresenta uma métrica associada ao comportamento do inseto em comunicar proximidade da comida por meio de liberação de radiação bioluminescente.

O FA é um algoritmo relativamente novo em muitas categorias de algoritmos inteligentes e demonstra eficiência, além disso, a otimização do Firefly parece ser mais promissora pelo fato de lidar com funções multimodais de forma natural e eficiente (YANG et

al, 2018). Mesmo com a crescente evolução desses algoritmos, são poucos os estudos envolvendo a aplicação do FA à quimiometria. Buscando a utilização do algoritmo na resolução de problemas como a sobreposição de respostas espectrais e a otimização de condições de trabalho em matrizes complexas, o presente estudo tem como objetivo desenvolver o FA para análise discriminante por mínimos quadrados parciais (FA-PLS-DA) e avaliar seu desempenho em diferentes bancos de dados para otimização de problemas envolvendo classificação.

2 OBJETIVOS

2.1 Objetivo Geral

Propor o algoritmo Firefly para seleção de variáveis em análise discriminante linear por mínimos quadrados parciais e avaliar seu desempenho em diferentes conjuntos de dados.

2.2 Objetivos Específicos

- ✓ Desenvolver o algoritmo FA-PLS-DA para classificar amostras em diferentes bancos de dados;
- ✓ Construir os modelos quimiométricos de classificação por FA-PLS-DA e PLS-DA utilizando dados brutos e pré-processados com suavização por Savitzky-Golay, correção de espalhamento multiplicativo (MSC), padronização normal de sinal (SNV), correção de linha de base e derivações por Savitzky-Golay;
- ✓ Comparar a capacidade de classificação em diferentes conjuntos de dados do FA-PLS-DA em relação ao PLS-DA utilizando os parâmetros de qualidade especificidade, sensibilidade e taxa de classificação.

3 REFERENCIAL TEÓRICO

As técnicas analíticas modernas como ressonância magnética nuclear, espectroscopias no Infravermelho/Visível/Ultravioleta, espectroscopia de massa, cromatografias, entre outras, fornecem um grande volume de dados. O tratamento destes dados pode ser uma tarefa difícil, principalmente em relação as amostras complexas (LAVIGNE, 2000). Nessa perspectiva, a quimiometria surgiu para fornecer soluções para esses problemas permitindo extrair o máximo de informação de uma grande quantidade de dados que passaram a ser gerados com a evolução da instrumentação.

Ao avaliar um conjunto amostral, diferentes analitos de interesse ou características intrínsecas do objeto/sistema de estudo influenciam nos resultados a serem obtidos, dessa maneira as observações são realizadas em condições semelhantes. Essas observações podem ser parâmetros físicos e/ou químicos, que variam conforme os objetivos definidos pelo analista (LUCIANO, 2021). Na quimiometria, essas observações são conhecidas como variáveis.

3.1 Variáveis e suas classes

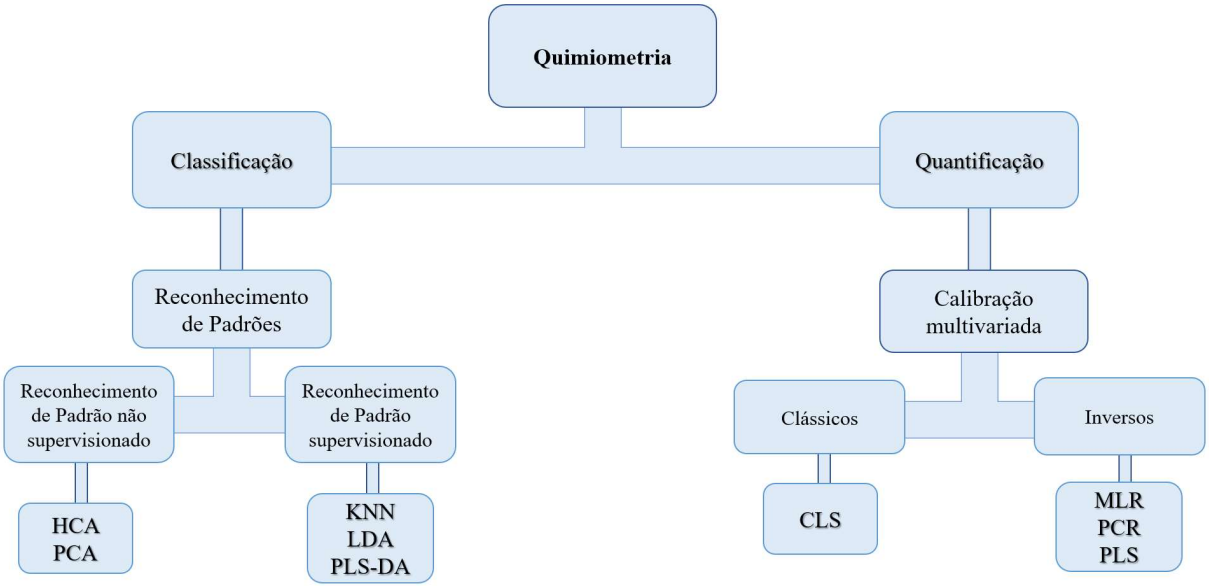
As variáveis podem ser classificadas como qualitativas ou quantitativas (FREEMAN; PISANI; PURVES, 2007). De acordo com Luciano (2021) as variáveis qualitativas fornecem dados categóricos e descrevem uma qualidade ou característica das amostras, ou seja, não são numéricas. Por outro lado, as variáveis quantitativas fornecem dados discretos e contínuos e medem uma quantidade ou dados numéricos em cada unidade experimental. As variáveis quantitativas podem ser de dois tipos: discretas ou contínuas. Segundo Freedman (2009), quando os valores possíveis são poucos e relativamente separados, a variável é discreta, isso quer dizer que os valores diferem por valores fixos, caso contrário, são contínuas.

A partir de 1960 avanços instrumentais e computacionais permitiram obter um grande conjunto de dados experimentais, mas este excesso de informações não podiam ser adequadamente processado e interpretado. Neste sentido, ferramentas estatísticas e matemásticas começaram a ser implementadas e criadas para tratar dados químicos, sendo esta ciência denominada de Quimiometria que permite extrair informações de interesse presentes nos dados químicos obtidos a partir de respostas instrumentais (FERREIRA, 2015).

As técnicas quimiométricas podem ser divididas em calibração multivariada e reconhecimento de padrões (Figura 1). Os métodos de classificação ou reconhecimento de padrões são utilizados para identificar semelhanças e diferenças entre as amostras comparando-

as entre si que são refletidas nas medidas utilizadas para caracterizá-las (BEEBE; PEEL; SEASHOLTZ, 1998).

Figura 1 – Fluxograma das técnicas quimiométricas



Fonte: Adaptado de Beebe; Peel; Seasholtz (1998).

3.2 Métodos de reconhecimento de padrão

Os métodos de reconhecimento de padrão têm como intuito encontrar uma métrica para identificar relações de similaridade e de dissimilaridade entre amostras. O interesse ao utilizar tais métodos consiste na medição de uma grande quantidade de variáveis que podem possuir relação às amostras pelas quais se pretende classificar (BRERETON, 2009).

Introduzido na Química na década de 70, os métodos de reconhecimento de padrões são divididos em: métodos de reconhecimento de padrão supervisionado e não supervisionado. No reconhecimento de padrão supervisionado as amostras são classificadas de acordo com as semelhanças e diferenças, ou seja, são admitidos critérios de discriminação das amostras. Essas informações a respeito das classes é que supervisionam o desenvolvimento de critérios de discriminação a serem utilizados posteriormente para realizar o reconhecimento de novas amostras (MILLER; MILLER, 2018; FERREIRA, 2015). Segundo Brereton (2009) os métodos são baseados na suposição de que quanto mais amostras se assemelham entre si em relação às variáveis medidas, mais próximas elas estarão no espaço multidimensional gerado por tais variáveis.

Os métodos de reconhecimento de padrão supervisionados mais utilizados na literatura são: k-ésimo vizinho mais próximo (*k*-NN do inglês *Kth Nearest Neighbor*), Modelagem

Independente Suave por Analogia de Classe (SIMCA, do inglês *Softh Independent Modeling of Class Analogy*) e métodos de Análise Discriminante (DA, do inglês *Disciminant Analysis*). Nos últimos podem ser destacados Análise Discriminante Linear (LDA, do inglês *Linear Discriminant Analysis)* e a Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA, do inglês *Pastial Least Squares – Discriminant Analysis*) (FERREIRA, 2015).

3.2.1 Análise discriminante linear

A análise discriminante linear foi proposta por Fisher e Mahalanobis em 1936, sendo considerado o método de reconhecimento de padrões supervisionado mais antigo e mais estudado. Este algoritmo, processo sistemático ou uma ferramenta para a resolução de problemas de otimização (SZWARCFITER; MARKEZON, 2015), consiste em uma técnica linear, siginificando que os limites de decisão que separam as classes no espaço multidimensional das variáveis são superfícies lineares (limite ou hiperplanos). Modo geral, o hiperplano é calculado de forma que a variação entre as classes seja maximizada e a variação dentro das classes individuais sejam minimizadas. Essa técnica tem sido extensamente empregada para resolução de problemas em uma gama de amostras destacando-se às relacionadas a alimentos, como em casos de adulteração, classificação da origem, grau de envelhecimento, dentre outros (MARINI, 2013).

Segundo Miller (2018), na análise discriminante linear (LDA) o objetivo é encontrar uma função discriminante linear Y que é uma combinação linear das variáveis medidas originais, como indicado na equação 1:

$$\mathrm{Y} = a_1 X_1 + a_2 X_2 + \cdots a_n X_n \qquad (1)$$

Pode-se inferir pela equação que as medidas originais de cada objeto são combinadas em um único valor de Y de modo que os dados sejam reduzidos de n dimensões para uma única. Os coeficientes dos termos são escolhidos de maneira com que Y diferencie ao máximo as classes, e assim, objetos pertencentes à mesma classe terão valores de Y próximos (MILLER, 2018). Em termos gráficos, o resultado do tratamento dos dados por LDA é representado por uma equação de reta que separa dois grupos de amostras em um espaço bidimensional.

3.2.2 Análise discriminante por mínimos quadrados parciais – PLS-DA

O algoritmo PLS, bastante empregado em calibração multivariada, foi introduzido para aplicações em reconhecimentos de padrão, como PLS-DA, com sucesso na modelagem de conjunto de dados em diversas situações que envolvem conjuntos de dados de alta dimensão, sendo possível utilizá-lo para predição e modelagem descritiva, bem como seleção de variáveis discriminantes. O algoritmo PLS-DA combina a redução da dimensionalidade dos dados e análise discriminante envolvendo duas etapas principais sendo elas: (i) a construção de componentes PLS (redução da dimensão) e (ii) construção do modelo de previsão (análise discriminante). Nesse sentido, o PLS-DA irá lidar com variáveis de saída catecóricas que são recodificadas em variáveis contínuas (LEE; LIONG; JEMAIN, 2018).

Os modelos de classificação baseados em PLS-DA podem ser descritos como uma relação entre uma matriz multivariada **X** independente e um vetor de respostas que admite os valores discretos de classes às quais as amostras pertencem. O método utiliza uma regressão que opera uma decomposição linear das marizes **X** e **Y** para calcular os parâmetros do modelo (RUSCHEL, 2017). Segundo Brereton (2009), o PLS-DA é mais adequado para modelos de duas classes e podem responder a perguntas sobre componentes que pertencem ou não a uma classe amostral.

Uma das vantagens do PLS-DA consiste na flexibilidade ao trabalhar com bancos de dados que possuem um número de variáveis muito maior do que o número de amostras, além disso, apresenta capacidade de classificação semelhante ao LDA (BRERETON, 2009).

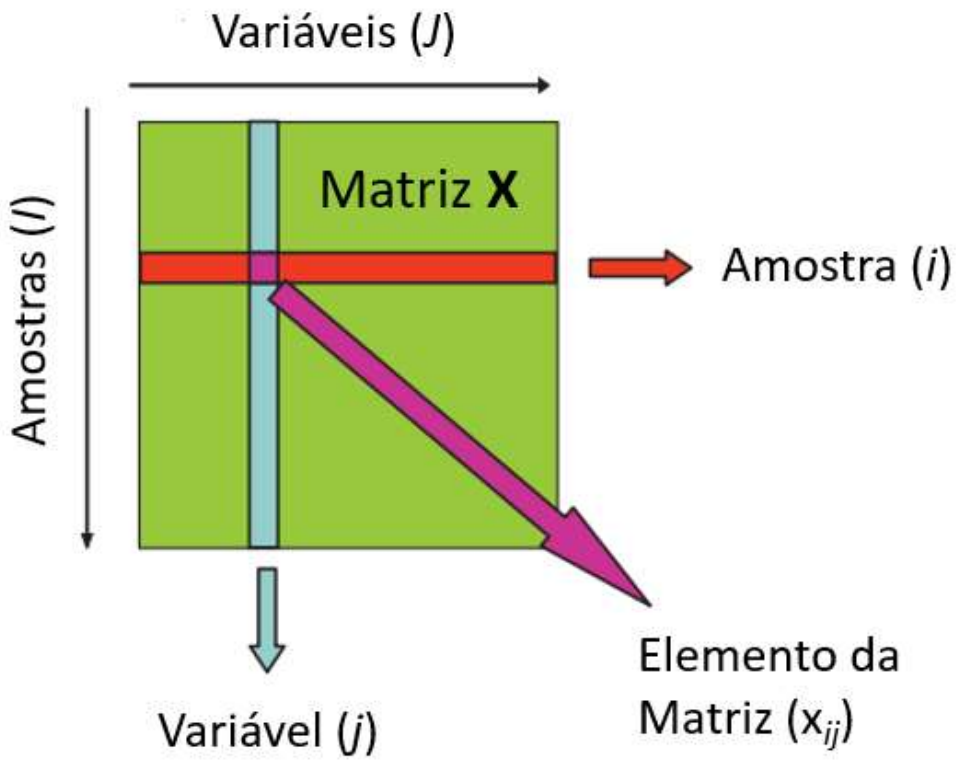
3.2.3 Regressão por mínimos quadrados parciais

Assim como a análise discriminante linear, o PLS em sua forma básica é um método linear que fornece resultados idênticos à LDA (BRERETON, 2009). Bastante utilizado em regressão por mínimos quadrados parciais (PLSR) é uma técnica para dados multivariados, sendo considerada uma ferramenta importante quando há conhecimento parcial dos dados, pois usa características típicas de um conjunto de dados para estabelecer a relação entre estes e os analitos de interesse, os modelos PLS geralmente são robustos, desde que as amostras futuras apresentem características semelhantes aos dados originais, já que diferenças espectrais podem afetar o desempenho (BRERETON, 2018).

O PLSR transforma as variáveis de resposta em um número reduzido de combinações lineares, que busca estabelecer uma relação entre a matriz de dados, como, por exemplo, os espectros (matriz **X**), as variáveis são dispostas nas colunas (cada resposta instrumental a um comprimento de onda específico ocupará uma célula) e as amostras serão dispostas em linhas

(Figura 4), com a concentração de um determinado analito (matriz **Y**). Ambas as matrizes são decompostas em novas variáveis denominadas variáveis latentes, fatores ou componentes principais, a partir de duas matrizes de variações: Pesos e *scores* (N. MILLER; C. MILLER; D. MILLER, 2018).

Figura 2 – Organização matricial dos dados instrumentais



Fonte: Adaptado de BRERETON (2018).

Os espectros brutos, por exemplo, considerados como combinações lineares das variações dos espectros (pesos/*loadings*) em que os *scores* representam a contribuição de cada resposta obtida representada pelas Equações:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{R} \tag{2}$$

$$\mathbf{Y} = \mathbf{C}\mathbf{Q}^{\mathrm{T}} + \mathbf{E} \tag{3}$$

Onde **X** e **Y** consistem nas variáveis independentes e dependentes decompostas em duas matrizes: **T** e **U** a matriz de *scores*, **P** e **Q** a matriz de *loadings* e **R** e **E** os resíduos, que consiste em uma matriz de erro, a partir disso, o modelo final é obtido, na qual obtemos a melhor relação

linear entre os *scores* das duas matrizes (dependentes (**Y**) e independentes (**X**)) (BRERETON, 2018).

$$\mathbf{C} = \mathbf{B_w T} + \mathbf{F} \quad (4)$$

$$\mathbf{Y} = \mathbf{B_w T Q^T} + \mathbf{G} \quad (5)$$

Onde $\boldsymbol{B_w}$ consiste no coeficiente ajustado, **F** e **G** os resíduos das matrizes de scores e da concentração respectivamente.

A regressão por mínimos quadrados parciais estende o conceito do modelo inverso trocando as variáveis originais por um subconjunto incompleto das variáveis latentes dos dados originais (MARTENS e NAES, 1989). Segundo Lavigne (2000) a PLSR tornou-se um método padrão bastante utilizado para calibração multivariada devido à qualidade dos modelos de calibração produzidos. Constituindo um método clássico de regressão linear, a PLSR tem sido empregada em análises espectrais buscando uma nova dimensão projetada pelos preditores e uma variável dependente **Y**, quanto menor o número de preditores menor será o risco pela multicolinearidade entre estes (HU et al, 2017).

Antes de sua utilização, o modelo empírico obtido deve ser testado para verificar a capacidade de prever as classes em diferentes amostras e para isso é utilizado um conjunto de amostras externa denominado de ‘conjunto de teste’ ou ‘conjunto de validação’ (FERREIRA, 2015), para isso as amostras podem ser selecionadas utilizando diferentes algoritmos, entre os mais utilizados estão o *Kennard-Stone*.

Como foi observado, para as técnicas de classificação são atribuídos valores discretos às variáveis, já para as técnicas de quantificação, como é o caso da calibração multivariada, são consideradas variáveis contínuas. No entanto, nem sempre todas as variáveis apresentam correlações com o objeto de estudo, a partir disso, as técnicas de seleção de variáveis são consideradas ferramentas promissoras para trabalhar com dados que apresentam grande dimensionalidade buscando resultados mais precisos.

3.3 Seleção de variáveis

A seleção de variáveis consiste em uma técnica que visa eliminar variáveis não informativas e ruídos que influenciam na qualidade de um modelo e tem se tornado relevante para fins industriais e acadêmicos. Devido ao elevado custo de instrumentação e tempo

operacional de equipamentos empregados em análises é possível utilizar parte dos dados obtidos para obtenção de modelos mais parcimoniosos, além de melhorar a exatidão e reduzir a demanda computacional (SUN et al, 2018).

A seleção de variáveis consiste em um conjunto de técnicas bastante importante principalmente no contexto das espectroscopias de Infravermelho/Visível/Ultravioleta como também em cromatografias, pois tais instrumentos geram uma grande quantidade de variáveis, que podem conter informações não relacionadas às propriedades de interesse. A técnica consiste em manter apenas uma parte das medições originais selecionando comprimentos de onda ou regiões que contém informações relevantes para o problema de interesse (BRERETON, 2018).

O objetivo da seleção de variáveis, portanto, é reduzir a dimensionalidade das matrizes buscando remover variáveis irrelevantes e/ou não confiáveis identificando as variáveis que melhor se correlacionam para uma predição adequada da variável de resposta (GALVÃO et al, 2005).

É possível classificar as técnicas de seleção de variáveis conforme a presença de variáveis aleatórias e diferentes modelos podem ser aplicados para problemas de otimização (CONCEIÇÃO, 2014). Realizar a seleção de variáveis significa encontrar um subconjunto da matriz **X** que melhor se correlaciona com o vetor resposta y, sem que informações relevantes sejam eliminadas.

3.3.1 Métodos de seleção de variáveis determinísticos

Os modelos determinísticos são aqueles em que será obtida correspondência biunívoca entre variáveis de entrada e saída, ou seja, um único resultado será obtido independente de quantas vezes o algoritmo for executado (BISCUOLA, 2011).

Alguns estudos mostram a aplicação de modelos determinísticos para seleção de variáveis nas quais possuem informações importantes para a quantificação de um analito (DA COSTA et al, 2020). Para classificação multivariada esses métodos também são utilizados, como por exemplo o Algoritmo de projeções sucessivas SPA-LDA, que possui diferentes aplicações (SOUTO et al, 2010; DE ALMEIDA et al, 2014; LEI; LIN; SUN, 2019; YUAN et al, 2018). E o PLS-DA que também seleciona variáveis na qual apresentam informações redundantes para resolução de problemas de multicolinearidade, é um exemplo de método determinístico que utiliza a seleção de variáveis (LEE; LIONG; JEMAIN, 2018).

3.3.2 Métodos de seleção de variáveis estocásticos

A palavra “estocástico” significa aleatório. Algoritmos estocásticos consistem em modelos matemáticos que incorporam processos probabilísticos/heurísticos. Segundo Barbosa (2017, p. 26) “a palavra heurística tem origem no nome grego *heuriskein* que significa descobrir”. Para o autor a heurística é definida como um conjunto de métodos e regras que podem levar à resolução de diversos tipos de problemas. Espera-se que um bom algoritmo apresente tempo de execução aceitável e forneça soluções ótimas para o problema de interesse.

Os modelos estocásticos fornecem diferentes resultados a partir de decisões baseadas na probabilidade. Eles consistem em um conjunto de variáveis aleatórias similares em um espaço amostral comum (LAW, 2015).

A viabilidade dos modelos heurísticos pode variar, mas geralmente fornecem boas soluções, ou seja, são métodos meta-heurísticos que utilizam combinações com escolhas aleatórias a partir de um conhecimento prévio do problema e encontrarem dentro do espaço de busca, soluções eficientes (BARBOSA, 2017).

Algoritmos metaheurísticos são algoritmos de otimização que tentam melhorar a qualidade das soluções iterativamente com algumas propriedades de aleatoriedade. A meta-heurística diz respeito a um método computacional interativo onde é possível modificar operações para busca de soluções aceitáveis de um dado problema que pode ser representado em um computador com uma demanda computacional razoável, garantindo a otimização das respostas obtidas (GOLDBARG, 2016).

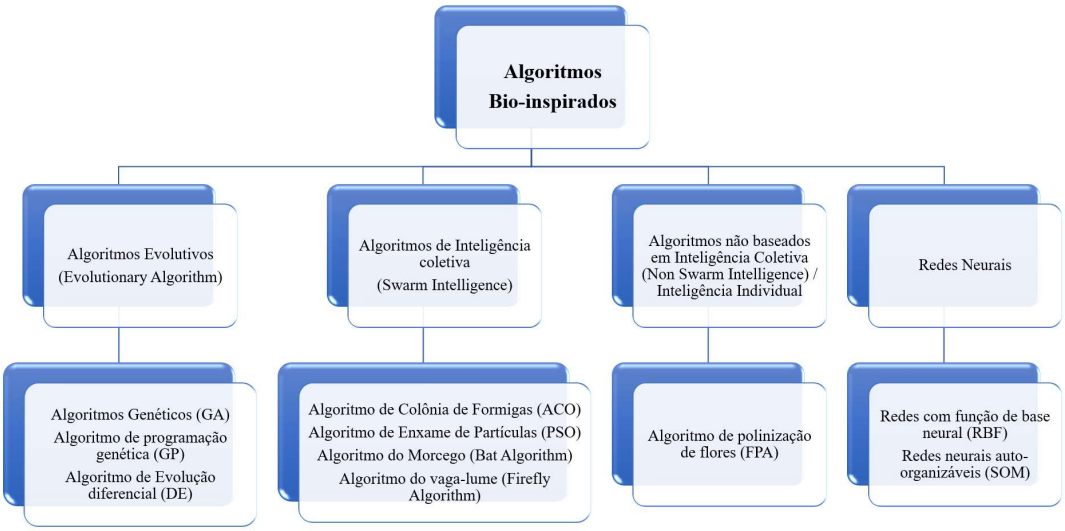
Diferentes métodos meta-heurísticos têm demonstrado boas soluções para diversos problemas de otimização, como os algoritmos bio-inspirados, que são baseados em princípios e comportamentos observados na natureza. Estes algoritmos bio-inspirados podem ser divididos em categorias, sendo elas: algoritmos evolutivos associados a processos biológicos, redes neurais, algoritmos de inteligência individual e os algoritmos de inteligência coletiva inspirandos no padrão comportamental de animais (Figura 3) (DEL SER et al, 2019). Alguns algoritmos bioinspirados podem ser citados: algoritmo genético (GA), algoritmo de enxame de gatos (CSO), de colônia de formigas (ACO), o algoritmo de vagalumes (FA), entre outros (BOZORG-HADDAD, 2018).

O desenvolvimento desses algoritmos para otimização e desempenho dos instrumentos tem sido observado em estudos quimiométricos aliados a espectroscopia na região do infravermelho próximo e médio e UV-vis (NAZI & LEARDI, 2012; POLO-CORPA et al 2009).

3.3 Algoritmos Evolutivos

Após o fim da II Guerra Mundial a primeira linguagem funcional baseada em funções foi lançada. Isso influenciou uma diversidade de trabalhos e seu desenvolvimento direcionou pesquisas em várias áreas de inteligência artificial, comprovação de teoremas, cálculo simbólico, entre outras utilizadas no contexto acadêmico e na indústria. Em 1950 empresas comerciais começaram a desenvolver linguagens de máquina para diferentes aplicações (MANZANO, 2020). O desenvolvimento de algoritmos tornou-se importante para problemas a serem solucionados.

Figura 3 – Categorias dos algoritmos bio-inspirados



Fonte: Adaptado de YADAV & VISHWAKARMA (2020); DEL SER et al (2019); MESQUITA (2018).

Neste sentido, algoritmos evolutivos baseados em técnicas inspiradas na biologia como mutação, seleção natural, hereditariedade e recombinação genética passaram a ser utilizados pela possibilidade de se adaptar a alterações das condições iniciais e se moldar em realidades mais complexas (DEL SER et al, 2019).

Esses algoritmos consistem em métodos de otimização que podem evoluir após diversas iterações buscando melhores respostas a uma solução inicial baseados em observações naturais. Este desenvolvimento de raciocínio é baseado no conceito Darwiniano em que soluções fracas enfrentam extinção enquanto soluções melhores se combinam para fornecer novas que podem melhorar a convergência para uma ótima solução. Assim, simulações de interações biológicas como feromônios humanos e respostas do sistema imunológico podem ser usadas como

ferramenta para descrever um protocolo de otimização, dentre outras possibilidades (MAROUANI; AL-MUTIRI, 2022).

Um exemplo importante para algoritmos baseados em evolução é o algoritmo genético (AG) que surgiu no início da década de 1960 e mostrou resultados promissores devido a robustez e desempenho satisfatório em relação aos métodos existentes. A técnica foi desenvolvida por John Holland e colaboradores, baseados na teoria da evolução de Charles Darwin e seleção natural de sistemas biológicos em que cada espécie evolui mediante a competição entre indivíduos pela reprodução e sobrevivência (BOZORG-HADDAD, 2018; GOLDBERG; RICHARDSON, 1987).

Desde então surgiram uma série de técnicas de otimização baseadas em inteligência coletiva inspirada na natureza, denominados de algoritmos bioinspirados (BOZORG-HADDAD, 2018). Alguns destes algoritmos bioinspirados são baseados em otimização por inteligência de partículas (PSO, do inglês *Particle Swarm Optimization*) como o caso de colônias de formigas (ACO, do inglês *Ant Colony Optimization*).

Os algoritmos de inteligência coletiva têm recebido grande atenção nos últimos anos, sendo aplicados para resolução de problemas complexos ou aperfeiçoar sistemas já existentes, envolvendo estudo de diferentes fenômenos biológicos na construção de modelos computacionais e algoritmos potenciais (GOEDERT; PAULA FILHO; BLANCO, 2017). Alguns estudos evidenciaram a efetividade na resolução de problemas de otimização dos algoritmos PSO frente aos algoritmos genéticos devido à rapidez e melhor convergência das estimativas. São exemplos de algoritmos PSO, a retropropagação em redes neurais artificiais (BP-ANN) (ZANDBAAF; MOHAMMAND; MAJID, 2022; BIN, 2021), sendo estes baseados em algoritmos genéticos (GA-PLSR) (SULAN, et al, 2020); modelos híbridos de redes neurais-vagalumes (ANN-FA) (SHARABIANI et al, 2020), entre outros que se mostraram mais parcimoniosos por buscarem modelos empregando uma quantidade menor de variáveis latentes e figuras de mérito mais adequadas.

Além da aplicação na calibração multivariada, tais algoritmos também têm sido adaptados para seleção de variáveis em métodos de reconhecimento de padrão como a análise discriminante por mínimos quadrados parciais (PLS-DA) e floresta aleatória (do inglês, *Random florest*) (BROUGHTON-NEISWANGER et al, 2020; ISMAIL et al, 2017) e o algoritmo genético mencionado anteriormente (GA-PLS-DA) (RAMADAN, et al, 2006; NASCIMENTO et al, 2022).

A utilização de novos algoritmos configura-se, portanto, numa busca pela resolução de problemas de maneira mais eficiente e neste sentido, uma possibilidade interessante é o algoritmo Firefly (ATTIA et al, 2017).

3.4 Algoritmos de inteligência coletiva: Firefly

O algoritmo Firefly, ou algoritmo do vagalume (FA) é um algoritmo bioinspirado relativamente novo (YANG et al, 2009) em que os vagalumes são considerados unissexuais e todos são atraentes entre si (TILAHUN; ONG, 2012). Este algoritmo se baseia no padrão de bioluminescência dos vagalumes para comunicação, atração de presas e parceiros potenciais.

No Firefly duas questões são consideradas, sendo elas: a intensidade do brilho emitido pelo vagalume que diminui à medida que a distância do objeto de interesse aumenta e a atratividade que é determinada pela tendência ao vagalume se aproximar de outros indivíduos com maior brilho. Esses fatores fazem com que a resposta de ação dos vagalumes está associada à função que se pretende otimizar (YANG; DEY; FONG, 2020).

Esse algoritmo tem a seguinte sequência de ações: 1. Os vagalumes (idenrtificadores de maior correlação entre variáveis de entrada e saída) podem ser atraídos por cada um dos indivíduos; 2. Esta atratividade é proporcional e relativa ao brilho emitido por cada vagalume; 3. À medida que a distância entre estes aumenta o brilho diminui; 4. Se não existir vagalumes mais brilhantes que os demais o movimento ocorrerá de forma aleatória (LINDFIELD; PENNY, 2017).

Considerando o princípio físico em que a intensidade da luz I é inversamente proporcional ao quadrado da distância d, esse método foi utilizado para definir uma função apropriada para a distância entre dois vagalumes (MACH; RONOH; LANGAT, 2023) em que a luz passa por um meio com coeficiente de absorção de luz γ que é indicado pela equação:

$$I_{(r)} = I_0 e^{-\gamma d^2} \tag{6}$$

Onde I_0 é a intensidade no ponto de origem. A intensidade pode ser calculada por:

$$I_{(r)} = \frac{I_0}{1+ \gamma d^2} \tag{7}$$

Segundo Tilahun e Choon Ong (2012) tal equação é empregada por ser computacionalmente mais fácil de calcular, da mesma forma a atratividade pode ser calculada por:

$$A_{(r)} = \frac{A_0}{1+ \gamma d^2} \tag{8}$$

Sendo o brilho proporcional ao valor da função objetivo, ou seja, às propriedades de luz intermitente da população dos vagalumes, o movimento deles é transformado em etapas conforme os mesmos se movimentam em uma região definida por:

$$\boldsymbol{X}_i^{(t+1)} = \boldsymbol{X}_i^{(t)} + A_0 e^{-\gamma d^2{}_{i,j}} \left(\boldsymbol{X}_i^{(t)} - \boldsymbol{X}_j^{(t)} \right) + \alpha r_i^t \tag{9}$$

Em que t é o número de interações ou gerações do processo, d é a distância entre os pares de vagalumes i e j, *r* é um vetor de números extraídos de uma distribuição gaussiana ou distribuição uniforme no tempo t, A_0 é a atratividade e se A_0= 0 o movimento aleatório torna-se simples, γ um parâmetro definido pelo usuário (geralmente 1) e α o parâmetro de randomização (LINDFIELD; PENNY, 2017).

Podem ser admitidos valores de $A_0 = 1$ e $\alpha \in [0, 1]$ e o termo de randomização pode ser facilmente estendido para uma distribuição normal N (0, 1) ou outras distribuições. O parâmetro γ caracteriza a variação da atratividade e seu valor é crucialmente importante para determinar a velocidade da convergência e como o algoritmo FA se comporta. Em teoria, $\gamma \in [0, \infty)$, mas na prática, $\gamma = 1$ é determinado pelo comprimento característico Γ do sistema a ser otimizado (YANG, 2010).

Os vagalumes irão mudar sua posição de forma sistemática ou aleatória para otimizar sua função de aptidão, até que a população se reúna em torno do mais brilhante. Dessa forma, três parâmetros são usados: atração (determinado pela diferença das intensidades de luz), randomização e absorção (YANG; KARAMANOGLU, 2013). Quando os parâmetros de absorção flutuam de 0 ao infinito, o valor dos parâmetros de atratividade muda. O movimento dos vagalumes parece ser aleatório no caso de convergência infinita, de acordo com os autores, quando o movimento aleatório é associado ao parâmetro de aleatoriedade do princípio da distribuição gaussiana é definido como zero, ele se comporta como se estivesse produzindo um número a partir de um intervalo (MACH; RONOH; LANGAT, 2023).

Como o vagalume mais brilhante apresentará a melhor solução global, ao movimentar-se aleatoriamente seu brilho pode diminuir dependendo da direção e reduzir o desempenho do algoritmo naquela interação específica. Então, para possibilitar que o vagalume mais brilhante

se mova apenas na direção que seu brilho melhore (TILAHUN; CHOON ONG, 2012). Assim, o movimento do vagalume mais brilhante será descrito por:

$$\boldsymbol{X}_i^{(t+1)} = \boldsymbol{X}_i^{(t)} + \alpha U \tag{10}$$

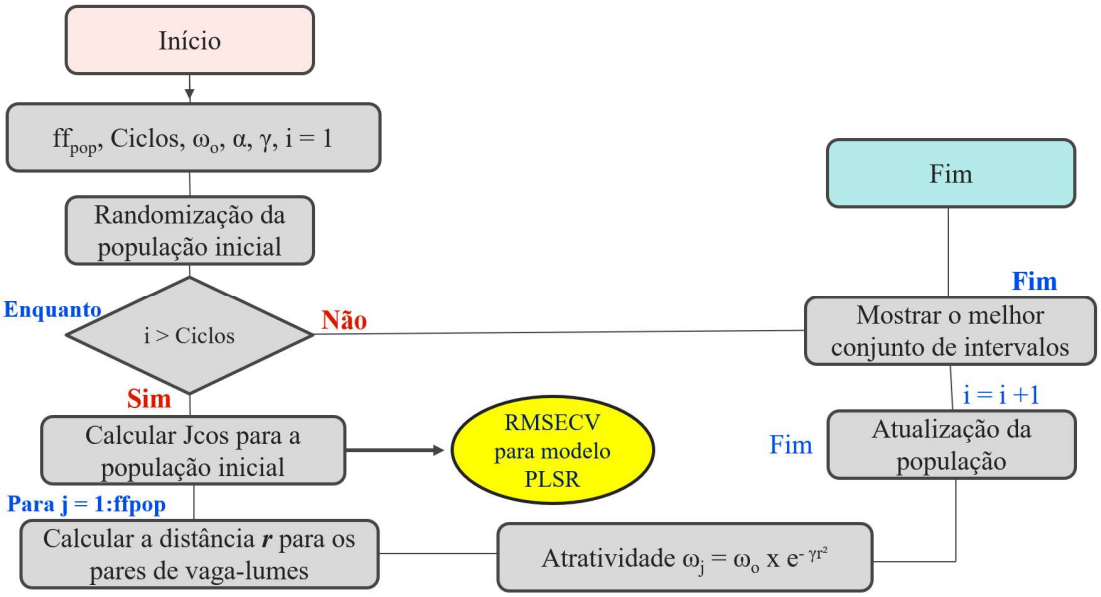
Ao otimizar o algoritmo, para determinar a direção do movimento do vagalume mais brilhante, serão gerados aleatoriamente vetores unitários e em seguida, uma direção, U, será escolhida entre as direções geradas aleatoriamente nas quais o brilho do vagalume mais brilhante aumenta. Caso contrário o vagalume permanecerá na posição atual (TILAHUN; CHOON ONG, 2012). O valor da atratividade também pode ser diferente de 1 para cada vagalume, sendo possível atribuir fontes de atratividade diferentes que dependem da intensidade do brilho do vagalume, dependente da função objetivo, conforme dado na equação:

$$A_{(0)} = \frac{I'_0}{I_0'} \tag{11}$$

Em que I'_0 é a intensidade em d = 0 para o vagalume mais brilhante e I_0 a intensidade para o vagalume que irá ser atraído por este, I_0 precisa ser diferente de zero. Se for adotado $A_{(0)}= I'_0$ e a intensidade for grande, o movimento do vagalume atraído por este será longo e por isso é interessante ajustar $A_{(0)}$ dependendo do espaço da solução, sendo ela diretamente proporcional a intensidade (TILAHUN; CHOON ONG, 2012).

Nesse contexto, Oliveira e colaboradores (2021) desenvolveram um algoritmo firefly para seleção de variáveis em PLSR para quantificação de algumas propriedades em solo. Foram assumidos os seguintes valores para os critérios de otimização: tratividade $A_{(0)}$ em d = 0 de 0,97, α como parâmetro de randomização sendo de 0,2 e o brilho, ou coeficiente de absorção de luz, γ, igual a 1. Após definir a população de vagalumes e a quantidade de ciclos em que o processo será repetido e, portanto, em cada ciclo o brilho de cada vagalume poderá ser alterado até que se obtenha um resultado convergente à solução otimizada do problema de seleção de variáveis. Assim, o brilho em cada ciclo para cada vagalume será proporcional à capacidade de produzir um bom condicionamento que será dado pela raiz quadrada do erro médio quadrático da validação cruzada (RMSECV). O modelo PLSR será construído a partir do vagalume com maior brilho quando o número de ciclos tiver sido alcançado. Nesta condição final o vagalume será um conjunto de variáveis como resposta (Figura 4).

Figura 4 – Algoritmo Firefly proposto para seleção de variáveis em PLSR



Fonte: Adaptado de OLIVEIRA et al (2021)

Este algoritmo forneceu bons resultados para as amostras utilizadas no estudo de Oliveira e colaboradores (2021) sendo considerada uma abordagem promissora para resolver problemas de seleção de intervalo usando PLSR visto que obteve resultados iguais ou melhores aos algoritmos PLS, iPLS e iSPA-PLS em termos de figuras de mérito selecionadas.

3.4.1 Algoritmo FA-PLS-DA proposto

Inicialmente, os parâmetros de entrada de dados selecionados são: as variáveis independentes do conjunto de treinamento (Train), o vetor coluna contendo os valores referentes às classes do conjunto de treinamento (Group_Train), a matriz das variáveis independentes do conjunto de validação (Val), o velor coluna contendo os valores referentes as classes do conjunto de validação (Group_Val), a matriz das variáveis independentes do conjunto de predição (Test), o vetor coluna contendo os valores referentes as classes do conjunto de predição (Group_Test), a quantidade de intervalos que o espectro será dividido, o número máximo de intervalos que devem ser selecionados (I_max), a quantidade de vagalumes (ff_{pop}), a quantidade de ciclos (*cycles*), a atratividade em d = 0 (w = 0,97), o coeficiente de absorção de luz (γ = 1), o parâmetro de randomização, ou seja, a percentagem aleatória no movimento dos vagalumes, (α = 0.2).

Através do brilho do vagalume mais brilhante (x_i), proporcional ao valor da função objetivo, que é obtido ao final do processo de busca, o algoritmo seleciona a região que apresenta as variáveis dos dados multivariados, assim, cada vagalume possui um conjunto de variáveis selecionadas na população inicial e o movimento deles é transformado em etapas conforme eles se movimental numa região definida por:

$$\boldsymbol{X}_j = \boldsymbol{X}_i + w_0 e^{-\gamma r^2{}_{i,j}} + \alpha r_1^{intervals} - \frac{1}{2} ones(1, intervals) \quad (12)$$

De modo que se adota que $0 \leq \alpha \leq 1$ e para uso prático é admitido que $w_0 = 1$ (TILAHUN; CHOON ONG, 2012). Após calcular o valor de X_j e armazenar as posições dos vagalumes (variáveis selecionadas), o menor valor obtido irá determinar a posição inicial do vagalume com maior brilho. Em seguida, novas posições são geradas para cada vagalume a partir de uma busca aleatória e a atualização só ocorre ao encontrar o melhor conjunto de intervalos, ou seja, caso o custo calculado para as novas posições dos for menor que o da iteração anterior e que, consequentemente, as variáveis selecionadas levam a um menor erro de classificação.

Por fim, a combinação das variáveis selecionadas (intervalos) na matriz é empregada na construção de um modelo PLS-DA usando a validação cruzada para determinar o número de variáveis latentes, sendo possível a obtenção de melhores subconjuntos de variáveis para melhorar a discriminação de classes de amostras.

Entretanto, nem sempre é possível utilizar os dados originais ou brutos devido a possíveis variabilidades nas medidas que podem tornar os modelos construídos não adequados. Inserem-se no rol de possíveis fenômenos que diminuem a qualidade dos modelos processos de espalhamentos da radiação que atinge a amostra, presença de ruídos instrumentais, dentre outros. Assim, pré-processamento de dados espectrais pode ser uma alternativa para garantir resultados melhores em termos das métricas das figuras de mérito.

3.5 Pré-processamento dos dados espectrais

A seleção de pré-processamento de dados deve estar relacionada ao objetivo químico ou físico da análise, ou seja, o tipo de amostra e/ou analito de interesse, já que podem influenciar nos modelos quimiométricos construídos (BRERETON, 2018).

Os espectros NIR, por exemplo, constituem combinações de características de absorção e dispersão, em que a absorção que é traduzida na presença de bandas em comprimentos de onda específicos e o espalhamento mediante efeitos aditivos modificando a linha de base. Nesse contexto, alguns pré-processamentos são empregados para correção de dispersão e alterações na linha de base, deslocamento de bandas (MISHRA et al, 2020).

O pré-processamento quimiométricos de dados espectrais é realizado visando de melhorar a capacidade preditiva de modelos quimiométricos empregando dados espectrais obtidos, essas abordagens são usadas para remover efeitos a variabilidade ou efeitos indesejados do sinal, para que as informações úteis sejam relacionadas às propriedades químicas de interesse, ou seja, o pré-processamento de dados reduz informações de dispersão dos espectros que levam a modelos insatisfatórios (MISHRA et al, 2021).

Entre os diferentes pré-processamentos utilizados estão incuídas a correção de espalhamento multiplicativo (MSC, do inglês *Multiplicative Scatter Correction*), derivação e suavização pelo método de Savitzky-Golay, além de correções de deslocamento ao longo da linha de base.

O método MSC é bastante utilizado na correção da linha de base. Tal método assume o comprimento de onda da luz espalhada como dependente da luz espalhada e a absorvida, podendo ser separados (SOUZA, et al. 2012). Segundo Ferreira (2015) para correção por MSC é feita tratando cada espectro *i* como vetor-colun ($\mathbf{x_i}$) a que pode ser considerado como uma função linear do espectro médio dado por:

$$x_i = a_i 1 + b_i x_m + e_i \tag{13}$$

Segundo Ferreira (2015) os parâmetros a_i e b_i são obtidos através de uma regressão de cada espectro i no espectro médio. Em seguida, após subtrair a_i de cada comprimento de onda do espectro original, o espectro corrigido por MSC é obtido, quando dividido por b_i é obtida a correção dos efeitos multiplicativos.

Por outro lado, problemas instrumentais ou de amostragem podem causar um deslocamento do espectro, que é chamado de deslocamento da linha de base. Esse deslocamento pode ser corrigido tirando as derivadas do espectro, aumentado a relação sinal/ruído (FERREIRA, 2015).

Ainda de acordo com Ferreira (2015) a suavização é utilizada para evidenciar a presença de pequenos picos que não são perceptíveis numa linha de base em que o ruído é alto. O filtro

digital de Savitzky-Golay é um algoritmo de suavização e filtragem dos dados, nele está presente as derivadas do sinal.

$$x_i = \frac{1}{2m+1} \cdot \sum_{j=-m}^{m} \omega_j \cdot x_i + j \quad (14)$$

Esse filtro possui coeficientes advindos de um ajuste linear não-ponderado de mínimos quadrados, a partir de um polinômio de determinada ordem no ponto central da janela (NISHIDA, 2017). De acordo com Ferreira (2015) a média ponderada entre os pontos da janela é usada para gerar um ponto central, para isso, pesos são obtidos ajustando-se um polinômio à janela e o valor do polinômio ajustado substitui o ponto central da janela que gera uma resposta é suavizada.

3.6 Figuras de mérito ao avaliar a eficiência dos algoritmos

Entre as figuras de mérito mais utilizadas nos métodos de classificação, como na modelagem PLS-DA, consiste na taxa de erro de classificação, na sensibilidade, na especificidade e na taxa correta de classificação ou exatidão (LEE; LIONG; JEMAIN, 2018).

Para avaliar a taxa de erros durante a classificação os resultados são organizados na forma de uma matriz chamada de matriz de confusão, em que as classes verdadeiras são representadas em colunas *versus* as classes estimadas pelo modelo construído nas linhas, sendo assim, as amostras corretamente classificadas estarão dispostas na diagonal principal (FERREIRA, 2015). Tais figuras de mérito são expressas em porcentagem e são definidas pelas equações a seguir:

$$\text{SENSIBILIDADE} = \frac{\text{VP}}{\text{VP+FN}} \quad (15)$$

Em que VP são os verdadeiros positivos e FN os falsos negativos. A sensibilidade é a medida do quão bem o modelo é capaz de classificar uma amostra que pertence a uma classe (FERREIRA, 2015). A sensibilidade é definida para cada classe *C* como a porcentagem de amostras dessa classe que são reconhecidas corretamente, ou seja, a taxa de verdadeiros positivos (RODIONOVA; POMERANTSEV, 2020).

A especificidade descreve a habilidade que o modelo tem de identificar amostras não pertencentes a uma classe e que foram corretamente classificadas. Essa métrica consiste na

razão dos verdadeiros negativos (não pertencentes à classe em que estão) pela soma dos verdadeiros negativos e falsos positivos, isto é, amostras classificadas incorretamente em uma dada classe.

$$\text{ESPECIFICIDADE} = \frac{\text{VN}}{\text{FP+VN}} 100 \tag{16}$$

Outro parâmetro utilizado consiste na taxa correta de classificação, ou exatidão, que consiste na proporção de resultados verdadeiros entre o número total de amostras (GRESSLING, 2021), sendo definida por:

$$\%\text{TCC} = \frac{\text{VP+VN}}{\text{VP+FN+FP+V}} 100 \tag{17}$$

Segundo Ferreira (2015) quando o modelo prevê corretamente todas as amostras, tanto a sensibilidade quanto a especificidade serão iguais a 1 ou 100%. Ou seja, a taxa correta de classificação informa a quantidade de amostras foram de fato classificadas adequadamente, atribuindo o mesmo peso para ambos os erros.

4 METODOLOGIA

O desempenho do algoritmo proposto FA-PLS-DA foi avaliado a partir de três bancos de dados, sendo eles: 4.1 Discriminação de amostras de saliva de pacientes infectados por SARS-CoV-2 empregando FTIR; 4.2 Classificação *in-situ* do leite de cabra em termos de adulteração pela adição de leite de vaca; 4.3 Classificação de aceites de oliva de diferentes localidades empregando FTIR; e 4.4 Banco de dados simulado para verificação do funcionamento do algoritmo. Nas seções 4.1 a 4.4 será descrito o conjunto de dados e o procedimento quimométrico utilizado para cada um dos problemas estudados.

4.1 Classificação de pacientes infectados por SARS-CoV-2 a partir de espectro no infravermelho de amostras de saliva

O primeiro conjunto de dados analisado consistiu-se de espectros ATR-FIR na faixa de 4.000 a 650 cm^{-1}, perfazendo 1798 variáveis, de 104 amostras de saliva para avaliar a presença ou ausência de SARS-CoV-2. Foram 44 amostras de saliva fornecidas por pacientes saudáveis (sem a presença do vírus) e 60 amostras provenientes de pacientes com o SARS-CoV-2. As classes foram denominadas, respectivamente, de $COVID^{POS}$ e $COVID^{NEG}$.

Os espectros das amostras foram pré-processados por correção de espalhamento multiplicativo (MSC, do inglês *Multiplicative Scatter Correction*), variação normal padrão (SNV, do inglês *Standard Normal Variate*), correções de linha de base *offset* e derivação pelo método de Savitzky-Golay (empregando primeira derivada, polinômio de segundo grau e janela de 15 pontos), além da suavização (empregando polinômio de segundo grau e janela de 15 pontos) utilizando o *Data Hand Gui* (GOMES, 2013).

Em seguida, os conjuntos de dados, brutos e pré-processados, foram separados em conjunto de treinamento e de teste utilizando o algoritmo *Kennard-Stone* (KENNARD; STONE, 1969) em ambiente Matlab® 2010b. Foram selecionadas 75% das amostras para o conjunto de treinamento e 25% para o de teste, conforme descrito na Tabela 1. O modelo foi validado utilizando validação cruzada completa.

Tabela 1 - Partição das amostras de COVID-19

Classe	Treinamento	Teste
$COVID^{POS}$	45	15
$COVID^{NEG}$	33	11
Total	78	26

Fonte: Elaborada pela autora, 2023.

As amostras de treinamento foram utilizadas para construção dos modelos quimiométricos e sua validação utilizando os algoritmos PLS-DA e FA-PLS-DA, ambos em ambiente Matlab® 2010b. Em seguida as amostras do conjunto de teste foram classificadas pelo modelo proposto e matrizes de confusão para ambos os conjuntos de amostra, treinamento e teste, foram apresentados.

4.2 Classificação *in-situ* do leite de cabra contendo leite de vaca como adulterante

O segundo banco de dados utilizados consiste em espectros NIR na faixa de 900-1650 nm, perfazendo 210 variáveis, de 192 amostras de leite de cabra. Esses dados foram obtidos de amostras coletadas em microrregiões do estado da Paraíba em outubro de 2019, totalizando 6 lotes de leite de cabra e 4 lotes de leite de vaca. Após a adulteração as amostras foram homogeneizadas e medidas em um espectrofotômetro NIR portátil (PEREIRA et al, 2021).

Foram obtidas 54 amostras de leite de cabra não adulteradas, e 138 amostras adulteradas pela adição de leite de vaca em diferentes níveis de adulteração (PEREIRA et al, 2021). As classes foram denominadas, respectivamente, de LC (leite de cabra) e LA (leite adulterado). As amostras foram adulteradas com 5 a 75 g de leite de vaca por 100 gramas de leite de cabra, variando de 5 em 5g/100g da amostra. Segundo os autores, é possível diferenciar a composição química das amostras de leite de cabra e de vaca devido a alta variabilidade dos espectros puros que são observados devido a influências como alimentação, sistema de produção, entre outros.

Posteriormente, os dados foram pré-processados empregando diferentes pré-processamentos nos dados sendo eles: suavização e derivação pelo método de Savitzky-Golay empregando polinômio de segundo grau e janela de 7 pontos, MSC, SNV e correção de deslocamento ao longo da linha de base. Em seguida, o algoritmo KS foi utilizado para a partição das amostas (Tabela 2) sendo utilizadas 75% das amostras para treinamento e 25% para teste.

Tabela 2 - Partição das amostras de leite de cabra

Classe	Treinamento	Teste
LC	40	14
LA	103	35
Total	143	49

Fonte: Elaborada pela autora, 2023.

As amostras de treinamento também foram utilizadas para construção dos modelos quimiométricos e sua validação utilizando os algoritmos PLS-DA e FA-PLS-DA, ambos em ambiente Matlab® 2010b.

4.3 Classificação de azeites de oliva de diferentes localidades empregando FTIR

Outro banco de dados utilizado consiste de amostras de azeite de oliva extra-virgem provenientes de quatro países europeus em que foram medidos espectros na região do infravermelho próximo com transformada de Fourier. O banco de dados foi obtido da International Olive Oil Council in Madrid no estudo realizado por Tapp, Defernez e Kemsley (2003) frente ao PLS-DA.

A faixa espectral foi de 800-4000 cm^{-1}, perfazendo 570 variáveis, em 120 amostras. Os azeites de oliva extra-virgem pertencem a diferentes grupos sendo eles: GR – Amostras da Grécia (20); IT – Amostras da Itália (34); PO – Amostras de Portugal (16); ES – Amostras da Espanha (50).

Os dados brutos e pré-procesados empregando diferentes pré-processamentos como: SNV e correção de deslocamento ao longo da linha de base foram particionados utilizando o algoritmo KS para a partição das amostas (Tabela 3) buscando utilizar 75% para as amostras de treinamento e 25% para as de teste.

Tabela 3 - Partição das amostras de azeite de oliva

Classe	Treinameno	Teste
GR	15	5
IT	25	9
PO	12	4
ES	37	13
TOTAL	89	31

Fonte: Elaborada pela autora, 2023.

4.4 Dados simulados

Por fim, foi utilizado um banco de dados elaborado para simular um sistema de classificação para verificação do desempenho do algoritmo. Este banco de dados foi construído pelo somatório do perfil gaussiano em que as bandas estão deslocadas uma da outra, mas mantendo um grau de sobreposição e adicionado ruído, que é determinado por:

$$f(x) = \mathrm{ae}^{\frac{-(\mathrm{x}-\mathrm{b})^2}{\mathrm{c}}} \tag{18}$$

Em que a, b e c são constantes associadas a altura, centro e largura das bandas respectivamente. O banco de dados compreendeu 135 observações e foi gerado usando quatro fatores (A1, A2, A3 e A4) com 600 variáveis. Esses fatores foram combinados de forma a gerar três classes com as seguintes composições: Classe 1 (fatores A1, A3 e A4), Classe 2 (fatores A2, A3 e A4) e Classe 3 (fatores A1, A2, A3 e A4). A variabilidade dentro da classe foi simulada misturando os fatores usando coeficientes aleatórios extraídos de uma distribuição gaussiana com média unitária e desvio padrão de 0,3. Além disso, ruído gaussiano com média zero e desvio padrão de 0,001 foi adicionado a todas as respostas.

Para partição das amostras também foi utilizado o algoritmo KS para a partição das amostas (Tabela 4) sendo utilizadas 66,67% das amostras para treinamento e 33,33% para teste.

Tabela 4 - Partição – Banco de dados simulados

Classe	Treinameno	Teste
1	30	15
2	30	15
3	30	15
TOTAL	90	45

Fonte: Elaborada pela autora, 2023.

Assim como nos bancos de dados mencionados anteriormente, as amostras de treinamento foram utilizadas para validação cruzada para determinar o número ideal de variáveis latentes utilizadas para construção dos modelos. Os resultados desse banco de dados também foram comparados avaliando as matrizes de confusão.

O banco de dados simulados foi utilizado buscando avaliar regiões das Gaussianas em que as variáveis selecionadas pelo algoritmo apresentem informações relevantes para que possam ser classificadas as diferentes classes.

5 RESULTADOS E DISCUSSÃO

5.1 Algoritmo firefly em análise discriminante linear por mínimos quadrados parciais – FA-PLS-DA

O algoritmo Firefly proposto por Oliveira e colaboradores (2021) foi adaptado para uma nova aplicação em a análise discriminante linear (LDA). Foram assumidos os mesmos valores de atratividade $A_{(0)}$ quando d = 0, de 0.97, percentagem aleatória no movimento dos vagalumes α sendo de 0.2, o coeficiente de absorção de luz, γ, igual a 1, a quantidade de ciclos utilizada foi 50 e a população de vagalumes (ff_{pop} = 50). Considerando que o brilho é proporcional à capacidade de produzir um bom condicionamento pelo RMSECV, os modelos foram construídos considerando apenas os intervalos selecionados pelo vagalume de maior brilho.

A validação do modelo PLS-DA foi realizada com as matrizes de treinamento utilizando a validação cruzada buscando determinar a quantidade ideal de variáveis latentes/fatores levando em consideração a menor taxa de erro.

O conjunto de variáveis foi dividido em intervalos não sobrepostos, nessa etapa os espectros dos dados brutos e pré-processados foram divididos em 20, 15, 10 e 5 intervalos.

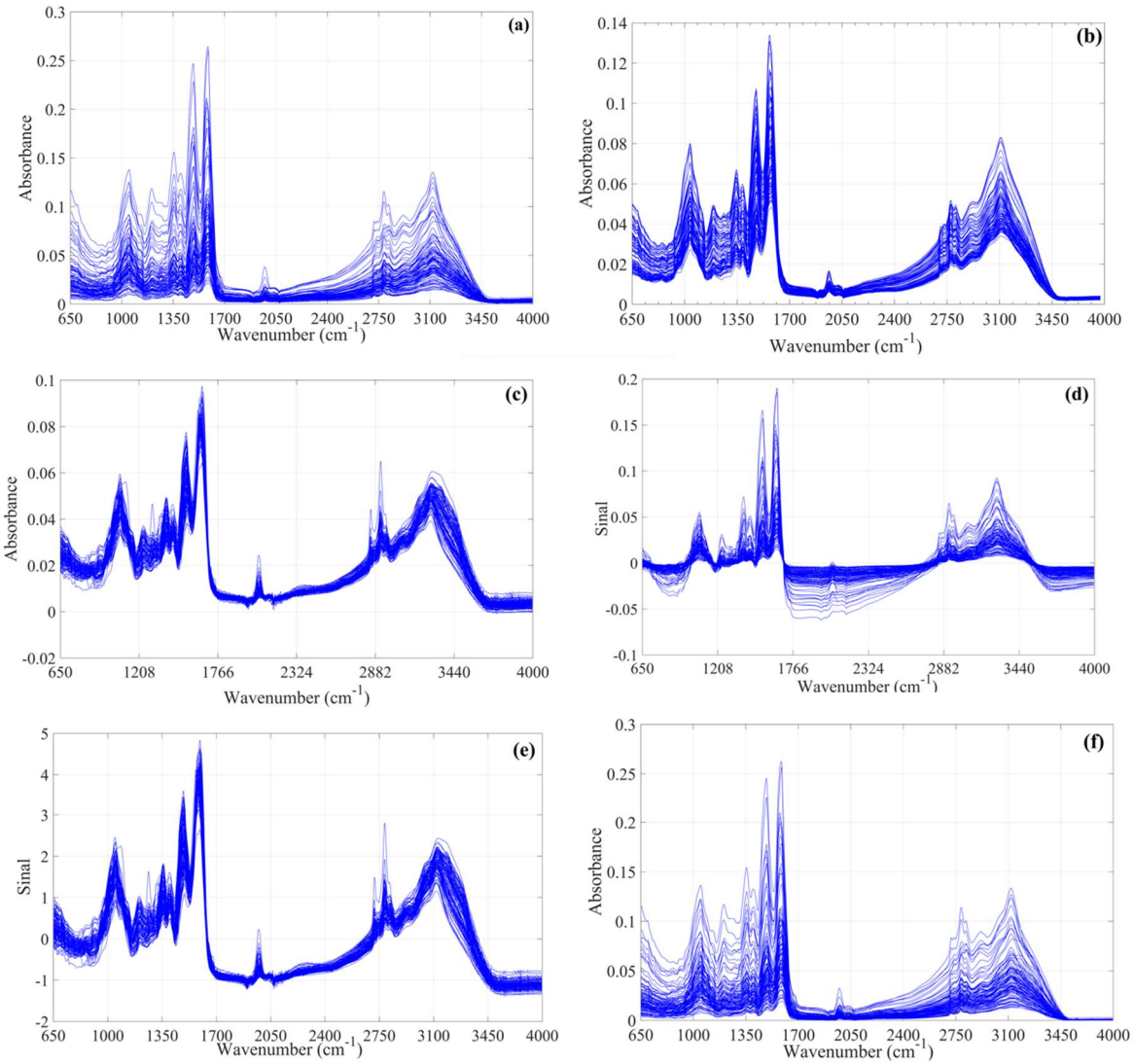
Para cada intervalo de variáveis, um vetor coluna das variáveis de treinamento foi tomado como elementos representativos de cada respectivo intervalo e reunido em uma nova matriz com *n* dimensões que é utilizada para combinação de intervalos do modelo construído. Para avaliar a capacidade preditiva do algoritmo proposto foram utilizados os parâmetros de desempenho: especificidade, sensibilidade e taxa correta de classificação.

As seções a seguir apresentam os resultados da aplicação do algoritmo proposto FA-PLS-DA e sua comparação com o algoritmo PLS-DA nos conjuntos de dados utilizados.

5.2 Dinstinção entre pacientes infectados ou não por SARS-CoV-2 a partir de espectro no infravermelho de amostras de saliva

A **Figura 5** apresenta o perfil espectral no infravermelho dos dados brutos e pré-processados das amostras de saliva de participantes para avaliação das respostas patofisiológicas à infecção por SARS-CoV-2.

Figura 5 - Dados espectrais das amostras de COVID-19 (a) Brutos e pré processados por (b) Suavização pelo método de Savitzky-Golay (w15pol2); (c) MSC; (d) Baseline – Linear; (e) SNV; (f) Baseline – Offset.



Fonte: Elaborado pela autora, 2023.

Os espectros foram pré-processados com o intuito de remover informações instrumentais ou físicas que não estão associadas à informação química de interesse, devido à variabilidade entre as amostras e até mesmo ao ruído instrumental. Em seguida as amostras foram particionadas e foi avaliada a eficiência do algoritmo proposto frente ao PLS-DA. De acordo com as métricas de avaliação de desempenho, os espectros pré-processados empregando a suavização pelo método de Savitzky-Golay (janela de 15 pontos e polinômio de segunda ordem) mostraram melhor desempenho para o PLS-DA e FA-PLS-DA utilizando janelas de 10 intervalos.

A eficiência dos modelos de classificação é medida em função da capacidade que ele possui de classificar corretamente as amostras em suas respectivas classes, para a classificação das amostras de COVID-19 a PLS-DA foram selecionadas 4 variáveis latentes. Para as amostras

do conjunto de validação o método obteve 98,72% de taxa de classificação correta, sendo uma amostra da classe 1 (controle/COVID^{POS}) classificada como classe 2 (COVID^{NEG}), a especificade das amostras classificadas como COVID^{POS} foi de 96,67%, sendo a especificidade relacionada a capacidade do modelo de classificar como não pertencente a classe modelada as amostras não pertencentes a mesma. A amostra que foi classificada como COVID^{POS} pertencente a classe 2 (COVID^{NEG}) mostrou bandas reduzidas de sacarídeos (ribose) que é uma característica das amostras de COVID^{POS}, além de bandas com maior absorção na região alifática que são características às amostras COVID^{POS}, o que pode ter levado a sua atribuição à classe 1 na construção do modelo PLS-DA já que tais variáveis não foram selecionadas para construção do modelo FA-PLS-DA.

Para o algoritmo proposto nesse estudo, o FA-PLS-DA selecionou 3 variáveis latentes, um número menor com relação ao PLS-DA que selecionou 4 LVs. Além disso, o algoritmo mostrou um desempenho superior com relação a taxa de classificação correta (%TCC) de 100% para as amostras nas etapas de validação cruzada e teste. Já o PLS-DA obteve TCC 98,72% apresentando menor especificidade da classe 1 (96,97%) e menor sensibilidade da classe 2 (96,97%), como é possível observar na Tabela 5.

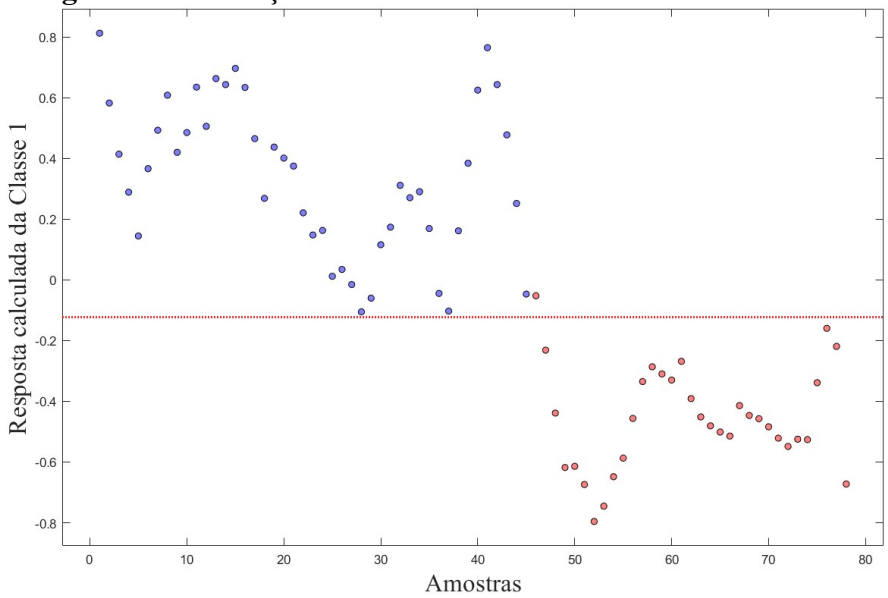
Tabela 5 - Resultados de classificação de COVID-19

PLS-DA (SG Smothing_15w_2pol[a]) LV = 4					
		Validação cruzada		**Teste**	
Classe verdadeira		**1**	**2**	**1**	**2**
Classe atribuída	**1**	45	0	15	0
	2	1	32	0	11
Sensibilidade (%)		100	96,97	100	100
Especificidade (%)		96,97	100	100	100
TCC (%)		98,72		100	
FA-PLS-DA (SG Smothing_15w_2pol _10) LV = 3					
		Validação cruzada		**Teste**	
Classe verdadeira		**1**	**2**	**1**	**2**
Classe atribuída	**1**	45	0	15	0
	2	0	33	0	11
Sensibilidade (%)		100	100	100	100
Especificidade (%)		100	100	100	100
TCC (%)		100		100	

Fonte: Elaborado pela autora, 2023.

Na **Figura 6** podemos observar o resultado da validação cruzada do modelo PLS-DA construído com 4 variáveis latentes, sendo as amostras em azul as da Classe 1 (controle/COVID^{POS}) e as amostras em vermelho da Classe 2 (COVID^{NEG}).

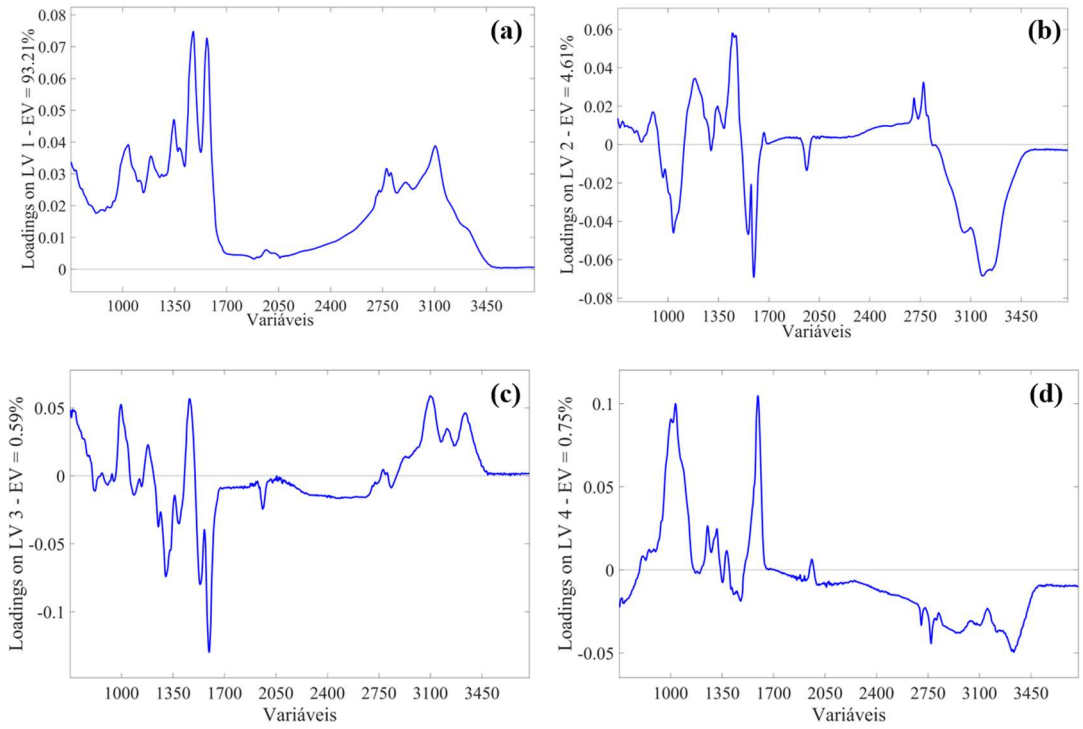
Figura 6 – Validação das amostras externas no modelo PLS-DA



Fonte: Elaborado pela autora, 2023**.**

De acordo com a **Figura 7**, podemos observar que as 4 variáveis latentes trazem informação para o modelo, em que variáveis que apresentam valores superiores a 1 representam as regiões espectrais mais significativas.

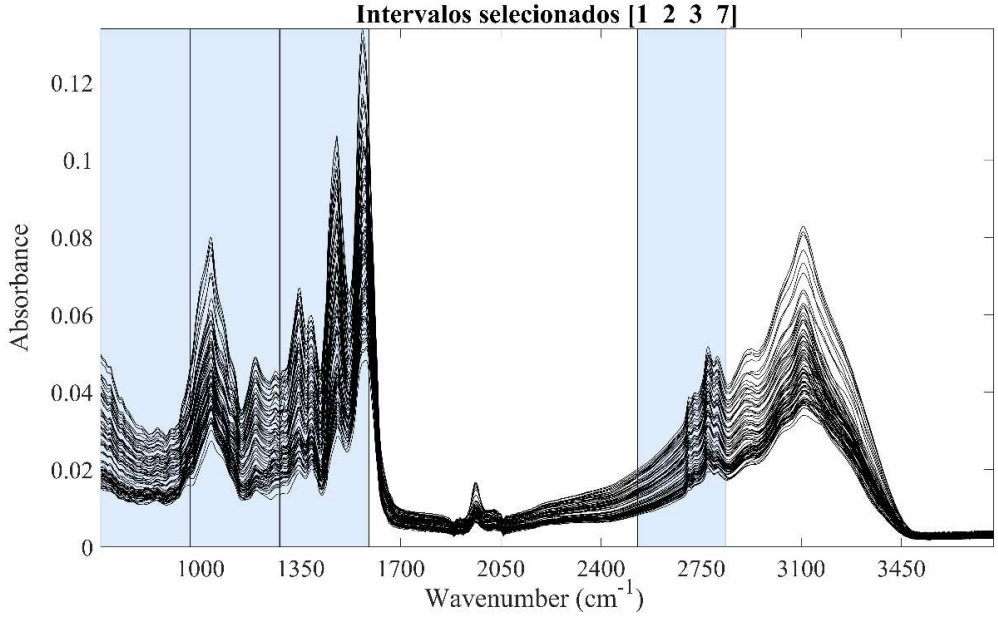
Figura 7 – Loadings nas variáveis latentes (a) LV – 1; (b) LV – 2; (c) LV – 3; (d) LV – 4.



Fonte: Elaborado pela autora, 2023**.**

O algoritmo FA-PLS-DA selecionou 4 intervalos empregando a suavização pelo método de Savitzky-Golay como pré-processamento e janelas de 10 intervalos, o subconjunto de variáveis selecionadas corresponde aos intervalos 1, 2, 3 e 7 respectivamente, sendo selecionadas 714 variáveis espectrais de um total de 1798 variáveis. De acordo com Kzmer e colaboradores (2022) diferenças significativas são observadas nas regiões alifáticas (intervalo 7) e amidas I, II e III (intervalos 1-3) (**Figura 8**), já que mudanças na entrutura de proteínas são dadas pelo aumento da intensidade na região da amida II em amostras infectadas pelo vírus SARS-CoV-2 ($COVID^{POS}$).

Figura 8 - Intervalos selecionados pelo FA-PLS-DA das amostras de COVID-19



Fonte: Elaborado pela autora, 2023**.**

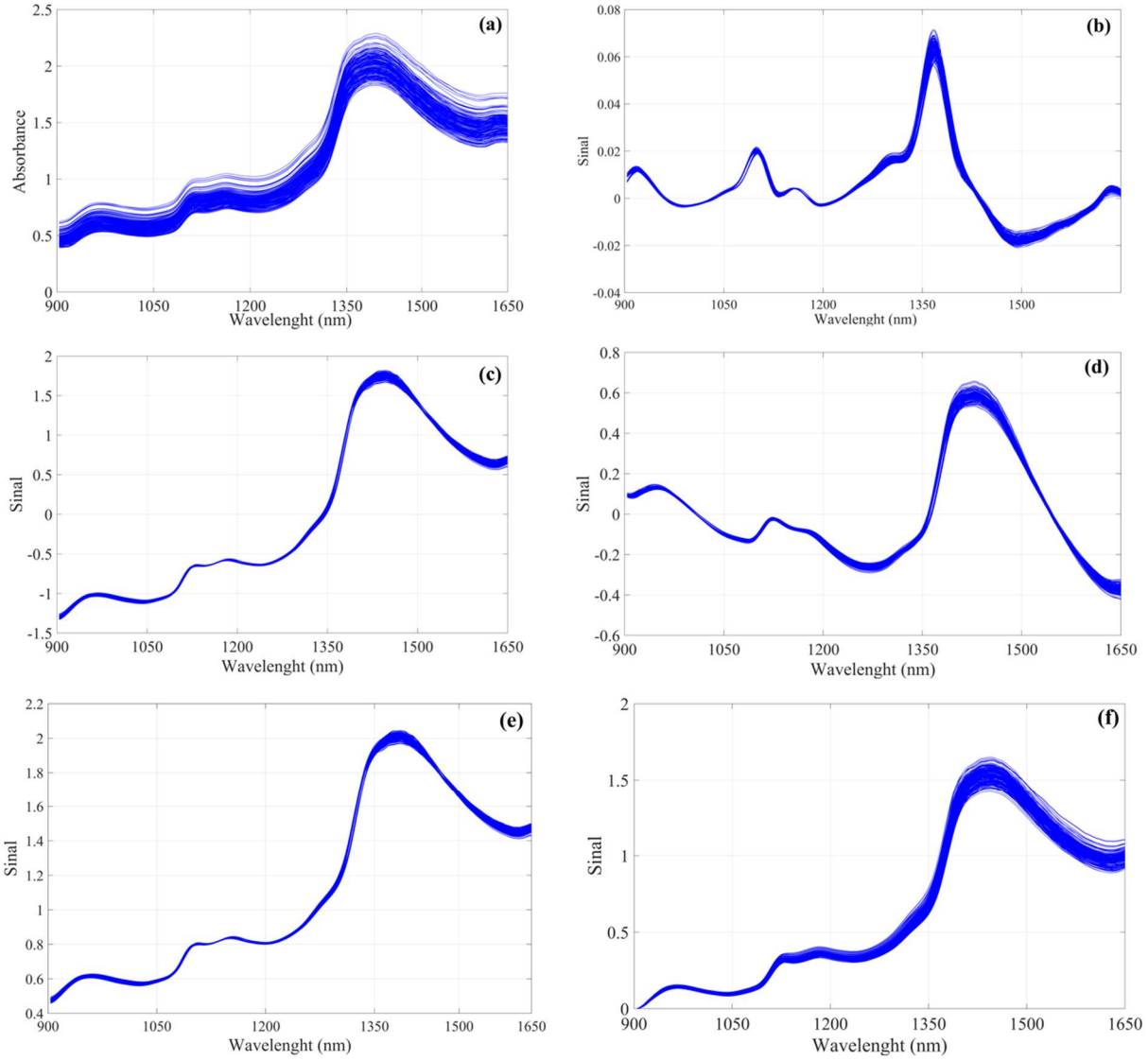
O algoritmo FA-PLS-DA apresentou um bom desempenho como o PLS-DA, no entanto, o FA-PLS-DA mostrou-se mais parcimonioso, empregando menor quantidade de variáveis latentes, alcançados resultados superiores de TCC (%) e menos susceptível a falsos positivos, favorecendo o desenvolvimento de aparelhos dedicados a este tipo de análise.

5.3 Classificação *in-situ* do leite de cabra

A **Figura 9** apresenta os espectros NIR das amostras de leite que compreendeu 192 observações e 204 variáveis, é observada uma similaridade entre o perfil espectral das amostras desse banco de dados obtido do estudo realizado por Pereira e colaboradores (2021) levando a

necessidade de selecionar regiões espectrais que possam discriminar as amostras de leite de cabra das amostras adulteradas.

Figura 9 - Espectros NIR das amostras de leite de cabra (a) brutos; e pré-processados por: (b) Derivação pelo método de Savitzky-Golay (w7pol2der2); (c) Suavização pelo método de Savitzky-Golay (w7pol2) + SNV; (d) Suavização pelo método de Savitzky-Golay (w7pol2) + Baseline (Linear); (e) Suavização pelo método de Savitzky-Golay (w7pol2) + MSC; (f) Suavização pelo método de Savitzky-Golay (w7pol2) + Baseline – Offset.



Fonte: Elaborado pela autora, 2023**.**

Como foi observado um deslocamento da linha de base e inclinação dos espectros em relação ao zero no eixo das ordenadas, os espectros foram pré-processados empregando métodos de correção de deslocamento de linha de base associada a suavização pelo método de Saviztky-Golay empregando janela de 7 pontos e polinômio de segunda ordem. A derivação pelo método de Saviztky-Golay também foi utilizada empregando primeira derivada e janela de 7 pontos, polinômio de segunda ordem também com o intuito de corrigir tendências na linha de base.

Os resultados empregando correção de linha de base associada a suavização pelo método de Savitzky-Golay (janela de 7 pontos e polinômio de segunda ordem) mostraram melhor desempenho e, portanto, foram selecionados para comparação entre os métodos utilizados.

Na tabela 6, é possível observar as figuras de mérito obtidas para ambos os métodos a classe 1 refere-se às 54 amostras de leite puro de cabra particionadas (sendo 40 utilizadas para treinamento e 14 para o conjunto de teste), já a classe 2 refere-se às 138 amostras de leite de cabra adulteradas também particionadas (103 utilizadas para o conjunto de treinamento e 35 no conjunto de teste).

O algoritmo proposto, FA-PLS-DA utilizando janelas de 5 intervalos selecionou 7 variáveis latentes com taxa correta de classificação de 97.20% e 97.96% para as amostras do conjunto de validação e de teste, respectivamente. Além disso, a especificidade e a sensibilidade do algoritmo proposto superaram as do método PLS-DA como é possível observar na Tabela 6, em que 7 amostras do conjunto de validação foram classificadas de forma incorreta.

Tabela 6 - Resultados de classificação – leite de cabra

PLS-DA (SG Smothing_7w_2pol + Linear) LVs = 16							
		Validação cruzada			**Teste**		
Classe verdadeira		**1**	**2**	**3**	**1**	**2**	**3**
Classe atribuída	**1**	40	0	0	14	0	0
	2	0	103	0	0	35	0
Sensibilidade (%)		100	100		100	100	
Especificidade (%)		100	100		100	100	
TCC (%)		100			100		
FA-PLS-DA (SG Smothing_7w_2pol + Linear_10) LVs = 7							
		Validação cruzada			**Teste**		
Classe verdadeira		**1**	**2**	**3**	**1**	**2**	**3**
Classe atribuída	**1**	35	5	0	13	1	0
	2	1	102	0	1	34	0
Sensibilidade (%)		85,50	99,03		92,86	97,14	
Especificidade (%)		99,07	87,80		97,22	93,33	
TCC (%)		95,80			95,92		

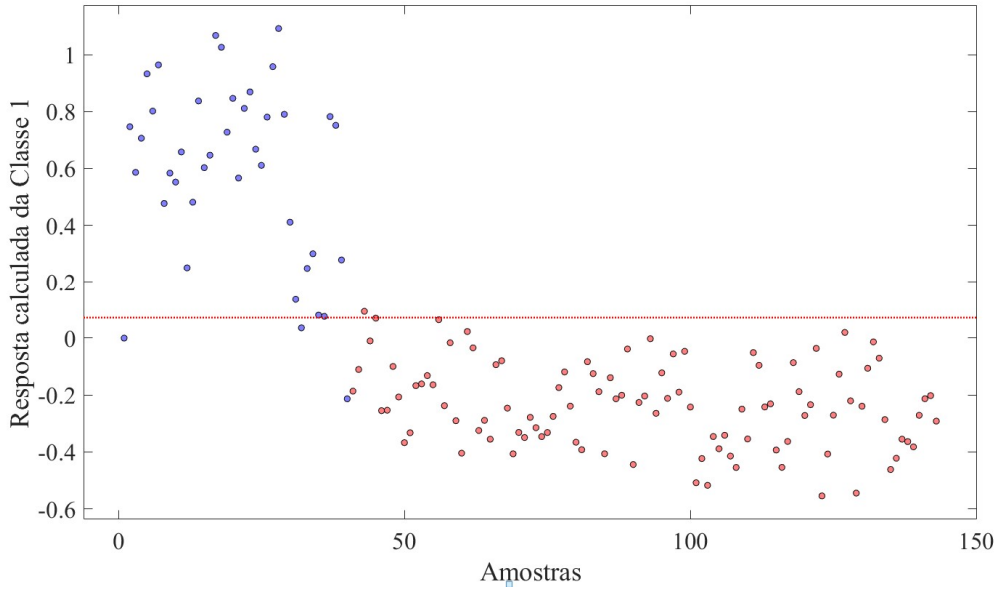
Fonte: Elaborado pela autora, 2023

No conjunto de teste o PLS-DA classificou adequadamente todas as amostras, contudo empregou uma quantidade superior de variáveis latentes com relação ao algoritmo FA-PLS-DA.

O algoritmo proposto classificou cinco amostras de leite de cabra puro como adulteradas e uma amostra de leite de cabra adulterada como leite puro no conjunto de validação cruzada. Obtendo uma taxa de classificação correta de 95.80%. Já na etapa de teste, o FA-PLS-DA classificou corretamente 34 amostras adulteradas, obtendo 97.22% de especificidade para as amostras não adulteras e uma taxa de classificação correta de 95.92%. O FA-PLS-DA empregou uma quantidade menor de variáveis latentes que o PLS-DA. Isto pode estar relacionado ao fato de que o algoritmo estocástico é mais parcimonioso frente ao PLS-DA forçando uma quantidade menor de variáveis latentes para a construção do modelo.

Na **Figura 10** é possível observar a validação das amostras externas no modelo PLS-DA, sendo as amostras em azul pertencentes à Classe 1 e as amostras em vermelho pertencentes à Classe 2. Mesmo atingindo TCC de 100% nas etapas de validação e predição, é possível observar que algumas amostras se encontram ao limiar de classificação definido pelo PLS-DA, isso pode refletir no desempenho da classificação, o que não ocorre, provavelmente devido ao uso de uma grande quantidade de variáveis latentes (LV = 16).

Figura 10 – Validação das amostras externas no modelo PLS-DA

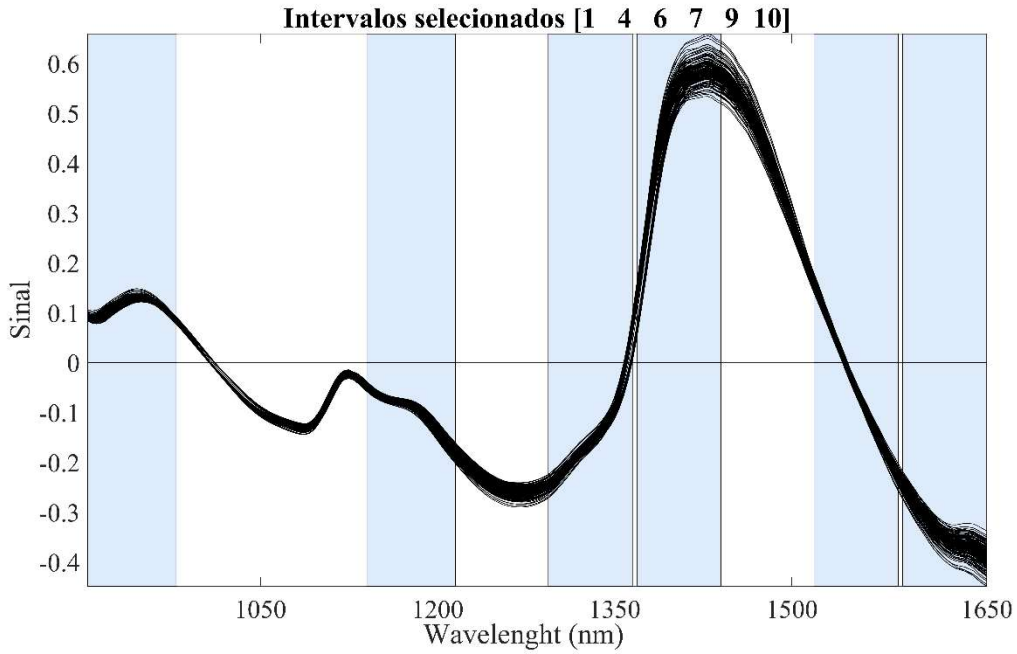


Fonte: Elaborado pela autora, 2023**.**

Buscando verificar a eficiência do FA-PLS-DA em selecionar variáveis informativas que explicam resultados melhores com relação ao PLS-DA, as variáveis selecionadas para a

construção do modelo FA-PLS-DA foram avaliadas de acordo com o estudo realizado por Pereira e colaboradores (2021). Foram selecionadas cerca de 122 variáveis espectrais, dos intervalos 1, 4, 6, 7, 9 e 10 (**Figura 11**).

Figura 11 - Intervalos selecionados pelo FA-PLS-DA das amostras de leite de cabra empregando suavização pelo método de Savitzky-Golay utilizando janela de 7 pontos e polinômio de segundo grau associada a correção de linha de base (Linear) e 10 janelas de intervalos.

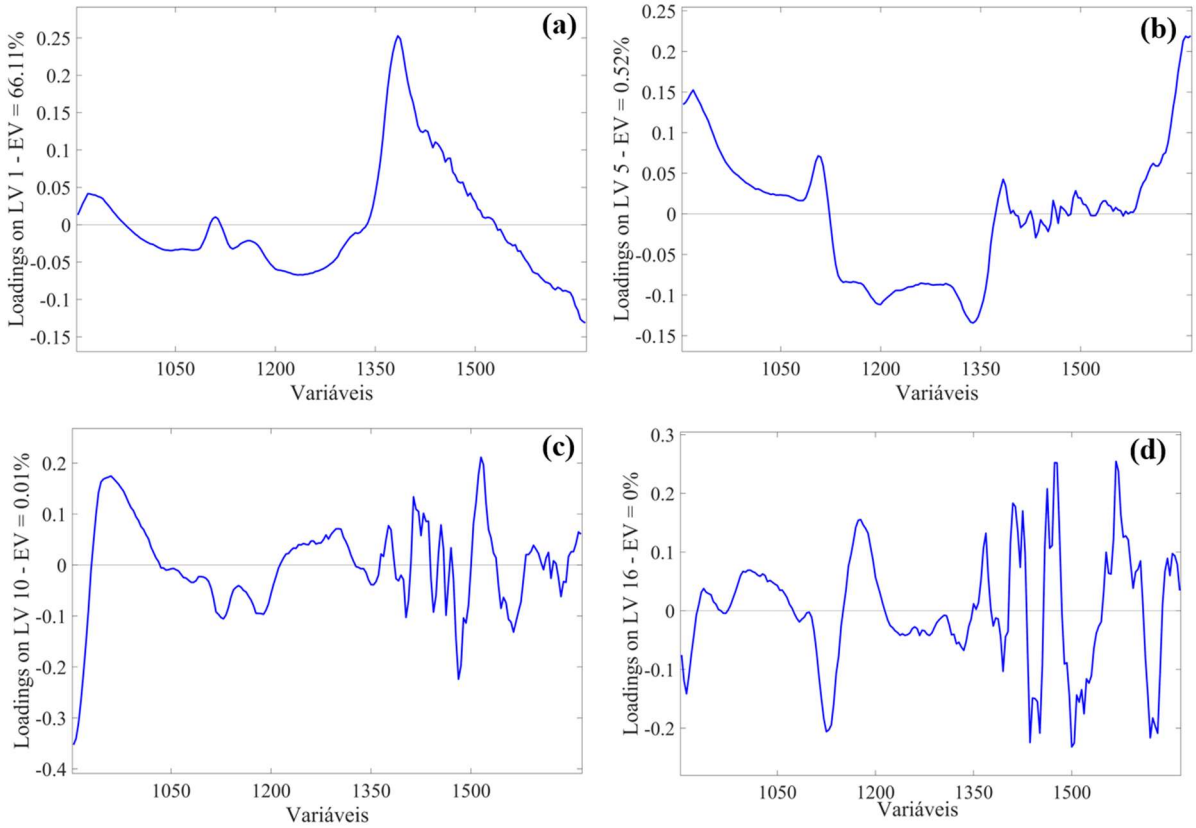


Fonte: Elaborado pela autora, 2023**.**

A banda em torno do intervalo 7 selecionado (1450 nm) pelo algoritmo proposto corresponde a região do primeiro harmônico da vibração de estiramento O-H da água. Já as bandas do intervalo 4 selecionado (1100-1200 nm) consistem na região espectral que está correlacionada com informações de proteínas e gorduras, em que há uma pequena banda correspondente ao segundo harmônico de -CH da gordura. Já o intervalo 10 (cerca de 1630 nm) corresponde a primeira combinação harmônica N-H estiramento/amida II ocorre provavelmente pela presença de proteína (PEREIRA et al, 2021).

Na **Figura 12** é possível observar os Loadings nas variáveis latentes, em que a partir da LV – 10 foi observada uma contribuição mínima, o que corrobora que o modelo PLS-DA empregando 16 LVs está sujeito a overfitting.

Figura 12 – Loadings nas variáveis latentes (a) LV – 1; (b) LV – 5; (c) LV – 10; (d) LV – 16.



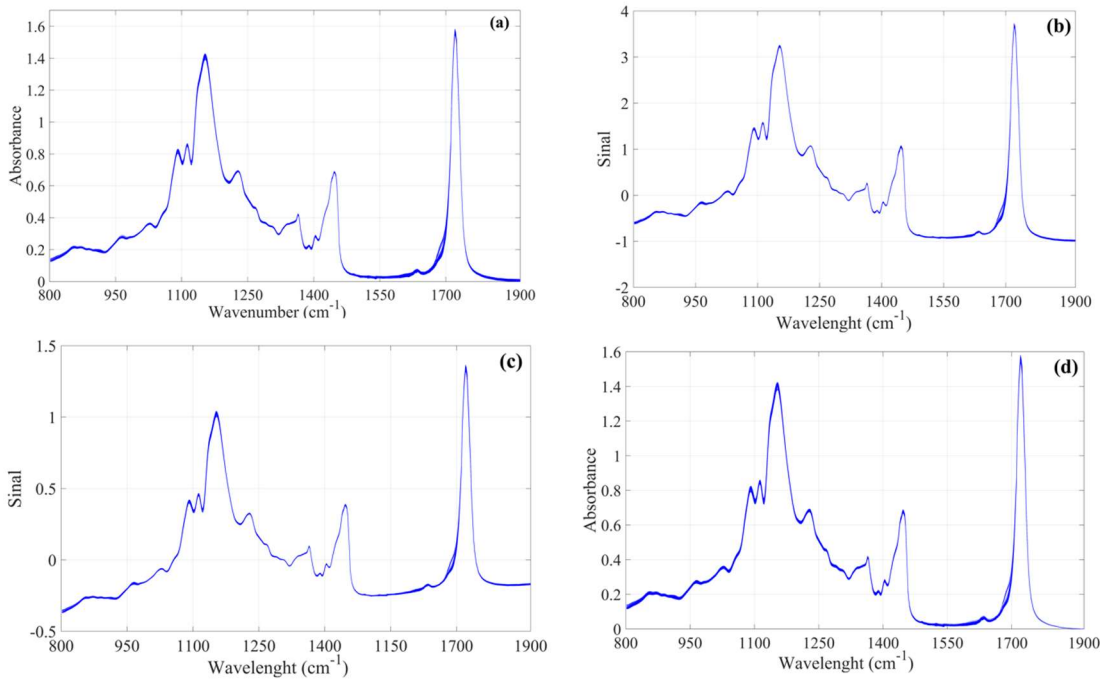
Fonte: Elaborado pela autora, 2023.

5.4 Classificação das amostras de azeite de oliva extra-virgem

O terceiro conjunto de dados utilizado consistiu em espectros FTIR (**Figura 13**) de amostras de azeite de oliva extra virgem de diferentes localidades provenientes do Conselho Oleícola Internacional de Madrid de um estudo realizado por Henri; Defernez e Kemsley (2003) que compreendeu 120 observações e 570 variáveis.

As amostras foram provenientes de quatro países europeus na faixa 800 cm^{-1} a 4000 cm^{-1}, ou seja, na região MID-IR/NIR. Foi observada uma similaridade entre o perfil espectral das amostras desse banco de dados, além disso, os espectros apresentaram baixo ruído. Os espectros apresentaram um leve deslocamento de linha de base em regiões específicas, além de bandas de absorção próximas levando a necessidade de selecionar regiões espectrais que possam discriminar as amostras de azeite de oliva.

Figura 13 - Dados de Azeite de oliva extra virgem (a) Brutos e pré-processados por (b) SNV; (c) Baseline – Linear; (d) Baseline – Offset.



Fonte: Elaborado pela autora, 2023**.**

Os espectros foram pré-processados empregando métodos de correção de deslocamento de linha de base, SNV e também derivação pelo método de Saviztky-Golay empregando janela de 7 pontos e polinômio de segunda ordem com o intuito de corrigir tendências na linha de base para comparação entre o algoritmo proposto e o PLS-DA, além disso, os dados brutos também foram utilizados para tal finalidade.

Diferente dos bancos de dados anteriores, esse conjunto possui 4 classes distintas, na tabela 7, é possível observar as figuras de mérito obtidas para ambos os métodos em que foi empregada a correção de deslocamento ao longo da linha de base, tal pré-processamento aplicado a esse conjunto apresentou resultados superiores compaados aos demais.

Tabela 7 - Resultados de classificação – Banco de dados Azeite de oliva.

(continua)

PLS-DA (Linear) LV = 16									
		Validação-cruzada				Teste			
Classe verdadeira		**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Classe atribuída	**1**	15	0	0	0	5	0	0	0
	2	0	25	0	0	0	9	0	0
	3	0	0	12	0	0	0	4	0
	4	0	0	0	37	0	0	0	13
Sensibilidade (%)		100	100	100	100	100	100	100	100
Especificidade (%)		100	100	100	100	100	100	100	100
TCC (%)		100				100			

Tabela 7 - Resultados de classificação – Banco de dados Azeite de oliva.

(conclusão)

PLS-DA (Linear) LV = 9 LVs

		Validação-cruzada				Teste			
Classe verdadeira		**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Classe atribuída	**1**	15	0	0	0	5	0	0	0
	2	0	25	0	0	0	9	0	0
	3	0	0	12	0	0	0	4	0
	4	0	0	2	35	0	0	0	13
Sensibilidade (%)		100	100	100	94,59	100	100	100	100
Especificidade (%)		100	100	97,40	100	100	100	100	100
TCC (%)		97,75				100			

FA-PLS-DA (Linear_20) LV = 9

		Validação-cruzada				Teste			
Classe verdadeira		**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Classe atribuída	**1**	15	0	0	0	5	0	0	0
	2	0	25	0	0	0	9	0	0
	3	0	0	12	0	0	0	4	0
	4	0	0	0	37	0	0	0	13
Sensibilidade (%)		100	100	100	100	100	100	100	100
Especificidade (%)		100	100	100	100	100	100	100	100
TCC (%)		100				100			

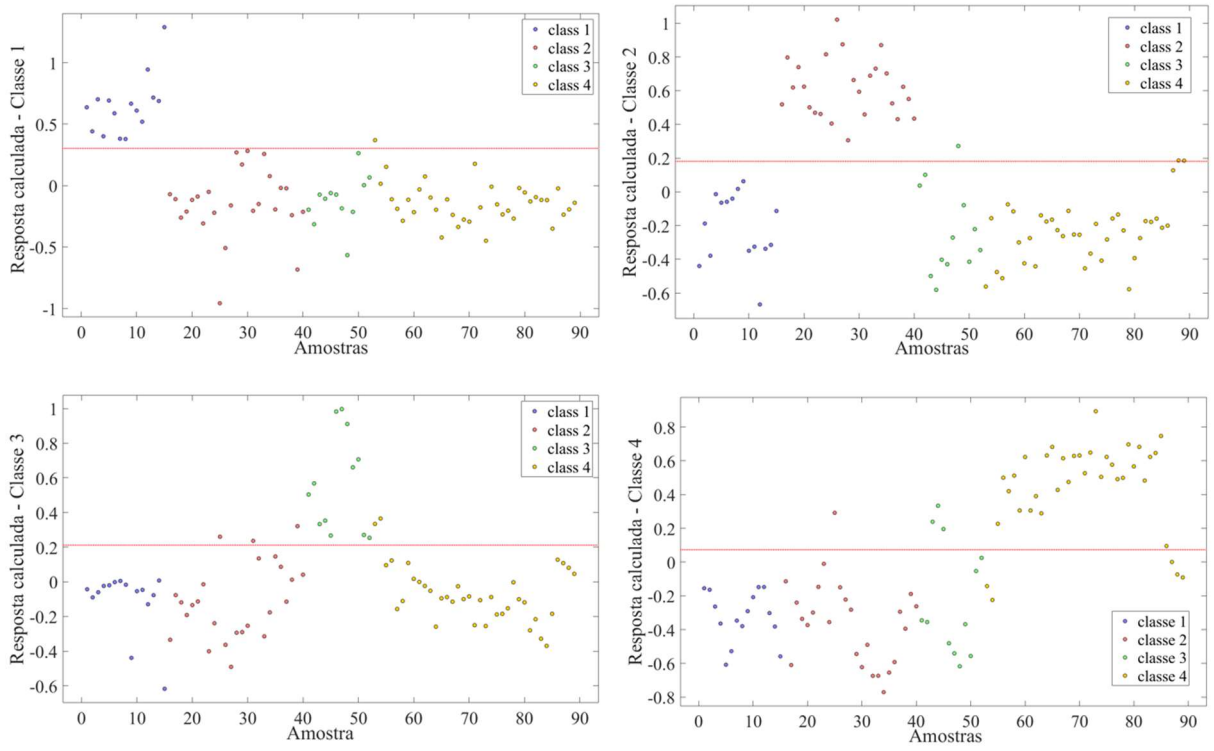
Fonte: Elaborado pela autora, 2023

O algoritmo proposto, FA-PLS-DA utilizando janelas de 20 intervalos selecionou 9 variáveis latentes com taxa correta de classificação de 100% para as amostras do conjunto de validação e de teste, o algoritmo PLS-DA também apresentou o mesmo percentual, entretanto, para alcançar a mesma taxa correta de classificação empregou 16 variáveis latentes. Para verificar a influência do número de variáveis latentes na taxa correta de classificação, o PLS-DA foi utilizado empregando a mesma quantidade de variáveis do que o FA-PLS-DA (LVs = 9). E foi observado um decréscimo da taxa de classificação correta para o PLS-DA indicando que o numero de variaveis latentes é determinante para a precisão do modelo.

A sensibilidade do algoritmo PLS-DA foi de 94,59% para a classe 4 (Amostras da Espanha) em que duas amostras foram classificadas como classe 3 (Amostras de Portugal), levando a especificidade dessa classe também ser inferior frente ao FA-PLS-DA (97,40%). Portanto, considerando um menor número de variáveis latentes empregado e a taxa de classificação correta alcançada é possível observar que o FA-PLS-DA mostrou ser mais parcimonioso e superou o método PLS-DA para o conjunto de treinamento.

Na **Figura 14** é possível observar que duas amostras da Classe 4 se encontram próximas às amostras pertencentes a Classe 3, o que reflete no desempenho do PLS-DA utilizando 9 variáveis latentes.

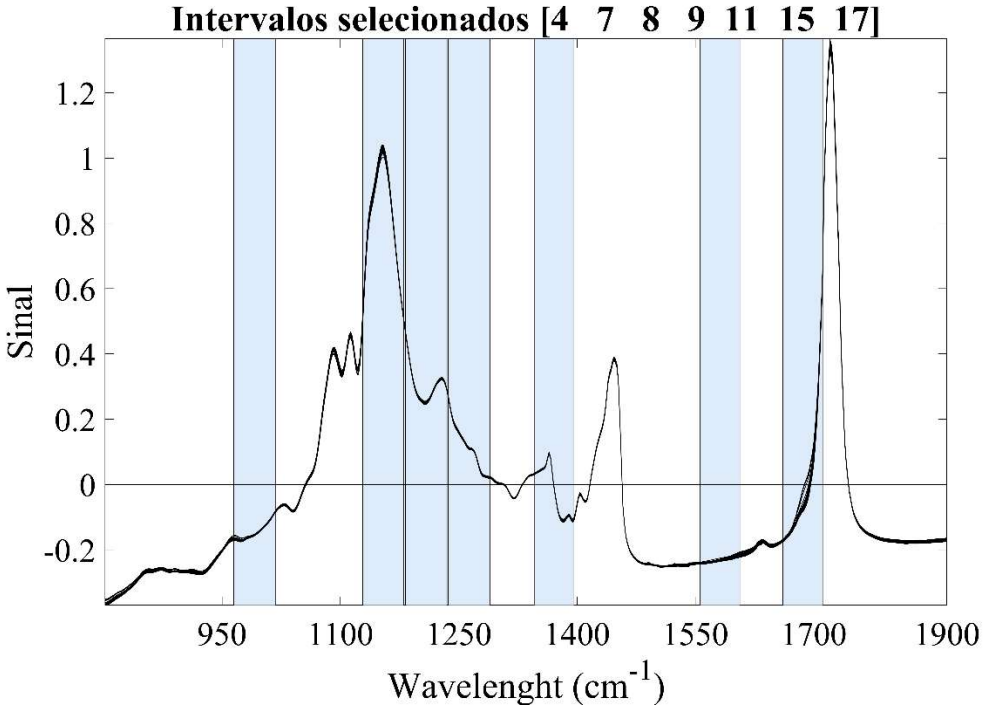
Figura 14 – Validação das amostras externas no modelo PLS-DA



Fonte: Elaborado pela autora, 2023

Para verificar a eficiência do FA-PLS-DA em selecionar variáveis informativas que explicam os resultados obtidos, as variáveis selecionadas para a construção do modelo FA-PLS-DA foram avaliadas de acordo com o estudo de Henri; Defernez; Kemsley (2003). O algoritmo proposto selecionou os intervalos 4, 7, 8, 9, 11, 15 e 17 (**Figura 15**).

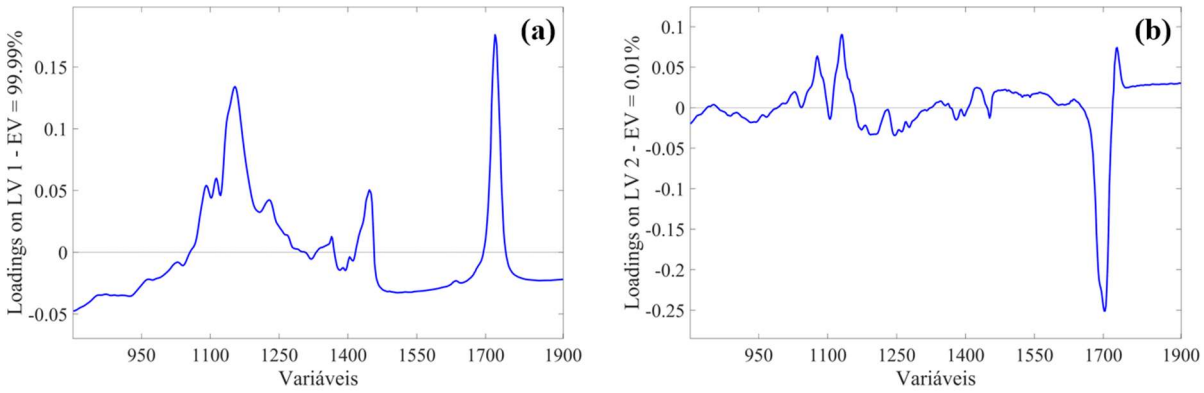
Figura 15 – Intervalos selecionados pelo FA-PLS-DA das amostras de azeite de oliva



Fonte: Elaborado pela autora, 2023

As diferenças entre as classes de azeite de oliva em diferentes localizações geográficas podem estar relacionadas a diferentes fatores como composição de ácidos graxos entre eles o oleico, palmítico e linoleico que são espectralmente diferentes e visíveis na região do infravermelho, além do conteúdo de compostos fenólicos (HENRI; DEFERNEZ; KEMSLEY, 2003). O intervalo 4 (região espectral 969 cm^{-1}) na qual as amostras da classe 1 (Grécia) apresentam uma banda característica nessa faixa específica selecionada. Os intervalos 7-9 selecionados também apresentam bandas na região 1128 cm^{-1} em que é possível discriminar a classe 3 (Amostras de Portugal). Os loadings nas variáveis latentes podem ser observados na **Figura 16**.

Figura 16 – Loadings nas variáveis latentes (a) LV – 1; (b) LV – 2.



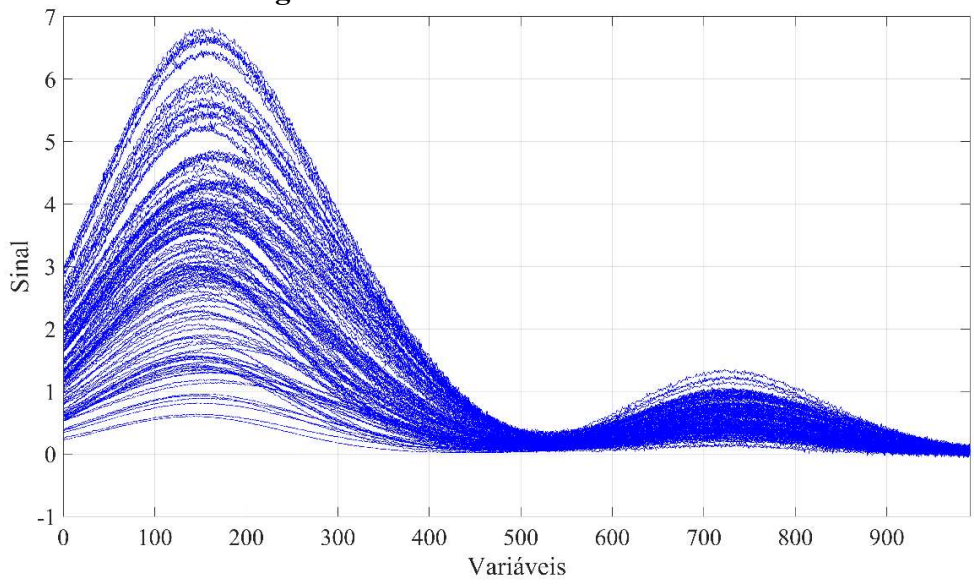
Fonte: Elaborado pela autora, 2023

A Figura 16 mostra os loadings das variáves latentes 1 e 2 aponta as variáveis influentes na formação de agrupamentos, tais variáveis foram selecionadas pelo algoritmo proposto.

5.5 Classificação – banco de dados simulado

O desempenho do algoritmo FA-PLS-DA foi avaliado utilizando um banco de dados simulado. Afim de verificar o funcionamento do algoritmo os espectros foram construídos como soma do perfil de gaussianas contendo duas bandas, em que a altura, posição e largura são constantes do perfil da gaussiana e relacionada a presença da informação associada à uma determinada classe. Tal banco de dados possui 135 observações e 991 variáveis como é possível observar na **Figura 17**.

Figura 17 – Banco de dados simulado



Fonte: Elaborado pela autora, 2023

O algoritmo proposto, FA-PLS-DA foi testado utilizando janelas de 5, 10, 15 e 20 intervalos, em ambos os casos o algoritmo mostrou ser mais parcimonioso frente ao PLS-DA. Ademais, o FA-PLS-DA apresentou melhores resultados empregando janela de 20 intervalos em que a taxa correta de classificação superior para o conjunto de teste de 82.22%, com especificidade de 83.87%, 97.22% e 93.55% para classe 1, 2 e 3 respectivamente (Tabela 8).

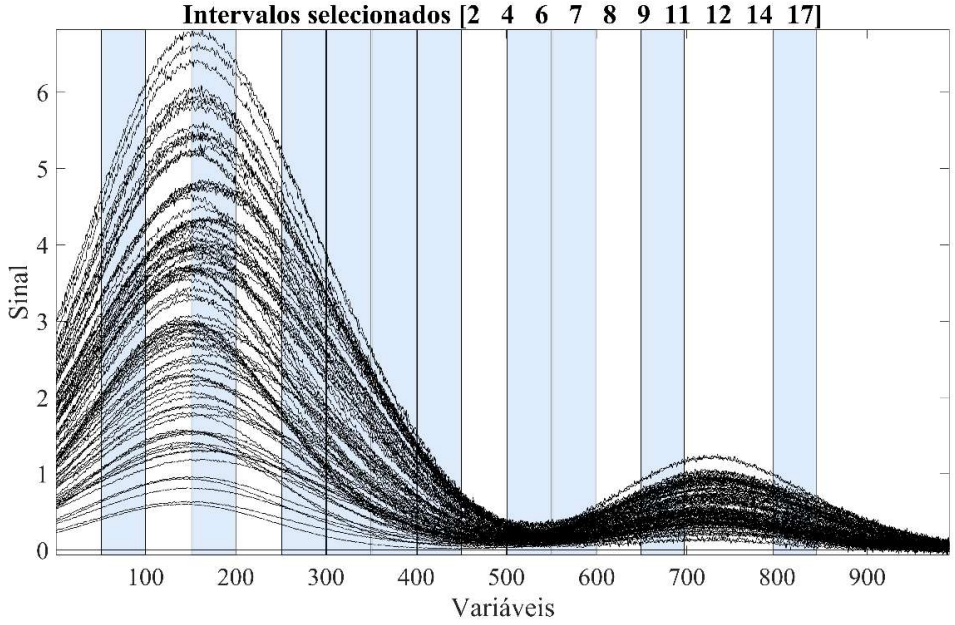
Table 8 - Resultados de classificação – Dados Simulados

PLS-DA (Simulados 2) LV = 8								
		Validação-cruzada				Teste			
Classe verdadeira		**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Classe atribuída	**1**	30	0	0	0	14	1	0	0
	2	0	30	0	0	5	9	1	0
	3	0	0	30	0	1	2	12	0
Sensibilidade (%)		100	100	100		93,33	60	80	
Especificidade (%)		100	100	100		80	90	96,67	
TCC (%)		100				77.78			
FA-PLS-DA (Simulados 2_20) LV = 7									
		Cross-validation				**Test**			
Classe verdadeira		**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Classe atribuída	**1**	30	0	0	0	14	1	0	0
	2	0	30	0	0	4	9	2	0
	3	0	0	30	0	1	0	14	0
Sensibilidade (%)		100	100	100		93,33	60	93,33	
Especificidade (%)		100	100	100		83,87	97,22	93,55	
TCC (%)		100				82,22			

Fonte: Elaborado pela autora, 2023

A **Figura 18** mostra os intervalos selecionados pelo algoritmo proposto.

Figura 18 – Intervalos selecionados pelo FA-PLS-DA do banco de dados simulado



Fonte: Elaborado pela autora, 2023

Dado que a posição das bandas está associada a região onde diferentes constituintes das amostras correspondentes às classes, o algoritmo FA-PLS-DA selecionou intervalos 2,4, 6-9, 14 e 17 como regiões em que apresentam informações mais relevantes.

6 CONCLUSÕES

O presente estudo propõe a adaptação do Algoritmo *Firefly* utilizado para seleção de variáveis empregando a análise discriminante por mínimos quadrados parciais para otimização de diferentes problemas de classificação. O algoritmo denominado FA-PLS-DA foi avaliado em três conjuntos de dados envolvendo espectros NIR e um conjunto de dados com informação simulada. O desempenho do algoritmo proposto foi comparado com resultados obtidos da Análise discriminante linear por mínimos quadrados parciais (PLS-DA).

O algoritmo FA-PLS-DA apresentou desempenho equivalente ao PLS-DA utilizando diferentes pré-processamentos nos bancos de dados utilizados. Foi possível observar que os modelos construídos com as variáveis selecionadas pelo algoritmo proposto apresentam vantagens com relação ao PLS-DA por ser mais parcimonioso, empregando um número menor de variáveis latentes para a construção dos modelos. Ademais, a eficiência do FA-PLS-DA foi corroborada ao avaliar os intervalos espectrais selecionados pelo algoritmo, em que convergiram com as regiões espectrais que apresentavam relação à resposta esperada.

Dado que o algoritmo proposto superou o PLS-DA ao selecionar subconjuntos de variáveis mais informativas buscando melhorar o desempenho da classificação, o FA-PLS-DA mostra-se um algoritmo potencial a ser utilizado em matrizes complexas para resolução de problemas envolvendo classificação.

REFERÊNCIAS

ALLEGRINI, F.; & OLIVIERI, A. C. **Figures of Merit.** Reference Module in Chemistry, Molecular Science and Chemical Engineering. 2nd edition: Chemical and Biochemical Data Analysis. Elsevier, 2019. Doi:10.1016/b978-0-12-409547-2.14612-8

ANDERSEN, C.M.; BRO, R. Variable selection in regression — a tutorial. **Journal of Chemometrics**, v. 24, n. 11-12, p. 728–737, 2010. DOI: 10.1002/cem.1360

ATTIA, K. A. M. et al. 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, n. 2017, p. 117-123, 2017. DOI: 10.1016/j.saa.2016.07.016

BARBOSA, C. E. M. **Algoritmo bioinspirado para solução de problemas de otimização**. Dissertação (Mestrado). Universidade Federal de Pernambuco. Programa de Pós-graduação em Ciência da Computação, Recife, 2017.

BARBOSA, M. F. et al. Fast determination of Biodiesel content in commercial diesel/biodiesel blends by using digital images and multivariate calibration. **Analytical Science,** v. 33, n. 11, p. 1285-1289, 2017. DOI:10.2116/analsci.33.1285

BEEBE, K. R. **Chemometrics: A practical guide**. Wiley: New York, 1998.

BIN, G. Study of PLSR-BP model for stability assessment of loess slope based on particle swarm optimization**. Scientific Reports,** v. 11, n. 1, p. 17888-17888, 2021. DOI:10.1038/s41598-021-97484-0

BISCUOLA, V. B. **Modelo matemático híbrido determinístico-estocástico para a previsão da macroestrutura de grãos bruta de solidificação.** Tese. (Doutorado em Engenharia). Universidade de São Paulo. São Paulo, 2011.

BOZORG-HADDAD, O. **Advanced Optimization by Nature-Inspired Algorithms.** Studies in Computational Intelligence, v. 720. Springer Nature, Singapore: 2018.

BRERETON, R. G. **Chemometrics:** data driven extraction for science. 2ª ed. Wiley: Hoboken, New Jersey, 2018.

BRERETON, R. G. **Chemometrics for Pattern Recognition**. Wiley: United Kingdom, 2009.

BROUGHTON-NEISWANGER, L. E. et al. Pharmacometabolomics with a combination of PLS-DA and random forest algorithm analyses reveal meloxicam alters feline plasma metabolite profiles. **Journal of Veterinary Pharmacology and Therapeutics**, v. 0, p. 1-11, 2020. DOI:10.1111/jvp.12884

CAO, M. D. et al Predicting long-term survival and treatment response in breast cancer patients receiving neoadjuvant chemotherapy by MR metabolic profiling**. NMR in Biomedicine**, v. 25 n. 2, p. 369–378, 2011. DOI:10.1002/nbm.1762

CHEN, D.; GRANT, E. Evaluating the validity of spectral calibration models for quantitative analysis following signal preprocessing. **Analytical and Bioanalytical Chemistry**, v. 404, p. 2317–2327, 2012. DOI: 10.1007/s00216-012-6364-1

CONCEIÇÃO, C. S. R. **Modelos determinísticos e Estocásticos Aplicados ao Cálculo de Provisões para Sinistros.** Dissertação (Mestrado em Matemática e Aplicações). Faculdade de Ciências e Tecnologia. Universidade Nova de Lisboa, 2014.

CORGOZINHO, C. N. C.; BARBEIRA, P. J. S. Quantification of biodiesel in biodiesel-diesel blends using spectrofluorimetry and multivariate calibration. **Journal of Analytical Chemistry**, v. 70, n. 8, p. 936–942, 2015. DOI:10.1134/s1061934815080067

DA COSTA, L. G. et al. Quantificação do teor de biodiesel de crambe em misturas com diesel utilizando espectroscopia mir e seleção de variáveis. **Química Nova**, v. 43, n. 6, p. 723-728, 2020. DOI:10.21577/0100-4042.20170554

DE ALMEIDA, V. E. et al. Using color histograms and SPA-LDA to classify bacteria. **Analytical and Bioanalytical Chemistry**, v. 406, n. 24, p. 5989–5995, 2014. DOI:10.1007/s00216-014-8015-1

DEL SER, J. et al. Bio-inspired computation: Where we stand and what's next. **Swarm and Evolutionary Computation**, v. 48, p. 220-250, 2019. DOI:10.1016/j.swevo.2019.04.008

DINIZ, P. H. G. D; PISTONESI, M. F.; ARAÚJO, M. C. U. Using iSPA-PLS and NIR spectroscopy for determination of total polyphenols and moisture in commercial tea samples. **Analytical Methods**, v. 7, n. 8, p. 3379-3384, 2015. DOI:10.1039/c4ay03099k

FERREIRA, M. M. C. et al. Quimiometria: calibração multivariada, um tutorial. **Química Nova**, v. 22, n. 5, 1999. DOI: 10.1590/S0100-40421999000500016

FERREIRA, M. M. C. **Quimiometria**: conceitos, métodos e aplicações. Campinas: Editora Unicamp, 2015.

FREEMAN, D.; PISANI, R.; PURVES, R. **Statistics.** 4ª ed. W. W. Norton & Company, Inc: New York, 2007.

FREEDMAN, D. A. **Statistical Models: Theory and Practice.** Revised Edition. Cambridge University Press: New York, 2009.

GALVÃO, R. K. H. et al. A method for calibration and validation subset partitioning. **Talanta,** v. 67, n. 2005, p. 736-740, 2005. DOI: 10.1016/j.talanta.2005.03.025

GARCIA, M.B.E.O.; DIAS, B.C.; GOMES A. A. Exploring estimated hydrocarbon composition via gas chromatography and multivariate calibration to predict the pyrolysis gasoline distillation curve. **Fuel,** v. 303, 121298, 2021. DOI:10.1016/j.fuel.2021.121298

GLOVER, F; KOCHENBERGER, G. A. **Handbook of Metaheuristics**. Kluwer Academic Publishers: New York, 2003.

GOEDERT, M. L; PAULA FILHO, P. L.; BLANCO, D. R. Computação Natural: conceitos e aplicações da computação inspirada na natureza. **Espacios**, v. 38, n. 34, p. 31, 2017.

GOLDBARG, E. **Otimização combinatória e meta-heurística:** algoritmos e aplicações. 1a ed. Elsevier: Rio de Janeiro, 2016.

GOLDBERG, D.E.; RICHARDSON, J. **Genetic Algorithms with Sharing for Multimodal Function Optimization.** Genetic Algorithms and their Applications: Proceedings of the Second International Conference on Genetic Algorithms, p. 41-49, 1987.

GOMES, A. A. et al. The successive projections algorithm for interval selection in PLS. **Microchemical Journal,** v. 110, n. 2013, p. 202–208, 2013. DOI:10.1016/j.microc.2013.03.015

GONTIJO, L. C. **Uso de espectrometria no infravermelho médio, calibração multivariada e seleção de variáveis por intervalos na quantificação de biodieseis em misturas com diesel**. Tese (Doutorado). Programa de Pós-graduação em Química. Instituto de Química. Universidade Federal de Uberlândia. Uberlândia, 2016.

GRESSLING, T. **Data Science in Chemistry:** Artificial Inteligence, Big Data, Chemometrics and Quantum Computing wih Jupyter**.** Walter de Gruyter GmbH: Berlin/Boston, 2021

HEINZE, G.; WALLISCH, C.; DUNKLER, D. Variable selection – A review and recommendations for the practicing statistician, **Biometrical Journal**, v. 60, p. 431-449, 2018 DOI:10.1002/bimj.201700067

ISMAIL, S. et al. Discriminative Analysis of Different Grades of Gaharu (Aquilaria malaccensis Lamk.) via 1H-NMR-Based Metabolomics Using PLS-DA and Random Forests Classification Models. **Molecules**, v. 22 n. 10, p. 1612, 2017. DOI:10.3390/molecules22101612

MILLER, J. N.; MILLER, J. C.; MILLER, R. D. **Statistics and Chemometrics for Analytical Chemistry**. 7ª ed. Harlow. Pearson Education Limited: United Kingdom, 2018.

JINGJING SUN, et al. An efficient variable selection method based on random frog for the multivariate calibration of NIR spectra. **Royal Society Chemistry Advances**, v. 10, p. 16245-16253, 2020. DOI:10.1039/D0RA00922A

KELTON, W. D.; SADOWSKI, R. P.; ZUPICK, N. B. **Simulation with Arena**. 6a ed. McGraw-Hill Education: New York, 2015.

KENNARD, R.W.; STONE, L. A. Computer Aided Design of Experiments, **Technometrics,** v. 11, n. 1969, p. 137–148, 1969. DOI: 10.1002/0471667196.ess0035.pub2

KHALILIYAN, H. et al. Direct Quantification of Lignin in Liquors by High Performance Thin Layer Chromatography-Densitometry and Multivariate Calibration, **American Chemical Society Sustainable Chemistry & Engineering,** v. 2020, n. 8, p. 16766-16774, 2020. DOI:10.1021/acssuschemeng.0c03950

KREPPER, G. et al. Determination of fat content in chicken hamburgers using NIR spectroscopy and the Successive Projections Algorithm for interval selection in PLS regression (iSPA-PLS). **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,** v. 189, n. 2018, p. 300–306, 2018. DOI: 10.1016/j.saa.2017.08.046

LAVIGNE, B. K. Chemometrics. **Analytical Chemistry,** v. 72, n. 12, p. 91-97, 2000. DOI:10.1021/a1000016x.

LAW, A. M. **Simulation Modeling and Analysis.** 5a ed. McGraw-Hill Education: Tucson, Arizona, USA, 2015.

LEI, T.; LIN, X.-H.; SUN, D.-W. Rapid classification of commercial Cheddar cheeses from different brands using PLSDA, LDA and SPA–LDA models built by hyperspectral data. **Journal of Food Measurement and Characterization**. v. 13, n. 4, p. 3119-3129, 2019. DOI:10.1007/s11694-019-00234-0

LINDFIELD, J.; PENNY, J. Chapter 5 – The Firefly Algorithm. **Introduction to Nature-Inspired Optimization**, 2017, p. 85-100. DOI:10.1016/B978-0-12-803636-5.00005-0

LUCIANO, G. **Statistical and multivariate analysis in material science**. 1ª ed. National Research Council of Italy, Institute for the study of macromolecules. CRC Press, Taylor & Francis Group: Genova, Italia, 2021.

MANZANO, J. A. N. G. **Algoritmos Funcionais**: Introdução minimalista à lógica de programação funcional pura aplicada à teoria dos subconjuntos. Rio de Janeiro: Alta Books, 2020.

MAROUANI, H; AL-MUTIRI, O. Optimization of Reliability–Redundancy Allocation Problems: A Review of the Evolutionary Algorithms. **Computers, Materials & Continua**, v. 71, n. 1, p. 537-571, 2022. DOI:10.32604/cmc.2022.020098

MARTENS, H.; NAES, T. **Multivariate Calibration**. John Wiley and Sons: Chichester, 1989.

MESQUITA, E. M. **Estudo comparativo de meta-heurísticas aplicadas ao controle preditivo baseado em modelo**. Dissertação (Mestrado). Programa de pós-graduação em sistemas mecatrônicos. Universidade de Brasília. Brasília, 2018.

MACH, J. B.; RONOH, K. K.; LANGAT, K. Improved spectrum allocation scheme for TV white space networks using a hybrid of firefly, genetic, and ant colony optimization algorithms. **Heliyon,** v. 9, n. 2023, e13752, 2023. DOI: j.heliyon.2023.e13752

MARINI, F. **Chemometrics in Food Chemistry**. Chapter 1. Data Handling in Science and ecnology, v. 28. Elsevier's Science & Technology Rights: United Kingdom, 2013.

MISHRA, P. et al. Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction. **Talanta,** v. 229, n. 2021, p. 122303, 2021. DOI:10.1016/j.talanta.2021.122303

MOHAMMADREZA, K. K.; MINA, S.; MAHSA, M. Quality classification of gasoline samples based on their aliphatic to aromatic ratio and analysis of PONA contente using genetic algorithm based multivariate techniques and ATR-FTIR spectroscopy. **Infrared Physics & Technology**, v. 126, p. 104354, 2022. DOI:10.1016/j.infrared.2022.104354

MORAIS, C. da S. et al. Aplicação de calibração multivariada em dados de espectroscopia UV-Visível para previsão da acidez total em vinhos. **Revista Brasileira de Pesquisa em Alimentos.** v. 6, n. 1, p. 70 – 79, 2015. DOI:10.14685/rebrapa.v6i1.193

MORETTIN, P. A.; BUSSAB, W. O. **Estatística Básica**. 9ª ed. São Paulo: Saraiva, 2017.

NASCIMENTO, M. H. C. et al. Noinvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis. **Analytical Chemistry,** v. 94, n. 5, p. 2425-2433, 2022. DOI: 10.1021/acs.analchem.1c04162

NAZI, A.; LEARDI, R. Genetic algorithms in chemometrics. **Journal of Chemometrics,** v. 2, p. 345–351, 2012. DOI:10.1002/cem.2426

NIST**. National Institute of Standards and Technology**. 2002. Standard reference materials -SRM 2709, 2710 and 2711. Addendum Issue Date: January 18 2002.

NØRGAARD, L. et al. Interval Partial Least-Squares Regression (iPLS): 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

OBIORA, S.C. et al. Potentially harmful elements and their health implications in cultivable soils and food crops around lead-zinc mines in Ishiagu, Southeastern Nigeria. **Journal of Geochemical Exploration**, v. 204, n. 2019, p. 289-296, 2019. DOI:10.1016/j.gexplo.2019.06.011

OLIVIERI, A. C. **Introduction a Multivariate Calibration:** A practical Approach. Springer, 2018.

OLIVIERI, A. C.; ESCANDAR, G. M. Analytical Figures of Merit. **Practical Three-Way Calibration**, p, 93–107, 2014. DOI:10.1016/b978-0-12-410408-2.00006-5

PASQUINI, C. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, n. 198, 2003. DOI:10.1590/S0103-50532003000200006

PAVIA, D. L. et al. **Introdução a espectroscopia**. 5ª ed. São Paulo: Cengage Learning, 2015.

PEREIRA, E. V. dos S.; et al. In-situ authentication of goat milk in terms of its adulteration with cow milk using a low-cost portable NIR spectrophotometer**. Microchemical Journal** 163 (2021) 105885. Doi:10.1016/j.microc.2020.105885

POLO-CORPA, M. J. et al. Curve fitting using heuristics and bio-inspired optimization algorithms for experimental data processing in chemistry. **Chemometrics and Intelligent Laboratory Systems,** v. 96, n. 1, p. 34-42, 2009. DOI:10.1016/j.chemolab.2008.11.004

QUILTY, J.; ADAMOWSKI, J.; BOUCHER, M.A. A stochastic data-drivenensemble forecasting framework forwater resources: A case study usingensemble members derived from adatabase of deterministicwavelet-based models. **Water Resources Research**, v. 55, p. 175–202, 2019. DOI:10.1029/2018WR023205

RAMADAN, Z. et al. Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. **Talanta**, v. 68, n. 5, p. 1683-1691, 2006. DOI:10.1016/j.talanta.2005.08.042

RAMBO, M. K. D. et al. Predição por calibração multivariada dos parâmetros de qualidade de biomassas de café. **Ciência e Natura**, v.37, n. 2, p. 374 – 380, 2015. DOI:10.5902/2179460X17124

ROCHA, W. F. C. et al. Laser-driven calorimetry and chemometric quantification of standard reference material diesel/biodiesel fuel blends. **Fuel**, v. 281, n. 118720, 2020. DOI:10.1016/j.fuel.2020.118720

RODIONOVA, O. Y.; POMERANTSEV, A. L. Chemometric tools for food fraud detection: The role of target class in non-targeted analysis. **Food Chemistry,** v. 317, p. 12644-126448, 2020. DOI: 10.1016/j.foodchem.2020.126448

RUSCHEL, C. F. C. **Aplicação de ferramentas quimiométricas e técnicas espectroscópicas na análise de combustíveis**. Tese (Doutorado em Química). Instituto de Química. Universidade Federal do Rio Grande do Sul, 2017.

RUIZ-PÉREZ, D. et al. So you think you can PLS-DA? **BioMed Central Bioinformatics**, v. 21, n. 2, 2020. DOI: 10.1186/s12859-019-3310-7

SÆBØA, S. et al. ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS. **Journal of Chemometrics**, v. 20, p. 54-62, 2008. DOI:10.1002/cem.1101

SHARABIANI, V. R. et al. Non-Destructive Prediction of Titratable Acidity and Taste Index Properties of Gala Apple Using Combination of Different Hybrids ANN and PLSR-Model Based Spectral Data. **Plants**, v. 9, n. 12, p. 1718, 2020. DOI:10.3390/plants9121718

SALDANHA, R. B. et al. Physical–Mineralogical–Chemical Characterization of Carbide Lime: An Environment-Friendly Chemical Additive for Soil Stabilization, **Journal of Materials in Civil Engineering**, v. 30, n. 6, p. 06018004, 2018. Doi:10.1061/(ASCE)MT.1943-5533.0002283

SANTOS, G. R. et al. Avanços analíticos baseados em modelos de calibração de primeira ordem e espectroscopia uv-vis para avaliação da qualidade da água: uma revisão - parte 1. **Quimica Nova**, v. 45, n. 3, p. 314-323, 2022. DOI:10.21577/0100-4042.20170830

SMITH, B. C. **Fundamentals of Fourier Transform Infrared Spectroscopy.** 2a edition. CRC Press: New York, 2001.

SOUTO, U. T. C. P. et al. UV–Vis spectrometric classification of coffees by SPA–LDA. **Food Chemistry**, v. 119, n. 1, p. 368–371, 2010. DOI: 10.1016/j.foodchem.2009.05.078

SSEGANE, H. et al. Advances in variable selection methods II: Effect of variable selection method on classification of hydrologically similar watersheds in three Mid-Atlantic ecoregions. **Journal of Hydrology,** v. 438–439, p. 26–38, 2012. DOI:10.1016/j.jhydrol.2012.01.035

SULAN, Z. et al. A hyperspectral GA-PLSR model for prediction of pine wilt disease. **Multimedia tools and applications**, v. 79, v. 23-24, p. 16645-16661, 2020. DOI:10.1007/s11042-019-07976-5

SUN, X. et al. Variable selection and chemometric models for discriminating symptomatic gout based on a metabolic target analysis. **Journal of Chemometrics**, v. 32, n. 5, e2984, 2018. DOI:10.1002/cem.2984

SZWARCFITER, J. L; MARKEZON, L. **Estruturas de dados e seus algoritmos**. 3ª ed. Rio de Janeiro: LTC, 2015.

TILAHUN, S. L.; ONG, H. C. Modified Firefly Algorithm. **Journal of Applied Mathematics,** V. 2012, Article ID 467631, 12 p. Doi:10.1155/2012/467631

TIMM, N. H. **Applied multivariate analysis**. Springer: New York, 2002.

VALDERRAMA, P.; BRAGA, J. W. B.; POPPI, R. J. Estado da arte das figuras de mérito em calibração multivariada. **Quimica Nova**, v. 32, n. 5, p. 1278-1287, 2009. DOI: 10.1590/S0100-40422009000500034

VARMUZA, K.; FILZMOSER, P. **Introduction to Multivariate Statistical Analysis in Chemometrics**. New York: CRC Press, 2009.

VIEIRA, L. S. et al. Building robust models for identification of adulteration in olive oil using FT-NIR, PLS-DA and variable selection. **Food Chemistry**, v. 345, n. 128866, 2021. DOI:10.1016/j.foodchem.2020.128866

WOODRUFF, L. G; et al. Critical mineral resources of the United States—Economic and environmental geology and prospects for future supply: U.S. **Geological Survey Professional Paper**, v. 1802, p. T1–T23, 2017. DOI:10.3133/ pp1802T.

YADAV, A.; VISHWAKARMA, D. K. A comparative study on bio-inspired algorithms for sentiment analysis. **Cluster Computing**, v. 23, p. 2969-2989, 2020. DOI:10.1007/s10586-020-03062-w

YANG, X.; DEY, N.; FONG, S. **Applications of Firefly Algorithm and its Variants**: Case studies and New Developments. Springer. 2020.

YANG, X. et al. Swarm Intelligence: Past, Present and Future, **Soft computing**, Berlin, Germany, v. 22, p. 5923-5933, 2018. DOI:10.1007/s00500-017-2810-5

YANG, X.; KARAMANOGLU, M. Swarm Intelligence and Bio-Inspired Computation: An Overview. **Swarm Intelligence and Bio-Inspired Computation**: Theory and Applications, v. 2013, p. 3-23, 2013. DOI: 10.1016/b978-0-12-405163-8.00001-6

YATES, R. D; GOODMAN, D. J. **Probabilidade e processos estocásticos:** uma introdução amigável para engenheiros eletricistas e da computação / Roy D. Yates, David J. Goodman; tradução Daniel Vieira. 3. ed. Rio de Janeiro: LTC, 2017.

YUAN, X. et al. Hyperspectral Imaging and SPA–LDA Quantitative Analysis for Detection of Colon Cancer Tissue. **Journal of Applied Spectroscopy**, v. 85, n. 2, p. 307–312, 2018. DOI:10.1007/s10812-018-0649-x

ZANDBAAF, S. et al. Infrared spectroscopy and chemometric approach for identifying morphology in embryo culture médium samples. **Infrared Physics & Tecnology**, v. 106, n. 2020, p. 103284, 2020. DOI: 10.1016/j.infrared.2020.103284

ZANDBAAF, S.; MOHAMMMAD, R. K. K.; MAJID, G. A. Genetic algorithm based artificial neural network and partial least squares regression methods to predict of breakdown voltage for transformer oils samples in power industry using ATR-FTIR spectroscopy. **Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy,** v. 273, p. 120999, 2022. DOI:10.1016/j.ssa.2022.120999