UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA
LABORATÓRIO DE QUIMIOMETRIA DO TRIÂNGULO

BALTAZAR VASCO SITOE

APLICAÇÃO DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS PARA O CONTROLE DE QUALIDADE DE BIODIESEL DE PINHÃO MANSO E DE MORINGA EM MISTURAS COM DIESEL USANDO ESPECTROMETRIA NO INFRAVERMELHO MÉDIO

UBERLÂNDIA
2019

BALTAZAR VASCO SITOE

APLICAÇÃO DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS PARA O CONTROLE DE QUALIDADE DE BIODIESEL DE PINHÃO MANSO E DE MORINGA EM MISTURAS COM DIESEL USANDO ESPECTROMETRIA NO INFRAVERMELHO MÉDIO

Tese apresentada ao Programa de Pós-Graduação em Química do Instituto de Química da Universidade Federal de Uberlândia, como requisito para obtenção de título de Doutor em Química.

Área de Concentração: Química Analítica

UBERLÂNDIA
2019

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

S623 Sitoe, Baltazar Vasco, 1981-
2019 Aplicação de Métodos Quimiométricos de Seleção de Variáveis para o Controle de Qualidade de Biodiesel de Pinhão Manso e de Moringa em Misturas com Diesel Usando Espectrometria no Infravermelho Médio [recurso eletrônico] / Baltazar Vasco Sitoe. - 2019.

Orientador: Waldomiro Borges Neto.
Tese (Doutorado) - Universidade Federal de Uberlândia, Pós-graduação em Química.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.te.2019.2520
Inclui bibliografia.

1. Química. I. Borges Neto, Waldomiro ,1970-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Química. III. Título.

CDU: 54

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Química
Av. João Naves de Ávila, 2121, Bloco 5I - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4385 - www.cpgquimica.iq.ufu.br - cpgquimica@ufu.br

ATA

Programa de Pós-Graduação em:	Química				
Defesa de:	Tese de Doutorado Acadêmico, 98, PPQUI				
Data:	doze de dezembro de dois mil e dezenove	Hora de início:	[14:00]	Hora de encerramento:	[18:20]
Matrícula do Discente:	11713QMI013				
Nome do Discente:	Baltazar Vasco Sitoe				
Título do Trabalho:	Aplicação de métodos de seleção de variáveis para o controle de qualidade de biodiesel de pinhão manso e de moringa em misturas com diesel usando espectrometria no Infravermelho Médio				
Área de concentração:	Química				
Linha de pesquisa:	Espectroanalítica Aplicada				
Projeto de Pesquisa de vinculação:	Monitoramento de qualidade de biocombusíveis				

Reuniu-se no Auditório Prof. Manuel Gonzalo Hernández Terrones , Campus Santa Mônica, da Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Química, assim composta: Professores Doutores: Jefferson Luis Ferrari e Edson Nossol, do Instituto de Química da Universidade Federal de Uberlândia - IQUFU; Frederico Garcia Pinto, da Universidade Federal de Viçosa - UFV; Valmir Jacinto da Silva, da Universidade Estadual de Goiás - UEG e Waldomiro Borges Neto, orientador(a) do(a) candidato(a).

Iniciando os trabalhos o(a) presidente da mesa, Dr(a). Waldomiro Borges Neto, apresentou a Comissão Examinadora e o candidato(a), agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor(a) presidente concedeu a palavra, pela ordem sucessivamente, aos(às) examinadores(as), que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o(a) candidato(a):

Aprovado.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Doutor.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação

interna da UFU.

Documento assinado eletronicamente por **Waldomiro Borges Neto**, **Professor(a) do Magistério Superior**, em 13/12/2019, às 19:56, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Jefferson Luis Ferrari**, **Professor(a) do Magistério Superior**, em 13/12/2019, às 20:03, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Edson Nossol**, **Professor(a) do Magistério Superior**, em 16/12/2019, às 08:52, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Frederico Garcia Pinto**, **Usuário Externo**, em 16/12/2019, às 13:59, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Valmir Jacinto da Silva**, **Usuário Externo**, em 16/12/2019, às 14:30, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1753717** e o código CRC **FC1E3E03**.

Referência: Processo nº 23117.106632/2019-92 SEI nº 1753717

Aos meus queridos pais Vasco Julião Sitoe e Celeste Zava Chongo (Que Deus a tenha).

Aos meus irmãos Cristina, Pelágio, Alegria, Geraldo e Reginaldo (Que Deus o tenha).

Aos meus filhos Cristóvão, Arginina, Tânia, Wesley e Carmélia.

À minha amada esposa Crimilda José Ussivane.

AGRADECIMENTOS

A Deus Misericordioso, O Dono de tudo, que me fez chegar a este mundo.

À minha família, que soube entender a minha saída para longe da terra em busca de conhecimentos e para o bem do nosso futuro.

Ao meu orientador, Prof. Dr. Waldomiro Borges Neto, que pela simples comunicação me aceitou no seu grupo de pesquisa, pelos conhecimentos transmitidos, paciência que teve comigo em todos os momentos acadêmicos, que passo a passo esteve na orientação de todos os trabalhos até a produção desta Tese de Doutorado.

Aos colegas do Laboratório de Quimiometria do Triângulo, desta época de Doutorado: Ademar Domingos Viagem Máquina, Letícia Maria de Souza, Eloíza Guimarães, José Eduardo Buiatte, Raquel Boaventura de Moraes, Maria Teresa Carvalho Ferreira, Lucas Gustavo da Costa e Tathiana Dienifer Godinho, pelo convívio e ajuda nos trabalhos acadêmicos.

Ao Governo da República de Moçambique através do Ministério da Ciência e Tecnologia, Ensino Superior e Técnico Profissional – projeto HEST financiado pelo Banco Mundial processo número 24.06.2016, pelo suporte financeiro da bolsa de estudos.

Ao Governo da República Federativa do Brasil pelo acolhimento na Instituição de Ensino Superior através de acordos bilateral entre Moçambique e Brasil na área da educação.

À Universidade Federal de Uberlândia e aos seus Professores e funcionários que forneceram toda a estrutura física, intelectual e tecnológica para a realização deste trabalho, e pelo acompanhamento em todos os processos técnico-administrativos, principalmente ao Coordenador da Divisão de Apoio à Pós-Graduação, João Martins Neto, e à Secretária do Programa de Pós-Graduação em Química, Mayta Mamede Negreto Peixoto.

À Universidade Pedagógica – Moçambique, então dirigida pelo Magnífico Reitor Prof. Dr. Luís Jorge Manuel Teodósio António Ferrão, e à Drª. Maria Francisca Teodósio Ferrão pela ajuda no processo do meu afastamento para continuar com os meus estudos de Doutorado.

A todos os meus colegas de serviço e alunos que sempre trabalharam comigo desde 1999 até hoje em Moçambique, onde fui Professor nas Escolas Primárias de Matsilele (Chicualacuala) e de Tomanine (Guijá), nas Escolas Secundárias do Guijá (Guijá) e 7 de Abril (Chimoio), no Instituto Agrário de Lichinga (Lichinga) e da Universidade Púnguè pela convivência profissional e apoio moral durante os meus estudos.

Finalmente, agradeço aos demais que não os mencionei aqui, mas que direta ou indiretamente contribuíram e os que continuam contribuindo para o sucesso da minha vida em todos os âmbitos.

“Estudemos e Façamos dos Nossos Conhecimentos um Instrumento de Libertação do Povo, porque a Alfabetização é a Chave para o Desenvolvimento!”

(Samora Moisés Machel, 1977).

RESUMO

No Brasil, a adição de 10% (v/v) do biodiesel ao diesel era obrigatório desde 1º de março de 2018 até 31 de agosto de 2019. Das análises das misturas do biodiesel/diesel feitas pelo Programa de Monitoramento de Qualidade dos Combustíveis, verificou-se que o teor do biodiesel corresponde a 55,90% do total de não conformidades. Assim, o presente estudo teve como objetivo aplicar métodos quimiométricos de calibração multivariadas por Quadrados Mínimos Parciais (PLS), de classificação supervisionada por Análise Discriminante por Quadrados Mínimos Parciais (PLS-DA) por meio de seleção de variáveis por intervalos para quantificar teores do biodiesel metílico de pinhão manso e de moringa e classificar as suas amostras em misturas com diesel, usando espectrometria no infravermelho médio. Antes da construção de modelos, foi produzido biodiesel puro, B100; depois foram preparadas amostras das misturas biodiesel/diesel. Os espectros foram tirados por espectrômetro de infravermelho médio. Os modelos construídos a partir de métodos de seleção de variáveis que apresentaram, estatisticamente, menor valor de Erro Quadrático Médio de Previsão (RMSEP) e melhor classificação, em comparação aos modelos de espectro total (modelos globais), foram validados conforme parâmetros de mérito, estabelecidos para análises multivariadas. Os resultados dos modelos de quantificação, para cada tipo de biodiesel, demonstraram excelente correlação entre os valores medidos e previstos ($R^2 > 0,99$) na faixa de concentração usada (de 0,50 – 30% (v/v)), valores de RMSEP entre 0,11 e 0,25% (v/v) e os modelos não apresentaram erros sistemáticos. A eficiência dos modelos de classificação, também construídos por seleção de variáveis, foi avaliada com base nas figuras de mérito: sensibilidade, especificidade, acurácia e coeficiente de correlação de Matthews (CCM). Os valores de limites (*threshold*) foram determinados com base no teorema de Bayes, de forma a minimizar os falsos positivos e falsos negativos. A relação entre a especificidade e a sensibilidade também foram avaliadas graficamente usando as curvas Características de Operação do Receptor (ROC). Estes parâmetros apresentaram valores iguais a 1,0 para todos os modelos, o que representa uma classificação correta das amostras dos conjuntos de treinamento e de teste. Assim, os métodos desenvolvidos podem ser uma alternativa viável às análises para a determinação do teor do biodiesel e classificação de amostras de biodiesel metílico de pinhão manso e de moringa em misturas com diesel.

Palavras-Chave: Quimiometria. Calibração multivariada. Classificação supervisionada. iPLS e iPLS-DA. Teor do biodiesel.

ABSTRACT

In Brazil, the addition of 10% (v/v) of biodiesel to diesel was mandatory since March 1st 2018 until August 31st, 2019. From the analysis of biodiesel/diesel blends made by the Fuel Quality Monitoring Program, it was found that the biodiesel content corresponds to 55.90% of the total nonconformities. Thus, the present study aimed to apply methods to quantify contents of jatropha and moringa methyl biodiesel and to classify their samples in blends with diesel, using Partial Least Squares Multivariate (PLS) calibration methods and supervised classification by Partial Least Squares Discriminant Analysis (PLS-DA) through interval variable selection methods and the mid-infrared spectrometry technique. Before to model building, pure biodiesel, B100; Then samples of the biodiesel/diesel blends were prepared. The models built from variable selection methods that presented statistically lower Root Mean Square Error of Prediction (RMSEP) value and better classification, compared to the full spectrum models (global models), were validated according to established merit parameters for multivariate analysis. The results of the quantification models for each type of biodiesel show an excellent correlation between the mensured and predicted values ($R^2 > 0.99$) in the concentration range used (0.50 – 30% (v/v)), RMSEP values between 0.11 and 0.25% (v/v) and de models have no systematic errors. The efficiency of the classification models, also constructed by variable selection, was evaluated based on the figures of merit: sensitivity, specificity, accuracy and Matthews correlation coefficient (MCC). The threshold values were determined based on the Bayes theorem in order to minimize false positives and false negatives. The relationship between specificity and sensitivity was also graphically assessed using Receiver Operating Characteristics (ROC) curves. These parameters presented values equal to 1.0 for all models, which represents a correct classification of the samples of the training and test sets. Thus, the developed methods may be a viable alternative to the analysis for biodiesel content determination and classification of jatropha and moringa methyl biodiesel samples in diesel blends.

Keywords: Chemometric. Multivariate calibration. Supervised classification. iPLS and iPLS-DA. Biodiesel content.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT NBR	Norma Brasileira da Associação Brasileira de Normas Técnicas
ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
ASTM	Sociedade Americana de Testes e Materiais (do inglês, *american society of testing and materials*)
ATR	Reflectância Total Atenuada (do inglês, *attenuated total reflectance*)
AUC	Área Sob a Curva ROC (AUC – do inglês, *area under the ROC curve*)
B100	Biodiesel Puro
BX	Mistura Biodiesel/Diesel com X % de Biodiesel
BMM	Biodiesel Metílico de Moringa
BMPM	Biodiesel Metílico de Pinhão Manso
CNPE	Conselho Nacional da Política Energética
DD-SIMCA	Modelagem Independente Suave de Analogia de Classe Orientada a Dados (do inglês, *data driven soft independent modeling of class analogy*)
EN	Norma Europeia (do inglês, *european norm*)
FIR	Infravermelho distante (do inglês, *far-infrared*)
FT–MIR	Infravermelho Médio com Transformada de Fourier (do inglês, *fourier transform mid-infrared*)
HCA	Análise de Agrupamentos Hierárquicos (do inglês, *hierarchical cluster analysis*)
iPLS	Quadrados Mínimos Parciais por Intervalos (do inglês, *interval partial least squares*)
KNN	K-Vizinho mais Próximo (do inglês, *K-nearest neighbors*)
LDA	Análise Discriminante Linear (do inglês, *linear discriminant analysis*)
MIR	Infravermelho Médio (do inglês, *mid-infrared*)
NAS	Sinal Analítico Líquido (do inglês, *net analyte signal*)
NIR	Infravermelho Próximo (do inglês, *near-infrared*)
PCA	Análise de Componentes Principais (do inglês, *principal components analysis*)
PCs	Componentes Principais (do inglês, *principal components*)
PETROBRAS	Petróleo Brasileiro S. A.
PETROMOC	Petróleos de Moçambique
PLS	Quadrados Mínimos Parciais (do inglês, *partial least squares*)

PLS-DA	Análise Discriminante por Quadrados Mínimos Parciais (do inglês, *partial least squares – discriminant analysis*)
PNPB	Programa Nacional de Produção e Uso do Biodiesel
RANP	Resolução da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
ROC	Característica de Operação do Receptor (ROC – do inglês, *receiver operating characteristics*)
RMSEC	Erro Quadrático Médio de Calibração (do inglês, *root mean square error of calibration*)
RMSECV	Erro Quadrático Médio de Validação Cruzada (do inglês, *root mean square error of cross-validation*)
RMSEP	Erro Quadrático Médio de Previsão (do inglês, *root mean square error of prediction*)
SIMCA	Modelagem Independente Suave por Analogia de Classe (do inglês, *soft independent modelling of class analogy*)
SVM	Máquinas de Vetores de Suporte (do inglês, *support vector machines*)
VLs	Variáveis Latentes

SUMÁRIO

1. INTRODUÇÃO

Uma das necessidades fundamentais na vida humana é a produção de energia, pois ela é indispensável para a produção de alimentos, industrial, agrícola, para o transporte, bem como para a geração de eletricidade em centrais térmicas convencionais. O desenvolvimento econômico dinâmico no mundo contribuiu para o aumento da demanda energética. O rápido desenvolvimento das economias mundiais faz com que o consumo de energia aumente (BÓRAWSKI et al., 2019). A maior percentagem da produção mundial de energia é gerada a partir de combustíveis fósseis (não renováveis). Contudo, esses combustíveis são finitos e provocam poluição ao meio ambiente.

Para suprir a demanda energética e minimizar danos ambientais, a cadeia produtiva deve ser aprimorada visando desenvolver um mercado sustentável de bioenergia, com base em recursos provenientes da biomassa e preferencialmente não alimentares para reequilibrar o apoio à produção de biocombustíveis (ZIELIŃSKI et al., 2019). O grande desafio científico é buscar novas fontes de energias renováveis para o desenvolvimento sustentável nas áreas ambiental, social e econômica melhorar a cadeia produtiva e viabilizar a produção de coprodutos e combustíveis a partir de resíduos que possa contaminar o meio ambiente, como por exemplo, o óleo de fritura residual.

Os combustíveis renováveis são menos tóxicos pela ausência de compostos sulfurados e aromáticos voláteis, por emitirem menos materiais particulados e, em alguns casos, por serem biodegradáveis no meio ambiente. O biodiesel é um exemplo, já em aplicação, do emprego da biomassa para produção de energia. O biodiesel é um combustível produzido a partir da reação de esterificação e/ou transesterificação de gorduras animais ou de óleos vegetais com um álcool de cadeia curta, geralmente etanol (CH_3OH) ou metanol (C_2H_5OH), em meio ácido, básico ou enzimáticos. É constituído por uma mistura de alquilésteres de cadeia linear. O uso do biodiesel diminui o dióxido de carbono (CO_2) na atmosfera através ciclo de carbono; libera pouca quantidade de CO_2 durante a combustão nos motores, quando comparado com o óleo diesel; gera oportunidades de emprego e renda para a população rural na agricultura familiar; diminui a importação de combustíveis fósseis; diversifica a matriz energética por meio do uso de novas tecnologias de produção (GEISSDOERFER; VLADIMIROVA; EVANS, 2018; JAIN, 2019).

O biodiesel possui elevado número de cetano e ponto de fulgor, porque é constituído por longas cadeias lineares que são facilmente craqueadas sob pressão e aquecimento, apresentando, deste modo, melhor característica de ignição que o diesel. O número de cetano

de um combustível para máquinas que operam segundo o ciclo diesel é a qualidade de combustão medida entre a injeção de combustível e o início da combustão. Como as propriedades físico-químicas do biodiesel são semelhantes às do óleo diesel, o biodiesel pode ser usado na sua forma pura e é denominado B100, ou em misturas com o diesel, BX, onde X corresponde à proporção percentual volumétrica do biodiesel na mistura. A adição de biodiesel em diesel aumenta o número de cetano, pois o valor mínimo do número de cetano do biodiesel é de 47 (RYU, 2010), e o número de cetano do diesel S10 é 48, e 42 para diesel S500.

Embora o biodiesel forneça uma quantidade de energia cerca de 10% menor que o diesel, seu desempenho no motor é praticamente o mesmo no que diz respeito à potência e ao torque. Por apresentar maior viscosidade, o biodiesel proporciona maior lubricidade que o diesel mineral, logo, tem-se observado redução no desgaste das partes móveis do motor. O biodiesel possui estruturas moleculares mais simples que o seu precursor, os triglicerídeos. Por isso, sua viscosidade é comparativamente menor, apresentando maior eficiência de queima, reduzindo significativamente a deposição de resíduos nas partes internas do motor.

No entanto, comparativamente ao diesel, o biodiesel apresenta algumas desvantagens das quais se podem citar o menor valor calorífico; maiores ponto de fluidez e ponto de turvação (nuvem); quando usado em um motor de combustão interna, aumenta ligeiramente as emissões de NOx; é corrosivo contra cobre e latão; é higroscópico e apresenta baixa volatilidade em baixas temperaturas (JAIN, 2019).

Até o fim de 2018, cerca de 69,75% do biodiesel consumido no Brasil era produzido a partir do óleo de soja (ANP, 2018; APROBIO, 2018), pois o país possui o maior potencial agrícola de produção deste grão em quase todas as estações do ano, e o processo de refino de óleo está consolidado para atender à demanda de energia e de alimentos. No entanto, o consumo familiar interno do óleo de soja e a exportação deste grão e do seu farelo para a produção de carnes impactam a produção do biodiesel oriundo do óleo de soja. Por isso, é necessário encontrar outros tipos de matérias-primas para a produção do biodiesel que não participem diretamente da cadeia alimentar humana, e que possam ser produzidos de acordo com as condições dos solos e climáticas de regiões tropicais, como por exemplo, pinhão manso e moringa.

O pinhão manso (*Jatropha curcas* L.) é uma espécie que se tem destacado nos estudos e pesquisas para a produção de biodiesel. É uma cultura perene, resistente às secas, desenvolve-se em terras marginais, rochosas, salinas e desérticas. As suas sementes têm um teor de óleo situado entre 54 – 64% (m/m) (KAMEL et al., 2018), um alto teor de óleo para

a produção de biodiesel quando comparado ao teor do óleo nas sementes de soja que é de 20– 24% (m/m) (ALBRECHT et al., 2008).

A moringa (M*oringa oleífera Lam*) é uma espécie persistente, cultivada nas regiões semiáridas do norte e nordeste do Brasil, principalmente devido ao seu uso no tratamento de água para uso doméstico. A partir de suas sementes pode ser extraído cerca de 33 – 41% de óleo. Embora o seu óleo seja usado na cadeia alimentar em algumas regiões no país e, a ele sejam atribuídas propriedades nutricionais e medicinais, apenas uma pequena quantidade é usada em cosméticos, assim como no caso do pinhão manso o óleo extraído não é destinado ao consumo alimentar humano (DA SILVA et al., 2010; RASHID et al., 2008).

Devido a sua grande extensão territorial e áreas que ainda podem ser destinadas ao plantio de oleaginosas, o Brasil pode aumentar muito a quantidade de óleos disponível para a produção de biodiesel. Outro fator importante é viabilizar a coleta de materiais considerados resíduos como gorduras animais e óleos residuais de processos de frituras (DA SILVA CÉSAR et al., 2019). Por isso, o planejamento energético no Brasil é de fundamental importância uma vez que pode antecipar situações de escassez, mapear alternativas e sugerir estratégias auxiliando então na tomada de decisões de empreendedores e consumidores em relação ao desenvolvimento econômico e social e a sustentabilidade energética.

Atualmente, a Agência Nacional do Petróleo, Gás Natural e Bicombustíveis (ANP) estabeleceu, a partir de 1º de março de 2018, que as misturas biodiesel/diesel atualmente usadas no Brasil são B10, ou seja, misturas constituídas por $10 \pm 0{,}50\%$ (v/v) do biodiesel em $90 \pm 0{,}50\%$ (v/v) de diesel (BRASIL, 2016).

Em julho de 2019, a ANP apresentou resultados de análises em amostras de diesel comerciais em que 136 resultados de análises não atendiam os requisitos de conformidade aos parâmetros de qualidade estabelecidos pela agência. Dentre as não conformidades encontradas destacam-se o teor do biodiesel (55,90%), ponto de fulgor (19,10%) e enxofre (7,40%). Na categoria “outros”, 17,60%, foram agrupadas não conformidades de destilação (7,40%), aspecto, teor de água, contaminação total e cor (ANP, 2019).

Para garantir a qualidade das misturas biodiesel/diesel de acordo com os padrões e regulamentos de qualidade, as normas da ASTM (*American Society of Testing and Materials*) e da ABNT (Associação Brasileira de Normas Técnicas) recomendam o desenvolvimento de modelos de calibração multivariada usando quadrados mínimos parciais (PLS) e espectrometria no infravermelho médio (MIR) com reflectância total atenuada (ATR) horizontal, enquanto que a norma EN 14078 (*European Norm*) recomenda a mesma técnica, porém com a construção da curva de calibração somente com a região espectral referente à

vibração de estiramento da carbonila (C=O) presente nos ésteres (ABNT, 2008; ASTM STANDARD D7371, 2014).

Como se pode notar, os métodos citados nas normas, são quantitativos, porém métodos qualitativos, como Análise Discriminante por Quadrados Mínimos Parciais (PLS-DA) e sua seleção de variáveis, também podem ser usados no controle de qualidade do teor do biodiesel no diesel. Das técnicas de espectrometria vibracional usadas para a análise das misturas biodiesel/diesel, a espectrometria no infravermelho com transformada de fourier e reflectância total atenuada é bem sucedida, porque não é destrutiva, permite a determinação direta e rápida de várias propriedades sem pré-tratamento de amostras, com ferramentas de análise altamente reprodutíveis (OLIVEIRA et al., 2006). Por isso, o uso da espectrometria MIR aliada aos métodos quimiométricos tem sido aplicado em muitas matrizes complexas, incluindo o controle de qualidade do biodiesel. Modelos de calibração multivariada, como a regressão por Quadrados Mínimos Parciais (PLS) são aplicados para prever um ou vários parâmetros de um conjunto de dados multivariados. Esses métodos podem tratar conjuntos de dados mesmo quando o número de variáveis é muito maior que o número de amostras. No entanto, em algumas situações pode ser uma vantagem reduzir o número de variáveis para, entre outras coisas, obter (a) melhoria das previsões do modelo, (b) uma melhor interpretação ou (c) menores custos de medição.

A seleção de variáveis é usada para melhorar o desempenho do modelo e fornecer previsões mais eficientes. Com muitas variáveis irrelevantes, ruidosas ou não confiáveis, a remoção delas normalmente melhora as previsões e/ou reduz a complexidade do modelo. O aprimoramento de propriedades estatísticas também pode ser um motivo para a seleção de variáveis. Em algumas situações, o propósito da seleção de variáveis é obter um modelo que seja mais fácil de entender e interpretar. Por isso, modelos que usam o menor número possível de variáveis, geralmente, são desejados. Além disso, às vezes, o objetivo é usar medições obtidas de instrumentos de alta resolução para identificar as variáveis mais importantes que podem ser aplicadas, por exemplo, em um instrumento baseado em filtro. Isso é relevante para fins industriais *on-line* ou *in-line*, pois um instrumento com componentes de medição mais sofisticados pode reduzir os custos e o tamanho comparado aos de bancada, possibilitando a construção de medidores compactos e portáteis.

Outra importância da seleção de variáveis é reduzir o risco de sobreajuste no modelo construído e simplificação nos processamentos computacionais. Fazer a seleção de variáveis significa tentar inúmeras combinações de variáveis e selecionar as melhores (ANDERSEN; BRO, 2010; DU et al., 2004).

Esta tese está dividida em cinco capítulos, a saber:

1. A Introdução, como primeiro capítulo, apresenta uma breve descrição da demanda energética, a introdução do biodiesel na matriz energética brasileira e a necessidade do uso do biodiesel como fonte alternativa para substituir o óleo diesel, o uso de métodos quimiométricos de seleção de variáveis para o controle da qualidade das misturas biodiesel/diesel e os objetivos deste trabalho.

2. O segundo capítulo faz a Fundamentação Teórica dos principais conteúdos tratados no trabalho, como as matérias-primas e métodos usados na produção do biodiesel, tendência de Moçambique no cenário de biocombustíveis através da cooperação bilateral com o Brasil, vantagens e desvantagens no uso do biodiesel, biodiesel de pinhão manso e de moringa, parâmetros de qualidade do biodiesel, as aplicações mais frequentes na região do infravermelho médio (MIR) e as vantagens do uso da Espectrometria com Transformada de Fourier e com Reflectância Total Atenuada horizontal (FT–MIR–ATR) para o controle de qualidade de biocombustíveis, métodos quimiométricos (incluindo quadrados mínimos parciais – PLS, análise discriminante por quadrados mínimos parciais – PLS-DA e seleção de variáveis), revisão bibliográfica dos trabalhos desenvolvidos para o controle de qualidade das misturas biodiesel/diesel no Laboratório de Quimiometria do Triângulo (LQT).

3. O terceiro capítulo, Procedimento Experimental, apresenta descrições de: produção do biodiesel, preparação de amostras dos conjuntos de calibração/treinamento e de validação/teste com as respectivas concentrações, aquisição de dados espectrais na espectrometria MIR e uma síntese de análise quimiométrica usada na construção dos modelos.

4. O quarto capítulo, Resultados e Discussão, apresenta os resultados obtidos em cada um dos modelos quimiométricos construídos, também é feita uma análise e discussão detalhada dos resultados e das potencialidades do uso métodos quimiométricos (PLS, PLS-DA e seleção de variáveis) aliados à técnica de espectrometria MIR para o controle de qualidade do biodiesel estudado. São apresentados e caracterizados os espectros das amostras usadas em cada modelo. Ainda neste capítulo são apresentados resultados da validação dos melhores modelos de seleção de variáveis, segundo as recomendações das normas de análise multivariada.

5. Finalmente, o quinto capítulo, Conclusões, apresenta a síntese das principais conclusões em relação aos métodos aplicados conforme os objetivos do projeto do trabalho. As referências bibliográficas que serviram de suporte para a produção deste trabalho são apresentadas no final da tese.

1.1. Objetivos

1.1.1. Objetivo Geral

- Aplicar métodos quimiométricos de calibração multivariada, de classificação supervisionada e de seleção de variáveis por intervalos para quantificar teores do biodiesel metílico de pinhão manso e de moringa e classificar as suas amostras em misturas com diesel, usando espectrometria no infravermelho médio.

1.1.2. Objetivos Específicos

- Produzir biodiesel metílico de pinhão manso e de moringa;
- Construir e validar modelos de classificação por PLS-DA empregando a faixa espectral no infravermelho médio, compreendida entre 3150 – 680 cm^{-1}, para classificação do biodiesel metílico de pinhão manso e moringa, em misturas binárias com diesel na faixa de concentração de 0,50 – 30% (v/v).
- Construir e validar modelos de calibração multivariada por PLS empregando a faixa espectral no infravermelho médio, compreendida entre 3150 – 680 cm^{-1}, para quantificar o teor do biodiesel metílico de pinhão manso e de moringa, em misturas binárias com diesel na faixa de concentração de 0,50 – 30% (v/v).
- Avaliar a utilização dos métodos de seleção de variáveis por intervalos (iPLS-DA, iPLS, biPLS e siPLS) comparando com base nos valores de RMSEP e número de variáveis, suas eficiências com os respectivos modelos globais.
- Validar os modelos de seleção de variáveis por intervalos que apresentarem melhor eficiência.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Matérias-Primas Usadas na Produção do Biodiesel

O biodiesel é produzido a partir de várias matérias-primas, como óleos vegetais, gorduras animais ou óleos de frituras residuais. Ou seja, as fontes de ácidos graxos e de álcoois usados são diferentes. Por isso, cada tipo de biodiesel produzido apresenta diferentes características, como o número de cetano, teor de enxofre total, ponto de fulgor, viscosidade cinemática, entre outros. Em diversos países, existem uma maior dependência de algumas culturas, em virtude da viabilidade que elas possuem tendo em conta o fornecimento contínuo e de grande escala. Por exemplo, os Estados Unidos dependem principalmente da soja, a Europa depende da colza, a Malásia, a Índia, a Indonésia a China e a Tailândia dependem do pinhão manso (PANDEY et al., 2012).

Como na produção de biodiesel é possível utilizar uma gama de diferentes óleos vegetais, e que em sua grande maioria apresentam condições técnicas adequadas tanto para a produção quanto para o consumo, frequentemente são investigadas oleaginosas mais abundantes e adaptáveis às condições climáticas e do solo de uma região. Assim, a inserção do biodiesel na matriz energética representa uma possibilidade ao fortalecimento da agroindústria local, regional e internacional, e à geração descentralizada de energia, atuando como forte apoio à agricultura familiar e possibilitando o consórcio de culturas (DEMIRBAS, 2009).

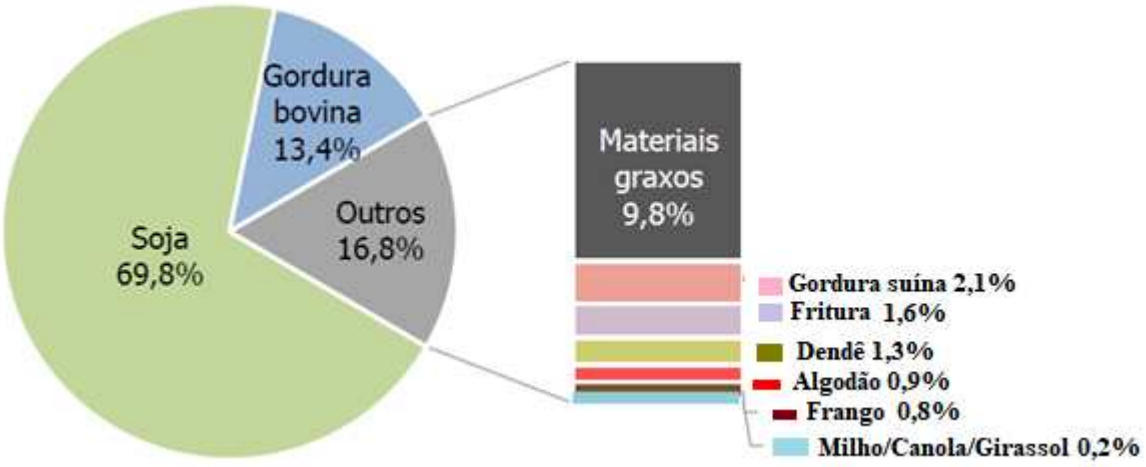
A escolha da matéria-prima para a produção do biodiesel é uma decisão estratégica agroindustrial e depende essencialmente da quantidade energética existente na matéria-prima por cada unidade de volume ou de massa produzida, da quantidade de energia gerada por unidade de energia gasta e do custo da unidade de energia produzida. Além disso, deve seguir os critérios específicos de qualidade estabelecidos para cada tipo de biodiesel, e que correspondem aos motores ou máquinas movidas por eles. Outro fator a considerar é o tipo de óleo, porque se apresentar um elevado teor de cadeias carbônicas saturadas será menos suscetível à oxidação, mas, por outro lado, solidifica-se a temperaturas mais altas. Já os óleos ricos em insaturações e poli-insaturações solidificam-se somente a temperaturas mais baixas, mas, em contrapartida, sofrem oxidação mais facilmente (ANP, 2018; BERGMANN et al., 2013).

No Brasil, o Programa Nacional de Produção e Uso do Biodiesel (PNPB) definiu que a cadeia produtiva do biodiesel seja direcionada para atender preferencialmente à produção agrícola com origem na agricultura familiar, ou seja, a introdução desse biocombustível na

matriz energética brasileira é realizada por uma política energética com forte aspecto social, cujo êxito depende da escolha da matéria-prima de maior viabilidade da região. Por isso, várias pesquisas têm sido desenvolvidas sobre a produção de biodiesel a partir de diferentes matérias-primas e o seu desempenho em motores e nas emissões de gases de escape (ANP, 2019).

Dentre as matérias-primas usadas na produção do biodiesel no Brasil, a soja é a mais utilizada, porque possui uma cadeia produtiva bem estabelecida, grande disponibilidade, com técnicas de cultivo e comercialização consagradas, embora com baixos índices de densidade energética, balanço energético e geração de emprego (ANP, 2018; APROBIO, 2018). Como ilustra a Figura 1, este insumo permaneceu no ano de 2018 como a principal matéria-prima para obtenção de biodiesel, com participação de 69,80% seguido pela gordura bovina com 13,40%. O complemento é representado por óleos de fritura, dendê, algodão, frango, milho, canola e girassol e outros materiais graxos não discriminados (MME, 2019).

Figura 1 – Participação de matérias-primas para a produção de biodiesel no Brasil



Fonte: ANP, 2019

O uso da soja como óleo comestível e a exportação de seus coprodutos como o farelo também consumido em alimentos e rações para animais podem impactar significativamente no preço do biodiesel comercializado. O governo brasileiro, visando minimizar esse impacto, busca por meio de incentivos fiscais e subsídios, por exemplo, do programa de agricultura familiar, compensar com indicadores sociais, exigindo que parte dos grãos de soja sejam adquiridos de pequenos produtores (DE OLIVEIRA; COELHO, 2017).

Considerando a extensão territorial, com áreas suficientes para produção de grãos, e condições geoclimáticas favoráveis, possibilita ao Brasil produzir diferentes espécies

oleaginosas como fonte de matérias-primas adequadas a produção de biodiesel, que possam minimizar os custos de produção e danos ambientais, que não participem direta ou indiretamente na cadeia alimentar humana, que se desenvolvam em terras agrícolas com o uso mínimo de insumos e de acordo com as condições climáticas e do tipo do solo, com produção de óleo igual ou superior ao da soja. Assim pode-se destacar as potencialidades de oleaginosas como o algodão, milho, girassol, moringa, crambe e canola, as espécies oriundas da região de cerrado como a macaúba, pinhão manso, palma, barú e buriti. Outra fonte alternativa e de grande apelo como solução a problemas ambientais relacionados a descartes de materiais residuais, consiste no uso de óleos residuais de frituras, bem como gorduras de origem animal como bovinos, suínos, aves, dentre outros (BERGMANN et al., 2013; MORTON, 1991; SCHUT et al., 2014).

O Brasil já tem uma larga experiência na área de biocombustíveis, incluindo o biodiesel, com vasta área territorial para a produção de matéria-prima e com o corpo técnico humano qualificado. A necessidade de aumentar a segurança energética e promover o desenvolvimento, especialmente nas áreas rurais, forçou muitos países em desenvolvimento no sul da África, como Moçambique a tomar várias ações para o desenvolvimento de várias infraestruturas e legislações para produção e uso de biocombustíveis líquidos (CUVILAS; JIRJIS; LUCAS, 2010). Moçambique está situado na África Austral em zonas tropical e subtropical, possui características edafoclimáticas semelhantes às do Brasil, altamente favoráveis à produção agrícola.

A cooperação Brasil-Moçambique na área dos biocombustíveis emergiu no cenário internacional da política externa dos dois países em 2007, com a assinatura do memorando de entendimento na área dos biocombustíveis. As motivações são, tanto de natureza conjuntural, quanto de semelhança geopolítica e geoestratégica, onde as potencialidades bioenergéticas de Moçambique, em conjunto com as tecnologias brasileiras na área, agregam além de lucros, uma maior representatividade de ambos no cenário energético internacional (BAMBO, 2014).

Os governos de Moçambique e do Brasil firmaram compromissos de desenvolverem projetos visando parcerias tecnológicas e comerciais para a produção de biocombustíveis, produção já consolidada no Brasil e com grande interesse ao Governo moçambicano. Nessa parceria, destaca-se ainda o início do funcionamento da usina de etanol em Moçambique da Empresa Petróleo Brasileiro S.A. (PETROBRAS) em parceria com a Empresa Petróleos de Moçambique (PETROMOC), S.A.R.L. O pinhão manso (*Jatropha curcas L.*) e a cana-de-açúcar (*Saccharum officinarum L.*) são as duas matérias-primas de biocombustíveis mais

proeminentes na África Subsaariana (onde Moçambique faz parte) para a produção de biodiesel e etanol, respectivamente. Destas culturas, a cana-de-açúcar é bem estabelecida e o açúcar derivado dela faz parte do importante mercado global há centenas de anos. Por outro lado, o pinhão manso é uma cultura oleaginosa introduzida recentemente, cujo óleo pode ser diretamente misturado ao diesel em pequenas quantidades ou transformado em biodiesel (MUDOMBI et al., 2018).

Desde os anos 2000, em Moçambique, o pinhão manso é cultivado como uma cultura de biocombustível, mas apenas em 2007 atingiu a idade de colheita em algumas das áreas em que foi introduzido. Sua produção foi incentivada tanto em pequenos produtores quanto em plantações em grande escala (VON MALTITZ; SETZKORN, 2013). Por isso, o Governo moçambicano disponibilizou uma vasta área para o cultivo do pinhão manso para atender projetos de produção do biodiesel. O primeiro projeto de biocombustível em Moçambique foi aprovado em outubro de 2007, seguido por três outros projetos maiores, formalmente aprovados pelo Governo, sendo um deles voltado à plantação de pinhão manso para futura produção de biodiesel. Até dezembro de 2008 haviam sido submetidas 17 propostas de investimentos relacionadas a biocombustíveis, para análise e aprovação governamental, 12 voltados ao biodiesel e 5 ao bioetanol; as áreas de cultivo para o biodiesel e para o bioetanol correspondem a 179 mil e 66 mil hectares, respectivamente (ALVES, 2014; SCHUT; SLINGERLAND; LOCKE, 2010).

Em Moçambique, as plantações de pinhão manso estabeleceram-se com apenas poucas informações disponíveis relacionadas à variedade de sementes, às boas práticas agrícolas, sistemas de produção, mercados e escala de operações. Contudo, apesar de o Governo ter distribuído sementes para a produção, não houve uma assistência técnica efetiva, e, consequentemente, a manutenção da cultura foi negligenciada, e muitas plantas morreram. Os poucos, os agricultores que produziram sementes de pinhão manso não sabiam o que fazer depois, pois não havia mercados e cadeias produtivas organizadas. No entanto, como a promoção dos biocombustíveis pelo então governo moçambicano havia atraído inúmeros investidores privados, bem como alguns projetos de desenvolvimento relacionados com biocombustíveis, recentemente está em curso a formação de recursos humanos nos países estrangeiros que vão atender a produção de matéria-prima e biodiesel e a aprendizagem de métodos de controle de qualidade.

Mais informações sobre o cenário do biodiesel em Moçambique podem ser encontradas na literatura (ALVES, 2014; SCHUT; SLINGERLAND; LOCKE, 2010; SLINGERLAND; SCHUT, 2014).

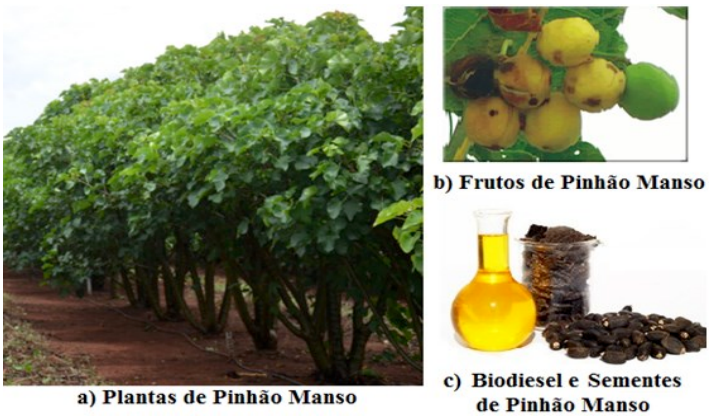
A planta de pinhão manso é um grande arbusto ou árvore resistente às terras secas, nativa e comumente encontrado e utilizado na maioria das regiões tropicais e subtropicais do mundo. Várias propriedades da planta, incluindo sua dureza, crescimento rápido, fácil adaptação e interesse comercial, resultaram em sua propagação muito além de sua distribuição original (KAMEL et al., 2018; KNOTHE; GERPEN; KRAHL, 2010).

A semente é relativamente grande, de cor cinza com uma película branca que envolve os embriões de onde se extrai o óleo que produz o biodiesel. As sementes e o óleo são tóxicos e cancerígenos devido à presença de ésteres do forbol, curcinas e flavonoides (RAKSHIT et al., 2008), e por isso não é usado na alimentação humana (JAIN, 2019). Em humanos, os relatos de toxicidade estão ligados a intoxicação aguda, por ingestão ou contato com as sementes e látex (cutânea e olhos), sendo que não há relatos de efeitos nas vias respiratórias.

O fruto de pinhão manso é capsular avóide achatado nas extremidades, inicialmente verde, passando a amarelo, castanho e por fim preto, quando atinge o estágio de maturação. O óleo de pinhão manso tem secagem lenta, é inodoro e incolor quando fresco, mas depois fica moderadamente amarelo à temperatura ambiente, apresenta baixa acidez, boa estabilidade oxidativa quando comparado ao óleo de soja. O teor de óleo das sementes de pinhão manso varia de 30 – 50% em massa e no próprio núcleo varia de 45 – 60% (PANDEY et al., 2012). A composição de ácidos graxos do óleo de pinhão manso consiste em ácidos palmítico, esteárico, oléico e linoléico. Contudo, a partir das propriedades deste óleo, prevê-se que o óleo seja adequado como combustível e para a produção do biodiesel.

A maior diferença entre o óleo de pinhão manso e o óleo diesel é a viscosidade. A alta viscosidade do óleo de pinhão manso pode contribuir para a formação de depósitos de carbono nos motores, a combustão incompleta do combustível e resultar na redução da vida útil de um motor. Por isso, ele é convertido em biodiesel através de reações químicas como a de transesterificação (PRAMANIK, 2003). A densidade e viscosidade do biodiesel do pinhão manso são estáveis durante o período de armazenamento, fazendo com que o biodiesel produzido satisfaça os padrões exigidos pela ANP. A Figura 2 apresenta as plantas, frutos, sementes e o biodiesel de pinhão manso.

Figura 2 – Plantas, frutos, sementes e biodiesel de pinhão manso



Fonte: O autor.

A produção da moringa está sendo desenvolvida nos distritos da Ilha de Moçambique e Mossuril, pela empresa MoSagri, de capital holandesa, numa extensão de mil hectares. A colheita ocorre com intervalos de 35 dias, sendo que por ano a produção média situa-se em 10 mil toneladas de folha daquela cultura. O distrito de Moamba, na província de Maputo, dispõe de uma unidade fabril de processamento de moringa para fins nutricionais e terapêuticos (BENTO VENÂNCIO, 2014; BURROWS et al., 2018).

A empresa Petróleo de Moçambique (PETROMOC) produziu em 2007 cerca de 500 mil litros de biodiesel na fase experimental, no âmbito do Projeto Ecomoz, Energias Alternativas Renováveis e em coordenação com a empresa PETROBRAS. Este Projeto é uma parceria entre quatro empresas moçambicanas, uma das quais a Bioenergia, que dispõe de uma unidade piloto de produção, com uma capacidade instalada para produzir cerca de 40 milhões de litros de biodiesel por ano na cidade de Matola e arredores de Maputo.

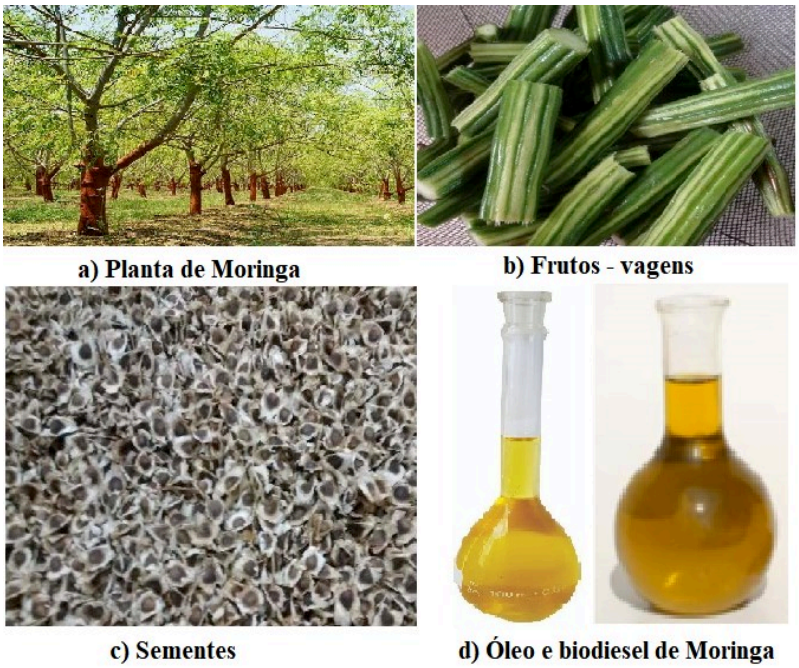
A moringa é uma espécie perene, nativa da região sub-Himalaias, persistente que pertence à família *Moringaceae*. É amplamente distribuído no Gana, na Índia, Moçambique, Egito, Filipinas, Ceilão, Tailândia, Malásia, Birmânia, Paquistão e Cingapura. Normalmente, esta planta cresce em clima tropical, pois é tolerante à seca e possui capacidade de crescer em solos pobres (HABTEMARIAM, 2017; MORTON, 1991). É também cultivado nas regiões semiáridas do nordeste do Brasil. A primeira introdução desta árvore no Brasil limitou-se a ornamentação nos parques públicos. Atualmente, em algumas regiões norte e nordeste é usado no tratamento de água para uso doméstico. Há muito tempo, povos no subcontinente indiano tem usado as vagens de moringa em sua alimentação. As folhas são usadas na alimentação na África Ocidental e em partes da Ásia. No Brasil é conhecida como cedro, lírio branco, quiabo de quina, acácia branca, árvore rabanete de cavalo e moringueiro

(FAHEY, 2005; PALIWAL; SHARMA; PRACHETA, 2011). A planta possui folhas grandes e flores brancas e perfumadas.

Os frutos da moringa são altamente perecíveis e não podem ser armazenadas por longo tempo em condições normais, pois se deterioram devido à ação microbiana em 8 a 10 dias após a descascarem. As sementes são aladas, com uma rica fonte de óleo e proteína e podem ser usadas para a purificação da água (HABTEMARIAM, 2017; MANI; JAYA; VADIVAMBAL, 2007). Elas contêm cerca de 33 – 41% de teor de óleo. O perfil de ácidos graxos do óleo de *Moringa oleífera Lam* é constituído de glicerídeos dos ácidos oleicos (76,0%), palmítico (6,50%), esteárico (5,70%), entre outros (ANWAR et al., 2007).

Em comparação com outros óleos, o óleo de moringa apresenta boa estabilidade oxidativa, embora o ponto de turvação seja bastante elevado; o seu biodiesel tem índice de cetano de aproximadamente 67, que está entre os mais altos reportados para biodiesel (RASHID et al., 2008). Assim, as propriedades acima mencionadas permitem o uso do seu óleo para a produção de biodiesel para substituir o óleo diesel. O biodiesel derivado do óleo de moringa apresenta características relevantes como, um elevado número de cetano e uma boa estabilidade oxidativa, e as demais propriedades satisfazem a qualidade do biodiesel, de acordo com as normas de qualidade (RASHID et al., 2008). As partes úteis da moringa são apresentadas na Figura 3.

Figura 3 – Planta, frutos, semente e biodiesel de moringa



Fonte: Adaptado de PALIWAL, 2011.

2.2. Produção do Biodiesel

Embora os óleos vegetais possuam alta viscosidade em relação ao diesel, antigamente eram usados diretamente em motores a diesel. Para reduzir a viscosidade dos óleos vegetais e permitir seu uso em motores a diesel comuns sem problemas operacionais, como depósitos em motores, alguns pesquisadores (KNOTHE; GERPEN; KRAHL, 2010; ZUÑIGA et al., 2011) estudaram quatro métodos para o uso do óleos vegetais como combustível: misturas de óleos vegetais com diesel, pirólise, microemulsão (misturas com solventes) e transesterificação. Somente a reação de transesterificação leva a produtos geralmente conhecidos como biodiesel, isto é, ésteres alquílicos de óleos e de gorduras.

Os ésteres mais comumente preparados são os metílicos, principalmente porque o metanol é o álcool mais barato, embora existam exceções em alguns países. No Brasil, por exemplo, onde o etanol é mais barato, os ésteres etílicos são usados como combustível. O metanol apresenta como vantagens maior rendimento de ésteres, menor consumo do álcool e maior eficiência energética. Além de óleos vegetais e gorduras animais, outros materiais, como óleos de fritura usados, também são adequados para a produção de biodiesel; no entanto, mudanças no procedimento de reação frequentemente precisam ser feitas devido à presença de água ou ácidos graxos livres nos materiais (COSTA NETO et al., 2000; LÔBO; FERREIRA; CRUZ, 2009).

Os ácidos graxos ou ácidos gordos são ácidos monocarboxilados de cadeia normal que apresentam o grupo carboxila (–COOH) ligado a uma longa cadeia alquílica, saturada ou insaturada. Geralmente, os ácidos graxos têm cadeia de números pares de átomos de carbono de 4 a 28. Os óleos e gorduras são constituídos por ácidos graxos na forma de mono, di e triglicerídeos. Se o óleo não for refinado ou se estiver deteriorado, uma quantidade de ácidos graxo libera-se dos triglicerídeos e forma ácidos graxos livres. A principal consequência disso é que o produto torna-se mais ácido. Um elevado índice de acidez indica, portanto, que o óleo ou gordura está sofrendo quebras em sua cadeia, liberando seus constituintes principais, os ácidos graxos. O índice de acidez corresponde à quantidade (em mg) de base (KOH ou NaOH) necessária para neutralizar os ácidos graxos livres presentes em 1 g de gordura (RAMOS et al., 2017).

Como a maioria dos óleos vegetais não refinados apresentam maior teor de ácidos graxos livres, para a produção do biodiesel, geralmente é usada a catálise dupla, isto é, catálise ácida (esterificação) seguida de uma catálise alcalina (transesterificação). Tendo em conta os aspectos e influências que cada tipo de catálise sofre nas diferentes condições, a catálise dupla mostrou-se como um método viável para a produção de biodiesel, de modo a

utilizar as vantagens de ambos os processos com o elevado rendimento em massa e a pouca interferência da catálise ácida frente a impurezas do óleo vegetal, e a elevada velocidade de conversão de ésteres na catálise alcalina.

2.3.1. Reação de Esterificação

Os ácidos graxos livres presentes nos óleos vegetais são removidos no processo de refino através de etapas de degomagem e neutralização. Eles desfavorecem a catálise no processo da produção do biodiesel, pois consomem maior quantidade de reagentes e catalisadores através de reações paralelas. Além disso, a execução industrial das etapas de refino do óleo eleva muito o custo da produção. Esta reação, catalisada por ácido, é realizada quando a quantidade de ácidos graxos livres no óleo é superior a 1%. Assim, na produção do biodiesel empregando óleos vegetais não refinados é necessária uma reação de esterificação para reduzir a viscosidade e o teor de ácidos graxos livres e aumentar o rendimento de ésteres alquílicos (MOSER, 2012).

Geralmente, a catálise ácida usa o ácido sulfúrico (H_2SO_4) como catalisador, tem uma reação mais lenta, exige temperaturas altas, maiores pressões e quantidades de álcoois em relação à catálise alcalina. Todavia, a catálise ácida apresenta a vantagem de não sofrer interferência com a presença de impurezas e da água na reação. Nesta catálise, o triglicerídeo ataca o hidrogênio do ácido e depois forma um composto intermediário tetraédrico pela reação com o álcool.

Os óleos vegetais são tratados nesta etapa para obter uma mistura com baixo teor de ácidos graxos livres, convertendo-os em ésteres metílicos ou etílicos. A alternativa da esterificação demonstrou ser viável pelo fato de que, no refino do óleo, os ácidos graxos livres são convertidos em sabões de ácidos graxos, enquanto que com a catálise ácida eles são convertidos em ésteres alquílicos derivados de ácidos graxos (RAMADHAS; JAYARAJ; MURALEEDHARAN, 2005). Os mecanismos da reação de esterificação podem ser encontrados na literatura (RAMOS et al., 2017; SITOE, 2016).

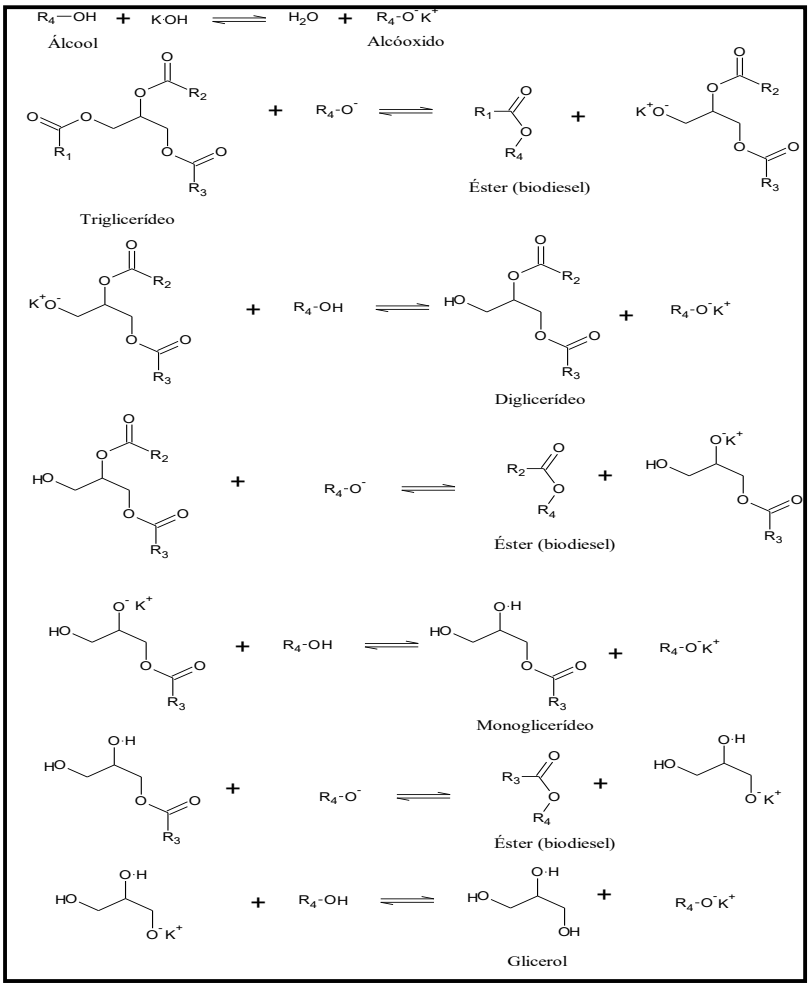
O emprego de óleos vegetais brutos é possível através desse método, sendo o tratamento dos óleos realizado pela catálise ácida. Assim, usa-se a matéria-prima mais barata e o processo mais flexível, podendo-se empregar diferentes oleaginosas e viabilizar economicamente a produção do biodiesel.

2.2.2. Reação de Transesterificação

A transesterificação é o método mais comum usado para converter óleos em biodiesel. Ela é realizada na presença de catalisadores básicos (RAMOS et al., 2017). Os catalisadores homogêneos alcalinos são os mais usados, pois são eficientes e promovem altos rendimentos, embora sejam mais sensíveis à água. Apesar de serem menos ativos, o NaOH e o KOH são de baixo custo, apresentam rendimentos satisfatórios durante a reação e têm sido amplamente empregados. Mas, como o NaOH é uma base mais forte que o KOH, diminui o rendimento e pode causar a reação de saponificação em vez da reação de transesterificação (SANTOS et al., 2014).

Devido à versatilidade e à facilidade do uso da catálise alcalina, esta vem sendo utilizada desde o início da produção e uso do biodiesel no Brasil. Contudo, esta rota de produção é mais empregada em óleos vegetais refinados ou degomados, e não em óleos com elevado teor de água ou ácidos graxos livres. A catálise alcalina é mais rápida que a catálise ácida, é utilizada quando a quantidade de ácidos graxos livres é inferior a 1%. Nesta etapa, ocorre a conversão de triglicerídeos em glicerol e éster alquílico, obtendo um biodiesel a partir dessas matérias-primas. A camada mais leve (biodiesel) é separada da fase mais pesada (glicerol) usando o funil de separação. O biodiesel é removido, lavado com água quente, seco sobre sulfato de sódio anidro e pode ser usado diretamente como combustível de motor em motores a diesel.

Figura 4 – Mecanismo de reação de transesterificação de triacilglicerídeos com metanol em meio alcalino para a produção do biodiesel



Fonte: Adaptado de RAMOS et al., 2017.

2.2.3. Purificação e Parâmetros de Qualidade do Biodiesel

Os processos de purificação do biodiesel, incluindo lavagem e secagem, são necessários, pois o biodiesel produzido contém pequenas impurezas, como glicerol livre, sabão, ácidos graxos livres, álcool, catalisadores, metais e glicerídeos. O excesso do álcool que não reagiu na produção do biodiesel apresenta riscos à segurança e pode corroer os componentes do motor; o catalisador residual (KOH) pode danificar os componentes do motor; e o sabão no biodiesel pode reduzir a lubrificação do combustível e causar superaquecimento, aceleração excessiva e dificuldades no injetor do motor e criar depósitos (NABI; RASUL, 2018).

O biodiesel é lavado com água morna a quente por pelo menos três vezes, o que ajuda na remoção do catalisador e também do álcool residual (se houver). Durante a lavagem, se a

camada de ésteres metílicos e água não se separar dentro de 20 a 30 min, isso indica que a conversão do óleo em biodiesel não foi realizada completamente. O mesmo processo é repetido até uma camada inferior apareça, como água transparente. Depois disso, o biodiesel é separado. No entanto, alguma porcentagem de água ainda está presente no biodiesel. Portanto, o biodiesel após a lavagem é aquecido a 100 ºC por 1 hora para evaporar toda a umidade. O sulfato de sódio anidro também pode ser utilizado para remover o resto da umidade. Assim, o biodiesel fica livre de todas as impurezas (LIN; CHEN, 2017; RAMADHAS; JAYARAJ; MURALEEDHARAN, 2005).

Entre os parâmetros de qualidade instituídos nas normas, encontram-se os que são provenientes da normatização do diesel mineral e os que foram originados de análises dos óleos vegetais, comumente utilizados na indústria oleoquímica. Parâmetros como viscosidade cinemática, ponto de fulgor e cinzas sulfatadas, embora tenham origem na normatização do diesel mineral, fornecem resultados bastante esclarecedores quanto à qualidade do biodiesel. Os parâmetros de qualidade visam fixar teores limites dos contaminantes que não venham causar possíveis degradações do biodiesel durante o processo de estocagem, ou prejudicar a qualidade das emissões da queima, bem como o desempenho, a integridade do motor e a segurança no transporte e manuseio. Vários testes estão disponíveis para caracterizar o biodiesel em relação ao óleo diesel, a fim de avaliar várias propriedades físicas, químicas e térmicas, como densidade, gravidade, viscosidade, ponto de inflamação, ponto de fluidez, valor calorífico, teor de umidade, valor de ácidos graxos livres, teor de éster, entre outras (MAHMUDUL et al., 2017).

Na Europa, os principais países produtores de biodiesel são Alemanha, Itália e França, sendo a Alemanha o maior consumidor. O combustível é usado basicamente em mistura, em proporções que vão de 5 a 30% (v/v). Entretanto, na Alemanha, o biodiesel pode ser comercializado puro (B100). Alguns países, como Brasil, Índia, Alemanha, Itália, França, Estados Unidos, entre outros, desenvolveram seus próprios padrões de biodiesel (APROBIO, 2018; KNOTHE, 2001; UMAR-GARBA; ALHASSAN; KOVO, 2006).

Como em qualquer outro produto que se pretende colocar no mercado, é essencial garantir a qualidade do combustível para uma comercialização e aceitação bem-sucedidas pelos consumidores. Por esse motivo, muitos países possuem legislação específica que estabelece características desejáveis para minimizar os problemas decorrentes do uso dos combustíveis. Na Alemanha, as especificações de qualidade do biodiesel devem estar de acordo com a norma local (DIN 51606) criada 1997 ou EN 14214, Comitê Europeu de Normalização.

Na Itália, o biodiesel não é utilizado para o transporte individual, mas sim pelo transporte público ou empresas de transporte privado e para o aquecimento residencial. As normas locais nacionais UNI 10946 e UNI 10947 são usadas para estabelecer os padrões de qualidade do biodiesel para automóveis e para aquecimento, respectivamente. Na França, o biodiesel é utilizado somente misturado, porém, até a proporção de B5 é comercializado como óleo diesel. Já o B30 é utilizado em frotas cativas, pois todos os ônibus urbanos utilizam a mistura diesel/biodiesel em uma faixa de 5 a 30% (v/v). O governo francês, através da Federação Francesa dos Produtores de Oleaginosas estabeleceu os parâmetros de qualidade do biodiesel produzido localmente (APROBIO, 2018).

O biodiesel austríaco é proveniente do óleo de colza e a produção é voltada mais para o consumo agrícola, onde os agricultores são organizados na forma de cooperativas rurais. A norma ON C1191 é usada desde 1997 para padronizar a qualidade do biodiesel produzido localmente. Nos Estados Unidos, a maior parte da produção é originada de óleo de soja e óleo residual de fritura. Geralmente, o padrão americano de controle de qualidade ASTM D6751 (*American Society of Testing and Materials*) é seguido por muitos países no mundo. A Índia é um dos países asiáticos que usam o biodiesel cuja matéria-prima mais usada é o pinhão manso proveniente das agriculturas familiar e industrial; neste país usa-se um padrão local e recente chamado BIS 15607 (*Bureau Indian Standards*) para padronizar a qualidade do biodiesel (KARMAKAR et al., 2018).

No Brasil, as especificações do B100, a ser misturado com o diesel mineral, são estabelecidas pela ANP, através da Resolução nº 07 de 2008 (RANP 07/08) que substituiu a Resolução nº 42 de 2004, tornando os critérios de avaliação da qualidade do biodiesel brasileiro mais restritivos. Os padrões de qualidades presentes nesta Resolução foram constituídos com base nas normas ASTM D6751 e EN 14214. Atualmente, as misturas biodiesel/diesel têm suas especificações estabelecidas pela resolução ANP 15/2006 (LÔBO; FERREIRA; CRUZ, 2009). A Tabela 1 apresenta alguns parâmetros de qualidade do biodiesel em alguns países.

Tabela 1 – Padrões de qualidade do biodiesel em alguns países

Parâmetros	Brasil ANP 07/2008	Áustria ON C1191	Índia BIS 15607	Norma Oficial Francesa	Alemanha DIN 51606	Itália UNI 10946	EUA ASTM D6751
Densidade a 15 ºC (g/cm^3)	0,86–0,90	0,85–0,89	0,87–0,89	0,87–0,89	0,87–0,89	0,86–0,90	0,88
Viscosidade 40 ºC (mm^2/s)	3,0– 6,0	3,5–5	1,9–6	3,5–5	3,5–5	3,5–5	1,96–6,0
Ponto de fulgor (ºC)	100	100	130	100	110	100	130
Ponto de fluidez (ºC)	---	---	---	10	---	1–5	15–18
Número de cetano	---	49 (mín)	40 (mín)	49 (mín)	49 (mín)	----	47 (mín)
Índice de acidez (mgKOH/g)	0,5 (máx)	0,8 (máx)	0,5 (máx)	0,5 (máx)	0,5 (máx)	0,5 (máx)	0,8 (máx)
Resíduos de carbono (%)	0,05 (máx)	0,05 (máx)	0,05(máx)	---	0,05 (máx)	---	0,05(máx)

Fonte: (APROBIO, 2018; JAIN, 2019; LÔBO; FERREIRA; CRUZ, 2009; MAHMUDUL et al., 2017)

Os padrões mencionados na Tabela 1 mostram que os valores limite de viscosidade, ponto de fulgor e índice de acidez no padrão ASTM são um pouco mais altos que os padrões DIN e ANP. O valor limite do resíduo de carbono é quase o mesmo para todos os padrões acima. Entre os parâmetros gerais do biodiesel, a viscosidade controla as características da injeção do motor a diesel. A viscosidade do éster metílico dos ácidos graxos pode atingir níveis muito altos e, portanto, é importante controlá-la dentro do nível aceitável para evitar impactos negativos no desempenho do sistema injetor de combustível. Portanto, as especificações de viscosidade propostas são praticamente as mesmas do diesel (PALOU et al., 2017).

2.3. Conformidade e não Conformidade do Teor do Biodiesel em Diesel

As propriedades físico-químicas do biodiesel como combustível são fortemente definidas pela estrutura molecular de suas espécies constituintes (ésteres alquílicos de ácidos graxos saturados, insaturados e grupos hidroxilas). Os combustíveis a biodiesel derivados de diferentes fontes podem ter perfis de ácidos graxos e propriedades significativamente variáveis. Essas diferenças de estruturas moleculares dos seus ésteres constituintes, presença de contaminantes oriundos da matéria-prima do processo de produção ou formadas durante a estocagem do biodiesel definem a qualidade do biodiesel. As estruturas moleculares dos ésteres variam tanto no tamanho da cadeia carbônica quanto na quantidade e posição de insaturações; ou mesmo devido à presença de agrupamentos na cadeia, a exemplo da

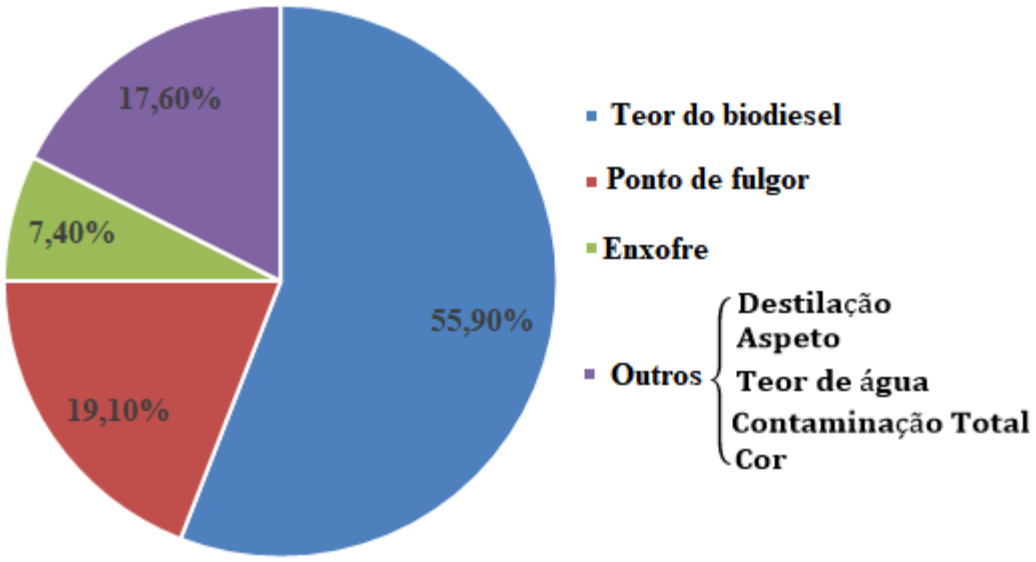
hidroxila ligada à cadeia carbônica do alquiléster derivado do ácido ricinoléico proveniente da mamona (PERDOMO; MILLÁN; ARAGÓN, 2014; SINGH; SINGH, 2010).

O fósforo, o enxofre, o cálcio e o magnésio também podem ser encontrados no biodiesel e são contaminantes procedentes da matéria-prima da produção do biodiesel. Eles podem estar presentes em maior ou menor quantidade, de acordo com a eficiência do processo de produção do biodiesel: glicerina livre, glicerídeos não reagidos, sabões, álcool residual, resíduos de catalisadores e água. A presença de água, peróxidos e ácidos carboxílicos de baixa massa molar podem provir da absorção de umidade e dos processos de degradação oxidativa durante o armazenamento do biodiesel.

O biodiesel usado na sua forma pura é denominado B100 e em misturas com o diesel BX, onde X corresponde à proporção percentual volumétrica do biodiesel na mistura. Embora a Lei nº 11.097/2005 estipulasse um cronograma de adição do mandatório iniciando em 2% de biodiesel no diesel (B2), em janeiro de 2008, e alcançando 5% somente no ano de 2013, o B5 foi antecipado para janeiro de 2010, por decisão de política pública, conforme autorizava a Lei. Esse teor se manteve até 2014, quando a Lei 13.033 definiu sua elevação para 6%, em julho, e 7%, em novembro daquele mesmo ano. Em 2016, a Lei 13.263 estabeleceu um cronograma de elevação do mandatório para 8%, 9% e 10%, em até 12, 24 e 36 meses, respectivamente, após sua promulgação. Dessa forma, em março de 2017 passou a vigorar o B8. A adição obrigatória foi alterada diretamente de 8% para 10% em março de 2018, por decisão do Conselho Nacional de Política Energética, conforme autorizava a Lei. Por isso, as misturas biodiesel/diesel atualmente usadas no Brasil são B10, ou seja, misturas constituídas por $10 \pm 0{,}50\%$ do biodiesel em $90 \pm 0{,}50\%$ de diesel a partir de 1º de março de 2018 (BRASIL, 2016).

Uma das vantagens do uso das misturas biodiesel/diesel é o aproveitamento da rede de distribuição de combustível já existente, e estas misturas podem ser usadas nos motores atuais sem qualquer modificação em uma porcentagem de até cerca de 30% (v/v) (MOFIJUR et al., 2014; TEOH et al., 2019). Das análises das misturas do biodiesel/diesel feitas pelo Programa de Monitoramento da Qualidade dos Combustíveis, verificou-se que o teor do biodiesel, ponto de fulgor e enxofre, que corresponderam, respectivamente, a 55,90%, 19,10% e 7,40% do total de não conformidades. Na categoria “outros”, 17,60%, foram agrupadas não conformidades de destilação (7,40%), aspecto, teor de água, contaminação total e cor (ANP, 2019), conforme se apresenta na Figura 5

Figura 5 – Índice de não-conformidade das misturas biodiesel/diesel



Fonte: ANP, 2019.

O programa de monitoramento da qualidade dos combustíveis foi instituído, pela ANP, visando a atender ao disposto no artigo 8º. da Lei nº. 9.478/1997, em particular os incisos que tratam da garantia de qualidade e do suprimento de combustíveis ao mercado nacional. Nas análises feitas pela ANP, as amostras são coletadas e levadas ao laboratório de análises, o que demandando tempo. Portanto, se faz necessário o uso de metodologias de análises rápidas e eficientes em que os resultados das análises sejam em tempo real. A espectrometria FT-MIR-ATR é uma técnica que não exige preparo de amostras, faz análise rápida e *in-situ*, mostrando-se assim como uma alternativa viável para o monitoramento de qualidade desse combustível. Além disso, esta técnica é utilizada como método padrão para quantificar o teor do biodiesel em misturas com diesel de acordo com a Norma Europeia (EN 14078) (PINHO et al., 2014), Associação Brasileira de Normas Técnicas (ABNT NBR 15568) (ABNT, 2008) e Sociedade Americana de Testes e Materiais (ASTM D7371) (ASTM STANDARD D7371, 2014).

2.4. Espectrometria no Infravermelho Médio

Espectrometria é um termo geral para a ciência que lida com a interação entre os vários tipos de radiação e matéria. Portanto, esses tipos de medidas só são possíveis se a interação entre fótons e matéria causar algum tipo de alteração em uma ou mais propriedades da amostra. A espectrometria no infravermelho fornece informações sobre estrutura molecular, níveis de energia e ligações químicas. Os compostos que possuem ligações

covalentes absorvem em várias frequências de radiação eletromagnética na região do infravermelho, representadas em um espectro eletromagnético (PAVIA et al., 2016).

O princípio da espectrometria no infravermelho é que a radiação eletromagnética pode interagir com as moléculas, levando-as da energia do estado fundamental para uma energia vibracional do estado excitado, na qual os movimentos dos átomos têm maior energia. No processo de absorção são absorvidas as frequências de radiação no infravermelho que equivalem às frequências vibracionais naturais da molécula em questão, e a energia absorvida serve para aumentar a amplitude dos movimentos vibracionais das ligações na molécula. Porém, nem todas as ligações em moléculas absorvem energia no infravermelho. Apenas as ligações em moléculas que têm um momento de dipolo que muda como uma função de tempo, ou seja, moléculas que têm momento de dipolo diferente de zero, são capazes de absorver radiação no infravermelho, como por exemplo, os modos vibracionais de $–CH_2$, de H_2O e algumas vibrações de CO_2. Ligações simétricas, como as do H_2 ou O_2, não absorvem radiação no infravermelho, porque elas não apresentam um dipolo elétrico que mude, ou seja, o seu momento dipolar é igual a zero (SILVERSTEIN et al., 2019).

A transição de estado segue regras da mecânica quântica e obedece aos níveis quantificados de energia; para uma visão do oscilador harmônico da vibração, a energia que leva uma molécula de um estado vibracional para um estado imediatamente superior é diretamente proporcional à frequência vibracional e é dada pela Equação 1:

$$E = hv = \frac{hc}{\lambda} \quad (1)$$

sendo E é a energia vibracional, h é a constante de Planck, v é a frequência da vibração. Portanto, o fóton que pode interagir com uma molécula deve ter uma energia mínima necessária para alterar seu estado vibracional; caso contrário, a transição não será alcançada.

A porção do espectro eletromagnético que pode levar a essas transições eletrônicas é a região do infravermelho. O espectro eletromagnético da região do infravermelho pode ser dividido em três regiões de acordo com o comprimento de onda da radiação, com diferentes aplicações para estudo espectrométrico, nomeadamente (SILVA et al., 2017):

- Infravermelho próximo (NIR), de 14.000 a 4.000 cm^{-1}, permite o estudo de sobretons e vibrações harmônicas ou combinadas;
- Infravermelho médio (MIR), de 4000 a 400 cm^{-1}, permite o estudo das vibrações fundamentais e da estrutura rotação-vibração de pequenas moléculas, e;
- Infravermelho distante (FIR), de 400 a 10 cm^{-1}, permite o estudo de vibrações de átomos de baixo peso molecular.

Atualmente, a espectrometria de infravermelho médio é amplamente utilizada para análise de biodiesel, devido às bandas de maior incidência espectral nesta região, bem como à maior intensidade e especificidade do sinal. A região entre 1500 – 800 cm^{-1} na região MIR é denominada "região da impressão digital" onde as absorções incluem as contribuições de vibrações interativas complexas que dão a cada composto uma informação única, ou seja, a região da impressão digital é a de identidade de cada substância que absorve no infravermelho.

Fazendo-se variar as distâncias percorridas pelos dois feixes, obtêm-se uma sequência de interferências construtivas e destrutivas e, consequentemente, variações de intensidade: interferograma. Uma transformada de Fourier converte o interferograma assim obtido, que está no domínio do tempo, para uma forma mais familiar de um interferograma no domínio de frequências – originando, assim, o espectro do infravermelho (SILVERSTEIN et al., 2019).

O desenvolvimento da espectrometria no infravermelho por Transformada de Fourier (FTIR) introduziu um método popular para a análise quantitativa de misturas complexas. Isso melhorou muito a qualidade e interpretação dos espectros e minimizou o tempo necessário para a obtenção de dados. A técnica FT-MIR é muito usada na caracterização e quantificação de componentes ou propriedades de biocombustíveis. Esses tipos de aplicações são muito comuns e uma grande quantidade de trabalhos foram publicados nos últimos anos (SILVA et al., 2017).

Outra grande característica da técnica FT-MIR é a dependência de concentração. A intensidade dos sinais está relacionada à característica da vibração da banda e ao número de moléculas necessárias para fazer a transição de estado (ou seja, concentração da molécula na matriz). Essa dependência pode ser definida pela lei de Lambert-Beer, que afirma que existe uma relação linear entre sinal e concentração. Portanto, isso pode ser usado para a estimativa da concentração de componentes presentes nas misturas ou para outras propriedades em diferentes sistemas (DUTTA, 2017).

2.4.1. Reflectância Total Atenuada

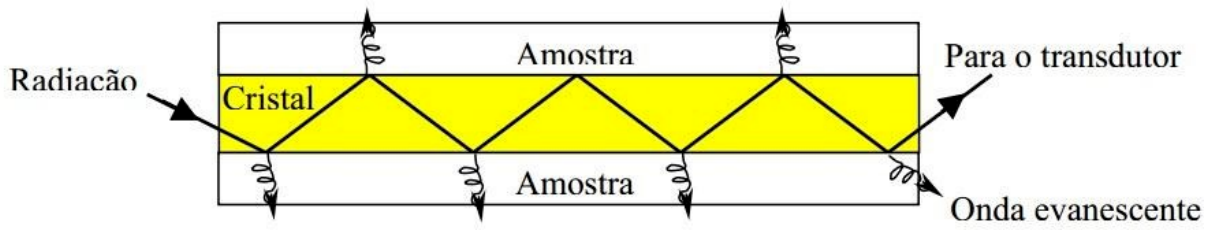
Os principais métodos de reflexão no infravermelho são: o de reflexão difusa, reflexão especular (ou externa), reflexão interna e o método de reflexão total atenuada. A reflectância total atenuada (ATR, do inglês, *Attenuated Total Reflectance*) tornou-se a técnica de amostragem mais amplamente praticada em espectrometria de infravermelho, pois ela requer pouca ou nenhuma preparação de amostra e facilmente podem ser obtidos resultados consistentes.

Ao contrário de outras técnicas de amostragem usadas na espectrometria de infravermelho, a radiação não é transmitida pela amostra; consequentemente, a amostra não precisa ser fina o suficiente para permitir a transmissão da radiação incidente, sem banda com uma absorvância maior que 2,0 UA. Além disso, as amostras podem ser medidas em seu estado puro e não requerem diluição para registrar o espectro. A morfologia física da amostra geralmente não é um problema, desde que seja possível manter uma área de contato suficiente entre a amostra e o elemento sensor. Mesmo amostras altamente irregulares, como pós e tecidos, às vezes fornecem bons espectros, mas pode ser difícil reproduzir a área de contato da amostra; portanto, a análise quantitativa pode ser difícil.

Para fins práticos, os materiais de reflexão interna devem ter o máximo de um ângulo crítico possível quando em contato com a amostra sob investigação. O cristal de ZnSe é mais usado entre os cristais de ATR disponíveis comercialmente. Ele é insolúvel em água, resistente a choques térmicos e pode ser aplicado a soluções ácidas, básicas e a solventes orgânicos com poucas restrições de pH que podem danificar o cristal. Este material suporta bem materiais líquidos, pastas, géis, filmes, polímeros, pós e sólidos macios. O cristal de ZnSe tem índice de refração de 2,40 e ângulo de incidência de 45º.

ATR é uma técnica pela qual a amostra é colocada em contato com um amostrador e um espectro é registrado como resultado desse contato, do índice de refração da amostra e da incidência de radiação. A radiação penetra determinada distância do meio mais denso (cristal de ATR) para o meio menos denso (amostra), ocorrendo uma reflexão. A fração do feixe incidente que é refletida aumenta com o ângulo de incidência, e quando excede o ângulo crítico (45º), a reflexão é total. No processo de reflexão, o feixe radiante comporta-se como se penetrasse um pouco no meio menos denso antes que a reflexão ocorra, conforme apresentado na Figura 6. A profundidade de penetração, que varia de uma fração de vários comprimentos de onda, depende do comprimento de onda do raio incidente, dos índices de refração dos meios materiais e do ângulo do feixe incidente em relação à interface.

Figura 6 – Representação da ATR



Fonte: O autor.

Essa radiação penetrante é denominada onda evanescente e pode ser parcialmente absorvida, colocando-se a amostra no meio mais denso denominado prisma ou elemento de reflexão interna. Se o meio menos denso absorve essa radiação, ocorre atenuação do feixe nos comprimentos de onda das bandas de absorção. Esse fenômeno é conhecido como Reflectância Total Atenuada. Os múltiplos pontos de ondas evanescentes ocorrem em cristais de ATR longos e finos, o que aumenta a sensibilidade do experimento (GRIFFITHS; DE HASETH, 2007).

Como os espectros MIR podem apresentar algumas semelhanças e complexidades, usando métodos quimiométricos é possível extrair informações muito importantes, não perceptíveis numa análise visual de registros espectrais da propriedade de interesse. Outra grande vantagem consiste na aplicação em análises de matrizes complexas, mesmo na presença de interferentes, o que seria impraticável para a análise univariada sem tratamento prévio da amostra.

2.5. Quimiometria

A Quimiometria é a parte da química que utiliza métodos matemáticos, estatísticos e de lógica formal para definir ou selecionar as condições ótimas de medidas e experiências, e permitir a obtenção do máximo de informações a partir da análise de dados químicos (BRUNS; FAIGLE, 1984). Com a sofisticação sempre crescente das técnicas instrumentais, impulsionada pela crescente onda de microprocessadores e microcomputadores no laboratório químico, tornam-se necessários tratamentos de dados mais complexos do ponto de vista matemático e estatístico, a fim de relacionar os sinais obtidos (intensidades das bandas espectrais, por exemplo) com os resultados desejados.

Muita ênfase tem sido dada aos sistemas multivariados, nos quais se pode medir muitas variáveis, simultaneamente (ou de forma sequencial, com grande eficiência) ao se estudar uma amostra qualquer. Nesses sistemas, a conversão da resposta instrumental no dado químico de interesse requer a utilização de métodos de estatística multivariada, álgebra matricial e análise numérica. Tais métodos constituem a melhor alternativa para a interpretação de dados para a aquisição do máximo de informações sobre o sistema (FERREIRA, 2015).

A princípio, os métodos quimiométricos foram desenvolvidos e empregados em outras áreas do conhecimento, como economia e psicologia, enquanto que as soluções dos problemas em química eram obtidas através de métodos clássicos de análises por via úmida ou instrumental. A Quimiometria é mais usada na química analítica devido ao avanço e ao

desenvolvimento da instrumentação química, associada a computadores. Porém, outras áreas em química vêm sendo também influenciados pela Quimiometria.

Os métodos espectrométricos comumente empregados em química orgânica e inorgânica tornam-se mais eficientes quando acoplados a métodos de matemática multivariada, o que permite a análise quantitativa de grande quantidade de dados. A Quimiometria é também aplicada nas áreas de alimentos, saúde, medicamentos, combustíveis, meio ambiente, dentre outras, atuando na cadeia produtiva, controle de qualidade, estudos de adulterações, quantificações, classificação de amostras, monitoramento do controle de qualidade, entre outras aplicações (TASIC et al., 2019).

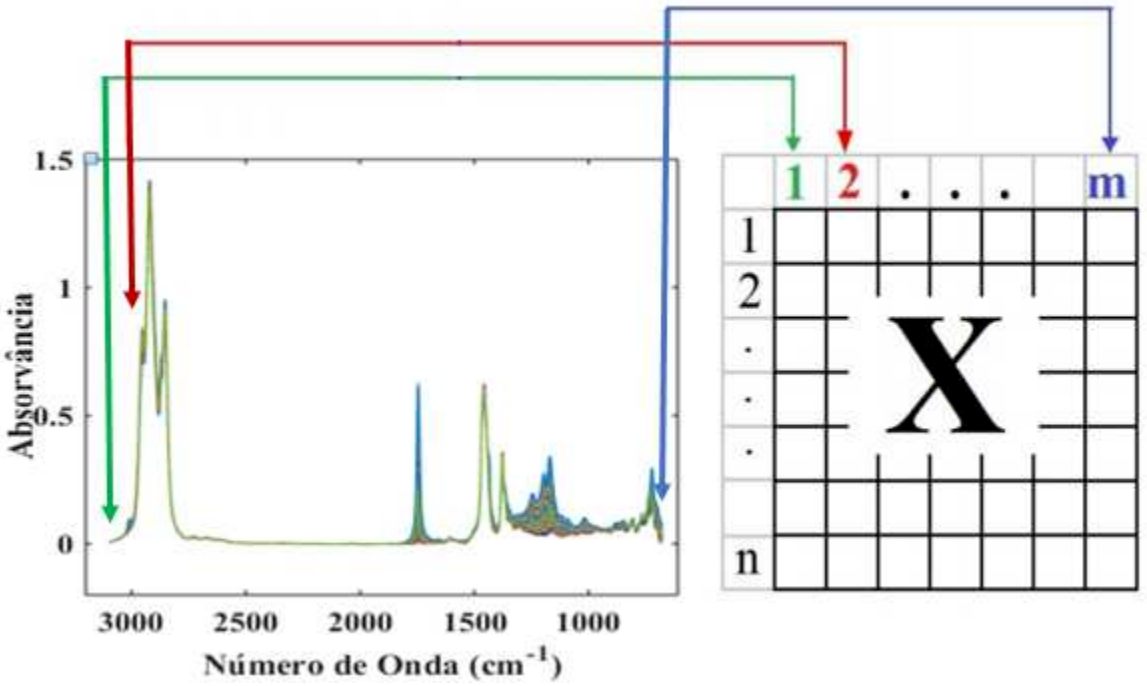
Geralmente, durante a obtenção de informações quantitativas e qualitativas a partir de espectros complexos, e sendo dados multivariados, é indispensável a utilização de métodos quimiométricos, desenvolvidos e disponibilizados em programas computacionais. Esses métodos, aliados aos avanços tecnológicos instrumentais, são os maiores responsáveis pela popularização e do uso da espectrometria vibracional (BRUNS; FAIGLE, 1984).

2.5.1. Organização e Tratamento de Dados

A organização de dados é importante para transformar o sinal analítico em um arquivo com números que formam o espectro, isto é, ao serem obtidos os espectros, estes devem ser exportados num formato numérico que possa ser lido por um programa específico. Para conjuntos com muitas variáveis, a alta dimensionalidade apresentada pode dificultar o tratamento dos dados, e uma ferramenta matemática que possibilite uma melhor visualização espacial dos mesmos, como por exemplo a Análise de Componentes Principais, torna-se de grande valia.

A análise multivariada requer a organização do conjunto de dados em estudo numa matriz **X** (*n* x *m*), onde as linhas (n) desta matriz representam o conjunto das amostras e as colunas (m), o conjunto das variáveis medidas, ou seja, resultados analíticos, como por exemplo os comprimentos de onda de dados espectrais (WOLD; ESBENSEN; GELADI, 1987). Uma matriz de dados **X**, contendo *m* medidas experimentais (variáveis) obtidas para *n* amostras, pode ser graficamente representada por *n* pontos num espaço *m*-dimensional ou, em outras palavras, é possível representar-se esta matriz espacialmente, onde cada variável medida corresponde a uma dimensão do espaço e cada amostra um ponto neste mesmo espaço, conforme apresentado na Figura 7.

Figura 7 – Representação da construção de uma matriz de dados **X**



Fonte: Adaptado de WOLD; ESBENSEN; GELADI, 1987.

Uma vez que os dados experimentais foram organizados, na forma da matriz, eles devem, se necessário, serem pré-tratados antes da análise quimiométrica. Esse é um procedimento muito importante em qualquer análise e, em geral, vários métodos são testados para garantir que o pré-tratamento mais adequado seja utilizado. O objetivo do pré-tratamento é reduzir as variações indesejáveis que não foram removidas durante a aquisição dos dados e que não serão eliminados naturalmente durante a análise, mas que podem influenciar os resultados finais (FERREIRA, 2015). No entanto, qualquer tipo de tratamento deve ser feito com cuidado e criteriosamente para não distorcer, comprometer ou até mesmo inutilizar os resultados da análise. Há dois tipos de pré-tratamentos: um dele é aplicado às amostras (às linhas na matriz de dados **X**), e o outro às variáveis (às colunas na matriz de dados **X**). O pré-tratamento aplicado às variáveis é também chamado de pré-processamento.

As amostras anômalas influenciam fortemente no modelo quimiométrico. Dessa forma, antes de desenvolver um modelo quimiométrico, é essencial encontrar e eliminar, ou corrigir as amostras anómalas. Outra circunstância onde se faz necessário tratar os dados seria no caso destes conterem informações correlacionadas, como ocorre, por exemplo, no caso de dados espectrométricos.

Os pré-processamentos das variáveis podem ser feitos basicamente de três maneiras: centrando na média, escalando pela variância ou auto-escalando (centralização dos dados na

média e posterior escalamento pela variância). É possível também proceder-se a pré-tratamentos nas amostras do conjunto de dados **X**, como por exemplo a correção da linha de base, aplicação da 1ª e 2ª derivadas, o alisamento e a normalização, entre outros, sendo que o uso de cada método deve ser avaliado previamente, dependendo do conjunto de dados em questão (WOLD; ESBENSEN; GELADI, 1987).

A centralização dos dados na média é convenientemente usada quando todas as variáveis forem medidas numa mesma unidade, possuindo uma mesma magnitude, como acontece normalmente no caso da espectrometria. Esse tipo de pré-processamento permite que a presença de ruídos não afete negativamente na análise. Neste tipo de pré-processamento, o centro da matriz de dados é levado à origem pela da subtração de cada elemento de cada coluna pela média da respectiva coluna, conforme a Equação 2 (FERREIRA, 2015):

$$\boldsymbol{X_{ij(cm)} = X_{ij} - \overline{X_j}} \quad (2)$$

onde $X_{ij(cm)}$ é o valor centrado na média para a variável *j* na amostra *i;* $\boldsymbol{X_{ij}}$ é o valor da variável *j* na amostra *i*; e $\boldsymbol{\overline{X_j}}$ é a média dos valores das variáveis na coluna *j*.

No escalamento pela variância, cada elemento de uma dada variável é dividido pelo desvio padrão dessa variável, levando dessa forma a variância à unidade. Esse tipo de escalamento conduz todos os eixos da coordenada ao mesmo comprimento, dando a cada variável a mesma influência no modelo (FERREIRA, 2015). O valor escalado pela variância para cada variável é dado pela Equação 3:

$$\boldsymbol{X_{ij(var)} = \frac{X_{ij}}{S_j}} \quad (3)$$

onde $\boldsymbol{X_{ij(var)}}$ é o valor escalado pela variância para a variável *j* na amostra *i*; $\boldsymbol{X_{ij}}$ é o valor da variável *j* na amostra *i*; e $\boldsymbol{S_j}$ é o desvio padrão dos valores da variável *j*.

O auto-escalamento é feito pela centralização dos dados na média e posterior escalamento pela variância, ou seja, o auto-escalamento implica subtrair de cada elemento de uma coluna da matriz de dados o valor médio da respectiva coluna e dividir o resultado pelo desvio padrão dessa coluna, de acordo com a Equação 4. As variáveis terão dessa forma média zero e um desvio padrão igual a um (1).

$$\boldsymbol{X_{ij(as)} = \frac{X_{ij} - \overline{X_j}}{S_j}} \quad (4)$$

Enquanto o centrar dados na média é mais aplicado para dados espectrométricos, o escalamento e auto-escalamento são métodos de pré-processamento mais utilizados quando se pretende dar o mesmo peso a todas as variáveis medidas, como na Análise de

Componentes Principais, que, por ser um método de quadrados mínimos, faz com que variáveis com alta variância possuam altos pesos.

De acordo com a área de estudo, os métodos quimiométricos são divididos em três grandes áreas: classificação, calibração multivariada e planejamento e otimização de experimentos. O planejamento e otimização de experimentos visa encontrar variáveis que mais afetam um determinado processo, assim como a interação entre elas, enquanto que a calibração multivariada estabelece um modelo matemático que relaciona uma série de medidas (químicas ou espectrais) realizadas em amostras com determinada (s) propriedade (s) de interesse (concentração) na amostra. A partir do modelo construído torna-se então possível a quantificação dessa (s) propriedade (s) em novas amostras onde elas são desconhecidas. A calibração univariada relaciona uma única resposta (variável) com essa propriedade de interesse, enquanto que na multivariada existe uma relação de um conjunto de resposta (FERREIRA et al., 1999).

Os métodos quimiométricos de classificação dividem-se em dois grupos, a saber: métodos de classificação não supervisionada ou conhecimento de padrões e métodos de classificação supervisionada ou reconhecimento de padrões. A Análise de Componentes Principais (PCA – do inglês, *Principal Component Analysis*) e Análise de Agrupamentos Hierárquicos (HCA – do inglês, *Hierarchical Clustering Analysis*) são métodos de classificação não supervisionada. Eles são usados para análise exploratória de dados. Quando não se possui um conjunto de treinamento, ou ainda se dispõe de informações sobre o sistema, para se prever o número de categorias esperado para um grupo de amostras, utiliza-se o conhecimento de padrões, que é uma forma de aprendizagem não supervisionada.

2.5.2. Análise de Componentes Principais

A base fundamental da maioria dos métodos modernos para tratamento de dados multivariados é a PCA, que consiste numa manipulação da matriz de dados com o objetivo de representar as variações presentes em muitas variáveis, através de um número menor de "fatores". Constrói-se um novo sistema de eixos (denominados rotineiramente de fatores, componentes principais ou ainda autovetores) para representar as amostras, no qual a natureza multivariada dos dados pode ser visualizada em poucas dimensões.

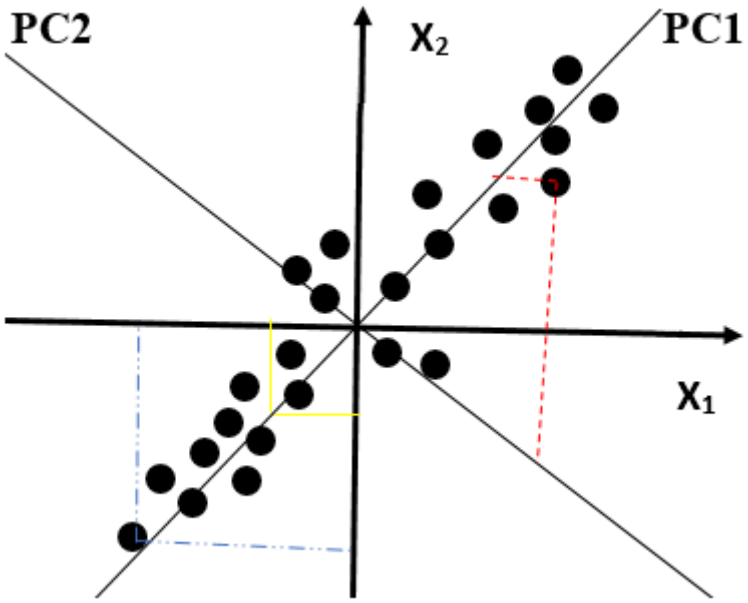
A Análise de Componentes Principais é um dos métodos bastante usado na análise exploratória, e pode ser considerado como método de projeção, uma vez que os dados são projetados em um espaço de dimensão menor. Agrupa as variáveis altamente correlacionadas numa nova variável. Permite a redução da dimensionalidade através da representação do

conjunto de dados em um novo sistema de eixos, denominadas componentes principais (PCs), permitindo a visualização da natureza multivariada dos dados em gráficos de bi- ou tri-dimensões. Variáveis que apresentam grande redundância entre si são colineares e a alta colinearidade é uma forte indicação de que é possível encontrar-se novas bases que melhor representem as informações presentes nos dados que aquela definida pelas medidas. A alta colinearidade entre as variáveis também implica em que os dados residem em um subespaço do espaço total definido pelas medidas.

Cada linha da matriz de dados é representada por um ponto no gráfico de PCA. Em termos geométricos a função das componentes principais é descrever a variação ou espalhamento entre os pontos usando o menor número possível de eixos. Isto é feito definindo novos eixos, componentes principais, que se alinham com os dados. A primeira componente principal, PC1, tem uma direção tal que descreve o máximo espalhamento das amostras, mais que qualquer uma das duas variáveis originais. Nota que a primeira componente principal, descreve maior variação dos dados que a segunda componente principal, PC2, que é ortogonal a PC1. A PC2 é estimada para descrever a máxima variação restante dos dados do modelo com duas PCs (WOLD; ESBENSEN; GELADI, 1987).

A Figura 8 apresenta uma representação gráfica de um conjunto de amostras no espaço bidimensional definido pelas duas variáveis, X_1 e X_2, e as respectivas componentes principais, PC1 e PC2. Pode-se dizer que a dimensionalidade intrínseca desse conjunto de dados é dois, pois são necessárias duas componentes principais para descrever todas as amostras. As novas coordenadas das amostras, no novo sistema de eixos das componentes principais mostradas pela linha tracejada vermelha são denominadas de escores (*scores*). Cada PC é construída pala combinação linear das variáveis ortogonais. Os coeficientes da combinação linear (o peso, ou quanto cada variável antiga contribui) são denominados de pesos (*loadings*) e representados pela linha amarela na Figura 8. Esses pesos variam entre +1 e –1 e são os cossenos dos ângulos entre a PC e os eixos das variáveis originais. As linhas tracejadas de cor azul representam as coordenadas de uma amostra em relação aos eixos originais.

Figura 8 – Representação gráfica de um conjunto de dados bidimensionais (X_1, X_2), mostrando os eixos das componentes principais (PC1, PC2), os pesos e os escores



Fonte: Fonte: Adaptado de WOLD; ESBENSEN; GELADI, 1987.

O novo conjunto de eixos de coordenadas no qual se projetam as amostras é muito mais informativo e, pelo fato de serem ordenados pela sua importância, é possível visualizar estas mesmas amostras num gráfico de baixa dimensionalidade. A projeção em uma base ortogonal pode ser feita, entre outros métodos, por meio da decomposição por valores singulares (*Singular Value Decomposition* – SVD). Nessa projeção, a matriz original $\boldsymbol{X}_{(n\times m)}$ é decomposta e então representada pelo produto de três novas matrizes, duas delas ortonormais (**T** e **P**) e uma diagonal (**S**), conforme a Equação 5:

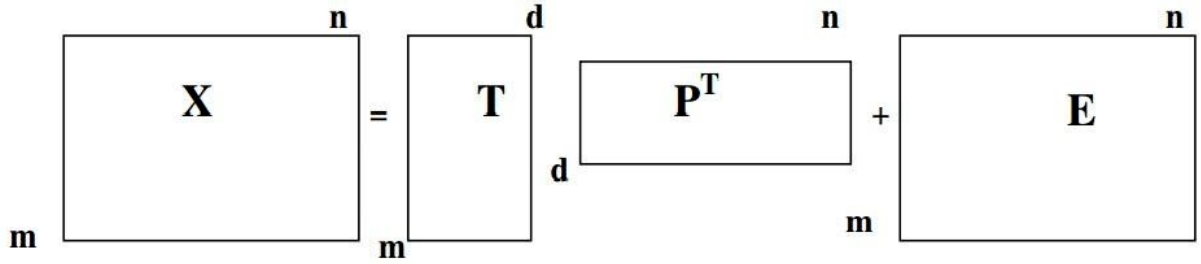
$$\mathbf{X} = \mathbf{TSP}^T \tag{5}$$

A matriz **S** é uma matriz diagonal com elementos diagonais não negativos arranjados em ordem decrescente. Os quadrados dos valores singulares correspondem aos autovalores da matriz $\mathbf{X^TX}$ e medem a importância das componentes principais individuais (cada valor singular representa a porcentagem de variância explicada em cada uma de suas respectivas componentes). As colunas de **P** são autovetores da matriz $\mathbf{X^TX}$, ou matriz dos pesos, abrangendo o espaço vetorial das colunas de **X**, enquanto que a matriz **T**, formada pelos autovetores da matriz $\mathbf{XX^T}$, abrange o espaço vetorial das linhas de **X**. O produto **TS** define as coordenadas das amostras na nova base que forma a matriz dos escores.

Matematicamente, na PCA, a matriz de dados **X** é decomposta em um produto de duas matrizes menores de escores (**T**) e de pesos (**P**), mais uma matriz de resíduos ($\mathbf{E_x}$), que indica a parte que não é modelada, de acordo com a Equação 6 e como mostra a Figura 9.

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E_x} \quad (6)$$

Figura 9 – Decomposição da matriz **X** nas matrizes de score, de pesos e de resíduos



Fonte: Fonte: Adaptado de WOLD; ESBENSEN; GELADI, 1987.

onde *d* é o número de PCs.

2.5.3. Calibração Multivariada

Calibração é o procedimento usado para encontrar um algoritmo matemático que produza propriedades de interesse a partir dos resultados registrados pelo instrumento. A absorbância e concentração são apenas exemplos das inúmeras possibilidades que este procedimento oferece para a obtenção de informações de interesse a partir de sinais instrumentais. O processo geral de calibração consiste de duas etapas: a modelagem, que estabelece uma relação matemática entre **X** e **y** nos conjuntos de calibração, e a validação, que otimiza a relação no sentido de uma melhor descrição de analito (s) de interesse. Uma vez concluída a calibração, o sistema está apto a ser utilizado para a previsão em outras amostras (FERREIRA et al., 1999).

Os modelos de calibração multivariada desempenham um papel importante em vários campos técnicos. Esses modelos não são aplicados apenas em particular nas indústrias química, petroquímica, farmacêutica, cosmética, de coloração, plásticos, papel, borracha e alimentos, mas também em pesquisas forenses, ambientais, médicas, sensoriais e de marketing. A espectrometria no infravermelho médio é cada vez mais usada na caracterização e controle de qualidade do biocombustíveis (SILVA et al., 2017). A calibração multivariada é então usada para desenvolver uma relação quantitativa, isto é, um modelo matemático, entre os espectros digitalizados, armazenados em uma matriz de dados **X** e as concentrações, armazenadas em uma matriz de dados **y**.

Vários métodos foram desenvolvidos para a construção de modelos de calibração multivariada. Os três métodos mais comuns são Regressão Linear Múltipla (MLR, do inglês *Multiple Linear Regression*), também conhecida como Quadrados Mínimos Ordinários (OLS – do inglês, *Ordinary Least Squares*), Regressão por Componentes Principais (PCR– do inglês, *Principal Components Regression*) e regressão por Quadrados Mínimos Parciais (PLS – do inglês, *Partial Least Squares*). Os métodos tradicionais de calibração, o método Clássico de Quadrados Mínimos (CLS – do inglês, *Classic Least Square*) e MLR, têm suas vantagens e desvantagens quando aplicados a problemas químicos. A CLS e MLR utilizam toda a informação contida na matriz de dados **X** para modelar a concentração, isto é, toda a informação espectral, incluindo informações irrelevantes (fazem pequena remoção de ruído). O CLS tem como principal problema a necessidade de se conhecer as concentrações de cada espécie espectrometricamente ativa no conjunto de calibração, o que em geral é impossível nos problemas práticos. Já o método MLR sofre do problema de colinearidade: o número de amostras (n) deve exceder o número de variáveis (m), que por sua vez devem fornecer predominantemente informação única. Tem-se, neste caso, a opção de selecionar um certo número de variáveis que seja menor que o número de amostras e que produzam informação "única", o que pode ser demorado e trabalhoso. Mais interessante, então, seria a utilização de algum método que, como o CLS, use o espectro inteiro para análise, e como o MLR, requeira somente a concentração do analito de interesse no conjunto de calibração (FABER; RAJKÓ, 2007; GELADI; KOWALSKI, 1986).

Dois métodos que preenchem estes requisitos são Regressão por Componentes Principais (PCR – do inglês, *Principal Components Regression*) e PLS. Estes dois métodos são consideravelmente mais eficientes para lidar com ruídos experimentais, colinearidades e não linearidades. Todas as variáveis relevantes são incluídas nos modelos via PCR ou PLS, o que implica que a calibração pode ser realizada eficientemente mesmo na presença de interferentes, não havendo necessidade do conhecimento do número e natureza dos mesmos. Os métodos PCR e PLS são robustos, isto é, seus parâmetros praticamente não se alteram com a inclusão de novas amostras no conjunto de calibração (FERREIRA et al., 1999).

Por esse motivo PCR e PLS são geralmente chamados de métodos de espectro total. O PCR e o PLS são capazes de lidar com um número arbitrariamente grande de variáveis espectrais, comprimindo a matriz de dados **X** em um número relativamente pequeno, denotado por A, do chamado t-escores, geralmente menos de dez. Em seguida, a matriz dos escores **T** de tamanho $n \times A$ substitui a matriz de dados originais, **X**, de tamanho $n \times m$ no passo subsequente da regressão, isto é, **y** é regredido para **T** em vez de **X**. A etapa de

regressão equivale a resolver um sistema de equações em que cada amostra representa uma Equação e cada *t*-score pode ser considerada desconhecida. Consequentemente, o estrito requisito matemático segue que o número de amostras deve exceder o número de t-escores, ou seja, $n > A$. Este requisito é facilmente cumprido na prática. O PCR constrói *t*-escores que descrevem sucessivamente a quantidade máxima de variações em **X** enquanto são ortogonais entre si. Já o PLS pode ser visto como um adicional da PCR porque os dados **y** contribuem para a construção dos *t*-score (FABER; RAJKÓ, 2007; WOLD et al., 1984).

Os métodos de espectro total são geralmente preferidos, uma vez que a etapa de seleção de comprimento de onda adicional necessária para a aplicação de MLR, para garantir que $n > m$, é problemática por si só. Além disso, a compressão em um pequeno número de *t*-escores atua como um filtro eficaz de ruído. PLS é atualmente o método padrão de calibração em Quimiometria, porque muitas vezes foi relatado exibir uma ligeira vantagem sobre a PCR no trabalho aplicado. Em especial, o método PLS tem se tornado uma ferramenta extremamente útil e importante em muitos campos da química, como a físico-química, a química analítica, a química medicinal, ambiental e ainda no controle de inúmeros processos industriais (DE ARAÚJO GOMES et al., 2013; SOUZA et al., 2013). Neste trabalho foi usado o PLS para quantificar o teor do biodiesel em diesel.

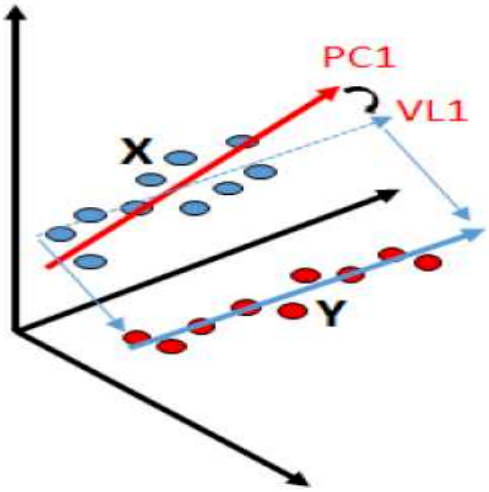
2.5.4. Quadrados Mínimos Parciais – PLS

A regressão por quadrados mínimos parciais (PLS) é o método utilizado na calibração multivariada que permite a construção de modelos de calibração para a previsão da concentração do analito de interesse em amostras de validação. O PLS descreve a situação real do modelo tendo em conta o maior número possível de variações. Neste método, a covariância nas medidas com as concentrações é utilizada em conjunto à variância em **X** para gerar **T** (escores de **X**), tirando vantagem assim da correlação existente entre os dados da matriz original e a variável dependente, **y**.

A decomposição da matriz **X** realizada pela PCA é feita de forma independente da matriz. Enquanto que para o método de regressão PLS a informação de **Y** é incorporada, de forma que cada PC do modelo sofra uma pequena modificação para buscar a máxima covariância entre **X** e **Y** (variável latente). Assim, o modelo PLS é obtido através de um processo iterativo, no qual se otimiza ao mesmo tempo a projeção das amostras sobre o (s) peso (s), para a determinação dos escores, e o ajuste por uma função linear dos escores da matriz **X** aos escores da matriz **Y** de modo a minimizar os desvios. Essa otimização simultânea ocasiona pequenas mudanças nas direções dos pesos, de modo que, rigorosamente

eles perdem a ortogonalidade, levando a pequenas redundâncias de informação. Porém, são essas pequenas redundâncias que otimizam a relação linear entre os escores, e estas distorções da ortogonalidade entre as componentes principais no PLS fazem com que não sejam mais componentes principais (que são ortogonais) e, sim, Variáveis Latentes (VLs), conforme representação observada na Figura 10 (GONTIJO, 2016; VANDEGINSTE et al., 1998).

Figura 10 - Rotação do eixo da PC1 para VL1



Fonte: GONTIJO, 2016.

No PLS, usa-se comumente o termo variável latente (VL) para designar as componentes principais. Isso se deve ao fato da construção das mesmas ser feita a partir de informações contidas no vetor das variáveis dependentes, **y**. O número de variáveis latentes que será utilizado é determinado durante o processo de validação cruzada (GELADI; KOWALSKI, 1986). Para a construção do modelo PLS global,usa-se uma matriz de dados espectrais $\boldsymbol{X}_{(n\times m)}$ e a outra matriz **Y**, definida com *k* colunas das concentrações do analito de interesse e com *n* linhas correspondente ao número de amostras.

O primeiro fator, neste caso chamado de variável latente 1, descreve a direção de máxima variância que também se correlaciona com a concentração. Estas variáveis latentes são na realidade combinações lineares das componentes principais calculados pelo método de Quadrados Mínimos Parciais Iterativos Não Lineares (NIPALS – do inglês, *Non-Iterative Partial Least Squares*), por exemplo. Usando este algoritmo, a matriz de dados espectrais $\boldsymbol{X}_{(n\times m)}$ e a outra matriz **Y**, definida com *k* colunas das concentrações do analito de interesse e com *n* linhas correspondente ao número de amostras são decompostas em escores e pesos. Finalmente, determina-se a correlação entre as variáveis latentes usando as componentes das

matrizes menores (GELADI; KOWALSKI, 1986; LINDGREN et al., 1994). As matrizes das variáveis independentes **X** e das variáveis dependentes **Y** são representadas por escores e pesos, de acordo com as equações 7 e 8:

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E}_X = \sum \boldsymbol{t_h}\, \boldsymbol{p_h^T} + \boldsymbol{E_X} \quad (7)$$

$$\mathbf{Y} = \mathbf{UQ^T} + \mathbf{E}_Y = \sum \boldsymbol{u_h}\, \boldsymbol{q_h^T} + \boldsymbol{E_Y} \quad (8)$$

sendo, **T** e **U** os escores para as duas matrizes de dados, **P** e **Q** os respectivos pesos, *h* é o número de VLs, $\mathbf{E_X}$ e $\mathbf{E_Y}$ são os respectivos resíduos, ou seja, as matrizes que contém a parte não modelada. Então, depois disso, é estabelecida uma nova relação linear entre os escores de **X** e de **Y** para cada VL (*h*), conforme a Equação 9 (MASSART et al., 1998):

$$\boldsymbol{u_h} = \boldsymbol{b_h t_h} + \boldsymbol{E} \quad (9)$$

em que $\mathbf{u_h}$ é uma matriz que contém as concentrações de todas as amostras, $\mathbf{t_h}$ é uma matriz de resposta (o espectro do infravermelho médio, neste caso) para as amostras de calibração, $\boldsymbol{E}$ é uma matriz que representa o ruído, e $\mathbf{b_h}$ é o vetor de coeficientes de regressão do modelo linear para cada LV, obtido através da Equação 10. O melhor modelo PLS deverá minimizar as matrizes de resíduos $\boldsymbol{E_X}$ e $\boldsymbol{E_Y}$ e ao mesmo tempo, obter uma relação linear entre **t** e **u**.

$$\mathbf{b_h} = \frac{\boldsymbol{u_h^T t_h}}{\boldsymbol{t_h^T t_h}} \quad (10)$$

Do mesmo modo como acontece no PCA, a variância explicada pela primeira VL será maior que a variância explicada pela segunda VL e a terceira VL explicará uma variância menor que a segunda VL, e assim sucessivamente até que o número de VLs seja definido e o algoritmo, geralmente, convergido rapidamente (GELADI; KOWALSKI, 1986).

Para a determinação do número correto de VL o método mais utilizado é o de validação cruzada, que se baseia na avaliação da magnitude dos erros de previsão de um dado modelo de calibração. Esta avaliação é feita pela comparação das previsões das concentrações previamente conhecidas. A validação cruzada pode ser realizada por *leave-one-out* (deixe uma de fora por vez), em que uma amostra do conjunto de calibração é deixada de fora no processo de construção do modelo e a seguir essa amostra é prevista pelo modelo construído (FERREIRA et al., 1999), ou ainda em blocos, ou seja, um número determinado de amostras é deixado de fora no processo de construção do modelo e a seguir essas amostras são previstas pelo modelo construído. Este último caso é chamado de *venetian blinds* (venezianas) (BALLABIO; CONSONNI, 2013). Em ambos os casos, o processo é repetido até que todas as amostras tenham sido previstas. Assim, o Erro Quadrático Médio de Validação Cruzada (RMSECV – do inglês, *Root Mean Square Error Cross Validation*) é calculado de acordo com a Equação 11:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n_c}(\hat{y}_{ci} - y_{ci})^2}{n_c}} \quad (11)$$

onde y_{ci} é o valor de medição de referência, $\hat{y}ci$ é o valor previsto para i-ésima amostra e n_c o número total de amostras do conjunto de calibração; este procedimento é repetido no conjunto de calibração, e o número de VL do modelo PLS incluídos no modelo é selecionado de acordo com o menor valor de RMSECV, tendo em conta a porcentagem de variância explicada nos blocos **X** (absorbância) e **y** (concentração).

O método padrão ASTM E1655-05, que relaciona o número de amostras nos conjuntos de calibração e validação, e o número de variáveis latentes usadas no modelo PLS, indica que o número mínimo de amostras para o conjunto de calibração deve ser igual a $6(k+1)$ para dados centrados na média e para o conjunto de previsão deve ser igual a $4k$, onde k é o número de VL (ASTM E1655-05, 2012).

O desempenho do modelo PLS global também é avaliado de acordo com o erro médio quadrático de previsão (RMSEP – do inglês, *Root Mean Square Error of Predition*) e os coeficientes de correlação (R). Os coeficientes de correlação entre os valores previstos e os valores medidos para o conjunto de calibração (R_c) foram calculados de acordo com a Equação 12:

$$R_c = \sqrt{1 - \frac{\sum_{i=1}^{n_c}(\hat{y}_{ci} - y_{ci})^2}{\sum_{i=1}^{n_c}(\hat{y}_{ci} - \bar{y}_{ci})^2}} \quad (12)$$

onde $\bar{y}_{ci}$ é a média de todos os valores de medição de referência de amostras no conjunto de calibração. O valor do RMSEP foi obtido através de amostras do conjunto de validação externa, utilizadas para a validação do modelo de calibração. O RMSEP e o coeficiente de correlação no conjunto de previsão (R_p) foram, respectivamente, calculados pelas Equações 13 e 14:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_p}(\hat{y}_{pi} - y_{pi})^2}{n_p}} \quad (13)$$

$$R_p = \sqrt{1 - \frac{\sum_{i=1}^{n_p}(\hat{y}_{pi} - y_{pi})^2}{\sum_{i=1}^{n_p}(\hat{y}_{pi} - \bar{y}_{pi})^2}} \quad (14)$$

onde n_p é o número de amostras do conjunto de previsão, y_{pi} é o valor de medição de referência da i-ésima amostra no conjunto de previsão e $\hat{y}_{pi}$ é o valor de estimação do modelo da i-ésima amostra no conjunto de previsão e $\bar{y}_{pi}$ é a média de todos os valores de medição de referência de amostras no conjunto de previsão (JIANG et al., 2012).

Os cálculos dos parâmetros acima citados são feitos para o número de VLs de 1, ..., *A* e os resultados são apresentados em um gráfico de RMSECV *vs* VLs. O melhor número de VLs que produziu um baixo erro de validação cruzada sem perda significativa da variância dos dados é escolhido. Mas esse número deve explicar o máximo da variância dos dados, minimizando a influência do ruído. Um modelo que contenha número de VLs inferior ao ideal resultará em subajuste e superior ao ideal em sobreajuste. A decisão sobre um número de VLs superior ao ideal ocorre com mais frequência ocasionando resultados satisfatórios para a previsão da propriedade de amostras do conjunto de calibração, porém, não adequados para a previsão de amostras de um conjunto validação ou previsão devido à informação não necessária incluída no modelo (BABYAK, 2004; FABER; RAJKÓ, 2007).

2.5.4.1. Identificação de Outliers

A detecção de *outliers* (amostras anômalas) é tão importante quanto a determinação do número de VLs que serão empregadas no modelo. Ao verificar a qualidade do conjunto de calibração, deve-se assegurar de que as amostras formam um conjunto homogêneo, removendo-se aquelas amostras que são solitárias. Para a detecção de *outliers*, usa-se duas grandezas complementares, que são *leverage* e resíduos de *Student*.

A *leverage* é uma medida da influência de uma amostra no modelo de regressão. Um valor de *leverage* pequeno indica que a amostra em questão influencia pouco na construção do modelo de calibração. Por outro lado, se as medidas experimentais de uma amostra são diferentes das outras do conjunto de calibração, ela provavelmente terá uma alta influência no modelo, que pode ser negativa. Em geral, estas amostras solitárias estão visíveis no gráfico de escores. A *leverage* pode ser interpretada geometricamente como a distância de uma amostra ao centro do conjunto de dados e é calculada segundo a Equação 15 (FABER; RAJKÓ, 2007; FERREIRA et al., 1999):

$$\boldsymbol{h_i = t_{A,i}^T \left(T_A^T T\right)^{-1} t_{A,i}} \tag{15}$$

onde $\boldsymbol{T}$ é a matriz de escores de amostras de calibração, $\mathbf{t}_i$ é o vetor de score da amostra *i*, e *A* é o número de VLs. Outra maneira bastante simples de se calcular *leverage* é medindo a distância de uma amostra ao centro do conjunto através do cálculo da distância Euclidiana no espaço das componentes principais, usando a matriz dos escores:

$$\boldsymbol{H = TT^T} \tag{16}$$

Os elementos da diagonal de $\boldsymbol{H}$, $\boldsymbol{h_i}$, estão diretamente relacionados com os valores da *leverage*. O valor limite dos dados de calibração centrados na média é dado pela Equação 17 (MARTENS; NAES, 1989).

$$h_{limit} = \frac{3(A+1)}{n_c} \tag{17}$$

em que A é o número de VLs. De acordo com ASTM standard, amostras com altos valores de *leverage* que o valor limite para dados centrados na média devem ser eliminados do conjunto de calibração e o modelo deve ser reconstruído (ASTM E1655-05, 2012). As amostras com $h_i > h_{limit}$ são consideradas "outliers".

A detecção de outliers baseada em resíduos não modelados em dados espectrais é feita através da comparação entre o desvio padrão total (s(e)) com o desvio padrão de uma amostra individual (s(e_i)), que se definem assim (WAHL et al., 2002):

$$s(e) = \sqrt{\frac{1}{v}\sum_{i=1}^{n_c}\left(\sum_{j=1}^{J}\left(x_{ij} - \hat{x}_{ij}\right)^2\right)}, s(e_1) = \sqrt{\frac{1}{v}\sum_{j=1}^{J}\left(x_{ij} - \hat{x}_{ij}\right)^2} \tag{18}$$

onde J é o número de variáveis espectrais, x_{ij} é o valor da absorvância da amostra i no comprimento de onda j, $\hat{x}_{ij}$ é o valor estimado da absorvância com A VLs. O número de graus de liberdade é dado por $v = n_c J - J - A(\max(n_c, J))$. Se a amostra tiver s($e_i$) >$n$s(e), quando n é uma constante que possa estar entre 2 a 3, ela é removida do conjunto de calibração (MARTENS; NAES, 1989).

Por fim, os *outliers* também podem ser detectados através dos resíduos não modelados em variáveis dependentes (conjunto de previsão), quando o conjunto de calibração já estiver construído. Esta detecção baseia-se na comparação do RMSEC (do inglês, *Root Mean Square Error of Calibration*) do modelo com o erro absoluto de amostras individuais. Se uma amostra apresentar uma diferença entre seu valor de referência (y_i) e o valor estimado ($\hat{y}_i$) maior que três vezes o RMSEC, ela é identificada como *outlier* e deve ser eliminada (MARTENS; NAES, 1989). O valor do RMSEC e o número de graus de liberdade (v) usados neste trabalho foram assim determinados:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_c}(y_i - \hat{y}_i)^2}{v}}, \qquad v = n_c - n_c\left(1 - \sqrt{\frac{MSEC}{MSECV}}\right) \tag{19}$$

Supondo-se que os resíduos de *Student* são normalmente distribuídos pode-se aplicar um teste-*t* como indicativo, para verificar se a amostra está ou não dentro da distribuição com um nível de confiança de 95%. Como os resíduos de *Student* são definidos em unidades de desvio padrão do valor médio, os valores além de ±2,5 são considerados altos sob as condições usuais da estatística.

A análise do gráfico dos resíduos de *Student vs leverage* para cada amostra é a melhor maneira de se determinar as amostras anômalas. Amostras com altos resíduos, mas com pequena *leverage* provavelmente têm algum erro no valor da concentração, que deve, de preferência, ser medida novamente. Outra opção será a exclusão de tal amostra do conjunto de calibração. Amostras com resíduo e *leverage* altos, simultaneamente, devem sempre ser excluídas e o modelo de calibração deve ser reconstruído (FERREIRA et al., 1999).

2.5.5. Métodos de Seleção de Variáveis por Intervalos

Como nem todo o sinal espectral contém informações úteis para a construção de um determinado modelo de regressão multivariado, visando melhorar o desempenho, a complexidade e robustez dos métodos de calibração, alguns procedimentos utilizam a seleção de regiões espectrais que associadas à propriedade em questão resultam em um modelo mais eficiente, evitando que variáveis desnecessárias interfiram na modelagem. Por isso, depois da construção do modelo geral de calibração multivariada, PLS global, em seguida, são construídos por meio de regressão multivariada modelos de métodos de seleção de variáveis por intervalos, nomeadamente quadrados mínimos parciais por intervalos (iPLS), quadrados mínimos parciais por exclusão (biPLS) e quadrados mínimos parciais por sinergismo (siPLS), usando regiões espectrais com variáveis importantes para o modelo.

2.5.5.1. Quadrados Mínimos Parciais por Intervalos – iPLS

O modelo do iPLS foi desenvolvido por (NØRGAARD et al., 2000). Ele consiste em dividir os espectros em subintervalos menores com igual peso. Posteriormente, são desenvolvidos modelos de regressão PLS para cada um dos subintervalos, usando o mesmo número de variáveis latentes do PLS global. A previsão de cada modelo local é comparada com o modelo PLS global através do parâmetro da validação cruzada, previsão, coeficiente de correlação e inclinação da reta do gráfico dos valores reais e previstos pelo modelo. A região com menor erro de validação cruzada é escolhida. As amostras *outliers* detectadas também são removidas antes da construção do modelo iPLS. Uma das principais vantagens deste método é a possibilidade de representar um modelo de regressão local em uma orientação gráfica, focando na escolha de melhores intervalos e permitindo uma comparação entre os modelos dos intervalos e o modelo do PLS global.

2.5.5.2. Quadrados Mínimos Parciais por Exclusão – biPLS

O método que se usa para a construção do modelo biPLS é o mesma do iPLS (NØRGAARD et al., 2000). Ele consiste na divisão dos espectros de amostras em intervalos equidistantes, mas agora, os modelos locais PLS são construídos a partir de um intervalo ou combinações de intervalos que apresentam uma melhor correlação entre as variáveis selecionadas e a resposta de interesse, quando comparado ao modelo PLS global. Os intervalos que apresentam maiores valores do RMSECV são eliminados e um novo modelo chamado biPLS é construído com outros intervalos. Assim, é obtido um valor de RMSECV para os restantes intervalos usados, o qual é também comparado com o modelo PLS global.

2.5.5.3. Quadrados Mínimos Parciais por Sinergia – siPLS

Este algoritmo é também uma expansão do iPLS (NØRGAARD et al., 2000). Primeiro, os espectros totais das amostras são divididos em vários subintervalos equidistantes; depois os modelos de regressão PLS são construídos através da permutação e combinação de todos os subintervalos possíveis nos quais os espectros foram divididos; a cada uma das combinações de intervalos obtém-se uma estimativa do valor de RMSECV, e a combinação que fornece menor valor de RMSECV é escolhida e é desenvolvido um novo modelo ideal chamado siPLS, o qual é também posteriormente comparado com o modelo PLS global.

2.5.5.4. Validação Analítica dos Modelos PLS

A validação analítica dos modelos PLS é feita através das figuras de mérito, a saber: seletividade, sensibilidade, sensibilidade analítica, limite de detecção, limite de quantificação, teste para erro sistemático (bias e t_{bias}) e SDV. O sinal analítico líquido (NAS – do inglês, *Net Analyte Signal*) é muito importante para determinar as figuras de mérito. O procedimento usado neste trabalho para calcular o NAS é o proposto por Lorber, e está descrito na literatura (LORBER, 1986; VALDERRAMA; BRAGA; POPPI, 2009). O sinal analítico líquido para o analito k (r_k^*) é definido como a parte do espectro que é ortogonal ao subespaço ocupado pelos espectros de todos os outros analitos e é dada pela seguinte Equação 20:

$$\mathbf{r}_k^* = \mathbf{P}_{\text{NAS},k}\boldsymbol{r} = [\mathbf{I} - \mathbf{R}_{-k}(\mathbf{R}_{-k})^+]\boldsymbol{r} \qquad (20)$$

onde $\mathbf{P}_{\mathrm{NAS},k}$ é a matriz de projeção ortogonal $\boldsymbol{J} \times \boldsymbol{J}$ que se projeta de um dado vetor sobre o espaço NAS, $\mathbf{r}$ é o espectro puro de uma dada amostra com *k* unidades de concentração do analito de interesse (por isso, a Equação anterior também pode ser escrita na forma $\mathbf{s}_k^*=\mathbf{P}_{\mathrm{NAS},k}\boldsymbol{s_k}$; $\boldsymbol{I}$ é uma matriz unitária $\boldsymbol{J} \times \boldsymbol{J}$; $\mathbf{R}_{-k}$ são espectros do branco (do diesel puro, neste caso) que fornecem todas as informações relacionadas com o interferente dentro das especificações de qualidade, organizados numa matriz de dimensão $j \times I$ com $\boldsymbol{I}$ espectros e *j* variáveis; $\boldsymbol{R}_{-\boldsymbol{k}}^{+}$ é o pseudo-inverso de Moore-Penrose de $\mathbf{R}_{-k}$, geralmente calculado pela decomposição de valores singulares usando fatores de *A*. Finalmente, a concentração de *k* na amostra desconhecida ($\mathbf{Y}_{\mathrm{un},k}$) a partir do espectro $\mathbf{r}$ é obtida por meio da Equação 21:

$$\mathbf{Y}_{\mathrm{un},k} = \frac{\mathbf{S}_k^{\mathrm{T}}\mathbf{P}_{\mathrm{NAS},k}\mathbf{r}}{\mathbf{S}_k^{\mathrm{T}}\mathbf{P}_{\mathrm{NAS},k}\mathbf{s}_k} = \frac{\mathbf{S}_k^{\mathrm{T}}\mathbf{P}_{\mathrm{NAS},k}\mathbf{P}_{\mathrm{NAS},k}\mathbf{r}}{\mathbf{S}_k^{\mathrm{T}}\mathbf{P}_{\mathrm{NAS},k}\mathbf{P}_{\mathrm{NAS},k}\mathbf{s}_k} = \frac{(\mathbf{S}_k^*)^{\mathrm{T}}\mathbf{r}_k^*}{\|\mathbf{S}_k^*\|^2} \qquad (21)$$

em que "|| ||" indica a norma euclidiana de um vetor; $\boldsymbol{S_k}$ é o espectro contendo o analito de interesse *k* em unidades de concentração e $\mathbf{S}_k^*$ é o seu correspondente NAS.

Precisão: A precisão pode ser expressa de maneira semelhante aos métodos univariados. Esta figura expressa o grau de concordância entre os resultados de uma série de medidas realizadas para uma mesma amostra homogênea em condições determinadas. Ela é calculada por uma estimativa do desvio padrão absoluto. Em calibração multivariada, a precisão do método é considerada em três níveis que são: a repetibilidade, precisão intermediária e reprodutibilidade.

Seletividade: O parâmetro de seletividade (sel) é a medida do grau de sobreposição entre o sinal da espécie de interesse e os interferentes presentes na amostra. Ela indica a parte do sinal total do espectro que é perdida devido a essa sobreposição espectral e pode ser definido no contexto multivariado usando os cálculos do NAS.

Sensibilidade: A sensibilidade (sens) corresponde à fração do sinal responsável pelo acréscimo de uma unidade de concentração à propriedade de interesse. Neste trabalho, a sensibilidade foi usada para medir a alteração da concentração do biodiesel em diesel.

Sensibilidade analítica: A sensibilidade analítica (γ) é outra Figura de mérito mais informativa, que apresenta a sensibilidade do método em termos da unidade de concentração que é utilizada, sendo definida como a razão entre a sensibilidade e o desvio padrão do sinal de referência. A diferença mínima de concentração, que é estatisticamente discernível por um método, pode ser expressa como inverso da sensibilidade analítica (γ^{-1}). Assim, o inverso da sensibilidade analítica estabelece uma diferença de concentração mínima que é

perceptível pelo método analítico, considerando o ruído experimental aleatório como a única fonte de erro, independentemente da técnica específica empregada.

Limite de detecção: O parâmetro de limite de detecção (LD) é o valor mínimo da concentração da espécie de interesse que pode ser detectada pelo método, mas não necessariamente quantificada. A estimativa para o ruído instrumental é obtida através de 9 espectros replicados do (s) interferente (s) puro (s) (sem o analito de interesse), seguindo as recomendações da IUPAC.

Limite de quantificação: O limite de quantificação (LQ) expressa o menor valor da concentração do analito de interesse que pode ser medida com uma incerteza máxima de 10%.

Linearidade: A qualidade da linearidade do modelo é estimada através dos *plots* dos valores dos resíduos e escores *vs* os valores de referência que devem apresentar comportamentos aleatórios e lineares, respectivamente. Qualitativamente, o gráfico dos resíduos para as amostras de calibração e de validação indica se os dados seguem um comportamento linear se a distribuição destes resíduos for aleatória. O ajuste do modelo, que está também ligada a linearidade, é estimado a partir da correlação entre os valores de referência e os valores estimados pelo modelo de calibração multivariada, para a propriedade de interesse.

Erros Sistemáticos (bias) e desvio padrão de validação: Os valores do *bias*, que se referem aos erros sistemáticos, são calculados a partir da diferença entre a média e os valores de referência, eles são os componentes de erro não aleatórios. Os valores de *bias* e do desvio padrão dos erros de validação (SDV – do inglês, *Standard Deviation Validation*) são calculados usando os dados das amostras de previsão. Posteriormente, de acordo com o padrão ASTM E1655-05, na abordagem deste parâmetro, deve-se usar um teste-*t* para amostras de validação com 95% de confiança e graus de liberdade iguais ao número de amostras de predição ($n_p - 1$). Este teste, t_{bias}, é aplicado com o objetivo de avaliar a existência de erro sistemático significativo para cada modelo.

Caso o valor do t_{bias} apresentar resultado maior do que o valor do $t_{crítico}$ para *n* graus de liberdade a 95% de confiança, isso é uma evidência de que erros sistemáticos presentes no modelo multivariado são significativos. No entanto, se o valor de t_{bias} apresentar valor menor do que $t_{crítico}$, então, os erros sistemáticos no modelo são desprezíveis (ASTM E1655-05, 2012).

Exatidão: É expressa pelos valores de RMSECV (Equação 11), RMSEP (Equação 13) e RMSEC (Equação 19). Entretanto estes valores são os parâmetros globais que incorporam tanto os erros sistemáticos quanto os aleatórios. Na comparação de dois

resultados entre os dois métodos distintos, a exatidão pode ser acessada por meio da comparação dos valores obtidos para a inclinação e o intercepto de uma reta ajustada entre os valores de referência e os estimados pelo modelo. Isso pode ser melhor observado através da elipse de confiança. Caso os intervalos de confiança para esses parâmetros contenham os seus valores esperados iguais a 1 e 0, para a inclinação e intercepto, respectivamente, os valores podem ser considerados estatisticamente equivalentes no nível de confiança considerado. Todavia, valores estimados para a inclinação e intercepto fora de seus intervalos de confiança indicam erros sistemáticos proporcionais e constantes, respectivamente. Por isso, neste trabalho foi usada para comparar os valores estimados e os previstos pelos modelos. Os cálculos dos valores previstos e reais da inclinação e intercepto para os modelos ideais e reais são descritos na literatura (GONZÁLEZ; HERRADOR; ASUERO, 1999). As equações usadas para o cálculo das figuras de mérito estão apresentadas na Tabela 2.

Tabela 2 – Equações de Figuras de Mérito aplicadas a validação multivariada dos modelos PLS

Figura de mérito	**Equação**	**Parâmetro**
1. Precisão	$\text{Precisão} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{y}_{ij} - \hat{y}_i)^2}{n(m-1)}}$	
2. Limite de detecção (%, v/v)	$\text{LD} = 3.3\delta_x \frac{1}{\text{sens}}$	
3. Limite de quantificação (%, v/v)	$\text{LQ} = 10\delta_x \frac{1}{\text{sens}}$	
4. Seletividade (%)	$sel = \frac{\|s_k^*\|}{\|s_k\|}$	
5. Sensibilidade% (v/v)	$sens = \frac{1}{\|\boldsymbol{b_k}\|}$	
6. Sensibilidade Analítica (%, v/v)	$\gamma = \frac{\text{sens}}{\|\delta_x\|}$	
7. Inverso da Sensibilidade Analítica (%, v/v)$^{-1}$	$\gamma^{-1} = \frac{1}{\gamma}$	
8. Testes para Erros Sistemáticos	$\text{bias} = \frac{\sum_{i=1}^{n_p}(y_{pi} - \hat{y}_{pi})}{n_p}$	Bias
	$\text{SDV} = \sqrt{\frac{\sum_{i=1}^{n_p}[(y_{pi} - \hat{y}_{pi}) - \text{bias}]^2}{n_p - 1}}$	Desvio padrão Graus de liberdade
	$t_{bias} = \frac{\|\text{bias}\|\sqrt{n_p}}{\text{SDV}}$	$t_{calculado}$ $t_{crítico}$

Fonte: Adaptado de VALDERRAMA, 2009.

em que $\boldsymbol{b_k}$ é o vector de regressão NAS, o qual pode ser obtido por qualquer método multivariado; o coeficiente 3,3 no cálculo do LD corresponde a 5% como probabilidades de falsos positivos e falsos negativos; $\|\delta_x\|$ é uma estimativa para o ruído instrumental; y_i

corresponde ao valor de referência para a amostra *i* e *n* é a quantidade de amostras e *m* é o número de replicadas usadas para o cálculo da precisão.

As figuras de mérito calculadas neste trabalho estão de acordo com normas específicas, como a ASTM, para validação em calibração multivariada a partir de espectrometria no infravermelho, e trabalhos científicos divulgados por meio de periódicos.

2.5.5.5. Comparação do Modelo PLS Global e de Seleção de Variáveis por Intervalos

Os valores do RMSEP obtidos por subconjunto de validação externa são empregados para avaliar o desempenho dos modelos iPLS, biPLS e siPLS em relação ao modelo PLS global. A significância estatística das diferenças entre os valores de RMSEP é avaliado usando um teste *F* a um nível de confiança de 95%. Os valores de *F* são calculados como a razão dos quadrados do maior e dos menores valores de RMSEP, conforme a Equação 22:

$$F_{calculado} = \left(\frac{RMSEP_1}{RMSEP_2}\right)^2 \quad (22)$$

onde $RMSEP_1$ é do PLS global e $RMSEP_2$ do modelo de seleção de variáveis por intervalos, sendo $RMSEP_1 > RMSEP_2$. O valor de $F_{\text{calculado}}$ para cada modelo é comparado com o valor da distribuição da estatística *F* (F_{tabelado}) com graus de liberdade igual ao número de amostras de previsão e um nível de significância de 5%. Nos modelos em que o valor de F_{tabelado} é inferior ao $F_{\text{calculado}}$, concluiu-se que não há evidência estatística de que os valores seguem a distribuição normal e, portanto, o método com $RMSEP_2$ apresenta melhor exatidão em relação ao PLS global (MARK; WORKMAN, 2003).

2.5.6. Métodos de Classificação não Supervisionada ou Conhecimento de Padrões

A Modelagem Independente e Suave por Analogia de Classe (SIMCA – do inglês *Soft Independent Modelling of Class Analogy*), Regra do Vizinho mais Próximo (KNN – do inglês, *K-Nearest Neighbor*), Análise Discriminante por Quadrados Mínimos Parciais (PLS-DA, do inglês *Partial Least Squares Discriminant Analysis*), Máquina de Aprendizagem Linear (LLM – do inglês,*Linear Learning Machine*), Análise Discriminante Linear (LDA – do inglês, *Linear Discriminant Analysis*) Cartas de Controle Multivariadas, entre outros, são métodos supervisionados, pois supervisionam o desenvolvimento dos critérios de discriminação das amostras. Nestes métodos, devem-se saber *a priori* quais amostras são semelhantes e quais são diferentes para encontrar os critérios de classificação. Os métodos

são usados para identificar as semelhanças e diferenças em diferentes tipos de amostras, comparando-as entre si. Eles fundamentam-se nas seguintes suposições:

- Amostras do mesmo tipo são semelhantes;
- Existem diferenças entre diferentes tipos de amostras;
- As semelhanças e diferenças são refletidas nas medidas utilizadas para caracterizar amostras.

O reconhecimento de padrões e sua área complementar de conhecimento de padrões são ramos da inteligência artificial. Se a distribuição da probabilidade estatística dos dados referentes a uma amostra é conhecida, pode-se trabalhar no modo parametrizado. O mais comum em problemas químicos é que essa distribuição não seja conhecida, o que exige a adição do reconhecimento e conhecimento de padrões não parametrizados. Havendo disponibilidade de um conjunto de treinamento, ou seja, um conjunto para o qual se conhece a categoria a que pertence cada amostra, é possível derivar regras de classificação, com base em medidas de variáveis relativas a cada espécie. A seguir, a partir de medidas relativas às amostras desconhecidas, pode-se classificá-las pelas regras estabelecidas ou a partir do conjunto de treinamento (BARKER; RAYENS, 2003; FERREIRA, 2015).

Métodos de classificação multivariada são métodos quimiométricos destinados a encontrar modelos matemáticos capazes de reconhecer a participação de cada amostra em sua classe apropriada, com base em um conjunto de medidas. Uma vez calibrado o modelo de classificação, é possível prever a associação de amostras desconhecidas a uma das classes definidas. Portanto, as técnicas de classificação (também conhecidas como reconhecimento supervisionado de padrões) lidam com respostas qualitativas, ou seja, definem relações matemáticas entre um conjunto de variáveis descritivas (por exemplo, medições químicas) e uma variável qualitativa, ou seja , a associação a uma categoria definida (BALLABIO; CONSONNI, 2013).

2.5.7. Análise Discriminante por Quadrados Mínimos Parciais – PLS-DA

A Análise Discriminante por Quadrados Mínimos Parciais (PLS-DA) é um método de classificação linear que combina as propriedades da regressão de Quadrados Mínimos Parciais (PLS) com o poder de discriminação de uma técnica de classificação. A PLS-DA baseia-se no algoritmo de regressão PLS, o qual já foi discutido nos subcapítulos anteriores. A principal vantagem do PLS-DA é que as fontes relevantes de variabilidade dos dados são modeladas pelas chamadas variáveis latentes, que são combinações lineares das variáveis originais e, consequentemente, isso permite a visualização gráfica e a compreensão das diferentes padrões e

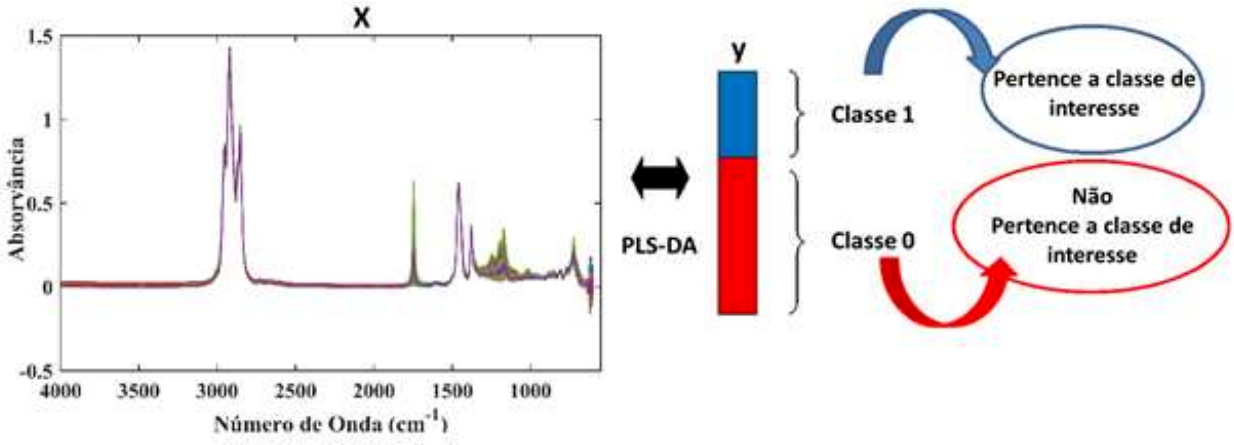
relações de dados por escores e pesos de VL. Os pesos são os coeficientes das variáveis nas combinações lineares que determinam os VLs e, portanto, podem ser interpretadas como a influência de cada variável em cada VL, enquanto escores representam as coordenadas das amostras no hiperespaço de projeção da VL.

Ao lidar com a PLS-DA, o vetor de classe (contendo conjuntos de amostras às *G* classes) é transformado em uma matriz fictícia **Y**, com *n* linhas (amostras) e *G* colunas (as informações da classe). Cada entrada y_{ig} de **Y** representa o conjunto da *i*-ésima amostra à *g*-ésima classe expressa com um código binário (1 ou 0). Portanto, o vetor de classe *n*- dimensional é transformado em uma matriz **Y** binária constituída por *n* linhas e *G* colunas (BARKER; RAYENS, 2003).

Existem duas variantes da PLS-DA, a PLS1-DA e a PLS2-DA, na primeira cada coluna de **Y** é modelada individualmente, caso se tenha 3 classes "a, b, c" são construídos 3 modelos, o primeiro modelo será construído utilizando o valor de **y** igual a 1 para a classe "a" e 0 para as demais classes, o segundo modelo será construído utilizando o valor 1 para a classe "b" e zero para as demais classes, o mesmo raciocino é empregado para a próxima classe. Já na PLS2-DA é calculado um único conjunto de escores e pesos para todas as colunas da matriz **Y**.

A Figura 11 ilustra como é feita a organização dos dados utilizados para a construção do modelo de classificação por PLS-DA. A matriz **X** é composta pelos espectros, onde as linhas representam as amostras e as colunas representam às absorbâncias para cada número de onda e o vetor **y** é construído com valores de 1 ou 0.

Figura 11 – Esquema da organização dos dados para a construção do modelo de classificação usando PLS-DA



Fonte: O autor.

Para fazer uma atribuição de classe, a probabilidade de uma amostra pertencer a uma classe específica pode ser calculada com base nos valores estimados da classe. Portanto, uma probabilidade é calculada para cada classe e a classificação das amostras é realizada escolhendo-se a classe que tem a maior probabilidade. Sob essa abordagem, as amostras são sempre classificadas em uma das classes. Matematicamente, empregando o algoritmo NIPALS, a decomposição da matriz **X** e do vetor **y** são realizadas conforme as Equações 6, 23 e 24, onde **T** são os escores, **P** e **q** os pesos, **W** são os pesos do PLS e $\mathbf{E_X}$ e $\mathbf{E_y}$ são os resíduos de **X** e **y** respectivamente (BALLABIO; CONSONNI, 2013).

$$\mathbf{X} = \mathbf{y}\mathbf{W}^T + \mathbf{E_X} \quad (23)$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^T + \boldsymbol{E}_\mathbf{y} \quad (24)$$

Os pesos **W** são calculados proporcionalmente a covariância entre os blocos **X** e **y**, conforme representado na Equação 25. Em seguida, estes pesos são normalizados utilizando a Equação 26.

$$\mathbf{W} = \mathbf{X}^T\mathbf{y}(\mathbf{y}^T\mathbf{y})^{-1} \quad (25)$$

$$\mathbf{W} = \frac{\mathbf{W}}{\sqrt{\mathbf{W}^T\mathbf{W}}} \quad (26)$$

A matriz de escores **T** é estimada pela combinação linear de **X** com a matriz de pesos **W** do PLS (Equação 27). Os escores são calculados simultaneamente com decomposição da matriz **X** e do vetor **y**, reunindo os vetores **W** que denotam as direções das variáveis latentes (GELADI; KOWALSKI, 1986).

$$\mathbf{T} = \mathbf{X}\mathbf{W} \quad (27)$$

Após o cálculo dos pesos **W** e dos escores **T**, são calculados os pesos **P** e $\boldsymbol{q}$, conforme representado nas Equações 28 e 29 respectivamente.

$$\mathbf{P} = \mathbf{X}^T\mathbf{T}\,(\mathbf{T}^T\mathbf{T})^{-1} \quad (28)$$

$$\boldsymbol{q} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y} \quad (29)$$

Em seguida é calculado a matriz de resíduos $\mathbf{E_X}$ e $\mathbf{E_y}$ e os valores de **X** e **y** são atualizados para o cálculo da próxima variável. Ao final do cálculo os coeficientes de regressão do modelo PLS são estimados através da Equação 30 e os valores obtidos de $\widehat{\boldsymbol{y}}$ previsto são estimados através da Equação 31.

$$\mathbf{b}_{PLS} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}q \quad (30)$$

$$\widehat{\boldsymbol{y}} = \mathbf{X}\mathbf{b}_{PLS} \quad (31)$$

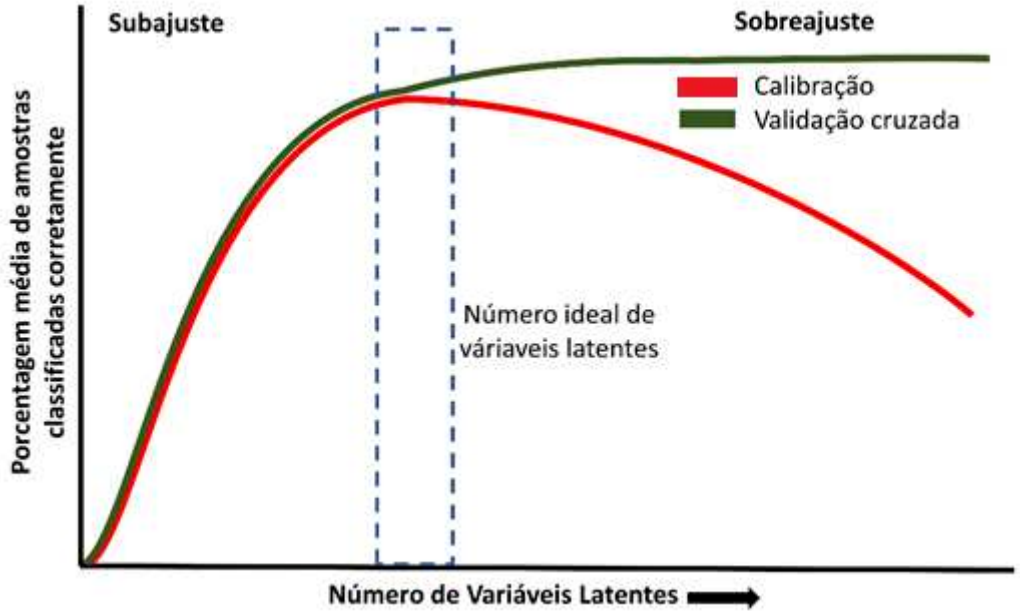
Assim como no PLS, uma das etapas principais para construir o modelo PLS-DA é a escolha correta do número de variáveis latentes, pois quando é escolhido um baixo número de variáveis latentes temos a falta de ajuste do modelo, ou seja, não são utilizadas todas as

informações úteis para construir o modelo. Todavia quando é escolhido um alto número de variáveis latentes tem-se o sobreajuste do modelo, onde são utilizadas informações não relacionadas a propriedade de interesse como por exemplo ruídos espectrais para a construção do modelo (BABYAK, 2004; FABER; RAJKÓ, 2007). Para selecionar o número ideal de VLs nos modelos PLS-DA, as amostras são geralmente divididas em grupos de validação cruzada, usando um dos procedimentos: venezianas ou blocos contínuos. A escolha do tipo adequado de grupos de validação cruzada depende basicamente de como as classes são distribuídas entre as amostras (BALLABIO; CONSONNI, 2013).

A prática comum é executar a validação por *leave-one-out*. No entanto, o *leave-one-out* geralmente superestima o poder preditivo de um modelo e, portanto, não pode retornar um número confiável de variáveis latentes. Consequentemente, boas estimativas fornecidas pelo *leave-one-out* parecem ser necessárias, mas não suficientes para ter um alto poder preditivo. Quando o conjunto de dados compreende algumas amostras, um número maior de grupos pode ser preferido para obter a maioria das amostras no conjunto de treinamento e algumas amostras selecionadas em cada grupo de validação; assim, uma pequena perturbação é produzida no modelo. Por outro lado, se o número de amostras no conjunto de dados for relativamente alto, é recomendável selecionar um pequeno número de grupos de validação cruzada para obter mais amostras selecionadas para testar o modelo e evitar a superestimação da capacidade preditiva do modelo (GOLBRAIKH; TROPSHA, 2002). Daí, o método de validação cruzada por blocos ganha vantagem.

Enquanto no PLS analisa-se os valores de RMSECV para definir o número ideal de variáveis latentes, no PLS-DA analisa-se a porcentagem de amostras classificadas corretamente em cada classe na validação cruzada. A Figura 12 ilustra de maneira genérica a região ideal para a escolha do número de variáveis latentes, de acordo com a porcentagem média de amostras classificadas corretamente.

Figura 12 – Representação genérica da escolha do número de variáveis latentes de acordos com a porcentagem média de amostras classificadas corretamente

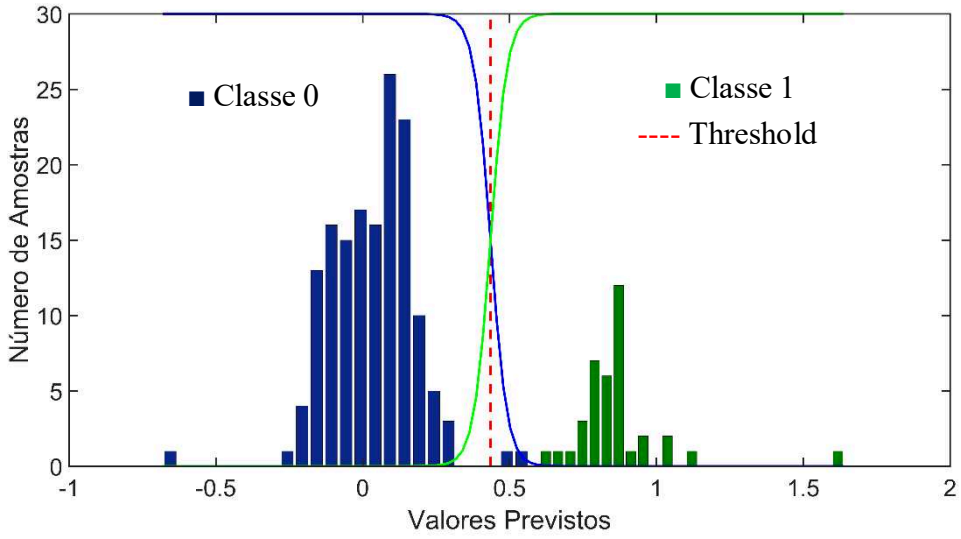


Fonte: O autor.

Os resultados obtidos pelo vetor $\hat{\boldsymbol{y}}$ (Equação 31) contêm a previsão da classe que pertence cada amostra. No entanto, devido às variações das amostras dentro de uma mesma classe, na prática os valores previstos não são exatamente 0 ou 1. Dessa forma é necessário estabelecer o valor limite (limiar/*threshold*) entre as classes empregando estatística bayesiana (BOQUÉ; FABER; RIUS, 2000).

Nesse caso, considera-se que os valores de $\hat{\boldsymbol{y}}$ previstos para cada classe (classe 0 e 1) seguem uma distribuição normal. O valor limite entre as classes é selecionado no ponto em que as duas distribuições se cruzam, como mostrado por linha tracejada a vermelha na Figura 13. Esse valor é estabelecido de forma a minimizar o número de resultados falsos positivos e falsos negativos. Ainda nesta Figura, as barras azuis (■) representam os valores de $\hat{\boldsymbol{y}}$ previstos *vs* o número de amostras da classe 0 e as barras verdes (■) representam os valores de $\hat{\boldsymbol{y}}$ previstos *vs* o número de amostras da classe 1.

Figura 13 – Distribuição dos valores de $\hat{y}$ previstos *vs* o número de amostras para determinação do valor limite entre as classes seguindo estatística bayesiana



Fonte: O autor.

Pode acontecer que os valores de classe estimados para uma amostra sejam maiores ou menores que todos os limites de classe definidos. Assim, se o valor for maior, a amostra seria atribuída a todas as classes e, consequentemente, seria uma amostra confusa. Por outro lado, se o valor for menor, a amostra não seria reconhecida como membro de nenhuma classe. Nos dois casos, a amostra pode ser definida como "não pertencente a nenhuma classe" (BALLABIO; CONSONNI, 2013).

Depois da construção dos modelos PLS-DA, faz-se a análise da Variável Importante na Projeção (VIP) com intuito de identificar quais variáveis dos espectros MIR foram mais relevantes para separação das classes das amostras, como também as que foram utilizadas em métodos de seleção de variáveis visando aumentar a eficiência do modelo ao eliminar variáveis não informativas. As variáveis com escores maiores que 1 são consideradas as principais variáveis do modelo. O parâmetro VIP expressa a importância de cada variável *j*, sendo refletida por *w* em cada variável latente (WOLD; SJOSTROM; ERIKSSON, 2002), conforme definido pela Equação 32:

$$\vartheta_j = \sqrt{\frac{p\sum_{a=1}^{A}[(q^2 at' ata)(waj/||wa||)^2]}{\sum_{a=1}^{A}(q^2 at' ata)}} \tag{32}$$

onde: ϑ_j é a média de VIP; q são os pesos de **Y**; **t** são os escores; $(waj/||wa||)^2$ representa a importância da variável *j-ésima*; $(q^2 at' ata)$ é a variância explicada por cada VL.Variável j pode ser eliminada se $\vartheta_j < u$ para alguns valores definidos do *threshold* $u \in [0, \infty)$. É geralmente aceite que uma variável seja selecionada se ϑ_j>1, mas um limite adequado do *threshold* entre 0,83 e 1,21 pode produzir variáveis mais relevantes (CHONG; JUN, 2005).

2.5.7.1. Análise Discriminante por Quadrados Mínimos Parciais de Intervalos – iPLS-DA

A PLS-DA também não é adequada para problemas de várias classes agrupadas ao longo de um eixo subjacente, uma vez que um plano discriminante não pode ser encontrado. Nesse caso, modelos PLS-DA locais podem representar uma melhor escolha para lidar com essas condições, pois a eficiência de um modelo de classificação combina todas as informações fornecidas pelas figuras de mérito, principalmente pela sensibilidade e especificidade. Quando o método é muito sensível para positivo gera muitos falsos positivos, e vice-versa. Dessa forma, para melhorar os resultados do modelo PLS-DA, faz-se a seleção de variáveis importantes por intervalos para desenvolver modelos chamados iPLS-DA. Os espectros totais das amostras são divididos em intervalos equidistantes através do algoritmo iPLS e então um modelo PLS-DA é desenvolvido para cada um dos subintervalos. Finalmente, o intervalo com os melhores resultados de classificação selecionado para construir o modelo de classificação.

Métodos baseados em intervalo, como iPLS, dividem o conjunto de dados em um número de intervalos. Existem muitas vantagens de usar métodos de intervalos: (*i*) uso de menos variáveis; (*ii*) menos interferências; (*iii*) obtenção de modelos mais locais; e (*iv*) métodos de intervalo geralmente levam a desempenho e interpretação aprimorados do modelo. De fato, uma das principais vantagens do uso dos métodos de classificação baseada em intervalos é que os resultados finais da classificação por iPLS-DA são efetivamente resumidos em um gráfico, fornecendo informações sobre o desempenho da classificação em função do intervalo espectral, quando comparado com o modelo de espectros totais, PLS-DA global (RINNAN; SAVORANI; ENGELSEN, 2018).

2.5.7.2 Validação Analítica dos Modelos PLS-DA

O desempenho do modelo de classificação é avaliado através da Tabela de contingência que mostra o número de amostras previstas em cada classe e seus respectivos valores de referência. A Tabela de contingência também é conhecida como Tabela de confusão. Uma amostra verdadeira positiva (VP) é aquela que pertence a classe 1 (classe de interesse) e que foi corretamente classificada pelo modelo como uma amostra que pertence mesmo a classe 1. Amostra falsa positiva (FP) é aquela que pertence a classe 0, mas que foi prevista como pertencente a classe 1. Uma amostra é chamada verdadeira negativa (VN) é aquela amostra da classe 0 e que foi corretamente classificada na sua respetiva classe pelo

modelo, e amostra falsa negativa (FN) é aquela que pertence a classe 1, mas que foi erradamente prevista na classe 0 pelo modelo.

Da Tabela de contingência é possível extrair as figuras de méritos utilizadas como parâmetros para avaliar o desempenho de modelos de classificação, como a sensibilidade, especificidade, acurácia e Coeficiente de Correlação de Matthews (CCM) (BALLABIO; CONSONNI, 2013). Estes parâmetros diferem dos parâmetros empregados em química analítica quantitativa e metrologia. Em problemas de classificação a sensibilidade é a habilidade de o modelo classificar corretamente as amostras positivas e é calculada empregando a Equação 33.

$$Sensibilidade = \frac{VP}{VP+FN} \quad (33)$$

A especificidade (ou seletividade) é capacidade de identificar ou quantificar o analito de interesse, inequivocamente, na presença de componentes que podem estar presentes na amostra, como impurezas, diluentes e componentes da matriz. Nos modelos PLS-DA, a especificidade é usada para identificar corretamente as amostras negativas, e é calculada através da Equação 34:

$$Especificidade = \frac{VN}{VN+FN} \quad (34)$$

A acurácia é um parâmetro estatístico que fornece um único valor para medir o desempenho do modelo de classificação. Seu cálculo é realizado através do número de amostras classificadas corretamente independentemente da classe dividido pelo número total de amostras, conforme mostra a Equação 35:

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN} \quad (35)$$

A taxa de confiabilidade e a acurácia (taxa de eficiência) são figuras de mérito globais que levam em conta tanto o FN quanto o FP. A taxa de confiabilidade é dada pela diferença da soma entre a sensibilidade e seletividade e 1 ($\text{sensibilidade} + \text{especificidade} - 1$). A acurácia e o CCM expressam através de seus valores o grau de eficiência do modelo construído. A eficiência é determinada pela média aritmética dos valores de sensibilidade e especificidade sendo que o valor 1 corresponde a uma eficiência de 100%, isto é, todas as amostras estão corretamente classificadas conforme sua classe de origem. Enquanto que o CCM é calculado conforme a Equação 36 e o resultados é dado por um único valor numérico entre −1 e +1, onde um valor de +1 corresponde a uma eficiência de 100%, valor igual a 0 uma classificação errada e igual a −1 uma classificação inversa.

$$CCM = \frac{(\mathrm{VP}\times\mathrm{VN}-\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{VP}+\mathrm{FP})(\mathrm{VP}+\mathrm{FN})(\mathrm{VN}+\mathrm{FP})(\mathrm{VN}+\mathrm{FN})}} \quad (36)$$

2.6. Aplicação da Espectrometria MIR e Métodos Quimiométricos no Controle de Qualidade das Misturas Biodiesel/Diesel

O controle de qualidade, não apenas de biodiesel, mas também de suas misturas com diesel cada vez mais usadas, é muito importante. O tipo (comprimento da cadeia, grau de insaturação e presença de outras funções químicas) e a concentração dos ácidos graxos utilizados na síntese do biodiesel, por exemplo, podem influenciar as propriedades do produto final e seus pré-requisitos em termos de armazenamento e sua tendência a oxidar. As matérias-primas utilizadas e o processo de obtenção de biocombustível podem ser decisivos em termos da introdução de alguns contaminantes nocivos no produto final.

Métodos cromatográficos e espectrométricos são as formas mais populares para análise de biocombustíveis. Para determinar o nível de alquilésteres de gorduras e óleos, adulterantes em misturas biodiesel/diesel, a cromatografia é vantajosa devido a alta sensibilidade, exatidão e precisão. Mas ela é desvantajosa por ser uma técnica cara, destrutiva e com etapas de preparação de amostras (BRANDÃO; BRAGA; SUAREZ, 2012; FERNANDES et al., 2013). Sendo a garantia de qualidade das misturas biodiesel/diesel um desafio para todos os órgãos de fiscalização e controle de qualidade, há necessidade do desenvolvimento de métodos que forneçam respostas rápidas para o controle da qualidade deste combustível. As técnicas espectrométricas vibracionais, no entanto, têm atributos que as tornam ideais para uso nesses tipos de medidas analíticas (BAETEN; DARDENNE, 2002). Esses atributos incluem o seguinte:

- É possível uma análise rápida e simples, qualitativa ou quantitativa;
- Normalmente, pouca ou nenhuma preparação de amostra é necessária e elas não usam solventes poluentes;
- É fácil encontrar padrões de calibração e métodos de análise apropriados, validados por métodos ou instituições reconhecidas;
- É possível fazer análises diretas e não invasivas;
- Normalmente, métodos não destrutivos podem ser empregados, com algumas exceções, como métodos de espectrometria atômica;
- Geralmente, a resposta observada é diretamente proporcional à concentração da espécie do sistema;
- A análise *in situ* pode ser realizada, por exemplo, no caso de monitoramento em tempo real das reações de produção de biocombustíveis;
- Elas possuem baixos custos em relação ao tempo de análise e reagentes.

Pelas razões acima, é relevante examinar os recentes avanços proporcionados pela introdução de métodos espectrométricos para análise do biodiesel. O foco deste subcapítulo é apresentar trabalhos publicados no Laboratório de Quimiometria do Triângulo (LQT) utilizando Espectrometria no Infravermelho Médio aliada a métodos multivariados para análise de amostras de biodiesel/diesel.

Até o presente momento no LQT, 17 publicações científicas foram produzidas, usando Espectrometria no Infravermelho Médio com Transformada de Fourier e acessório de Reflectância Total Atenuada Horizontal (FT-MIR-ATR) aliada aos métodos químiométricos qualitativos e quantitativos para análises de amostras de biodiesel/diesel. Os métodos quimiométricos mais usados foram PLS, PLS-DA e Cartas de Controle Multivariadas.

No trabalho publicado por (GUIMARÃES et al., 2015) foi utilizado o PLS para quantificar etanol em amostras de biodiesel etílico B100 provenientes de óleos de soja e de fritura usado. Os modelos obtidos foram eficientes na determinação da concentração de etanol que variaram entre 0,14 – 1,00% (m/m). Em ambos os modelos PLS, observou-se baixos valores de RMSEP (0,02% m/m), excelente correlação entre os valores medidos e preditos no conjunto de previsão, e não houve a presença de erros sistemáticos, conforme ASTM E1655. Além disso, os métodos foram validados de acordo com orientações internacionais e nacionais pela estimativa de figuras de mérito. Os mesmos autores também desenvolveram e validaram métodos utilizando PLS para quantificar etanol em biodiesel metílico (B100) provenientes de óleo de soja, e quantificar o óleo de soja como adulterante em misturas B5 de biodiesel metílico e etílico de soja no diesel (GUIMARÃES et al., 2014).

No trabalho publicado por (GONTIJO et al., 2014b) foi aplicada a espectrometria FT-MIR-ATR e PLS para quantificar biodiesel metílico e etílico de soja, na faixa de concentração de 1,00 – 30,00% (v/v) em amostras de misturas biodiesel/diesel, seguindo as orientações da ASTM E1655. Os modelos PLS apresentaram baixos valores de RMSEP 0,08% (v/v) e 0,10% (v/v), respectivamente, para os modelos do biodiesel metílico e etílico de soja. Estes mesmos autores (GONTIJO et al., 2014a), usando a mesma técnica, desenvolveram e validaram os modelos PLS para quantificar biodiesel metílico e etílico provenientes de óleo de fritura usado em amostras de misturas biodiesel/diesel empregando dados de Espectrometria MIR. Os valores da exatidão dos modelos PLS mostraram excelente desempenho com valores de RMSEC menor que 0,07% (v/v), RMSECV menor que 0,09% (v/v) e RMSEP menor que 0,09% (v/v). A exatidão foi comprovada pela avaliação da região de elipse de confiança. O método foi validado de acordo com as diretrizes internacionais e

nacionais pela estimativa de figuras de mérito, tais como exatidão, precisão, linearidade, seletividade, sensibilidade analítica, limites de detecção e quantificação e erro sistemático. Os modelos não apresentaram erros sistemáticos conforme a norma ASTM E1655.

(MÁQUINA et al., 2017a) desenvolveram método para quantificar o teor do biodiesel metílico de mafurra e classificar suas amostras de misturas biodiesel/diesel usando espectrometria FT-MIR com métodos de seleção de variáveis por siPLS e siPLS-DA. Posteriormente, (MÁQUINA et al., 2019a) desenvolveram métodos para quantificar o teor do biodiesel metílico de algodão e classificar suas amostras de biodiesel/diesel usando espectrometria no infravermelho médio associada a PLS e PLS-DA. O modelo PLS, desenvolvido para determinar o teor do biodiesel, foi validado com base em alguns valores de mérito: seletividade, sensibilidade, sensibilidade analítica, limite de detecção, limite de quantificação e teste de erro sistemático. O ajuste desse modelo também foi avaliado usando a correlação dos valores atuais e previstos dos conjuntos de calibração e previsão e foi observada uma alta correlação, com um coeficiente de correlação superior a 0,99 e erros relativamente baixos para os parâmetros. O monitoramento qualitativo foi realizado utilizando o modelo PLS-DA, cuja eficiência foi analisada com base em parâmetros de sensibilidade e especificidade. Esses parâmetros mostraram classificação 100% correta nas amostras utilizadas para calibração e previsão do teor do biodiesel no combustível B10 brasileiro para motores a diesel.

Os métodos de classificação por PLS-DA foram publicados em artigos científicos para identificação da matéria-prima de origem do biodiesel, ou seja, para identificar o tipo de óleo e álcool utilizado em sua produção (MÁQUINA et al., 2017b; MAZIVILA et al., 2015a); classificação do tipo de biodiesel tanto metílico quanto etílico provenientes de óleos de soja, óleo de fritura usado e pinhão manso em misturas B5 (MAZIVILA et al., 2015c); identificação da presença de adulterantes (óleos de soja, fritura usado, automotivo usado e gasolina) em misturas B5 composta de biodiesel metílico de soja (MAZIVILA et al., 2015b). Em todos os modelos PLS-DA os autores obtiveram 100% de classificação correta das amostras.

As cartas de controle multivariadas baseadas no sinal analítico líquido (*NAS*) foram usadas por (SITOE et al., 2016), monitorar teores do biodiesel, provenientes de pinhão manso e crambe, em misturas B7 com o diesel, enquanto que (MITSUTAKE et al., 2015; SITOE et al., 2017) avaliaram o controle de qualidade de misturas B5 e B7 formadas a partir de biodiesel de diferentes matérias-primas em misturas com diesel e adulteradas por óleos vegetais, óleo lubrificante automotivo residual, querosene e gasolina. Os resultados obtidos mostraram uma excelente distinção entre as amostras dentro e fora das especificações de

qualidade, com 91% e 100% de classificação correta. Já (BUIATTE et al., 2015) usaram PLS e cartas de controle multivariadas também baseadas no NAS para quantificar e monitorar o teor do biodiesel metílico de algodão em misturas com diesel na faixa de 1,00 – 30,00% (v/v). No modelo PLS, validado por figuras de mérito, os autores obtiveram valor de RMSEP igual a 0,05% (v/v); todas as amostras conforme e não-conformes do biodiesel metílico de algodão foram corretamente classificadas pelas cartas de controle multivariadas.

Usando métodos de seleção de variáveis e FT-MIR-ATR, (SITOE et al., 2019) selecionaram regiões espectrais e desenvolveram métodos para quantificar o teor do biodiesel metílico de pinhão manso em misturas com diesel. O grupo do Laboratório de Quimiometria do Triângulo fechou o ano de 2019 com a publicação de um artigo sobre a análise de espectros de ^{1}H RMN (Ressonância Magnética Nuclear de Hidrogênio) de misturas de diesel/biodiesel de crambe usando ferramentas quimiométricas para avaliar a autenticidade de uma mistura de biodiesel padrão brasileiro (MÁQUINA et al., 2019b).

Para fins de quantificação ou classificação, muitos algoritmos quimiométricos foram aplicados a dados da espectrometria vibracional para transformar os dados em informações quantitativas mais relevantes. Vários outros trabalhos na área de biocombustíveis utilizaram algoritmos diferentes para a avaliação quantitativa de diferentes respostas ou para controle e classificação da qualidade, como regressão linear multivariada (MLR), regressão por componentes principais (PCR), máquinas de vetores de suporte (SVM), rede neural artificial (RNA), análise discriminante por quadrados mínimos parcial (PLS-DA), análise de componentes principais (PCA), K-vizinhos mais próximos (KNN), análise discriminante linear (LDA), análise discriminante quadrática (QDA), conforme se pode observar nesta referência (SILVA et al., 2017) sobre avanços na aplicação de técnicas espectrométricas na área de biocombustíveis nas últimas décadas.

3. PROCEDIMENTO EXPERIMENTAL

Neste trabalho o biodiesel metílico de pinhão manso (BMPM) e o biodiesel metílico de moringa (BMM) foram utilizados para a preparação de amostras. Para a produção do biodiesel, foram utilizados reagentes químicos com alto grau de pureza laboratorial que incluem ácido sulfúrico (99,9%), álcool metílico (> 99%), hidróxido de potássio (95%) e água destilada. Os B100 usados para a preparação das amostras foram produzidos por via metílica. Todos os experimentos foram conduzidos no Laboratório de Quimiometria do Triângulo (LQT) do Instituto de Química da Universidade Federal de Uberlândia (Uberlândia, Minas Gerais, Brasil).

3.1. Produção do Biodiesel

Os óleos de pinhão manso e de moringa têm valor de ácidos graxos livres maior que 3% e teor de umidade maior que 1% em massa, que levam à saponificação e hidrólise do óleo, reduzindo a taxa de produção de biodiesel. Por isso, o processo de produção do biodiesel foi realizado em duas etapas. O primeiro passo foi a esterificação ácida, na qual o teor de ácidos graxos livres reduziu para um nível mínimo adequado; o segundo passo foi a transesterificação catalisada por bases.

A primeira etapa da produção dos B100 foi a catálise ácida (esterificação), realizada a fim de melhorar o processo de conversão dos óleos brutos e aumentar o rendimento de ésteres metílicos. Nesta reação foi usando CH_3OH e H_2SO_4 concentrado como catalisador. Um misturador agitador magnético de placa quente digital e um reator de vidro com fundo plano de dois tubos de 500,00 mL foram conectados a um condensador de refluxo resfriado a água usado para a esterificação ácida de óleos brutos.

Uma quantidade de 100,00 g do óleo foi pré-aquecida a 110 °C por 30 min para obter óleo livre de umidade. Uma mistura de 60% (m/m) de metanol e 1% (m/m) de catalisador (no óleo) foi preparada e adicionada ao óleo. A mistura foi vigorosamente agitada a 200 rpm e depois aquecida a 50 °C por 1 h. Após a reação de esterificação, a mistura heterogênea resultante foi transferida para um funil de separação e permaneceu em repouso durante 1 h para separar o excesso de álcool, ácido sulfúrico e impurezas presentes na camada superior.

A mistura resultante (fase inferior) foi transferida para um balão de fundo redondo para a destilação a vácuo usando o evaporador rotatório a 100 mmHg, 90 rpm e 80 ºC durante 1 hora, com o objetivo de remover o álcool metílico em excesso. Assim, o óleo esterificado ficou adequado para a transesterificação com catalisador básico (REDDY et al., 2017).

Na etapa da transesterificação (catálise básica), o óleo resultante da esterificação foi submetido a uma reação com o metanol, e com hidróxido de potássio (KOH) como catalisador. A proporção molar do óleo e álcool foi de 6:1 e 1% (m/m) do KOH em relação a massa do óleo. O catalisador foi inicialmente dissolvido em metanol e a mistura resultante foi adicionada ao óleo. Esta reação de transesterificação foi mantida a 65 ºC durante 2 h e 400 rpm de velocidade de agitação. Após a reação, a mistura monofásica resultante foi depositada em um funil de separação e mantido em repouso por 24 h para separar o glicerol (camada inferior) do biodiesel (camada superior).

A camada inferior que continha impurezas e glicerol foi guardada em recipiente para posterior tratamento, enquanto que a camada superior (biodiesel) foi submetida a lavagens sucessivas com água destilada aquecida a 90º C e depois seco usando um rota-evaporador durante 1 h a 88 rpm para remover o excesso de água e metanol (DA SILVA et al., 2010). Finalmente, a secagem foi complementada utilizando o sulfato de sódio e em seguida filtrado.

O esquema geral das etapas da produção do biodiesel por reações de esterificação e de transesterificação encontram-se na literatura (SITOE, 2016). Para se obter maior representatividade e aumentar a robustez dos modelos quimiométricos, foram preparados 16 lotes de B100 para cada tipo de óleo.

3.2. Preparação de Amostras

Foram utilizados 12 diferentes lotes de diesel puro para preparar 124 misturas de biodiesel/diesel de pinhão manso e 110 misturas biodiesel/diesel de moringa. As amostras de diesel foram fornecidas pela empresa TRANSPETRO S.A.. Todas as amostras dos conjuntos de calibração/teste e de validação/teste foram preparadas na faixa de concentração de 0,50 – 30% (v/v).

Normalmente, para a construção dos modelos de classificação PLS-DA, as amostras são divididas em duas classes: classe 1, denominada classe de interesse (B10, ou seja, amostras com 10% (v/v) de biodiesel e mais 90% (v/v) de diesel, considerando a variação percentual de ±0,50% (v/v) permitida pela ANP e classe 0 (outras amostras diferentes de B10, ou seja, amostras BX). Mas, neste trabalho, as amostras da classe 0 (BX) foram divididas em duas subclasses de acordo com a faixa do teor do biodiesel. Assim, no total foram usadas três classes, nomeadamente, (i) classe 1 constituída por todas as amostras BX ($BX, 0{,}50\% \leq X < 9{,}50\%$); (ii) classe 2 é a classe de interesse constituída por todas as amostras B10; e (iii) classe 3 constituída por todas as amostras BX ($\mathrm{BX}, 10{,}50\% < X \leq 30\%$). A opção de usar duas classes para as amostras que não são de interesse é devido à

natureza linear do modelo PLS-DA, o que poderia dificultar ou mesmo atrapalhar o desempenho do modelo caso estas fossem agrupadas em uma única classe (RINNAN; SAVORANI; ENGELSEN, 2018).

Na preparação de cada amostra, primeiro pesou-se a massa do biodiesel e depois acrescentava-se o diesel até a concentração pretendida; em seguida, as amostras foram homogeneizadas em um agitador *Vortex* por um minuto. As frações mássicas (%, m/m) dos componentes das amostras foram convertidas em frações volumétricas (%, v/v) por meio da densidade, com objetivo de expressar os resultados, para teor de biodiesel, na mesma unidade de medida referenciada nas normas EN 14078 e ABNT NBR 15568 (PINHO et al., 2014). Do total de amostras de cada modelo, pelo menos 60% foram selecionadas para o conjunto de calibração/treinamento e os restantes 40% foram utilizadas no conjunto de validação/teste. A seleção das amostras para cada conjunto foi realizada empregando o algoritmo Kennard-Stone (KENNARD; STONE, 1969). As quantidades de amostras usadas para modelos de calibração e de classificação são apresentadas nas Tabelas 3 e 4.

Tabela 3 – Números de amostras usadas nos conjuntos de calibração e de previsão na construção de modelos de quantificação

Modelo	Nº de Amostras	Conjunto de Calibração	Conjunto de previsão
BMPM	74	48	26
BMM	70	45	25

Fonte: O autor.

Tabela 4 – Números de amostras usadas nos conjuntos de treinamento e teste na construção de modelos de classificação

		Conjunto de Treinamento			Conjunto de Teste		
Modelo	**Nº de Amostras**	**Classe 1**	**Classe 2**	**Classe 3**	**Classe 1**	**Classe 2**	**Classe 3**
BMPM	124	20	42	20	10	22	10
BMM	110	18	37	18	9	19	9

Fonte: O autor.

3.3. Obtenção dos Dados Espectrais de Infravermelho Médio

Os dados espectrais MIR das amostras de todos os modelos foram adquiridos em quintuplicatas, usando um espectrômetro Perkin Elmer, modelo Spectrum Two (Waltham, MA, EUA) com Transformada de Fourier, equipado com um acessório amostrador ATR

horizontal (Pike Technologies, Fitchburg, WI, EUA). Nesse acessório, o cristal de ZnSe tem uma espessura de 4 mm e 80 mm de comprimento; o ângulo de incidência é de 45º e um valor de 10 reflexões. A Figura 14 apresenta o espectrômetro de infravermelho médio usado no LQT para a aquisição de espectros.

Todas as amostras usadas eram líquidas. Antes de medir o espectro de cada amostra, o acessório de ZnSe era limpado, usando álcool isopropílico, depois era feito o *background* usando o cristal sem amostra. Depois disso, cada espectro foi registrado na faixa de 4000 – 600 cm^{-1}, com uma resolução de 4 cm^{-1} e 16 varreduras.

Figura 14 - Espectrômetro do infravermelho médio (a) e acessório amostrador de ZnSe (b)

Fonte: O autor.

3.4. Análises Quimiométricas

Os modelos foram construídos utilizando MATLAB versão 7.5 (The Mathworks Inc., Natick, MA, USA) e PLS_Toolbox 8.7.0 (Eigenvectors Inc.). Os espectros de todos os modelos foram corrigidos pelo método da correção de linha de base nas faixas de 4000 – 3200 cm^{-1} e de 2550 – 1850 cm^{-1}; o corte de região não informativas foi de 4000 – 3150 cm^{-1} e o corte da região com ruído foi feita nas faixas de 680 – 600 cm^{-1}.

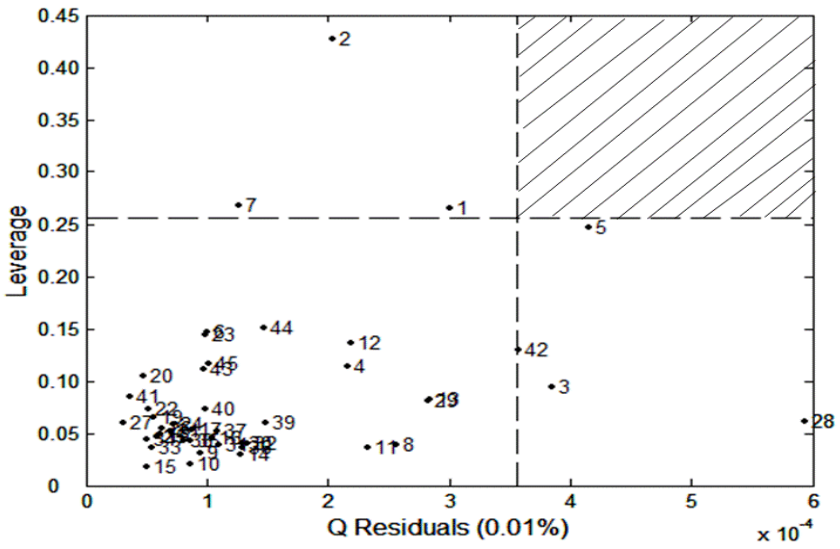
Todos os modelos de calibração multivariada foram obtidos por meio de validação pelo procedimento de venezianas para definir o número de variáveis latentes (VLs). Neste processo, as amostras foram divididas em 12 blocos com 2 amostras por bloco. O número de VLs que apresentou o menor valor de RMSECV foi selecionado, em comparação conjunta da porcentagem de variância explicada em **X** (matriz de espectros) e em **y** (concentração).

Nos modelos de classificação a escolha de variáveis latentes foi feita com base na porcentagem de amostras classificadas corretamente.

A capacidade preditiva dos modelos de calibração foi avaliada pelo erro quadrático médio de predição (RMSEP) obtido para o conjunto de validação externa, enquanto que para modelos de classificação, a capacidade preditiva foi avaliada através da Tabela de contingência de onde foram obtidas as figuras de mérito. As amostras do branco (diesel puro) de cada modelo de calibração foram usadas para a determinação das figuras de mérito. A validação externa do desempenho de cada modelo foi feita usando outras amostras de biodiesel/diesel com o teor do biodiesel na faixa de 1,00 – 29,50% (v/v).

A análise de amostras anómalas (*outliers*) foi feita através do gráfico de influência (*leverage*) versus (*vs*) resíduos de *Student* (*Q Residuals*), a nível de significância de 5%. Amostras fora do limite de confiança de 95% (região achurada na Figura 15) são consideradas *outliers*. Além disso, amostras *outliers* também foram detectadas a partir da comparação entre valores de erros absolutos nas amostras individuais em relação ao valor de erro quadrático médio de calibração (RMSEC), ou seja, as amostras com uma diferença entre o valor de referência e o valor estimado maior que três vezes o valor do RMSEC foram excluídas.

Figura 15 – Representação gráfica de *Q residuals vs Leverage* na identificação de *outliers*



Fonte: GONTIJO, 2016.

Os modelos de seleção de variáveis foram desenvolvidos através da subdivisão dos espectros totais em subintervalos menores, com o mesmo peso de 4, 8, 12 e até 32 faixas (conforme descrito no item 2.5.5, página 57), usando rotinas desenvolvidas para o ambiente

computacional MATLAB. Posteriormente, modelos de regressão PLS e PLS-DA foram desenvolvidos para cada um dos subintervalos usando o mesmo número de VLs. Para os resultados de saída dos algoritmos iPLS, biPLS e siPLS, foi selecionado e construído um modelo PLS em cada subintervalo ou combinação de intervalos que apresentaram menor valor de RMSECV se comparado com o valor de RMSECV do modelo PLS global. E os modelos de classificação por iPLS-DA, foram selecionados os que apresentaram melhores resultados de classificação.

Como os dados espectrais são altamente correlacionados, neste trabalho foram usadas janelas de variáveis em vez de fazer uma seleção de variáveis em cada variável individualmente, porque o algoritmo iPLS fornece uma imagem geral do conjunto de dados, incluindo partes que contêm informações, interferências e ruídos relevantes (ANDERSEN; BRO, 2010).

4. RESULTADOS E DISCUSSÃO

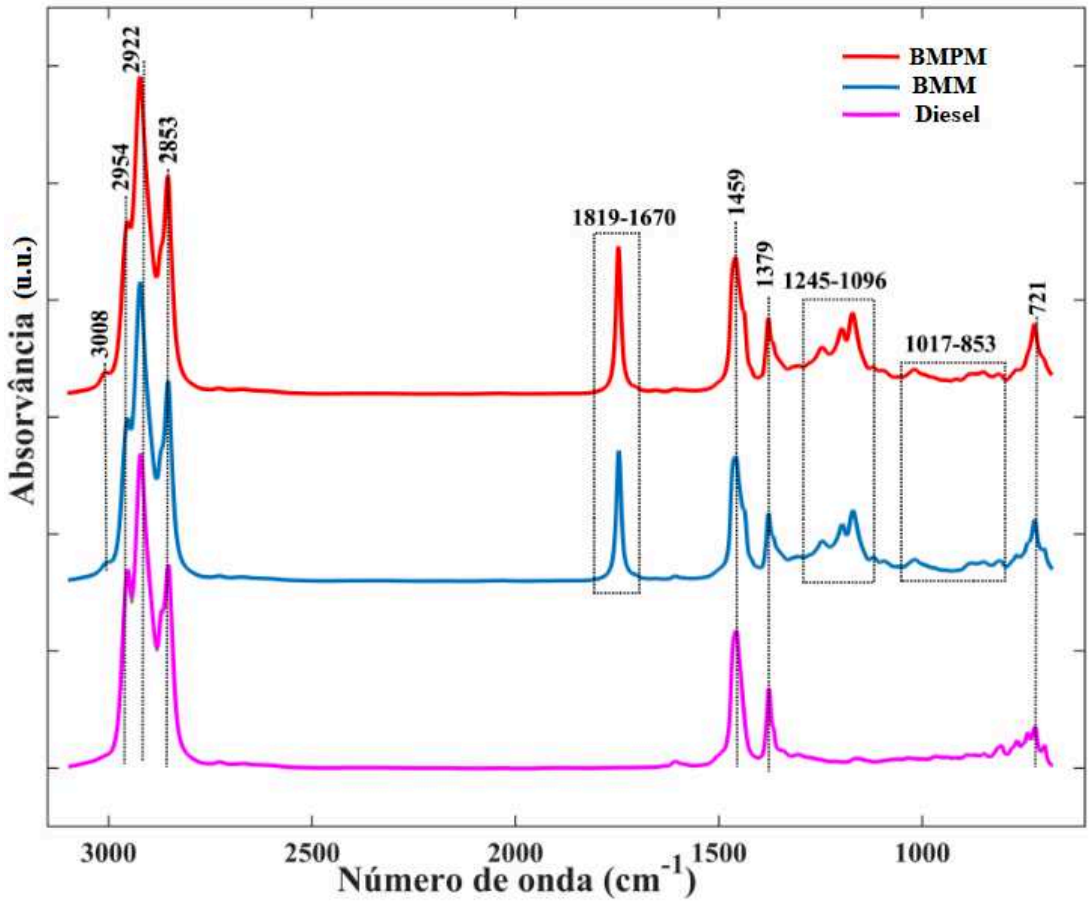
4.1. Caracterização dos espectros MIR das misturas B10 e BX

A Figura 16 apresenta os espectros do diesel puro, das misturas biodiesel/diesel (B10) de moringa (BMM) e de pinhão manso (BMPM) na faixa de 3150 – 680 cm^{-1}. Nota-se que os espectros dos dois tipos de biodiesel apresentam grande semelhança entre si. Tanto o espectro do diesel como os espectros das misturas B10 apresentam, essencialmente absorções correspondentes aos modos vibracionais característicos de alcanos normais, com picos em 2954, 2922 e 2853 cm^{-1} referentes aos estiramentos assimétricos de $–CH_3$, simétrico e assimétrico de $–CH_2$, respectivamente. As bandas de pequena intensidade presentes em 3008 cm^{-1} apenas nas amostras de B10 representam o estiramento simétrico e assimétrico das ligações =C–H em grupos olefínicos ($\upsilon_s C–H_{olef}$ e $\upsilon_{as} C–H_{olef}$) para os derivados insaturados.

As bandas de absorção das ligações C=O e C–O são as mais relevantes nas amostras de biodiesel; na região 1819 – 1679 cm^{-1} com picos máximos em 1742 cm^{-1} e em 1745 cm^{-1} ocorre a absorção do estiramento da ligação carbonila, [υ(C=O)], de ésteres alifáticos originários de biodiesel, enquanto que as bandas vizinhas entre 1245 – 1096 cm^{-1} (com pico máximo em 1170 cm^{-1}) referem-se à deformação axial assimétrica da ligação C–O do grupo C–O(OR) dos mesmos ésteres alifáticos (CASTILHO-ALMEIDA et al., 2012; SILVERSTEIN et al., 2019).

As bandas muito próximas e de média intensidade presentes em 1459 e 1379 cm^{-1} são, respectivamente atribuída à deformação angular assimétrica da ligação C–H do tipo tesoura do grupo metil, $–CH_3$, e também à deformação angular da ligação C–H, mas simétrica do tipo tesoura de $–CH_2–$. Na região de baixa frequência, a partir de 1095 cm^{-1} até próximo de 853 cm^{-1} vibrações significantes da deformação fora do plano são atribuídas às ligações δ(C=CH), δ(C–H olefina), e ω(C–H olefina) dos grupos olefinas de derivados insaturados. A banda de deformação angular do tipo balanço, fina e intensa em torno de 721 cm^{-1} é devido ao modo de deformação angular assimétrica no plano de $\rho(CH_2)$ dos derivados não saturados. Ela aparece entre as bandas de absorção para estiramento de ligação –C–C–, geralmente designado como vibração esquelética (que são modos vibracionais relacionados as variações de energia encontradas para as diferentes funções na região da impressão digital do espectro) (MCMURRY, 2017). O diesel não apresenta moléculas de oxigênio na sua constituição, por isso que os seus espectros não apresentam bandas de ligação C=O e C–O.

Figura 16 – Espectros MIR de diesel puro (—), B10 de moringa (—) e B10 de pinhão manso (—)

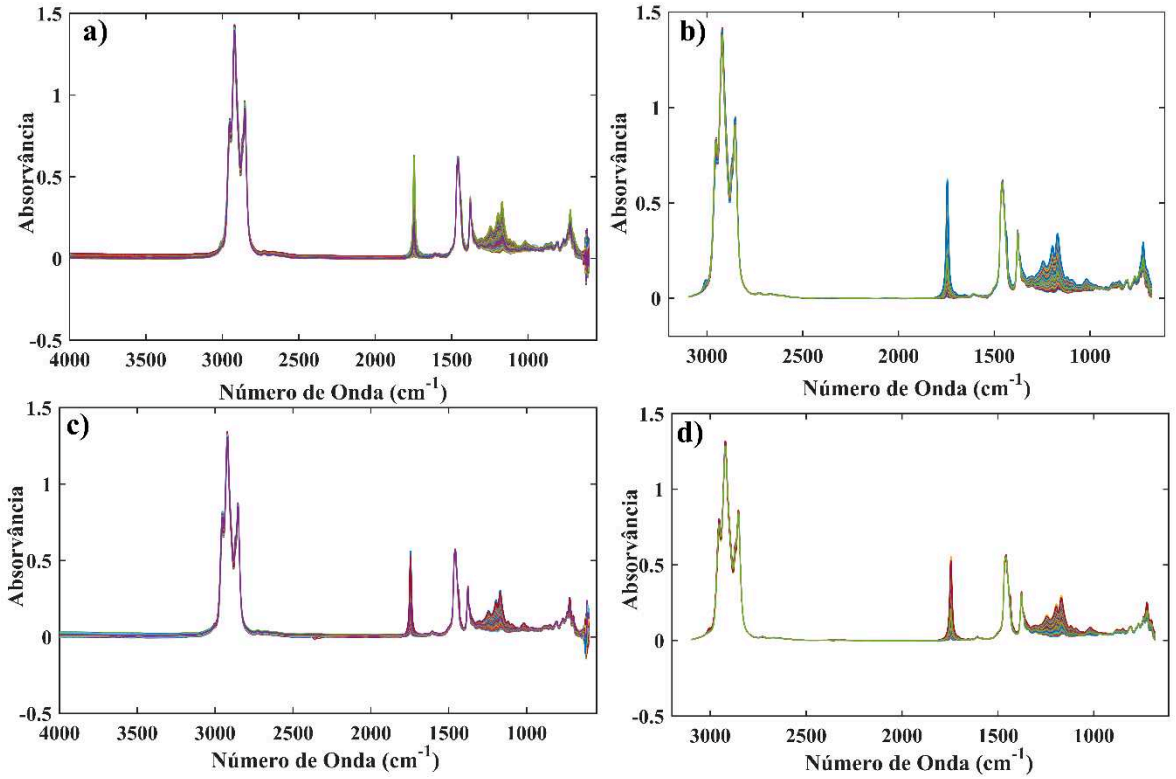


Fonte: O autor.

A Figura 17 apresenta todos os espectros MIR de amostras usadas para a construção dos modelos de calibração/treinamento e validação/teste na faixa de 0,50 – 30% (v/v) do teor do biodiesel no diesel. As regiões de maiores variações ocorrem na região de impressão digital (entre 1300 – 1000 cm^{-1}) e na região da carbonila, em torno de 1819 – 1679 cm^{-1}. Isso porque estas regiões correspondem a ligações de carbono e oxigênio, presentes no biodiesel. No entanto, também são verificadas pequenas variações nas regiões do diesel, uma vez que com o aumento na concentração de biodiesel, ocorre uma diminuição na concentração do diesel. Essas pequenas variações devem-se a diferentes matérias-primas usadas para a produção do biodiesel, e também devido a diferença do teor do biodiesel em diesel.

Deve-se ressaltar que, devido a grande sobreposição de sinais, uma análise visual ou mesmo um método univariado (como a intensidade de um pico em específico) se torna praticamente impossível diferenciar um espectro do outro. Por isso, nestes casos, é necessário usar métodos multivariados para extrair a informação e fazer as devidas análises.

Figura 17 − Espectros MIR das amostras usadas para a construção dos modelos de calibração/treinamento e validação/teste: (a) e (c) são espectros originais e (b) e (d) são espectros do BMPM e BMM, respectivamente, pré-tratados pela correção da linha de base



Fonte: O autor.

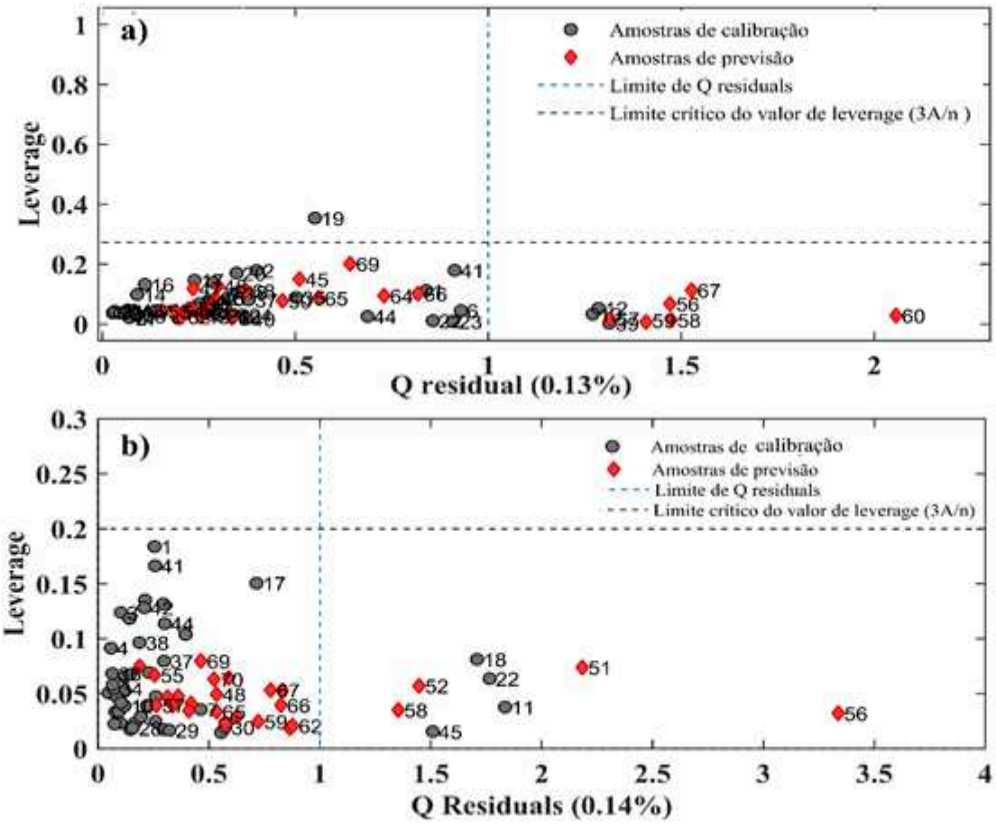
4.2. Construção de Modelos PLS Globais para Quantificar o Teor do BMPM e do BMM em Misturas com Diesel

Os espectros totais apresentados na Figura 17 (b) e (d) foram usados para a construção dos modelos do BMPM e BMM. O primeiro passo na construção de modelos PLS foi a identificação de possíveis amostras anômalas uma vez que estas amostras podem afetar a qualidade final dos modelos. As Figuras 18 (a) e (b) apresentam o gráfico de *Q residual vs leverage* para os 2 modelos PLS. As linhas verticais representam os limites de resíduos, a 95% confiança, enquanto que as linhas horizontais representa a influência crítica do valor de *leverage*, definido como $\frac{3A}{n}$, onde A é o número de VLs e n o número de amostras.

Verifica-se que em todos os modelos algumas amostras apresentam altos valores de resíduos, mas com baixos valores de *leverage*; por exemplo, a amostra (19) do modelo do BMPM apresenta alto valor de *leverage*, mas com valor de *Q residual* dentro do limite determinado. Contudo, apesar de algumas amostras tanto de calibração como de validação

estarem longe do centro do modelo, como não apresentam simultaneamente altos valores de *leverage* e de *Q residual*, significa que os modelos não têm presença de amostras anômalas.

Figura 18 – Gráfico de *Q residual vs leverage* para os modelos PLS global de (a) BMPM e (b) BMM com as amostras de calibração (•) e de validação (♦) para análise de amostras anômalas a 95% de confiança



Fonte: O autor.

Como nenhuma amostra se comportou como anômala em ambos os modelos, em seguida foram construídos os modelos de calibração tendo em conta o número de amostras exigidas pela norma ASTM E1655-05, que estabelece que no conjunto de calibração o número mínimo de amostras deve ser $6(k+1)$ para os dados centrados na média e, para o conjunto de previsão, $4k$ (ASTM E1655-05, 2012). O valor de k refere-se ao número de VLs escolhido para construção do modelo PLS. Os números de amostras entre parênteses na Tabela 5, tanto no conjunto calibração quanto no de previsão, referem-se ao número mínimo de amostras exigidas pela norma ASTM E1655-05, tendo em conta o número de VLs usados em cada modelo. O número de VLs foi escolhido de acordo com o menor valor do RMSECV.

Tabela 5 – Número de amostras nos conjuntos de calibração e previsão, número de VLs, variância capturada em cada bloco e valores de exatidão para os modelos PLS global

Modelo PLS*	Nº. amostras de calibração	Nº. amostras de validação	VLs	Variância em **X** (%)	Variância em **y** (%)	RMSECV (%)	RMSEC (%)	RMSEP (%)
BMPM	48 (30)	26 (16)	4	99,78	99,81	0,39	0,29	0,43
BMM	45 (24)	25 (12)	3	99,89	99,98	0,13	0,12	0,21

*BMPM=biodiesel metílico de pinhão manso; BMM=biodiesel metílico de moringa

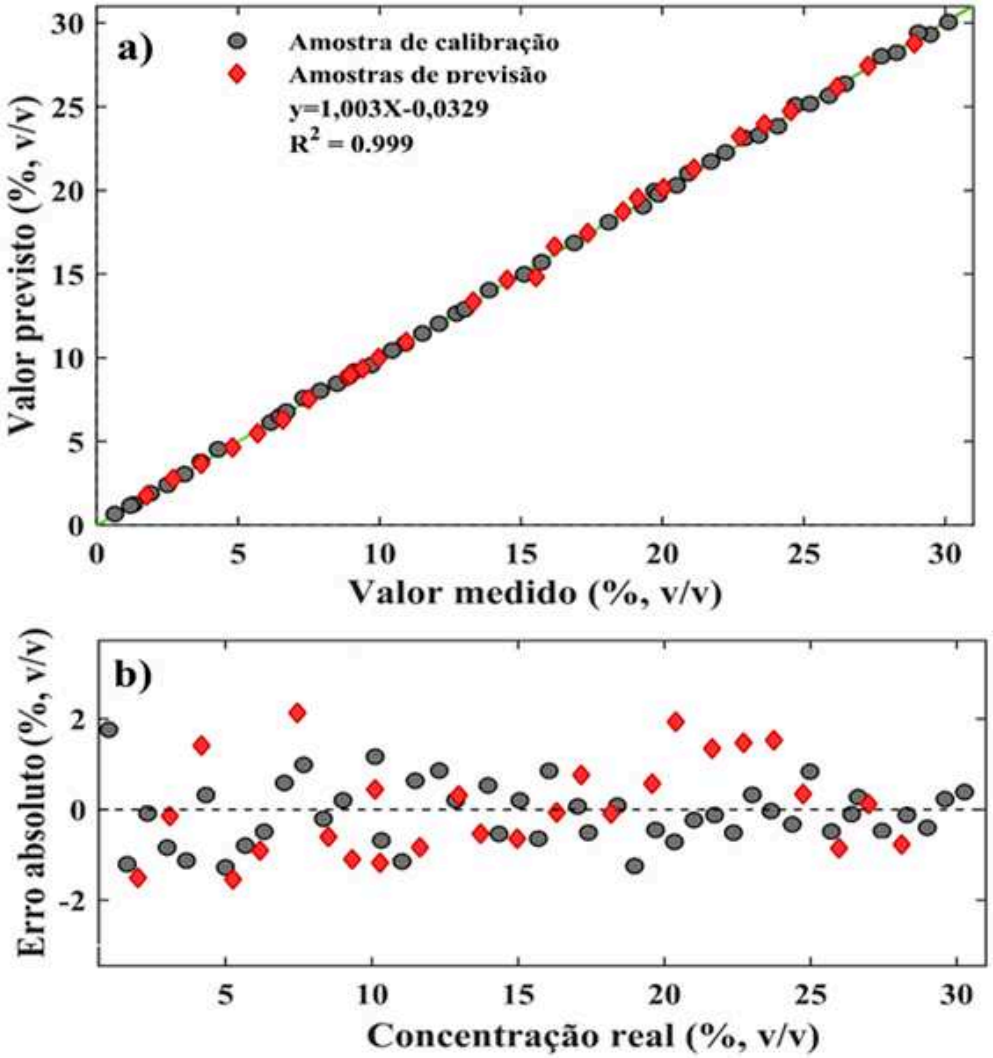
Fonte: O autor.

O modelo PLS global do BMPM foi construído com 4VLs e o do BMM com foi construído com 3 VLs. Avaliando a exatidão dos modelos através dos valores de RMSEC, RMSECV e RMSEP, pode-se observar que em todos os modelos os valores de erros foram satisfatórios, ou seja, estão dentro da exatidão requerida pela norma ABNT NBR 15568 (ABNT, 2008), a qual especifica que o valor máximo de RMSEP permitido é de 0,1% (v/v) e 1% (v/v) para os modelos construídos de 0,00 – 8,00% (v/v) e de 8,00 – 30,00% (v/v), respectivamente. Além disso, como estas amostras foram preparadas com a adição de 10% (v/v) de biodiesel ao diesel, seria aceitável uma exatidão próxima de 1% (v/v). Também pode-se observar que os valores de RMSEC e RMSEP, apresentam uma pequena diferença entre eles indicando que não há sobreajuste ou subajuste dos modelos.

A linearidade dos modelos globais foi verificada com base no gráfico dos valores reais de concentração de biodiesel *vs* valores previstos pelos modelos em avaliação conjunta com o gráfico de resíduos. Assim, conforme apresentado nas Figuras 19 e 20, as amostras de calibração e de previsão estão próximas à regressão e bem distribuídas ao longo da reta. De acordo com os valores dos coeficientes de correlação (R^2), observa-se que o ajuste para os modelos PLS tem uma boa correlação entre os dados de referência e os dados calculados pelos modelos PLS, uma vez que quanto mais próximo de 1 para a inclinação, e mais próximo de 0 para o intercepto, melhor é a concordância entre os valores.

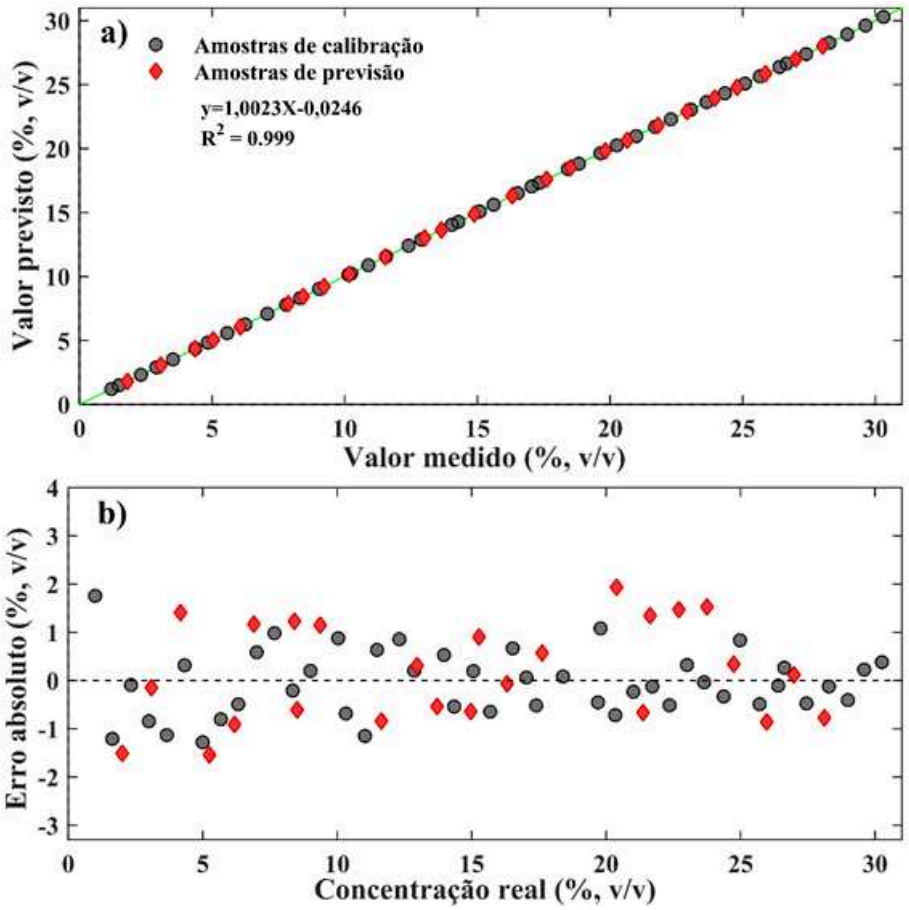
A linearidade é também evidenciada pela distribuição aleatória dos resíduos. As Figuras 19 e 20 também apresentam os gráficos de resíduos para os modelos. Verifica-se que em todos os modelos os resíduos estão distribuídos aleatoriamente ao redor da linha no valor zero. Portanto, os modelos PLS apresentaram comportamento linear ao longo da faixa de trabalho de 0,50 – 30% (v/v).

Figura 19 – (a) Gráfico de ajuste de valores medidos *vs* valores previstos, e (b) gráfico de resíduos do modelo PLS global do BMPM



Fonte: O autor.

Figura 20 – (a) Gráfico de ajuste de valores medidos *vs* valores previstos, e (b) gráfico de resíduos do modelo PLS global do BMM



Fonte: O autor.

A Tabela 6 apresenta os resultados da análise dos erros sistemáticos. Segundo a norma ASTM E1655-05, a realização do teste-*t* tem o objetivo de avaliar se o "*bias*" incluso no modelo é ou não significativo. Em todos os modelos PLS global verificou-se que $t_{calculado}$ é menor do que $t_{crítico}$, para graus de liberdade (GL) igual ao número de amostras de previsão. Logo, conclui-se que todos os modelos não apresentaram erros sistemáticos e estão de acordo com as recomendações da norma.

Tabela 6 – Valores de $t_{crítico}$ e $t_{calculado}$ para avaliação da presença de erros sistemáticos nos modelos PLS globais

Modelo	**BMPM**	**BMM**
Graus de liberdade	26	25
$t_{crítico}$	2,056	2,060
$t_{calculado}$	0,6863	0,6411

Fonte: O autor.

4.3. Construção de Modelos de Seleção de Variáveis por Intervalos (iPLS, biPLS e siPLS) para Quantificar o Teor do BMPM no Diesel

Para a construção de modelos de seleção de variáveis por iPLS, biPLS e siPLS, os espectros foram divididos em 4, 8, 12, 16 até 32 intervalos equidistantes. Todos os modelos apresentaram melhores resultados quando os espectros foram divididos em 24 faixas, ou seja, antes e depois de 24, os erros são maiores quando comparados aos erros do modelo PLS global. Para cada método, a (s) faixa (s) espectral (is) usada (s), o número de variáveis (NV), o número de VLs e a exatidão através dos valores de RMSEC, RMSECV e RMSEP estão apresentados na Tabela 7. Nestes dados, pode-se verificar que em todos os modelos os valores de erros foram satisfatórios, ou seja, estão dentro da exatidão requerida pela norma ABNT NBR 15568, a qual especifica que o valor máximo de RMSEP permitido é de 1% (v/v) para modelos construídos com a faixa da concentração do teor do biodiesel em diesel entre 0,50 – 30% (v/v).

Tabela 7 - Resultados dos modelos PLS global, iPLS, biPLS e siPLS para a quantificação do teor do biodiesel metílico de pinhão manso em mistura com diesel

Modelo	Intervalos	NV	VLs	RMSEC	RMSECV	RMSEP	Valor de F*	Valor de F**
Global PLS	Todos	2421	4	0,29	0,39	0,43	---	3,21
$iPLS_{24}$	[1]	101	4	0,22	0,25	0,24	3,21	---
$biPLS_{24}$	[1]	101	4	0,22	0,24	0,24	3,21	---
Si_4PLS_{24}	[1 17 18 19]	404	4	0,20	0,24	0,25	2,96	1,18

*Valor de $F_{calculado}$ em relação ao modelo de PLS global; **em relação ao menor valor de RMSEP do modelo de seleção de variáveis; [$F_{tabelado}$ (0,05,26,26) = 1,93].

Fonte: O autor.

Os resultados do modelo $iPLS_{24}$ e $biPLS_{24}$ tem os mesmos resultados pois eles foram construídos usando o mesmo intervalo quando os espectros totais foram divididos em 24 intervalos equidistantes. Os modelos de seleção de variáveis ($iPLS_{24}$, $biPLS_{24}$ e si_4PLS_{24}) apresentaram menores valores de RMSEP que do PLS global.

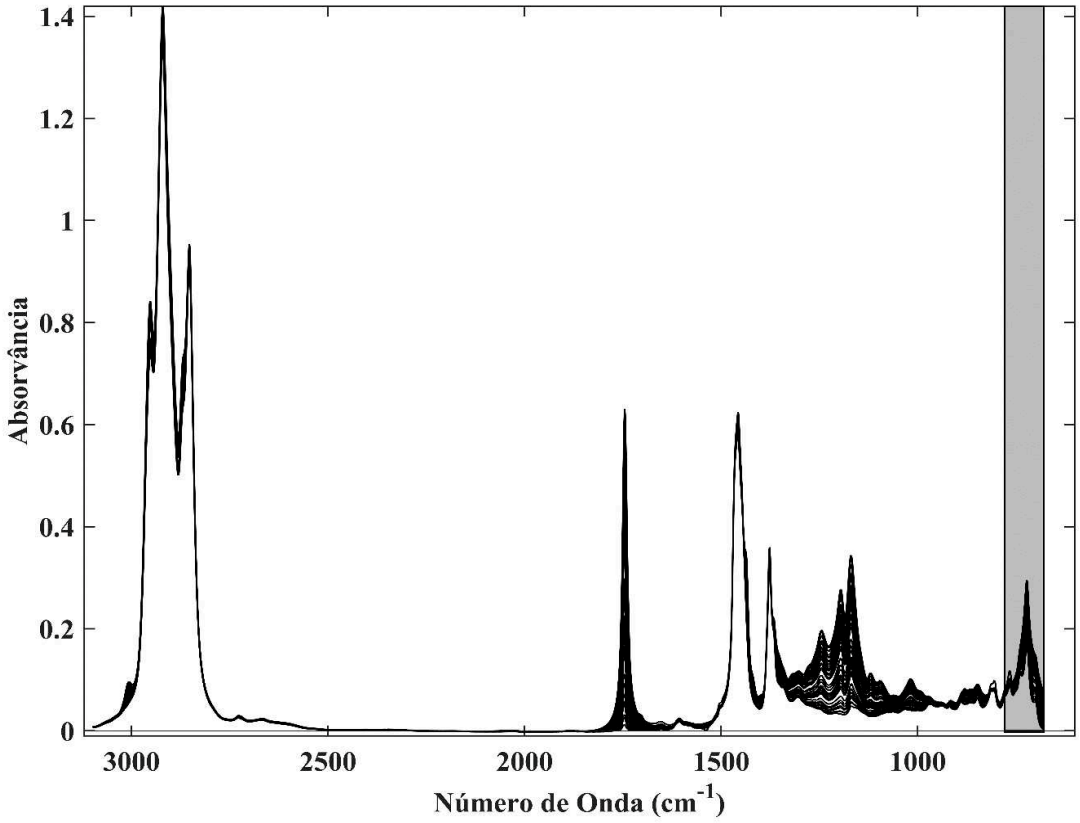
Foi aplicado o teste *F* para verificar se o RMSEP do PLS global foi ou não significativamente maior que cada modelo de seleção de variáveis, a 95% de confiança. De acordo com a comparação feita entre todos os modelos, verificou-se que os modelos de

seleção de variáveis apresentaram diferenças estatísticas significativas em relação ao modelo de PLS global, porque os valores de $F_{calculado}$ são maiores que $F_{tabelado}$. Consequentemente, os valores de RMSEP de todos os modelos de seleção de variáveis são melhores em relação ao do PLS global. O mesmo teste *F* também foi aplicado para comparar os modelos de seleção de variáveis ($iPLS_{24}$, $biPLS_{24}$ e si_4PLS_{24}) entre os seus valores de RMSEP. Verificou-se que não há uma diferença significativa entres os modelos, pois o valor de $F_{calculado}$ para o si_4PLS_{24} é menor que o $F_{tabelado}$.

Assim, comparativamente ao PLS global, todos os modelos de seleção de variáveis podem ser usados para a quantificação do teor do BMPM em diesel. Mas, entre os modelos de seleção de variáveis, o $iPLS_{24}$ e $biPLS_{24}$ apresentam melhor performance para a quantificação do teor do biodiesel metílico de pinhão manso. Neste trabalho foi escolhido o modelo de $iPLS_{24}$ para a quantificação do teor do biodiesel metílico de pinhão manso nas misturas com diesel.

O intervalo usado para a construção do modelo está apresentado na Figura 21, que corresponde à região entre 779 – 680 cm^{-1}, referente a absorção da ligação C–H da deformação angular assimétrica dos grupos metilenos ligados [$\rho–(CH_2)_n–$], relacionado ao biodiesel e ao diesel .

Figura 21 – Intervalo selecionado na construção do modelo $iPLS_{24}$ para quantificar o teor do BMPM em misturas com diesel

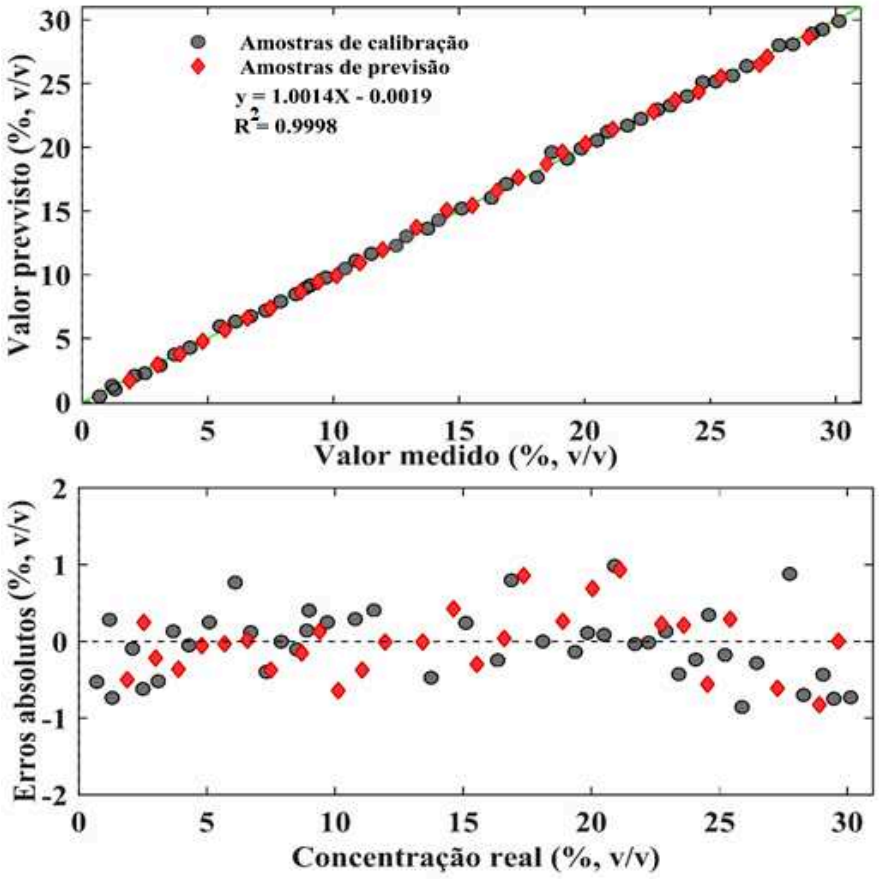


Fonte: O autor.

A avaliação do ajuste do modelo através das correlações dos valores medidos e os valores previstos dos conjuntos de calibração e de previsão é apresentada na Figura 22 (a). O modelo presenta um coeficiente de correlação acima de 0,9998, indicado que o modelo tem bom ajuste entre os valores medidos e os previstos. Este modelo é eficaz de quantificar o teor do biodiesel faixa de concentração de 0,50 – 30% (v/v) em misturas com diesel.

A linearidade do modelo também pôde ser avaliada por meio do gráfico dos erros sistemáticos. Na Figura 22b) observa-se que a distribuição dos erros de calibração e de previsão acontece de forma aleatória no modelo, o que significa que a curva está bem ajustada. Os pontos que apresentam algum desvio do comportamento linear não são tidos como indício de falta de linearidade dos dados, porque o modelo foi construído com 4 LV, e o coeficiente de correlação dos erros absolutos e valores medidos estão acima 0,99.

Figura 22 – (a) Gráfico de ajuste de valores medidos *vs* valores previstos, e (b) gráfico de resíduos do modelo $iPLS_{24}$ do BMPM



Fonte: O autor.

4.4. Construção de Modelos de Seleção de Variáveis por Intervalos (iPLS, biPLS e siPLS) para Quantificar o Teor do BMM no Diesel

Para a quantificar o teor do BMM, usando métodos de seleção de variáveis, os melhores resultados do iPLS, biPLS e siPLS foram obtidos quando os espectros totais foram divididos em 28, 32 e 24 intervalos equidistantes, respectivamente. Os indicadores da exatidão apresentam um nível aceitável de dispersão, ou seja, uma pequena diferença entre eles, o que significa que não há sobreajuste do modelo. Isso indica que os valores de RMSEC são uma boa estimativa do desvio padrão dos erros de previsão observados nos conjuntos de previsão. A Tabela 8 apresenta os resultados do PLS global e dos melhores modelos de seleção de variáveis por intervalos para quantificar o teor do biodiesel metílico de moringa.

Tabela 8 – Resultados dos modelos PLS global, iPLS, biPLS e siPLS para a quantificação do teor do biodiesel metílico de moringa em mistura com diesel

Modelo	Intervalos	NV	VLs	RMSEC	RMSECV	RMSEP	Valor de F*
Global PLS	Todos	2421	3	0,12	0,14	0,21	---
$iPLS_{28}$	[13]	87	3	0,10	0,13	0,11	3,64
$biPLS_{32}$	[1 2 3 5 6 9 12]	530	3	0,09	0,10	0,15	1,96
Si_4PLS_{24}	[5 6 8 11]	404	3	0,09	0,10	0,16	1,72

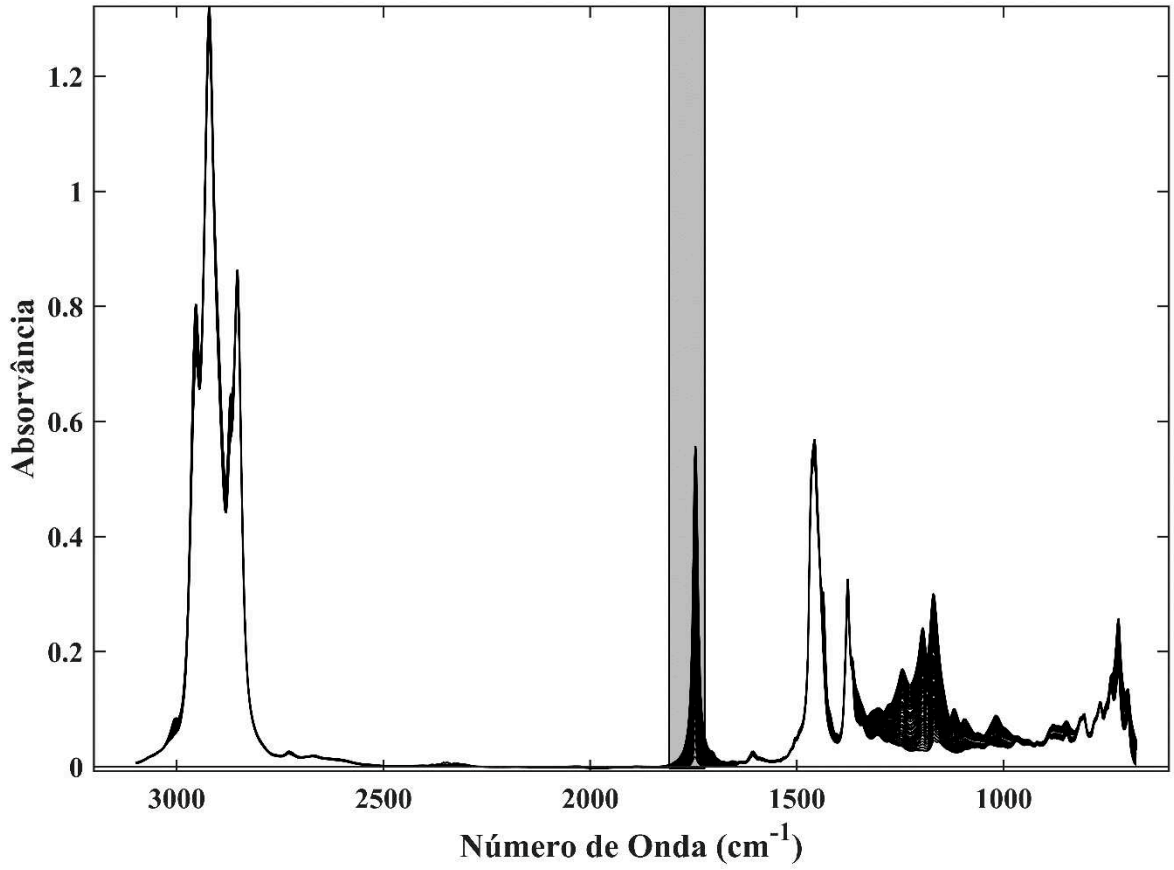
*Valor de $F_{calculado}$ em relação ao modelo de PLS global; [$F_{tabelado}$ (0,05,25,25) = 1,96].

Fonte: O autor.

Ainda na Tabela 8, verifica-se que apenas o modelo PLS_{28} é que apresentou diferença estatística em relação ao valor de RMSEP do modelo PLS global, porque o valor de $F_{calculado}$ é maior que $F_{tabelado}$. A Figura 23 ilustra o intervalo selecionado pelo algoritmo $iPLS_{28}$. Este intervalo corresponde à região de 1809 – 1723 cm^{-1}; é constituída por 87 variáveis e corresponde ao estiramento da ligação carbonila, [υ(C=O)], presente nos ésteres metílicos que constituem o biodiesel metílico de moringa.

Esta região é semelhante às regiões sugeridas pela norma (ABNT, 2008), para a quantificação do teor do biodiesel em diesel usando espectrometria MIR. É importante ressaltar que esta norma recomenda a utilização das regiões para quantificar qualquer tipo de biodiesel, mas aplicando os métodos de seleção de variáveis percebe-se até o momento que dependendo do tipo de biodiesel há sugestões de regiões ou intervalos diferentes indicados por este algoritmo, $iPLS_{28}$. Comparativamente aos métodos de seleção de variáveis para quantificar o teor do BMPM (Tabela 8), estes modelos têm maior exatidão, ou seja, com valores de RMSEC, RMSECV e RMSEP menores e bem próximos.

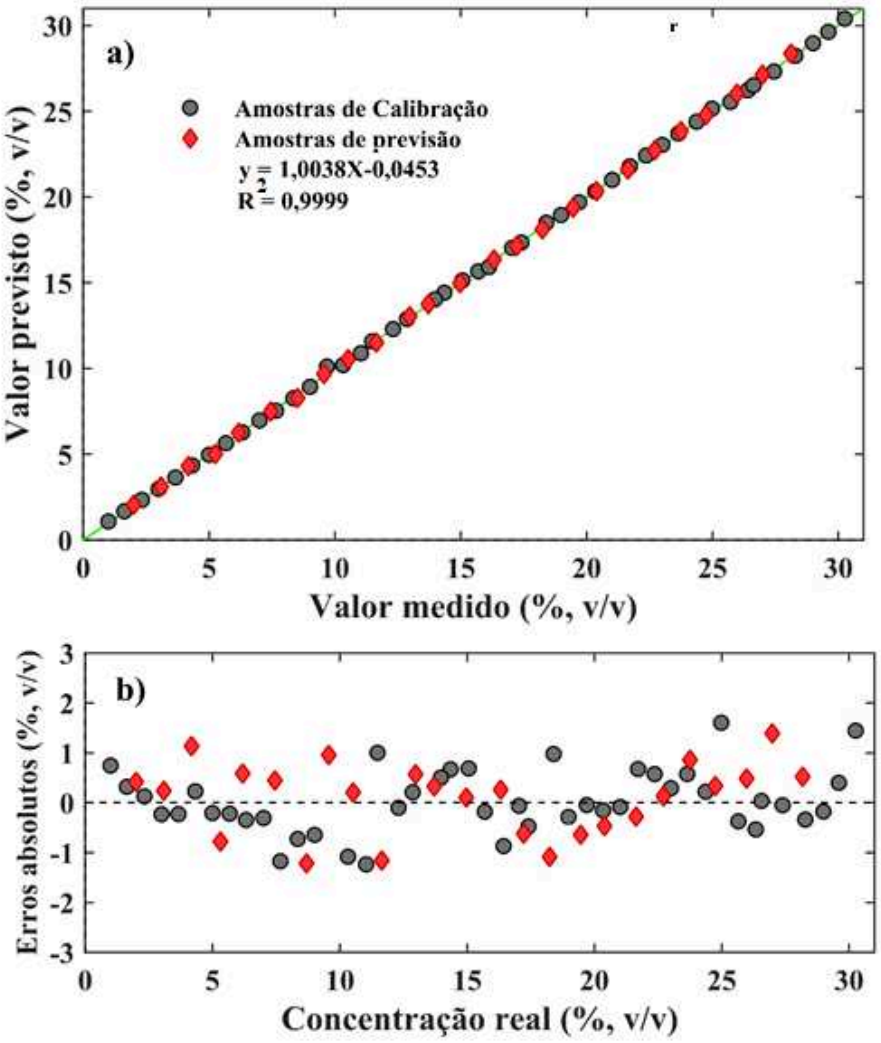
Figura 23 – Intervalo selecionado (1809 – 1723cm^{-1}) na construção do modelo iPLS$_{28}$ para quantificar o teor do BMM em mistura com diesel



Fonte: O autor.

O modelo apresenta uma boa linearidade com uma correlação entre os valores medidos e os previstos, conforme se pode notar na Figura 24 (a). O gráfico dos resíduos para as amostras de calibração e de validação apresentam uma distribuição aleatória, indicando que os dados seguem um comportamento linear (Figura 24 (b)). Por isso, o modelo iPLS$_{28}$ pode ser aplicado na quantificação de biodiesel metílico de moringa em misturas com diesel.

Figura 24 – (a) Gráfico de ajuste de valores medidos *vs* valores previstos, e (b) gráfico de resíduos do modelo $iPLS_{28}$ do BMM



Fonte: O autor.

4.5. Validação Analítica dos Métodos de Seleção de Variáveis do BMPM e do BMM

Para avaliar a capacidade de prever corretamente os valores das outras amostras, ou seja, para verificar a eficiência de cada modelo considerado melhor na seleção de variáveis para quantificar os BMPM e BMM, foram analisadas as figuras de mérito, anteriormente descritas no item 2.5.5.4, página 58. Os valores calculados para erros sistemáticos, seletividade, sensibilidade assim como as demais figuras de mérito são apresentados na Tabela 9.

Tabela 9 – Valores das figuras de mérito calculados dos modelos de $iPLS_{24}$ e $iPLS_{28}$ para quantificação do BMPM e BMM em misturas com diesel, respectivamente

Figura de mérito	**Parâmetro**	**Valor**	
		$iPLS_{24}$ (BMPM)	**$iPLS_{28}$ (BMM)**
Ajuste da linearidade	Inclinação	1,0014	1,0038
	Intercepto	-0,0019	-0,0453
	Coeficiente de correlação	0,9998	0,9999
Razão sinal-ruído	Máximo	100,79	254,02
	Mínimo	34,76	92,24
Limite de detecção/%(v/v)		0,13	0,30
Limite de quantificação/%(v/v)		0,40	0,41
Seletividade (%)		0,10	0,17
Sensibilidade/% (v/v)		0,01	0,01
Sensibilidade analítica/% (v/v)		24,98	25,52
Inverso da sensibilidade analítica (γ^{-1})/%(v/v)$^{-1}$		0,040	0,039
Erros sistemáticos	Bias	0,0502	0,0113
	SDV	0,2415	0,1122
	GL	26	25
	$t_{calculado}$	1,0390	0,5021
	$t_{crítico}$	2,060	2,056

Fonte: O autor.

Os modelos apresentaram baixos valores numéricos de sensibilidade, graças ao pré-processamento utilizado. De acordo com os valores calculados para o inverso da sensibilidade analítica, os modelos são capazes de distinguir amostras com diferença de concentração de pelo menos 0,04% (v/v) do teor do BMPM e de pelo menos 0,039% (v/v) do teor do BMM em diesel, e, como a menor diferença de concentração usada na construção dos modelos de calibração foi de 0,50% (v/v), então, os modelos conseguem fazer uma boa distinção entre as amostras. Em relação aos erros sistemáticos, verificou-se nos modelos que o valor de $t_{calculado}$ é menor que $t_{critico}$, o que significa que os erros sistemáticos presentes não são estatisticamente significativos, ou seja, são desprezíveis. Logo, pode-se dizer que a

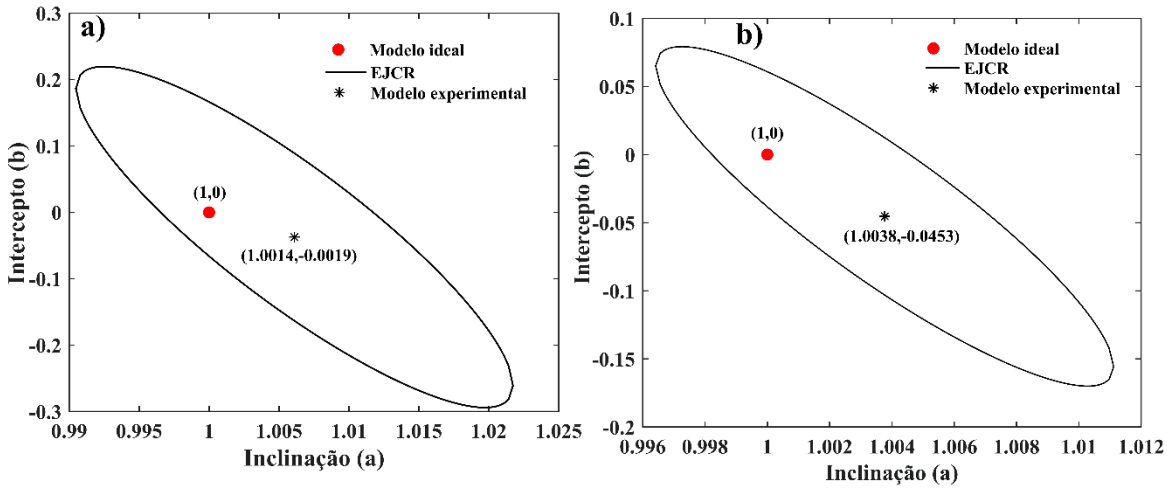
construção de cada modelo apresentado foi realizada de forma adequada desde a fase experimental até ao cálculo de figuras de mérito para a validação, conforme os procedimentos analíticos recomendados pela norma (ASTM E1655-05, 2012).

Os valores da razão sinal-ruído do modelo mostram o quanto o escalar *NAS* está acima do desvio padrão (SDV) da flutuação do sinal espectral. Os valores da seletividade representam a fração média do sinal espectral da amostra que foi retirada, por não ser ortogonal em relação à propriedade de interesse. Os modelos de BMPM e BMM apresentaram valores da seletividade na faixa de 0,10% e 0,17%, respectivamente.

A exatidão também foi analisada pela comparação dos valores obtidos para a inclinação e o intercepto de uma reta ajustada entre valores de referência, iguais a 1 e 0 (valores reais de proporção da espécie de interesse) e os valores de concentração estimados pelo modelo. Tal análise foi efetuada através da elipse de confiança (EJCR – do inglês, *elliptical joint confidence region*). Os modelos foram considerados estatisticamente aceitáveis, porque estão localizados dentro dos limites delimitados pela elipse, no nível de confiança a 95%, como se pode notar na Figura 25 (a) e (b).

O resultado ideal consistindo em inclinação ou declive (a = 1) e intercepto (b = 0) é mostrado por ponto (•), enquanto que o resultado do modelo real (experimental) correspondente a intercepto e inclinação indicados por asterisco (⁕). Se os valores estimados para a inclinação e intercepto estivessem fora de seus intervalos de confiança, dir-se-ia que há indícios de erros sistemáticos proporcionais e constantes, respectivamente.

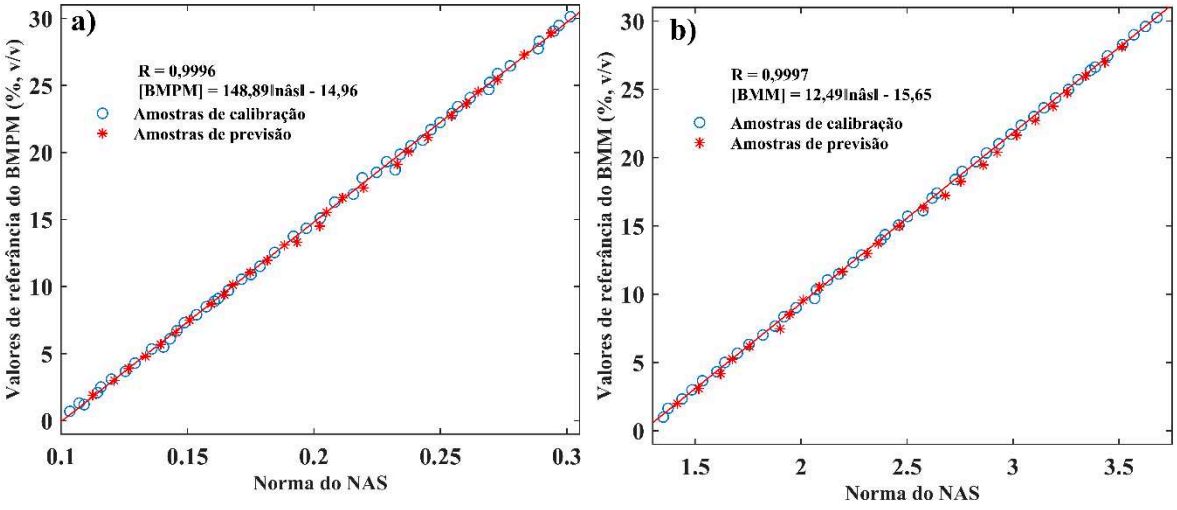
Figura 25 – Região de confiança da articulação elíptica (EJCR) para a inclinação e intercepção da regressão da concentração prevista *vs* os valores de referência dos modelos de (a) BMPM e (b) BMM.



Fonte: O autor.

Finalmente, utilizando o conceito de NAS, discutido anteriormente, foi possível apresentar o modelo multivariado de maneira mais simples por meio de uma curva pseudo-univariada. Esta representação da curva equivale a obter um sinal filtrado univariado, que apresenta uma relação linear com a concentração do analito e é particularmente útil na análise da rotina para o analista visualizar o modelo construído de maneira univariada. Os valores da norma do NAS sãos equivalentes aos sinais seletivos obtidos de cada vetor de espectro da amostra. A curvas pseudo-univariadas para os métodos propostos são mostradas na Figura 26 juntamente com a Equação de ajuste e o coeficiente de correlação.

Figura 26 – Curva de calibração pseudo-univariada. Plotagem das normas do NAS *vs* os valores de referência para as amostras de calibração e de previsão dos modelos do (a) BMPM e (b) BMM



Fonte: O autor.

4.6. Construção de Modelos PLS-DA Globais para Classificar Amostras de acordo com o Teor do BMPM e do BMM em Misturas com Diesel

Como se disse anteriormente, modelos PLS-DA usados neste trabalho para classificar o teor das amostras de biodiesel/diesel, foram utilizadas três (3) classes: a classe 1 refere-se às misturas BX cujo teor de biodiesel é abaixo de 10% (v/v); a classe 2 são as amostras B10 (classe de interesse); e a classe 3 são as misturas BX com teor de biodiesel acima de 10% (v/v).

As amostras anômalas (*outliers*) foram identificadas usando os procedimentos anteriormente descritos, e elas foram retiradas antes da construção de cada modelo. O número de amostras de treinamento e de teste, o número de VLs e a variância capturada nos blocos **X** e **y** de cada modelo são apresentados na Tabela 10.

Tabela 10 – Números de amostras e parâmetros usados na construção dos modelos PLS-DA globais do BMPM e BMM

	Nº. de amostras			**Variância acumulada**		**Erros**		
Modelo	Treinamento	Teste	VLs	Bloco **X**	Bloco **y**	RMSEC	RMSECV	RMSEP
BMPM	82	42	3	98,40%	72,49%	0,26	0,32	0,28
BMM	73	37	4	99,70%	78,28%	0,27	0,32	0,27

Fonte: O autor.

As amostras dos conjuntos de treinamento foram utilizadas para a construção do modelo e as amostras dos conjuntos de teste fora utilizada para a validação, isto é, para testar se o modelo construído consegue prever bem amostras desconhecidas. Nesse caso, observando o gráfico das estimativas dos modelos PLS-DA globais na Figura 27 e na Tabela de contingência (Tabela 11), verifica-se que:

- No modelo do BMPM, uma (1) amostra da classe 3 do conjunto de treinamento foi prevista erroneamente como não pertencente a esta classe, pertencendo assim a classe 1; porém, no conjunto de teste todas as amostras foram corretamente classificadas nas suas respectivas classes.
- No modelo do BMM, no conjunto de treinamento, uma (1) amostra da classe 2 e outra da classe 3 foram erroneamente classificadas como sendo das classes 1 e 2, respectivamente. Já no conjunto de teste, duas (2) amostras da classe 2 foram erroneamente classificadas como pertencente a classe 3.

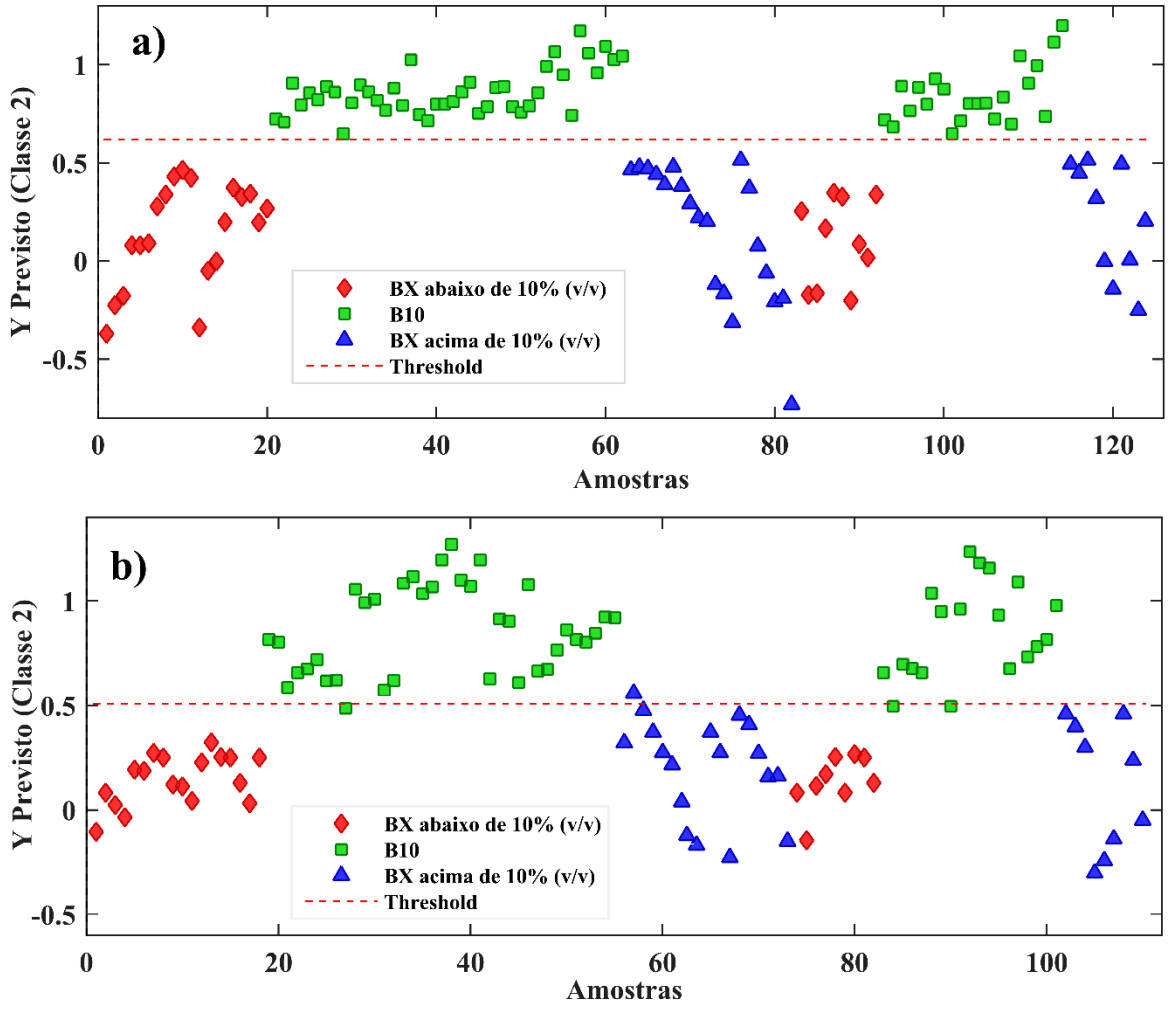
Na Tabela 11, pode-se notar que algumas amostras não foram corretamente previstas nas suas respectivas classes. O uso de espectros pode ter contribuído para os erros apresentados no modelo, porque os espectros totais apresentam algumas regiões não informativas.

Tabela 11 − Dados da Tabela de contingência para os resultados dos modelos PLS-DA globais para classificar as amostras do BMPM e BMM em misturas com diesel

	Conjunto de Treinamento						**Conjunto de Teste**					
	BMPM			BMM			BMPM			BMM		
	Classe atual			Classe atual			Classe atual			Classe atual		
	1	2	3	1	2	3	1	2	3	1	2	3
Previsto como 1	20	0	1	18	1	0	10	0	0	9	0	0
Previsto como 2	0	42	0	0	36	1	0	22	0	0	17	0
Previsto como 3	0	0	19	0	0	17	0	0	10	0	2	9

Fonte: O autor.

Figura 27 – Estimativas dos modelos PLS-DA globais para a classificação das amostras de B10 e BX do (a) BMPM e (b) BMM.

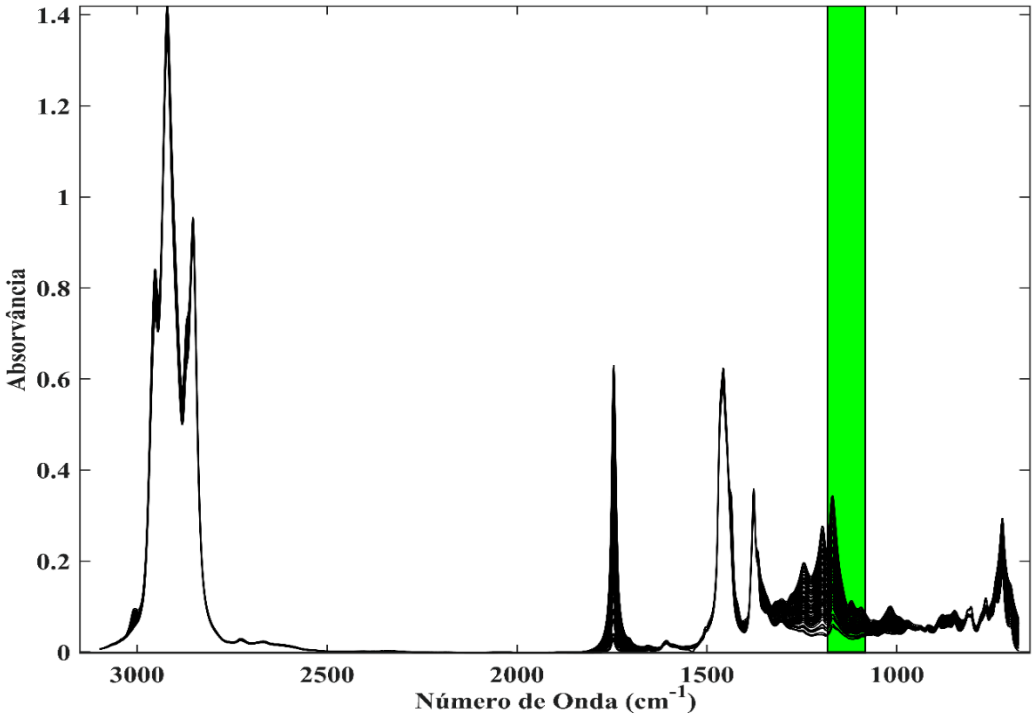


Fonte: O autor.

4.7. Construção de Modelos de Seleção de Variáveis por iPLS-DA para Classificar Amostras de acordo com o Teor do BMPM e do BMM em Misturas com Diesel

Para melhorar os resultados dos modelos PLS-DA globais, os modelos de seleção de variáveis por iPLS-DA foram construídos. Usando o algoritmo iPLS, os espectros totais do BMPM e do BMM foram divididos em 24 e em 28 regiões equidistantes, respectivamente, e então, em cada intervalo com menor valor do RMSECV do PLS-DA global um modelo PLS-DA foi desenvolvido. Finalmente, o intervalo com os menores erros e com melhores resultados de classificação foi selecionado para construir o modelo iPLS-DA. A faixa espectral selecionada e que apresentou melhores resultados para o iPLS-DA$_{24}$ é mostrada na Figura 28.

Figura 28 – Intervalo selecionado (1183 – 1083 cm^{-1}) na construção do modelo iPLS-DA_{24} para classificar amostras do BMPM em mistura com diesel

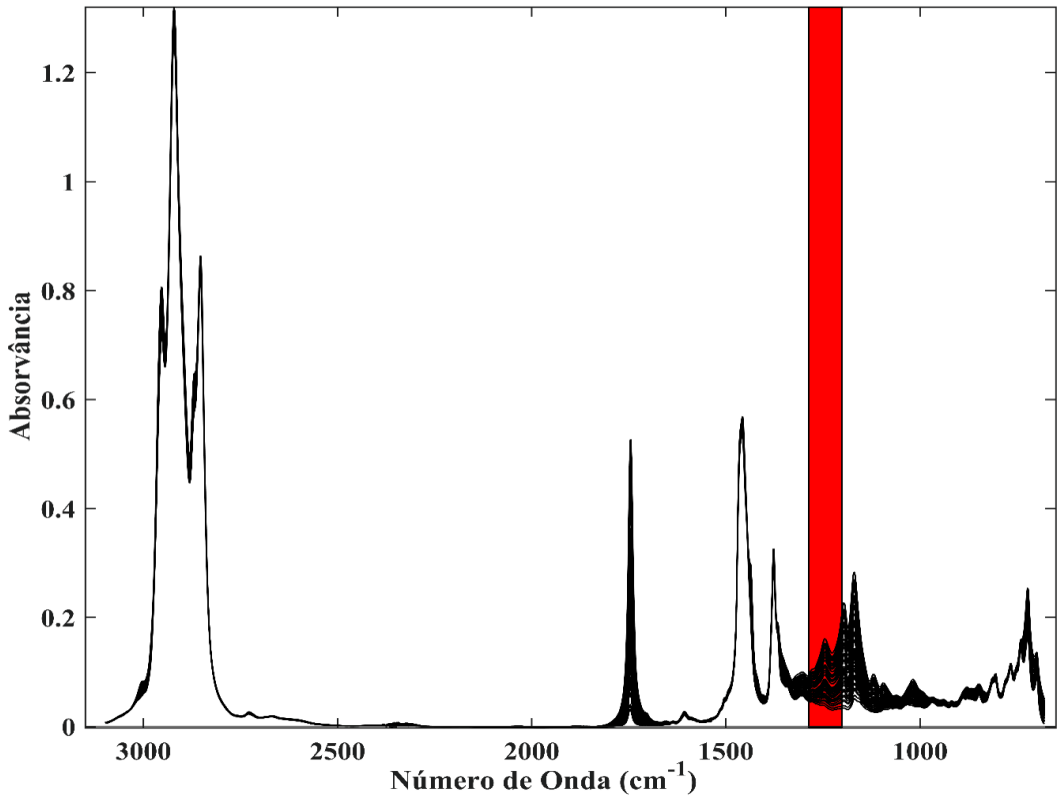


Fonte: O autor.

Esta região constituída por 101 variáveis foi usada para classificar as amostras de acordo com o teor do biodiesel metílico de pinhão manso no diesel. Ela corresponde à região entre 1183 – 1083 cm^{-1} (com o máximo de absorção em 1170 cm^{-1}) e refere-se à deformação axial simétrica da ligação C–O do grupo C–O(OR) dos ésteres metílicos alifáticos (SILVERSTEIN et al., 2019) presentes nas moléculas do biodiesel.

Na Figura 29, a região compreendida ente 1287 – 1201cm^{-1} é constituída por 87 variáveis e corresponde à vibração de deformação axial assimétrico do sistema da ligação C–O–C ligado ao grupo C–C(=O)–O também presente nas moléculas dos ésteres metílicos alifáticos (SILVERSTEIN et al., 2019). Este intervalo foi usado para classificar as amostras do biodiesel metílico de moringa em misturas com diesel.

Figura 29 – Intervalo selecionado (1287 – 1201cm^{-1}) na construção do modelo iPLS-DA$_{28}$ para classificar amostras do BMM em mistura com diesel



Fonte: O autor.

Numa simples análise visual, nas duas regiões selecionadas para a construção dos modelos, os espectros apresentam semelhanças entre eles e as bandas aumentam de intensidade à medida que o teor do biodiesel aumenta na amostra. Na seleção de variáveis por intervalos também foi feita a análise de *outliers* antes da construção dos modelos. O modelo de classificação do teor do BMPM (iPLS-DA$_{24}$) foi construído com 3VLs; as variâncias totais acumuladas nos blocos **X** e **y** foram 99,99% e 83,95%, respectivamente, enquanto que o modelo de classificação do BMM foi construído com 4 VL, e as variâncias acumuladas para os blocos **X** e **y** foram 99,70% e 89,10%, respectivamente. A Tabela 12 apresenta os dados da Tabela de contingência dos dois modelos construídos. Nessa tabela pode-se notar que todas as amostras foram corretamente previstas nas suas respectivas classes.

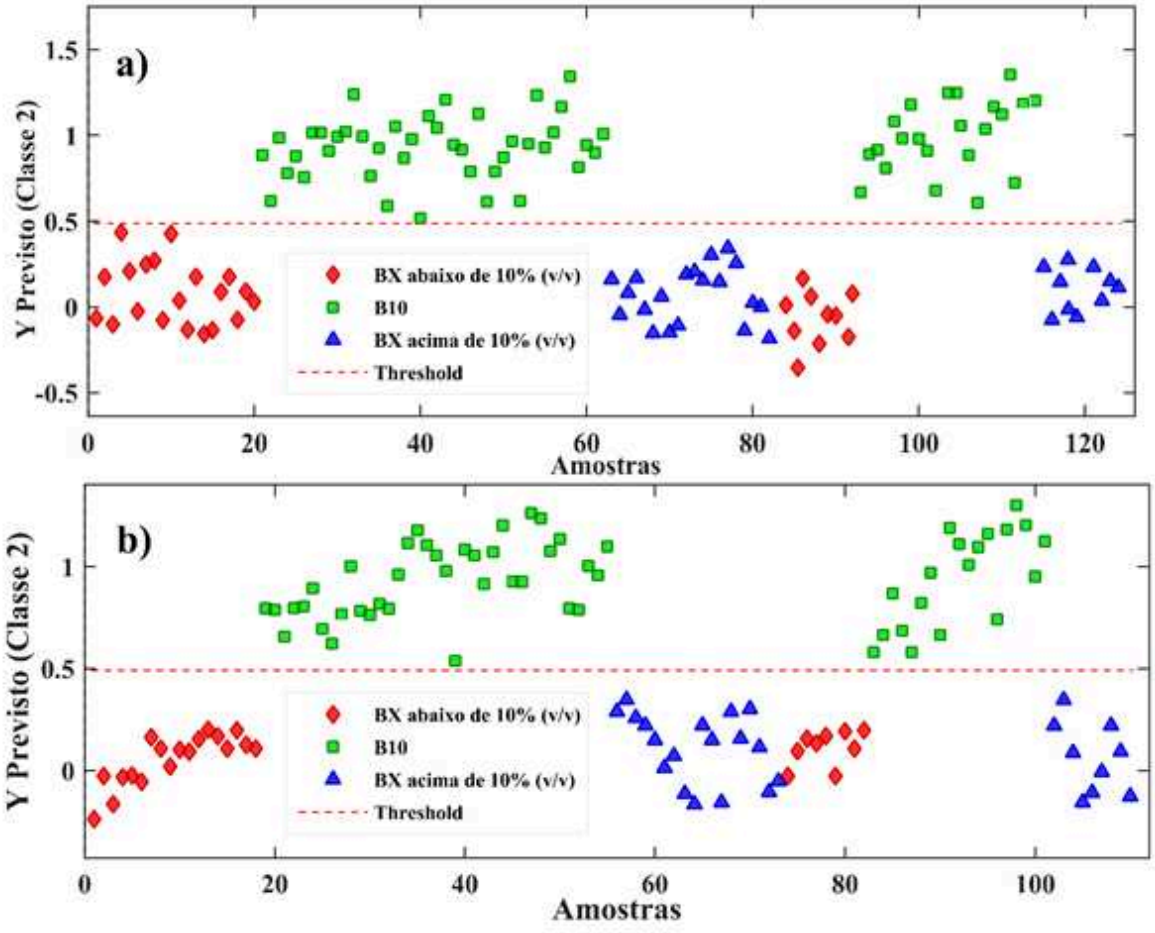
Tabela 12 − Dados da Tabela de contingência para os resultados dos modelos iPLS-DA_{24} iPLS-DA_{28} para classificar as amostras do BMPM e BMM em misturas com diesel

	Conjunto de Treinamento						**Conjunto de Teste**					
	BMPM			BMM			BMPM			BMM		
	Classe atual			Classe atual			Classe atual			Classe atual		
	1	2	3	1	2	3	1	2	3	1	2	3
Previsto como 1	20	0	0	18	0	0	10	0	0	9	0	0
Previsto como 2	0	42	0	0	37	0	0	22	0	0	19	0
Previsto como 3	0	0	20	0	0	18	0	0	10	0	0	9

Fonte: O autor.

Na Figura 30, observa que o valores de *threshold* determinados pelo teorema de Bayes para os modelos do (a) BMPM e (b) BMM são 0,4998 e 0,4999, respectivamente. Isso significa que todas as amostras em que o valor previsto, $\hat{\boldsymbol{y}}$, foi maior do que do *threshold* foram classificadas como verdadeiras positivas, e as que apresentaram valores do $\hat{\boldsymbol{y}}$ previsto menor que o *threshold* foram classificadas como verdadeiras negativas. As amostras que se encontram acima do *threshold* (■) são as de interesse (B10 de treinamento e de teste). As amostras da classe 1 representadas por (♦) e por (▲) são de BX (de treinamento e de teste) com teor de biodiesel abaixo e acima de 10% (v/v), respectivamente. Nos dois modelos, verificou-se que todas as amostras foras corretamente classificadas nas suas respectivas classes.

Figura 30 – Estimativas dos modelos iPLS-DA para a classificação das amostras de B10 e BX do (a) BMPM e (b) BMM



Fonte: O autor.

4.8. Validação Analítica dos Métodos de Seleção de Variáveis por iPLS-DA

Para avaliar a eficiência do modelo de classificação por seleção de variáveis, iPLS-DA_{24} e iPLS-DA_{28}, a acurácia, a sensibilidade, a especificidade e o CCM foram aplicadas a partir das respostas dos parâmetros verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo, conforme de se pode observar na Tabela 13. Os resultados para todas as 3 classes, tanto nos conjuntos de treinamento quanto nos conjuntos de teste, quando comparados com os modelos do PLS-DA globais (vide Tabela 13), apresentam baixos erros. Os seus valores das figuras de mérito são iguais a 1,0 em todas as classes. Estes resultados atestam a boa eficiência dos modelos de seleção de variáveis por iPLS-DA construídos para a classificação das amostras em relação ao teor do biodiesel em diesel.

Tabela 13 – Parâmetros de classificação obtidos pelos modelos iPLS-DA para o teor do BMPM e BMM em misturas BX

Parâmetros	Conjunto	Modelo do BMPM			Modelo do BMM		
		Classe 1	Classe 2	Classe 3	Classe 1	Classe 2	Classe 3
RMSEC	Treinamento	0,11	0,12	0,11	0,09	0,10	0,10
RMSECV	Treinamento	0,12	0,13	0,11	0,11	0,12	0,12
RMSEP	Teste	0,12	0,13	0,12	0,10	0.11	0,11
Especificidade	Treinamento	1,0	1,0	1,0	1,0	1,0	1,0
	Teste	1,0	1,0	1,0	1,0	1,0	1,0
Sensibilidade	Treinamento	1,0	1,0	1,0	1,0	1,0	1,0
	Teste	1,0	1,0	1,0	1,0	1,0	1,0
Acurácia	Treinamento	1,0	1,0	1,0	1,0	1,0	1,0
	Teste	1,0	1,0	1,0	1,0	1,0	1,0

Fonte: O autor.

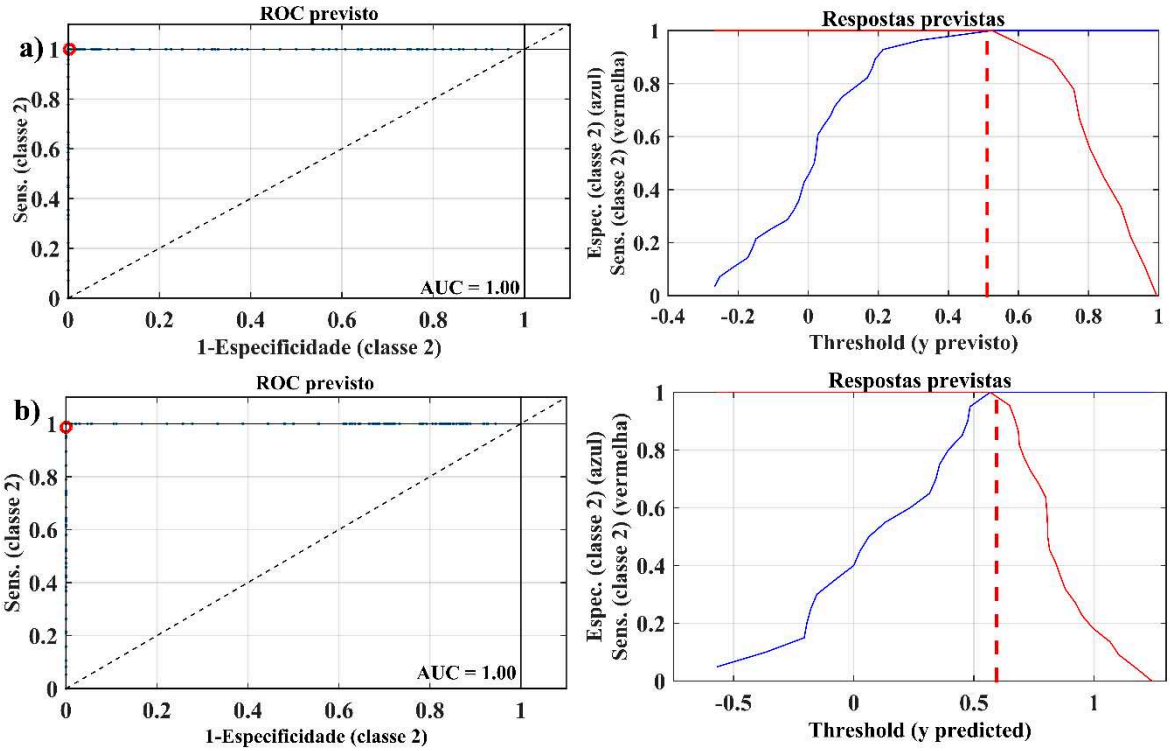
Outras ferramentas gráficas usadas para avaliar a capacidade de classificação dos modelos de seleção de variáveis são as curvas Característica de Operação do Receptor (ROC – do inglês, *Receiver Operating Characteristics*) apresentadas na Figura 31. Estas curvas, calculadas separadamente apenas para a classe de interesse, são gráficos da especificidade *vs* a sensibilidade. O número de variáveis usadas para a construção das curvas ROC é o mesmo usado para a construção de cada modelo. No modelo de classificação das amostras do BMPM foram usadas 101 variáveis e no modelo do BMM foram usadas 87 variáveis.

Como o eixo *x* é "1 – Especificidade", o valor ideal deste é igual a 0, enquanto para o eixo *y* o valor ideal é igual a 1 (sensibilidade). Portanto, um método de classificação perfeito fornece um ponto no canto superior esquerdo do espaço ROC, o que representa máxima sensibilidade e especificidade. Esse valor foi alcançado em todos os modelos.

A área sob a curva ROC (AUC – do inglês, *Area Under the ROC Curve*) mede a área bidimensional inteira debaixo da curva ROC (a integral) de (0,0) a (1,1) cujo valor varia entre 0 e 1. Como os modelos têm previsão de 100% correta, AUC é igual a 1. Ainda na Figura 31 pode observar-se outro gráfico que corresponde a variação do *threshold* para a classe de interesse. De acordo com o teorema de Bayes, este valor limite deve ser escolhido de forma a minimizar os falsos negativos e falsos positivos, isto é, sensibilidade e especificidade devem ser idealmente 1. A sensibilidade para valores de *threshold* baixos é sempre igual a 1, isso porque as amostras VP estarão acima deste limite. No entanto,

threshold muito baixo diminui a especificidade, isto é, aumenta o número de FP. Em casos de limites muito elevados, ocorre o inverso: as amostras VN estão abaixo do limite, mas aumentam-se os casos de FN e, portanto, diminui-se a sensibilidade. Assim, o valor de *threshold* ideal corresponde ao ponto onde a linha da especificidade (Espec.) (em azul) cruza com a linha da sensibilidade (sens.) (em vermelho).

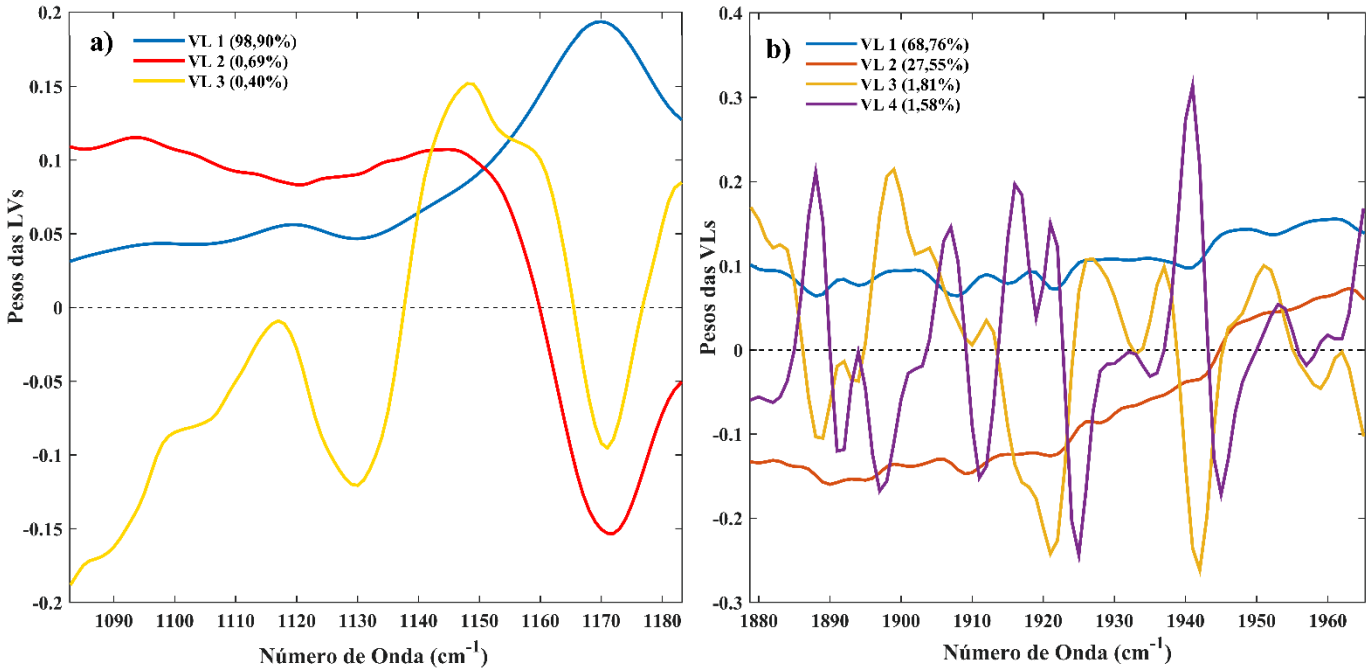
Figura 31 – Curvas ROC e relação entre especificidade (....) e sensibilidade (....) para a escolha do *threshold* da classe de interesse para os modelos do (a) BMPM e (b) BMM



Fonte: O autor.

A partir das informações fornecidas pelos espectros MIR foi possível analisar a importância das variáveis originais (os pesos das VLs) das bandas usadas no desenvolvimento dos modelos iPLS-DA. Os pesos das variáveis latentes mostram as VLs e as porcentagens das variáveis capturadas que influenciaram a discriminação das classes das amostras. Em cada modelo, a primeira VL tem maior porcentagem de variância capturada. Essas variáveis são de cada banda espectral usada para a construção de cada modelo de seleção de variáveis, conforme se pode notar na Figura 32 (a) para o modelo do BMPM e (b) para o modelo do BMM.

Figura 32 - Gráfico dos pesos das VLs *vs* o número de onda de espectros MIR usados para a construção dos modelos do (a) BMPM e do (b) BMM



5. CONCLUSÕES

A aplicação de métodos quimiométricos de calibração multivariada, de classificação supervisionada e de seleção de variáveis, permitiu a construção e desenvolvimento de modelos para a quantificação dos teores do biodiesel metílico de pinhão manso e de moringa e classificar as suas amostras biodiesel/diesel na faixa de concentração de 0,50 – 30% (v/v), dentro dos padrões exigidos pela ANP.

Para quantificar o teor do biodiesel usando métodos de seleção de variáveis (iPLS, biPLS e siPLS), foi possível dividindo os espectros totais em regiões equidistantes e construir um modelo PLS em cada subintervalo. Para quantificar o biodiesel metílico de pinhão manso e de moringa em misturas com diesel foram usadas variáveis das faixas de 779 – 680 cm^{-1} e de 1809 – 1723 cm^{-1}, respectivamente. Os modelos propostos para a quantificação foram validados conforme orientações da norma ASTM E1655-05 e atenderam às exigências da norma ABNT NBR 15568 em relação ao parâmetro de exatidão; têm excelente linearidade e não apresentaram erros sistemáticos. Fez-se o teste *F* para comparar o desempenho dos modelos PLS globais e os de seleção de variáveis; alguns métodos de seleção de variáveis apresentaram diferenças estatísticas significativas, uma vez que os valores de $F_{calculado}$ são maiores que $F_{tabelado}$. Os modelos que apresentaram os melhores resultados foram validados.

Para classificar amostras de acordo com os teores do biodiesel metílico de pinhão manso e de moringa em misturas com diesel foram usadas variáveis das faixas de 1183 – 1083 cm^{-1} e de 1287 – 1201cm^{-1}, respectivamente. Em cada tipo de biodiesel, o modelo de classificação por seleção de variáveis (iPLS-DA) classificou corretamente todas as amostras de acordo com o teor do biodiesel em diesel. Estes modelos de classificação foram validados usando parâmetros de qualidade como CCM, acurácia, sensibilidade e especificidade através de valores de falso positivo, falso negativo, verdadeiro positivo, verdadeiro negativo.

Assim, o uso de métodos quimiométricos aliada a técnica de espectrometria MIR apresenta-se como uma solução viável no que tange às análises na área de biocombustíveis. Esta aplicabilidade tem se dado graças às vantagens da técnica como: rapidez na obtenção de espectros, exigência de mínima preparação de amostras, uso de pouca quantidade da amostra, instrumentação com portabilidade, dentre outras. Assim, a técnica de espectrometria MIR aliada aos métodos de calibração multivariada e seleção de variáveis por intervalos, mostrou-se como ferramenta promissora e pode ser empregada para o controle quantitativo e qualitativo destes combustíveis do biodiesel metílico de pinhão manso e de moringa.

REFERÊNCIAS

ABNT - ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 15568**: Biodiesel - Determination of biodiesel content in diesel fuel oil by mid infrared spectroscopy. Rio de Janeiro, 2008. https://www.target.com.br/produtos/normas-tecnicas/40641/nbr15568-biodiesel-determinacao-do-teor-de-biodiesel-em-oleo-diesel-por-espectroscopia-na-regiao-do-infravermelho-medio

ALBRECHT, L. P. et al. Teores de óleo, proteínas e produtividade de soja em função da antecipação da semeadura na região oeste do Paraná. Bragantia, Campinas, v. 67, n. 4, p. 865-873, 2008. https://doi.org/10.1590/S0006-87052008000400008

ALVES, A. S. H. **A Sustentabilidade do Biodiesel em Moçambique :** uma análise integrada das dimensões institucional, social, alimentar, ambiental e energética. 2014. 293 f. Dissertação (Mestrado em Desenvolvimento Sustentável) – Centro de Desenvolvimento Sustentável, Universidade de Brasília, Brasília, 2014. https://repositorio.unb.br/handle/10482/16397

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

ANP – AGÊNCIA NACIONAL DO PETRÓLEO, GÁ NATURAL E BIOCOMBUSTÍVEIS (Brasil). **Percentual Obrigatório de Biodiesel passa para 10%**. Disponível em: <http://www.anp.gov.br/noticias/4333-percentual-obrigatorio-de-biodiesel-passa-para-10>. Acesso em: 28 dezembro. 2018.

ANP – Agência Nacional do Petróleo, Gá Natural e Biocombustíveis (Brasil). **Boletim de Monitoramento da Qualidade dos Combustíveis**. Disponível em: <http://www.anp.gov.br/publicacoes/boletins-anp/boletim-de-monitoramento-da-qualidade-dos-combustiveis>. Acesso em: 19 jul. 2019.

ANWAR, F. et al. Moringa oleifera: a food plant with multiple medicinal uses. **Phytotherapy Research**, v. 21, n. 1, p. 17–25, 2007. https://doi.org/10.1002/ptr.2023

APROBIO – Associação dos Produtores de Biocombustíveis do Brasil. **Matérias-primas alternativas batem recorde de participação no biodiesel**. Disponível em: <https://aprobio.com.br/2018/12/26/materias-primas-alternativas-batem-recorde-de-participacao-no-biodiesel/>. Acesso em: 17 jun. 2019.

ASTM – AMERICAN SOCIETY FOR TESTING AND MATERIALS: Standard Practices for Infrared Multivariate Quantitative Analysis. **ASTM E1655-05 International**, West Conshohocken, 2012. https://doi.org/10.1520/E1655-05R12

ASTM – AMERICAN SOCIETY FOR TESTING AND MATERIALS: Standard Test Method for Determination of Biodiesel (Fatty Acid Methyl Esters) Content in Diesel Fuel Oil Using Mid Infrared Spectroscopy (FTIR-ATR-PLS Method). **ASTM D7371 International**, West Conshohocken, 2014. https://doi.org/10.1520/D7371-12

BABYAK, M. A. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. **Psychosomatic Medicine**, v. 66, n. 3, p. 411–421, 2004. https://doi.org/10.1097/01.psy.0000127692.23278.a9

BAETEN, V.; DARDENNE, P. Spectroscopy: Developments in instrumentation and analysis. **Grasas y Aceites**, v. 53, n. 1, p. 45–63, 2002. https://doi.org/10.3989/gya.2002.v53.i1.289

BALLABIO, D.; CONSONNI, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. **Analytical Methods**, v. 5, n. 16, p. 3790–3798, 2013. https://doi.org/10.1039/c3ay40582f

BAMBO, T. F. **Cooperação Sul-Sul:** o acordo Brasil-Moçambique na área de biocombustíveis. 2014. 113 f. Dissertação (Mestrado em Ciências – Programa de Relações Internacionais) – Instituto de Relações Internacionais, Universidade de São Paulo, São Paulo, 2014.

https://doi.org/10.11606/D.101.2014.tde-26052014-111753

BARKER, M.; RAYENS, W. Partial least squares for discrimination. **Journal of Chemometrics**, v. 17, n. 3, p. 166–173, 2003. https://doi.org/10.1002/cem.785

BENTO VENÂNCIO. **Chegou a fábrica de processamento da moringa**. Disponível em: <https://www.jornaldomingo.co.mz/index.php/reportagem/4306-chegou-a-fabrica-de-processamento-da-moringa>. Acesso em: 17 set. 2019.

BERGMANN, J. C. et al. Biodiesel production in Brazil and alternative biomass feedstocks. **Renewable and Sustainable Energy Reviews**, v. 21, p. 411–420, 2013. https://doi.org/10.1016/j.rser.2012.12.058

BOQUÉ, R.; FABER, N. (KLAAS) M.; RIUS, F. X. Detection limits in classical multivariate calibration models. **Analytica Chimica Acta**, v. 423, n. 1, p. 41–49, 2000. https://doi.org/10.1016/S0003-2670(00)01101-6

BÓRAWSKI, P. et al. Development of renewable energy sources market and biofuels in The European Union. **Journal of Cleaner Production**, v. 228, p. 467–484, 2019. https://doi.org/10.1016/j.jclepro.2019.04.242

BRANDÃO, L. F. P.; BRAGA, J. W. B.; SUAREZ, P. A. Z. Determination of vegetable oils and fats adulterants in diesel oil by high performance liquid chromatography and multivariate methods. **Journal of Chromatography A**, v. 1225, p. 150–157, 2012. https://doi.org/10.1016/j.chroma.2011.12.076

BRASIL. Lei nº. 13.263, de 23 de março de 2016. Altera a Lei nº 13.033, de 24 de setembro de 2014, para dispor sobre os percentuais de adição de biodiesel ao óleo diesel comercializado no território nacional. **Diário Oficial [da] República Federativa do Brasil**, Poder Executivo,Brasília, DF, 24 de março de de 2016, v. 153, n. 57, p. 94.

BRUNS, R. E.; FAIGLE, J. F. G. Quimiometria. **Química Nova**, p. 84–98, 1984.

BUIATTE, J. E. et al. Qualitative and Quantitative Monitoring of Methyl Cotton Biodiesel Content in Biodiesel/Diesel Blends Using MIR Spectroscopy and Chemometrics Tools. **Journal of the Brazilian Chemical Society**, v. 27, n. 1, p. 84–90, 2015. https://doi.org/10.5935/0103-5053.20150251

BURROWS, J. et al. **Trees and Shrubs of Mozambique**. Collectors ed. Cape Town: Publishing Print Matters (Pty), 2018.

CASTILHO-ALMEIDA, E. W. et al. Estudo teórico e experimental de espectros infravermelho de ésteres de ácido graxo presentes na composição do biodiesel de soja. **Química Nova**, v. 35, n. 9, p. 1752–1757, 2012. https://doi.org/10.1590/S0100-40422012000900009

CHONG, I.-G.; JUN, C.-H. Performance of some variable selection methods when multicollinearity is present. **Chemometrics and Intelligent Laboratory Systems**, v. 78, n. 1–2, p. 103–112, 2005. https://doi.org/10.1016/j.chemolab.2004.12.011

COSTA NETO, P. R. et al. Produção de biocombustível alternativo ao óleo diesel através da transesterificação de óleo de soja usado em frituras. **Química Nova**, v. 23, n. 4, p. 531–537, 2000. https://doi.org/10.1590/S0100-40422000000400017

CUVILAS, C. A.; JIRJIS, R.; LUCAS, C. Energy situation in Mozambique: A review. **Renewable and Sustainable Energy Reviews**, v. 14, n. 7, p. 2139–2146, 2010. https://doi.org/10.1016/j.rser.2010.02.002

DA SILVA CÉSAR, A. et al. Competitiveness analysis of "social soybeans" in biodiesel production in Brazil. **Renewable Energy**, v. 133, p. 1147–1157, 2019. https://doi.org/10.1016/j.renene.2018.08.108

DA SILVA, J. P. V. et al. Moringa oleifera oil: Studies of characterization and biodiesel production. **Biomass and Bioenergy**, v. 34, n. 10, p. 1527–1530, 2010.

https://doi.org/10.1016/j.biombioe.2010.04.002

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

DE OLIVEIRA, F. C.; COELHO, S. T. History, evolution, and environmental impact of biodiesel in Brazil: A review. **Renewable and Sustainable Energy Reviews**, v. 75, p. 168–179, 2017. https://doi.org/10.1016/j.rser.2016.10.060

DEMIRBAS, A. Progress and recent trends in biodiesel fuels. **Energy Conversion and Management**, v. 50, n. 1, p. 14–34, 2009. https://doi.org/10.1016/j.enconman.2008.09.001

DU, Y. P. et al. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. **Analytica Chimica Acta**, v. 501, n. 2, p. 183–191, 2004. https://doi.org/10.1016/j.aca.2003.09.041

DUTTA, A. Fourier Transform Infrared Spectroscopy. In: **Spectroscopic Methods for Nanomaterials Characterization**. Elsevier, 2017. p. 73–93. https://doi.org/10.1016/B978-0-323-46140-5.00004-2

FABER, N. M.; RAJKÓ, R. How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative. **Analytica Chimica Acta**, v. 595, n. 1–2, p. 98–106, 2007. https://doi.org/10.1016/j.aca.2007.05.030

FAHEY, J. W. Moringa oleifera: A Review of the Medical Evidence for Its Nutritional, Therapeutic, and Prophylactic Properties. Part 1. **Trees for Life Journal**. v. 1, n. 5, p. 157-164, 2005. https://doi.org/10.1201/9781420039078.ch12

FERNANDES, D. D. S. et al. UV-Vis Spectrometric Detection of Biodiesel/Diesel Blend Adulterations with Soybean Oil. **Journal of the Brazilian Chemical Society**, v. 25, n. 3, p. 230–231, 2013. https://doi.org/10.5935/0103-5053.20130259

FERREIRA, M. M. C. et al. Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, v. 22, n. 5, p. 724–731, 1999. https://doi.org/10.1590/S0100-40421999000500016

FERREIRA, M. M. C. **Quimiometria: conceitos, métodos e aplicações**. Campinas, SP: Editora da Unicamp, 2015. 496 p. https://doi.org/10.7476/9788526814714

GEISSDOERFER, M.; VLADIMIROVA, D.; EVANS, S. Sustainable business model innovation: A review. **Journal of Cleaner Production**, v. 198, p. 401–416, 2018. https://doi.org/10.1016/j.jclepro.2018.06.240

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

GOLBRAIKH, A.; TROPSHA, A. Beware of q2! **Journal of Molecular Graphics and Modelling**, v. 20, n. 4, p. 269–276, 2002. https://doi.org/10.1016/S1093-3263(01)00123-1

GONTIJO, L. C. et al. Development and Validation of PLS Models for Quantification of Biodiesels Content from Waste Frying Oil in Diesel by HATR-MIR. **Revista Virtual de Química**, v. 6, n. 5, p. 1517–1528, 2014a. https://doi.org/10.5935/1984-6835.20140098

GONTIJO, L. C. et al. Quantification of soybean biodiesels in diesel blends according to ASTM E1655 using mid-infrared spectroscopy and multivariate calibration. **Fuel**, v. 117, n. PART B, p. 1111–1114, 2014b. https://doi.org/10.1016/j.fuel.2013.10.043

GONTIJO, L. C. **Uso da Espectrometria no Infravermelho Médio, Calibração Multivariada e Seleção de variáveis por Intervalos na Quantificação de biodieseis em Misturas com Diesel**. 2016. 104 f. Tese (Doutorado em Química) - Instituto de Química, Universidade Federal de Uberlândia, Uberlândia, 2016. https://repositorio.ufu.br/handle/123456789/17536

GONZÁLEZ, A. G.; HERRADOR, M. A.; ASUERO, A. G. Intra-laboratory testing of method accuracy from recovery assays. **Talanta**, v. 48, n. 3, p. 729–736, 1999. https://doi.org/10.1016/S0039-9140(98)00271-9

GRIFFITHS, P. R.; DE HASETH, J. A. **Fourier Transform Infrared Spectrometry**. Second ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., v. 42, 2007. 557 p. https://doi.org/10.1002/047010631X

GUIMARÃES, E. et al. Quantification of Ethanol in Biodiesels Using Mid-Infrared Spectroscopy and Multivariate Calibration. **Industrial & Engineering Chemistry Research**, v. 53, n. 35, p. 13575–13580, 2014. https://doi.org/10.1021/ie502067h

GUIMARÃES, E. et al. Infrared Spectroscopy and Multivariate Calibration for Quantification of Soybean Oil as Adulterant in Biodiesel Fuels. **Journal of the American Oil Chemists' Society**, v. 92, n. 6, p. 777–782, 2015. https://doi.org/10.1007/s11746-015-2656-x

HABTEMARIAM, S. Moringa stenopetala —Botanical and Ecological Perspectives. In: **The African and Arabian Moringa Species**. Elsevier, 2017. p. 3–12. https://doi.org/10.1016/B978-0-08-102286-3.00001-4

JAIN, S. The production of biodiesel using Karanja (Pongamia pinnata) and Jatropha (Jatropha curcas) Oil. In: **Biomass, Biopolymer-Based Materials, and Bioenergy**. Elsevier, 2019. p. 397–408. https://doi.org/10.1016/B978-0-08-102426-3.00017-5

JIANG, H. et al. Measurement of process variables in solid-state fermentation of wheat straw using FT-NIR spectroscopy and synergy interval PLS algorithm. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 97, p. 277–283, 2012. https://doi.org/10.1016/j.saa.2012.06.024

KAMEL, D. A. et al. Smart utilization of jatropha (Jatropha curcas Linnaeus) seeds for biodiesel production: Optimization and mechanism. **Industrial Crops and Products**, v. 111, p. 407–413, 2018. https://doi.org/10.1016/j.indcrop.2017.10.029

KARMAKAR, R. et al. Production of biodiesel from unused algal biomass in Punjab, India. **Petroleum Science**, 2018. https://doi.org/10.1007/s12182-017-0203-0

KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. **Technometrics**, v. 11, n. 1, p. 137–148, 1969. https://doi.org/10.1080/00401706.1969.10490666

KNOTHE, G. Analytical Methods Used in the Production and Fuel Quality Assessment of Biodiesel. **Transactions of the ASAE**, v. 44, n. 2, p. 193–200, 2001. https://doi.org/10.13031/2013.4740

KNOTHE, G.; GERPEN, J. VAN; KRAHL, J. **The Biodiesel Handbook**. United States of America: Elsevier, 2010. 286 p. https://doi.org/10.1016/C2015-0-02453-4

LIN, J.-J.; CHEN, Y.-W. Production of biodiesel by transesterification of Jatropha oil with microwave heating. **Journal of the Taiwan Institute of Chemical Engineers**, v. 75, p. 43–50, 2017. https://doi.org/10.1016/j.jtice.2017.03.034

LINDGREN, F. et al. Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms. **Journal of Chemometrics**, v. 8, n. 5, p. 349–363, 1994. https://doi.org/10.1002/cem.1180080505

LÔBO, I. P.; FERREIRA, S. L. C.; CRUZ, R. S. DA. Biodiesel: parâmetros de qualidade e métodos analíticos. **Química Nova**, v. 32, n. 6, p. 1596–1608, 2009. https://doi.org/10.1590/S0100-40422009000600044

LORBER, A. Error propagation and figures of merit for quantification by solving matrix equations. **Analytical Chemistry**, v. 58, n. 6, p. 1167–1172, 1986. https://doi.org/10.1021/ac00297a042

MAHMUDUL, H. M. et al. Production, characterization and performance of biodiesel as an alternative fuel in diesel engines – A review. **Renewable and Sustainable Energy Reviews**, v. 72, p. 497–509, 2017. https://doi.org/10.1016/j.rser.2017.01.001

MANI, S.; JAYA, S.; VADIVAMBAL, R. Optimization of Solvent Extraction of Moringa (Moringa Oleifera) Seed Kernel Oil Using Response Surface Methodology. **Food and Bioproducts Processing**, v. 85, n. 4, p. 328–335, 2007. https://doi.org/10.1205/fbp07075

MÁQUINA, A. D. V. et al. Fast quantitative and qualitative monitoring of mafurra biodiesel content using fourier transform mid-infrared spectroscopy, chemometric tools, and variable selection. **Energy and Fuels**, v. 31, n. 1, p. 571–577, 2017a. https://doi.org/10.1021/acs.energyfuels.6b02079

MÁQUINA, A. D. V. et al. Characterization of Biodiesel by Infrared Spectroscopy with Partial Least Square Discriminant Analysis. **Analytical Letters**, v. 50, n. 13, p. 2117–2128, 2017b. https://doi.org/10.1080/00032719.2016.1267186

MÁQUINA, A. D. V. et al. Quantification and classification of cotton biodiesel content in diesel blends, using mid-infrared spectroscopy and chemometric methods. **Fuel**, v. 237, p. 373–379, 2019a. https://doi.org/10.1016/j.fuel.2018.10.011

MÁQUINA, A. D. V. et al. Analysis of 1H NMR spectra of diesel and crambe biodiesel mixtures using chemometrics tools to evaluate the authenticity of a Brazilian standard biodiesel blend. Talanta, p. em prelo, 2019b. https://doi.org/10.1016/j.talanta.2019.120590

MARK, H.; WORKMAN, J. The F Statistic. In: MARK, H.; WORKMAN, J. (Eds.). **Statistics in Spectroscopy**. 2nd. ed. New York, USA: Elsevier, 2003. p. 205–211. https://doi.org/10.1016/B978-012472531-7/50065-2

MARTENS, H.; NAES, T. **Multivariate Calibration**. New York: John & Sons, 1989. https://doi.org/10.1002/0471667196.ess1105

MASSART, D. L. et al. **Handbook of Chemometric and Qualitimetrics, Part B**. Amsterdam: Elsevier, 1998. 713 p. https://doi.org/10.1080/00401706.2000.10486023

MAZIVILA, S. J. et al. Fast classification of different oils and routes used in biodiesel production using mid infrared spectroscopy and PLS2-DA. **Journal of the Brazilian Chemical Society**, v. 26, n. 4, p. 642–648, 2015a. https://doi.org/10.5935/0103-5053.20150020

MAZIVILA, S. J. et al. Fast Detection of Adulterants/Contaminants in Biodiesel/Diesel Blend (B5) Employing Mid-Infrared Spectroscopy and PLS-DA. **Energy & Fuels**, v. 29, n. 1, p. 227–232, 2015b. https://doi.org/10.1021/ef502122w

MAZIVILA, S. J. et al. Discrimination of the type of biodiesel/diesel blend (B5) using mid-infrared spectroscopy and PLS-DA. **Fuel**, v. 142, p. 222–226, 2015c. https://doi.org/10.1016/j.fuel.2014.11.014

MCMURRY, J. **Química Orgânica - Combo:** Tradução da 9ª edição norte-americana. 9ª ed. São Paulo: Cengage Learning, 2017. 1472 p.

MITSUTAKE, H. et al. Multivariate control charts based on NAS and mid-infrared spectroscopy for quality control of B5 blends of methyl soybean biodiesel in diesel. **Journal of Chemometrics**, v. 29, n. 7, p. 411–419, 2015. https://doi.org/10.1002/cem.2720

MME - Ministério de Minas e Energia (Brasil). **Análise de Conjuntura dos Biocombustíveis**. Disponível em: <http://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-402/Análise_de_Conjuntura_Ano 2018.pdf>. Acesso em: 17 set. 2019.

MOFIJUR, M. et al. Properties and use of Moringa oleifera biodiesel and diesel fuel blends in a multi-cylinder diesel engine. **Energy Conversion and Management**, v. 82, p. 169–176, jun. 2014. https://doi.org/10.1016/j.enconman.2014.02.073

MORTON, J. F. The horseradish tree, Moringa pterygosperma (Moringaceae)—A boon to Arid Lands? **Economic Botany**, v. 45, n. 3, p. 318–333, 1991. https://doi.org/10.1007/BF02887070

MOSER, B. R. Efficacy of specific gravity as a tool for prediction of biodiesel–petroleum diesel

blend ratio. **Fuel**, v. 99, p. 254–261, 2012. https://doi.org/10.1016/j.fuel.2012.04.050

MUDOMBI, S. et al. Multi-dimensional poverty effects around operational biofuel projects in Malawi, Mozambique and Swaziland. **Biomass and Bioenergy**, v. 114, p. 41–54, 2018. https://doi.org/10.1016/j.biombioe.2016.09.003

NABI, M. N.; RASUL, M. G. Influence of second generation biodiesel on engine performance, emissions, energy and exergy parameters. **Energy Conversion and Management**, v. 169, p. 326–333, 2018. https://doi.org/10.1016/j.enconman.2018.05.066

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. https://doi.org/10.1366/0003702001949500

OLIVEIRA, J. S. et al. Determination of methyl ester contents in biodiesel blends by FTIR-ATR and FTNIR spectroscopies. **Talanta**, v. 69, n. 5, p. 1278–1284, 2006. https://doi.org/10.1016/j.talanta.2006.01.002

PALIWAL, R.; SHARMA, V.; PRACHETA. A Review on Horse Radish Tree (Moringa oleifera): A Multipurpose Tree with High Economic and Commercial Importance. **Asian Journal of Biotechnology**, v. 3, n. 4, p. 317–328, 2011. https://doi.org/10.3923/ajbkr.2011.317.328

PALOU, A. et al. Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 180, p. 119–126, 2017. https://doi.org/10.1016/j.saa.2017.03.008

PANDEY, V. C. et al. Jatropha curcas: A potential biofuel plant for sustainable environmental development. **Renewable and Sustainable Energy Reviews**, v. 16, n. 5, p. 2870–2883, 2012. https://doi.org/10.1016/j.rser.2012.02.004

PAVIA, D. L. et al. **Introdução à espectriscopia:** Tradução da 5ª edição norte-americana. 2ª. edição ed. São Paulo: Cengage Learning, 2016. 700 p.

PERDOMO, F. A.; MILLÁN, B. M.; ARAGÓN, J. L. Predicting the physical–chemical properties of biodiesel fuels assessing the molecular structure with the SAFT−γ group contribution approach. **Energy**, v. 72, p. 274–290, 2014. https://doi.org/10.1016/j.energy.2014.05.035

PINHO, D. M. M. et al. Evaluating the use of EN 14078 for determination of biodiesel in diesel blends sold in the Brazilian market. **Fuel**, v. 136, p. 136–142, 2014. https://doi.org/10.1016/j.fuel.2014.07.043

PRAMANIK, K. Properties and use of jatropha curcas oil and diesel fuel blends in compression ignition engineRenewable Energy. **Renewable Energy**, v. 28, n. 2, p. 139–248, 2003. https://doi.org/10.1016/S0960-1481(02)00027-7

RAKSHIT, K. D. et al. Toxicity studies of detoxified Jatropha meal (Jatropha curcas) in rats. **Food and Chemical Toxicology**, v. 46, n. 12, p. 3621–3625, 2008. https://doi.org/10.1016/j.fct.2008.09.010

RAMADHAS, A.; JAYARAJ, S.; MURALEEDHARAN, C. Biodiesel production from high FFA rubber seed oil. **Fuel**, v. 84, n. 4, p. 335–340, 2005. https://doi.org/10.1016/j.fuel.2004.09.016

RAMOS, L. P. et al. Biodiesel: Raw Materials, Production Technologies and Fuel Properties. Revista Virtual de Química, v. 9, n. 1, p. 317-369, 2017. https://doi.org/10.21577/1984-6835.20170020

RASHID, U. et al. Moringa oleifera oil: A possible source of biodiesel. **Bioresource Technology**, v. 99, n. 17, p. 8175–8179, 2008. https://doi.org/10.1016/j.biortech.2008.03.066

REDDY, A. N. R. et al. Active Razor Shell CaO Catalyst Synthesis for Jatropha Methyl Ester Production via Optimized Two-Step Transesterification. **Journal of Chemistry**, v. 2017, p. 1–20, 2017. https://doi.org/10.1155/2017/1489218

RINNAN, Å.; SAVORANI, F.; ENGELSEN, S. B. Simultaneous classification of multiple classes in NMR metabolomics and vibrational spectroscopy using interval-based classification methods: iECVA vs iPLS-DA. **Analytica Chimica Acta**, v. 1021, p. 20–27, 2018. https://doi.org/10.1016/j.aca.2018.03.020

RYU, K. The characteristics of performance and exhaust emissions of a diesel engine using a biodiesel with antioxidants. Bioresource Technology, v. 101, n. 1, p. S78-S82, 2010. https://doi.org/10.1016/j.biortech.2009.05.034

SANTOS, D. Q. et al. Evaluation and Characterization of Biodiesels Obtained Through Ethylic or Methylic Transesterification of Tryacylglicerides in Corn Oil. **AIMS Energy**, v. 2, n. 2, p. 183–192, 2014. https://doi.org/10.3934/energy.2014.2.183

SCHUT, M. et al. Multi-actor governance of sustainable biofuels in developing countries: The case of Mozambique. **Energy Policy**, v. 65, p. 631–643, 2014. https://doi.org/10.1016/j.enpol.2013.09.007

SCHUT, M.; SLINGERLAND, M.; LOCKE, A. Biofuel developments in Mozambique. Update and analysis of policy, potential and reality. **Energy Policy**, v. 38, n. 9, p. 5151–5165, 2010. https://doi.org/10.1016/j.enpol.2010.04.048

SILVA, J. C. et al. Advances in the Application of Spectroscopic Techniques in the Biofuel Area over the Last Few Decades. In: **Frontiers in Bioenergy and Biofuels**. Rio de Janeiro, Brazil: InTech, 2017. p. 25–58. https://doi.org/10.5772/65552

SILVERSTEIN, R. M. et al. **Identificação Espectrométrica de Compostos Orgânicos**. 8ª. ed. Rio de Janeiro, Brasil: LTC, 2019. 455 p.

SINGH, S. P.; SINGH, D. Biodiesel production through the use of different sources and characterization of oils and their esters as the substitute of diesel: A review. **Renewable and Sustainable Energy Reviews**, v. 14, n. 1, p. 200–216, 2010. https://doi.org/10.1016/j.rser.2009.07.017

SITOE, B. V. **Controle de qualidade de biodieseis de Pinhão manso e Crambe usando espectrometria no infravermelho médio e cartas de controle multivariadas**. 2015. 109 f. Dissertação (Mestrado em Química) – Instituto de Química, Universidade Federal de Uberlândia, Uberlândia, 2016. https://repositorio.ufu.br/handle/123456789/17447

SITOE, B. V. et al. Quality Control of Biodiesel Content of B7 Blends of Methyl Jatropha and Methyl Crambe Biodiesels Using Mid-Infrared Spectroscopy and Multivariate Control Charts Based on Net Analyte Signal. **Energy & Fuels**, v. 30, n. 2, p. 1062−1070, 2016. https://doi.org/10.1021/acs.energyfuels.5b02489

SITOE, B. V. et al. Monitoring of biodiesel content and adulterant presence in methyl and ethyl biodiesels of jatropha in blends with mineral diesel using MIR spectrometry and multivariate control charts. **Fuel**, v. 191, p. 290–299, 2017. https://doi.org/10.1016/j.fuel.2016.11.078

SITOE, B. V. et al. Quantification of Jatropha methyl biodiesel in mixtures with diesel using mid-infrared spectrometry and interval variable selection methods. **Analytical Letters**, v. em prelo, 1 set. 2019. https://doi.org/10.1080/00032719.2019.1659805

SLINGERLAND, M.; SCHUT, M. Jatropha Developments in Mozambique: Analysis of Structural Conditions Influencing Niche-Regime Interactions. **Sustainability**, v. 6, n. 11, p. 7541–7563, 2014. https://doi.org/10.3390/su6117541

SOUZA, A. M. de et al. Experimento didático de quimiometria para calibração multivariada na determinação de paracetamol em comprimidos comerciais utilizando espectroscopia no infravermelho próximo: um tutorial, parte II. **Química Nova**, v. 36, n. 7, p. 1057–1065, 2013. https://doi.org/10.1590/S0100-40422013000700022

TASIC, L. et al. Peripheral biomarkers allow differential diagnosis between schizophrenia and bipolar disorder. **Journal of Psychiatric Research**, v. 119, p. 67–75, 2019.

https://doi.org/10.1016/j.jpsychires.2019.09.009

TEOH, Y. H. et al. Investigation on particulate emissions and combustion characteristics of a common-rail diesel engine fueled with Moringa oleifera biodiesel-diesel blends. **Renewable Energy**, v. 136, p. 521–534, 2019. https://doi.org/10.1016/j.renene.2018.12.110

UMAR-GARBA, M.; ALHASSAN, M.; KOVO, A. S. A Review of Advances and Quality Assessment of Biofuels. **Leonardo Journal of Sciences**, v. 5, n. 9, p. 167–178, 2006. http://ljs.academicdirect.org/A09/get_htm.php?htm=167_178

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

VANDEGINSTE, B. G. M. et al. **Handbook of Chemometrics and Qualimetrics: Part B**. 20B. ed. Amsterdam, The Netherlands: Data Handling in Science and Technology —volume 20B, 1998. 713p. https://doi.org/10.2307/1271476

VON MALTITZ, G. P.; SETZKORN, K. A. A typology of Southern African biofuel feedstock production projects. **Biomass and Bioenergy**, v. 59, p. 33–49, 2013. https://doi.org/10.1016/j.biombioe.2012.11.024

WAHL, F. et al. Methods for outlier detection in prediction. **Chemometrics and Intelligent Laboratory Systems**, v. 63, n. 1, p. 27–39, 2002. https://doi.org/10.1016/S0169-7439(02)00034-5

WOLD, S. et al. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. **SIAM Journal on Scientific and Statistical Computing**, v. 5, n. 3, p. 735–743, 1984. https://doi.org/10.1137/0905052

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 2, n. 1–3, p. 37–52, 1987.

WOLD, S.; SJOSTROM, M.; ERIKSSON, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In: **Encyclopedia of Computational Chemistry**. Chichester, UK: John Wiley & Sons, Ltd, 2002. p. 857–875. https://doi.org/10.1016/0169-7439(87)80084-9

ZIELIŃSKI, M. et al. Cavitation-based pretreatment strategies to enhance biogas production in a small-scale agricultural biogas plant. **Energy for Sustainable Development**. v. 49, p. 21–26, 2019. https://doi.org/10.1016/j.esd.2018.12.007

ZUÑIGA, A. D. G. et al. Revisão: propriedades físico-químicas do biodiesel. **Pesticidas: Revista de Ecotoxicologia e Meio Ambiente**, v. 21, p. 55–72, 2011. https://doi.org/10.5380/pes.v21i0.25939