ULISSES FERNANDO DE OLIVEIRA

APLICAÇÃO DA ESPECTROSCOPIA NA REGIÃO DO INFRAVERMELHO PRÓXIMO PARA ANÁLISE DE MATERIAIS VEGETAIS

Tese submetida à Universidade Federal de Viçosa, como parte dos requisitos do Programa de Pós-Graduação Multicêntrico em Química de Minas Gerais, para obtenção do grau de *Doctor Scientiae*.

Orientador: Reinaldo Francisco Teófilo

VIÇOSA - MINAS GERAIS
2020

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

O48a
2020

Oliveira, Ulisses Fernando de, 1986-

Aplicação da espectroscopia na região do infravermelho para análise de materiais vegetais / Ulisses Fernando de Oliveira. – Viçosa, MG, 2020.

93 f. : il. (algumas color.) ; 29 cm.

Orientador: Reinaldo Francisco Teófilo.

Tese (doutorado) - Universidade Federal de Viçosa.

Inclui bibliografia.

1. Espectroscopia de infravermelho. 2. Mínimos quadrados. 3. Análise multivariada. I. Universidade Federal de Viçosa. Departamento de Química. Programa de Pós-Graduação em Química. II. Título.

CDD 22. ed. 543.57

ULISSES FERNANDO DE OLIVEIRA

APLICAÇÃO DA ESPECTROSCOPIA NA REGIÃO DO INFRAVERMELHO PRÓXIMO PARA ANÁLISE DE MATERIAIS VEGETAIS

Tese submetida à Universidade Federal de Viçosa, como parte dos requisitos do Programa de Pós-Graduação Multicêntrico em Química de Minas Gerais, para obtenção do grau de *Doctor Scientiae*.

APROVADA: 26 de novembro de 2020

Assentimento:

Ulisses Fernando de Oliveira
Autor

Reinaldo Francisco Teófilo
Orientador

Aos meus pais

Maria Aparecida de Souza Oliveira

e José Irio de Oliveira

AGRADECIMENTOS

Primeiramente a Deus, por tudo que sou, por tudo que tenho, por tudo que vivo e por tudo que conquistei.

À minha família, pelos conselhos e ensinamentos. A meu pai, por todo seu esforço; à minha mãe, por seu carinho e amor incondicionais e aos meus queridos irmãos, companheiros pra toda a vida.

Aos amigos e familiares pela presença, pela convivência em todas as etapas da minha caminhada.

Ao Prof. Reinaldo F. Teófilo, pelos ensinamentos, orientação, oportunidades, apoio e amizade.

Aos amigos do MCDA (Multivariate Chemical Data Analysis Laboratory) pela ajuda, convivência e momentos de descontração.

A Embrapa mandioca e Fruticultura, em especial ao pesquisador Eder Jorge de Oliveira, pelo apoio e ajuda financeira ao trabalho.

Ao colégio de aplicação COLUNI, por me permitir e flexibilizar a realização desse trabalho.

Ao Laboratório de biotecnologia e melhoramento vegetal da UFV, em especial a técnica Andressa, pela ajuda na realização dos trabalhos.

Ao Laboratório de Celulose e Papel da Universidade Federal de Viçosa (LCP/UFV), em especial ao doutorando Thales Martins, pela utilização e ajuda na obtenção dos espectros NIR.

Ao Laboratório de Melhoramento de Oleaginosas da UFV, em especial do técnico Guilherme, pela ajuda na realização do trabalho.

À Universidade Federal de Viçosa, em especial o Departamento de Química, pela estrutura de poder realizar este curso e este trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

A todos aqueles que contribuíram direta ou indiretamente para a realização deste trabalho.

"Para realizar grandes conquistas, devemos não apenas agir, mas também sonhar; não apenas planejar, mas também acreditar."

(Anatole France)

BIOGRAFIA

ULISSES FERNANDO DE OLIVEIRA, filho de Maria Aparecida de Souza Oliveira e José Irio de Oliveira, nasceu na cidade de Dores do Turvo, estado de Minas Gerais, em 08 de abril de 1986.

Iniciou o curso de Química em maio de 2006 pela Universidade Federal de Viçosa (UFV), em Viçosa, MG, diplomando-se bacharel em julho de 2010. No mesmo ano, no mês de agosto, iniciou o curso de pós-graduação em Agroquímica, com área de concentração em Química Analítica, em nível de Mestrado, na mesma instituição, submetendo-se à defesa de dissertação em setembro de 2012. Em agosto de 2013 obteve o título de licenciatura em química também pela UFV. Em agosto de 2016 iniciou o curso de doutorado, também na UFV, no programa de pós-graduação Multicêntrico em Química de Minas Gerais, submetendo a defesa em novembro de 2020.

RESUMO

OLIVEIRA, Ulisses Fernando, D.Sc., Universidade Federal de Viçosa, novembro de 2020. **Aplicação da espectroscopia na região do infravermelho próximo para análise de materiais vegetais.** Orientador: Reinaldo Francisco Teófilo.

Neste trabalho foram construídos modelo de regressão e classificação a partir de espectros na região do infravermelho próximo (NIR) obtidos de materiais vegetais. O primeiro estudo teve como objetivo realizar a quantificação antecipada do teor de óleo usando espectros de NIR obtidos de frutos verdes de macaúba. A espectroscopia NIR e a regressão por quadrados mínimos parciais (PLS) provou ser útil na quantificação precoce do teor de óleo prevendo seu teor com vinte e cinco dias antes do máximo acúmulo de óleo no fruto. A previsão antecipada foi semelhante às previsões para frutos maduros. No segundo estudo, foi realizada a comparação entre dois instrumentos NIR, um de bancada (NIRB) e um portátil (NIRP). A comparação se deu para determinação da amilose aparente (AM) em amostras de fécula de mandioca, usando modelos de regressão e classificação multivariados. Ao contrário do NIRB, os modelos de regressão construídos a partir do instrumento NIRP não apresentaram bons ajustes. Apesar disso, um modelo PLS por análise discriminante (PLS-DA) pode ser aplicado na classificação entre concentrações maiores e menores ou iguais a 20%, obtendo resultados tão confiáveis quanto o NIRB na classificação. No último estudo, foram construídos modelos para a determinações da concentração dos ácidos graxos (AG) palmítico, esteárico, oleico e linoleico do óleo de pinhão-manso. Foram usados espectros do NIRB, NIRP e infravermelho médio (MIR). Os melhores resultados foram para o NIRB, com valores da raiz quadrada do erro quadrático médio de previsão (*RMSEP)* e do coeficiente de correlação (R_p), respectivamente iguais a 3,85 mg mL^{-1} e 0,87 para AG palmítico; 2,20 mg mL^{-1} e 0,87 para AG esteárico; 8,61 mg mL^{-1} e 0,86 para AG oleico; 15,75 mg mL^{-1} e 0,85 para AG linoleico. A aplicação do NIR a materiais vegetais teve um papel importante neste trabalho, obtendo métodos mais rápidos, baratos e não-destrutivos, possibilitando seleção de melhores materiais vegetais, auxiliando no melhoramento genético.

Palavra-chave: Materiais vegetais. Espectroscopia no infravermelho próximo. Regressão multivariadas. Quadrados mínimos parciais.

ABSTRACT

OLIVEIRA, Ulisses Fernando, D.Sc., Universidade Federal de Viçosa, November, 2020. **Application of spectroscopy in the near infrared region for analysis of plant products**. Advisor: Reinaldo Francisco Teófilo.

In this work, regression and classification models were built using spectra obtained from plant materials in the near-infrared (NIR) region. The first study aimed to perform the early quantification of ripe macaw fruits' oil content using NIR spectra obtained from unripe macaw fruits. NIR spectroscopy and partial least squares (PLS) regression proved useful in the early quantification of oil content by predicting its content twenty-five days before the fruit's maximum oil accumulation. The early quantification model was similar to ripe fruits. In the second study, two NIR instruments were compared, a benchtop (NIRB) and a portable (NIRP) one. The comparison was made to determine the apparent amylose content (AM) in cassava starch samples using both regression models or multivariate classification. Unlike the NIRB, the regression models built from the NIRP instrument did not show a suitable fit. However, the PLS discriminant analysis model (PLS-DA) was successfully applied to discriminate higher and lower or equal to 20% of amylose contents, obtaining reliable results as the NIRB in the classification. In the last study, PLS models were built to determine the concentration of palmitic, stearic, oleic, and linoleic fatty acids (AG) in jatropha oil. Spectra from NIRB, NIRP, and medium-infrared (MIR) instruments were used. The best results were for the NIRB, with root mean square error (RMSEP) and correlation coefficient of prediction (Rp) values equal to 3.85 mg mL^{-1} and 0.87 for palmitic AG; 2.20 mg mL^{-1} and 0.87 for stearic AG; 8.61 mg mL^{-1} and 0.86 for oleic AG; 15.75 mg mL^{-1} and 0.85 for linoleic AG. NIR's application to predict plant materials properties played an essential role in this work, obtaining faster, cheaper, and non-destructive methods, enabling the selection of better plant materials, and assisting in genetic breeding programs.

Keywords: Vegetable Materials. Near-infrared spectroscopy. Multivariate regression. Partial least squares.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

LISTA DE ABREVIAÇÕES

Abreviaturas	Termos em inglês	Termos em português
AM	Apparent Amylose	Amilose aparente
AutoOPS	Automatic ordered predictors selection	Seleção dos preditores ordenados automático
AutoiOPS	Automatic Interval ordered predictors selection	Seleção dos preditores ordenados automático em intervalos
FAO	Food and Agriculture Organization of the United Nations	Organização das Nações Unidas para Alimentação e Agricultura
FAR	Far-infrared spectroscopy	Espectroscopia no infravermelho distante
FeedOPS	Feedback ordered predictors selection	Seleção dos preditores ordenados por *feedback*
FeediOPS	Feedback Interval ordered predictors selection	Seleção dos preditores ordenados por *feedback* em intervalos
FN	False negative	Falso negativo
FP	False positive	Falso positivo
GC	Gas chromatography	Cromatografia a gás
MIR	Mid-infrared spectroscopy	Espectroscopia no infravermelho médio
NIR	Near-infrared spectroscopy	Espectroscopia no infravermelho próximo
nlvs	Number of latent variables	Número de variáveis latentes
nVars	Number of selected variables	Número de variáveis selecionadas
OC	oil content	Teor de óleo
OPS	Ordered predictors selection	Seleção dos preditores ordenados
PLS	Partial least squares	Quadrados mínimos parciais
PLS-DA	Partial least squares for discriminant analysis	Análise discriminantes por quadrados mínimos parciais

Abreviaturas	Termos em inglês	Termos em português
R_{cv}	Correlation coefficient of cross-validation	Coeficiente de correlação para validação
RMSECV	Root mean square error of cross-validation	Raiz quadrada do erro quadrático médio de validação cruzada
RMSEP	Root mean square error of prediction	Raiz quadrada do erro quadrático médio de previsão
$\mathbf{R_p}$	Correlation coefficient of prediction	Coeficiente de correlação de previsão
UV-Vis	Ultraviolet–visible spectroscopy	Espectroscopia no ultravioleta e visível

SUMÁRIO

INTRODUÇÃO GERAL

O Brasil é um dos maiores produtores agrícolas no mundo [1]. Segundo dados da Organização das Nações Unidas para Alimentação e Agricultura (FAO), o Brasil terminou o ano de 2016 com uma fatia de 5,7% do mercado global, abaixo apenas dos Estados Unidos, com 11%, e Europa, com 41% [2]. Tais resultados se devem ao grande investimento em pesquisas e uso de tecnologias na produção.

Com o aumento da produção agrícola e a necessidade de fornecer produtos com alta qualidade para atender as exigências internacionais, métodos analíticos simples e rápidos tem sido apresentado tanto na literatura como para mercado. Estes métodos tem como objetivo substituir as análises clássicas de produtos vegetais, as quais envolvem etapas morosas como extração e separação, principalmente em função da complexa composição desse tipo de matriz [3].

Os métodos alternativos de análise usam, na maioria das vezes, técnicas espectroanalíticas que atuam na região de transição eletrônica (180 – 780 nm) como também na região vibracional (780 - 10000 nm). Dentre as técnicas espectroanalíticas mais usadas, destaca-se a espectroscopia na região do infravermelho próximo (NIR), que compreende a região entre 780 e 2500 nm. A espectroscopia NIR tem sido muito usada pois permite que a luz penetre melhor na matéria e possui instrumentação mais resistente em condições adversas de laboratórios e fábricas que a espectroscopia no infravermelho médio. Além disso, espectrofotômetros portáteis tem se estabelecido, em virtude da sua mobilidade, oferendo medições com alta velocidade, robustez, estabilidade e baixo consumo de energia, sendo usado com sucesso em várias aplicações [4–6].

No entanto, dados espectrais fornecem um grande número de variáveis altamente correlacionadas, além de sobreposições de sinais e ruído experimental. Sendo assim, a análise de todo espectro tornou-se possível com o advento das análises multivariadas tais como a regressão por quadrados mínimos parciais (PLS) e de classificação, como a quadrados mínimos parciais por análise discriminante (PLS-DA) [7,8]. A obtenção de espectros diretamente sobre a amostra complexa e o uso de métodos multivariados nos tratamentos dos dados espectrais permitem obter informações químicas de maneira rápida, substituindo procedimentos morosos, que consomem muitos reagentes e são de

alto custo [9]. A aplicação de análises multivariadas sobre dados contendo informações químicas é conhecida como quimiometria.

Há diversos estudos utilizando a espectroscopia NIR e métodos quimiométricos na análise de produtos vegetais, como na classificação de variedades [10], suscetividade e resistência à *Diatraea saccharalis* [11], estimativa da cristalinidade da celulose da biomassa e previsão de lignina em amostras de cana de açúcar [12]; caracterização da qualidade de semente de mamona [13]; determinação de óleos essenciais em lavanda [14], determinação de carotenoides em amostras de mandioca [15]; determinação de forbol éster em amostras de pinhão-manso [16]; determinação de polifenóis e capacidade antioxidante do extrato líquido de *Brassica oleracea* [17]; dentre outras.

Neste contexto, este trabalho tem como objetivo desenvolver procedimentos rápidos e simples para análise de produtos vegetais empregando espectroscopia NIR e métodos quimiométricos. Os procedimentos desenvolvidos são para (1) quantificação do teor de óleo no mesocarpo do fruto da macaúba (*Acrocomia aculeata)*; (2) Comparara dois instrumentos NIR, bancada e portátil, na quantificação e classificação do teor de amilose aparente em fécula de mandioca (*Manihot esculenta Crantz*) e (3) quantificação dos ácidos graxos presentes no óleo de pinhão-manso (*Jatropha curcas* L.).

Referências

[1] R. de O. Bordonal, J.L.N. Carvalho, R. Lal, E.B. de Figueiredo, B.G. de Oliveira, N. La Scala, Sustainability of sugarcane production in Brazil. A review, Agron. Sustain. Dev. 38 (2018) 13. https://doi.org/10.1007/s13593-018-0490-x.

[2] FAO, Food and Agriculture Organization of the United Nations, (2016). http://www.fao.org/countryprofiles/index/en/?iso3=BRA (accessed September 20, 2019).

[3] C.S.W. Miaw, C. Assis, A.R.C.S. Silva, M.L. Cunha, M.M. Sena, S.V.C. de Souza, Determination of main fruits in adulterated nectars by ATR-FTIR spectroscopy combined with multivariate calibration and variable selection methods, Food Chem. 254 (2018) 272–280. https://doi.org/https://doi.org/10.1016/j.foodchem.2018.02.015.

[4] C. Malegori, E.J. Nascimento Marques, S.T. de Freitas, M.F. Pimentel, C. Pasquini, E. Casiraghi, Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms, Talanta. 165 (2017) 112–116. https://doi.org/https://doi.org/10.1016/j.talanta.2016.12.035.

[5] D.M. Friedrich, C.A. Hulse, M. von Gunten, E.P. Williamson, C.G. Pederson, N.A. O'Brien, Miniature near-infrared spectrometer for point-of-use chemical analysis, in: 2014: pp. 899203–899211. https://doi.org/10.1117/12.2040669.

[6] C.G. Pederson, D.M. Friedrich, C. Hsiung, M. von Gunten, N.A. O'Brien, H.-J. Ramaker, E. van Sprang, M. Dreischor, Pocket-size near-infrared spectrometer for narcotic materials identification, in: 2014: pp. 91010O-9101–11. https://doi.org/10.1117/12.2050019.

[7] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32. https://doi.org/https://doi.org/10.1016/j.aca.2010.03.048.

[8] E.I.. Balabin, R.M.aEmail Author, Safieva, R.Z.a, Lomakina, Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction(, Chemom. Intell. Lab. Syst. 88 (2007) 183–188.

[9] B. de Barros Neto, I.S. Scarminio, R.E. Bruns, 25 anos de quimiometria no Brasil, Quim. Nova. 29 (2006) 1401–1406. https://doi.org/10.1590/S0100-40422006000600042.

[10] A.J. Steidle Neto, D.C. Lopes, J. V. Toledo, S. Zolnier, T.G.F. Silva, Classification of sugarcane varieties using visible/near infrared spectral reflectance of stalks and multivariate methods, J. Agric. Sci. 156 (2018) 537–546. https://doi.org/10.1017/S0021859618000539.

[11] N. de A. Porto, J. V. Roque, C.A. Wartha, W. Cardoso, L.A. Peternelli, M.H.P. Barbosa, R.F. Teófilo, Early prediction of sugarcane genotypes susceptible and resistant to Diatraea saccharalis using spectroscopies and classification techniques, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 218 (2019) 69–75. https://doi.org/10.1016/j.saa.2019.03.114.

[12] Í.P. Caliari, M.H.P. Barbosa, S.O. Ferreira, R.F. Teófilo, Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods, Carbohydr. Polym. 158 (2017) 20–28. https://doi.org/10.1016/j.carbpol.2016.12.005.

[13] R. Gislum, P. Nikneshan, S. Shrestha, A. Tadayyon, L. Deleuran, B. Boelt, Characterisation of Castor (Ricinus communis L.) Seed Quality Using Fourier Transform Near-Infrared Spectroscopy in Combination with Multivariate Data Analysis, Agriculture. 8 (2018) 59. https://doi.org/10.3390/agriculture8040059.

[14] S. Lafhal, P. Vanloot, I. Bombarda, J. Kister, N. Dupuy, Chemometric analysis of French lavender and lavandin essential oils by near infrared spectroscopy, Ind. Crops Prod. 80 (2016) 156–164. https://doi.org/10.1016/j.indcrop.2015.11.017.

[15] H. Ceballos, F. Davrieux, E.F. Talsma, J. Belalcazar, P. Chavarriaga, M.S. Andersson, Carotenoids in Cassava Roots, in: Carotenoids, InTech, (2017). https://doi.org/10.5772/intechopen.68279.

[16] J. V. Roque, L.A.S. Dias, R.F. Teófilo, Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies, J. Braz. Chem. Soc. (2017). 1506-1516. https://doi.org/10.21577/0103-5053.20160332.

[17] I.R.N. de Oliveira, J. V. Roque, M.P. Maia, P.C. Stringheta, R.F. Teófilo, New strategy for determination of anthocyanins, polyphenols and antioxidant capacity of Brassica oleracea liquid extract using infrared spectroscopies and multivariate regression, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 194 (2018) 172–180. https://doi.org/10.1016/j.saa.2018.01.006.

CAPÍTULO 1

REVISÃO BIBLIOGRÁFICA

1.1. Infravermelho Próximo

A espectroscopia trata basicamente das interações envolvendo a radiação eletromagnética com a matéria. A energia eletromagnética pode ser ordenada de maneira contínua em função de seu comprimento de onda ou de sua frequência, sendo esta disposição denominada como espectro eletromagnético, que apresenta subdivisões de acordo com as características de cada região (Fig. 1.1).

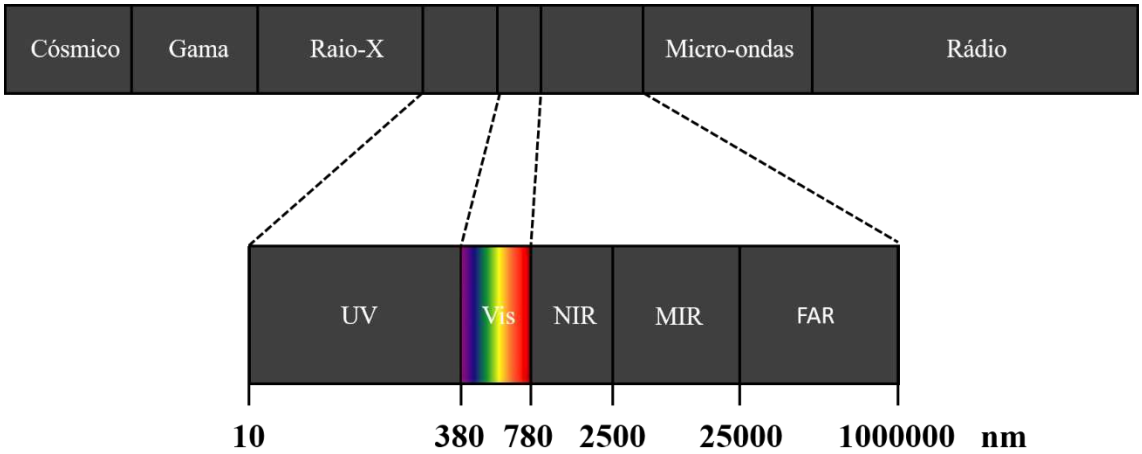

Figura 1.1. Espectro eletromagnético com destaque para as regiões UV-VIS-NIR-MIR-FAR.

Na Fig. 1.1, desataca-se do espectro eletromagnético a região do ultravioleta (UV), Visível (VIS), NIR, infravermelho médio (MIR) e infravermelho distante (FAR). De acordo com o valor de energia da radiação eletromagnética, as transições entre os estados podem ser de vários tipos, dos quais as principais são as transições eletrônicas, vibracionais e rotacionais [1]. O espectro na região do infravermelho é usualmente dividido em NIR, MIR e FAR. A região NIR mostrada na Fig. 1.2, compreende a faixa de comprimento de onda no intervalo de aproximadamente 780 a 2500 nm (12800 a 4000 cm^{-1}) [2].

As vibrações observadas na espectroscopia NIR se devem aos sobretons e combinações de ligações fundamentais do infravermelho médio (Fig. 1.2). Como pode ser observado, as ocorrências espectrais na região NIR provêm de ligações das moléculas em que o hidrogênio está envolvido, restringindo a ligações C-H, N-H, O-H e S-H [3]. Como essas ligações estão presentes na maior parte das moléculas, a utilização do NIR em análise qualitativa para identificação de compostos é bastante restrita. Esta é uma das diferenças entre o MIR e próximo, uma vez que no MIR a análise qualitativa é muito

utilizada devido as vibrações fundamentais características de diferentes grupos funcionais. Além disso, também é rara a observação de um comprimento de onda seletivo, que permita o desenvolvimento de um método de quantificação univariado. Por isso, a espectroscopia NIR ficou estagnada por um longo período de tempo, até que o desenvolvimento da quimiometria permitisse a aplicação quantitativa desta técnica [1].

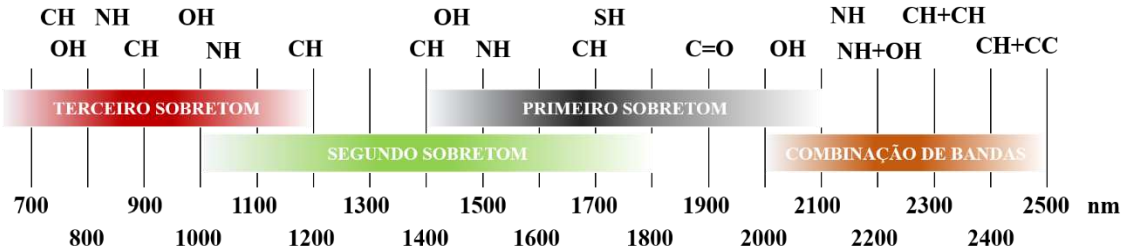

Figura 1.2. Região espectral de sobretons e combinação de bandas.

A espectroscopia NIR foi usada pela primeira vez em aplicações agrícolas por Norris [4] em 1964 para medir a umidade nos grãos. Desde então, tem sido utilizado para análises rápidas, principalmente de umidade, proteínas e teor de gordura de uma ampla variedade de produtos agrícolas e alimentares [5]. Sua ampla aceitação em diferentes campos se deve em virtude de suas vantagens em relação a outras técnicas analíticas. A mais importante delas é a capacidade de registrar espectros para amostras sólidas e líquidas sem nenhum pré-tratamento. Essa característica a torna especialmente atraente para a caracterização direta e rápida de produtos naturais ou sintéticos [6].

Além disso ela se apresenta como uma técnica simples, rápida, não destrutiva, com alto poder de penetração do feixe de radiação, com exigências mínimas no preparo de amostras e que proporciona análise de vários tipos de matrizes com níveis de exatidão e precisão que são comparáveis aos métodos de referência [7].

1.2. Regressão Multivariada

A regressão ou calibração multivariada é o método mais importante dentre os diversos métodos quimiométricos. Isso se deve ao fato de sua grande utilização e importância para o setor produtivo. Nela, são obtidos os chamados "*soft models*" ou modelos flexíveis, construídos a partir de dados empíricos, ao contrário dos "*hard models*" ou modelos rígidos, construídos com embasamento teórico a respeito do sistema de estudo [8,9].

A calibração refere-se à construção de modelo de regressão que relaciona respostas instrumentais à(s) propriedade(s) conhecida(s) de padrões. O principal objetivo do modelo construído é realizar predições das propriedades desconhecidas a partir de respostas obtidas de amostras [10].

O tipo de regressão pode ser classificado quanto a dimensão das respostas instrumentais obtidas para uma determinada amostra, podendo ser citada a de ordem zero, quando a resposta instrumental é um escalar; de primeira ordem, quando a resposta é um vetor; de segunda ordem, quando a resposta é uma matriz e terceira ordem, quando a resposta é expressa na forma de um tensor [11].

Os modelos de calibração multivariada mais usados atualmente são os de primeira ordem. Neste tipo de calibração, para cada amostra, um vetor contendo os sinais instrumentais é necessário para a construção do modelo (e.g., todas as absorbâncias de um espectro). Neste contexto, as respostas instrumentais obtidas são usualmente tratadas como variáveis independentes, matriz **X** e as propriedades de interesse são tratadas como variáveis dependentes, vetor **y**, caracterizando uma regressão inversa (Fig. 1.3). O vetor **b** é a solução do sistema linear e é também conhecido como coeficiente de regressão.

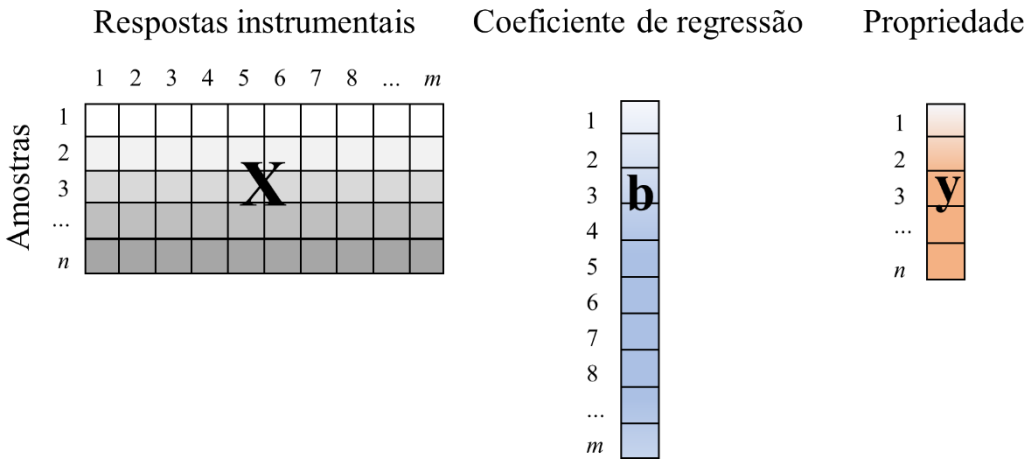

Figura 1.3. Representação das variáveis independentes (**X**) e variáveis dependentes (**y**)

A vantagem deste tipo de modelo de regressão, também chamada de vantagem de primeira ordem, seria a possibilidade de se analisar amostras complexas mesmo com a presença de diversos interferentes, desde que estes também estejam presentes durante a construção do modelo [12].

Dentre os vários métodos de regressão multivariada, a regressão por quadrados

mínimos parciais (PLS) tem se destacado como um dos métodos mais utilizados para dados químicos [13]. Mais detalhes sobre o método PLS são apresentados no item 1.4. desta tese.

1.3. Reconhecimento de Padrões

O primeiro *corpus* de reconhecimento de padrões começou no final da década de 1960 e a partir de então se tornou uma das áreas significativas da quimiometria [14]. Esta técnica surgiu para facilitar a análise de conjuntos de dados maiores, principalmente por causa de suas vantagens, como o poder de fornecer a definição de similaridade, dissimilaridade entre um grupo de dados e a representação gráfica dos resultados [15].

As técnicas de reconhecimento de padrões podem ser divididas em duas categorias: as não supervisionadas e as supervisionadas. No que diz respeito à reconhecimento não supervisionado, o padrão é determinado por uma "fronteira" de classe desconhecida, sem o conhecimento prévio do conjunto estudado. O reconhecimento de padrões supervisionado é um conjunto de técnicas nas quais é necessário um conhecimento prévio, ou seja, há necessidade de um conjunto de calibração, em que as classes das amostras precisam ser bem conhecidas para se construir o modelo de classificação. Em seguida, o desempenho do modelo é avaliado comparando as previsões de classificação com as classes verdadeiras das amostras de validação. Exemplos de métodos supervisionados são a modelagem independente por analogia de classe (SIMCA), *k*-vizinho mais próximo (*k*-NN), análise discriminante linear (LDA) e quadrados mínimos parciais por análise discriminante (PLS-DA) [14–16].

1.4. Quadrados Mínimos Parciais

A regressão por quadrados mínimos parciais (PLS) surgiu por volta do ano de 1975 com Herman Wold trabalhando com dados da área de econometria. Atualmente, a regressão por PLS tornou-se uma ferramenta padrão para modelagem de dados multivariados [17]. A regressão PLS está dentro da classe dos métodos de regressão inversa, sendo utilizado em dados altamente correlacionados, com pequenas não linearidades, ruídos experimentais, além de permitir a modelagem e predições com um

vetor de variável dependente (PLS1) ou uma matriz de variáveis dependentes (PLS2), ou seja, uma propriedade ou múltiplas propriedades simultaneamente [10].

O PLS é calculado por meio de algoritmos computacionais. Diversos algoritmos PLS estão disponíveis na literatura e em softwares, tais como NIPALS [18], NIPALS modificado [19], Kernel [20], SIMPLS [21] e PLS bidiagonal [9,22]. Martins et al. [23] testaram alguns algoritmos PLS e concluíram que todos atingem o mesmo resultado, porém o algoritmo bidiagonal é o que possui menor tempo de pré-processamento computacional e também é o mais intuitivo. Dessa forma, a regressão por meio do algoritmo bidiagonal foi escolhida para a realização desse trabalho.

O algoritmo bidiagonal realiza a decomposição da matriz **X** em três matrizes, conforme Equação 1.1, considerando a informação de **y** [24]:

$$\mathbf{y} \rightarrow \mathbf{X} = \mathbf{URV}^{\mathbf{t}} \tag{1.1}$$

em que **U** e **V** são ortonormais e **R** é bidiagonal. O produto **UR** é definida como matriz de *scores*, trazendo informações das linhas da matriz **X.** A matriz **V** é definida como matriz de *loadings* e traz informações das colunas da matriz **X**.

Então, uma nova matriz **X** é reconstruída a partir das matrizes decompostas trucadas no número de variáveis latentes (*nlvs*), conforme mostrado na Equação 1.2. O *nlvs* são variáveis que extraem o máximo da variância das variáveis originais em **X**.

$$\hat{\mathbf{X}} = \mathbf{U}_{nlvs}\mathrm{R}_{nlvs}\mathbf{V}_{nlvs}^{t} \tag{1.2}$$

Logo após, a pseudoinversa pode ser estimada, o coeficiente de regressão pode ser encontrado conforme Equação 1.3.

$$\hat{\mathbf{b}} = \mathbf{V}_{nlvs}\mathbf{R}_{nlvs}^{-1}\mathbf{U}_{nlvs}^{t}\mathbf{y} \tag{1.3}$$

A suposição básica de qualquer modelo PLS é que o sistema estudado é conduzido por um pequeno número de variáveis latentes. Assim, a escolha do *nlvs* é de extrema importância para evitar sub ajuste ou sobre ajuste do modelo. O sub ajuste acontece quando o modelo é insuficiente para explicar toda informação. O sobre ajuste ocorre com a inclusão de excesso de informação no modelo, que pode ser aleatória ou estar relacionada à presença de erros sistemáticos [25].

Para escolha do número de variáveis latentes utiliza-se normalmente a técnica de validação cruzada, que se baseia no procedimento de reamostragem. Um gráfico relacionando os erros nesta reamostragem versus *nlvs* é construído e o ponto com menor

erro é então selecionado. Normalmente, o cálculo do erro utilizado é a raiz quadrada do erro quadrático médio de validação cruzada (*RMSECV*), conforme apresentado na Equação 1.4.

$$RMSECV = \sqrt{\sum_{i}^{N} (\mathbf{y} - \hat{\mathbf{y}})^2 / N_{vc}} \qquad \textbf{(1.4)}$$

em que **y** é o valor observado (experimental), $\hat{\mathbf{y}}$ é o valor estimado pelo modelo, e N_{VC} é o número de amostras usadas na validação cruzada.

Os principais métodos de validação cruzada são: (1) *leave-one-out*, que remove uma amostra de cada vez; (2) blocos contíguos, que separa amostras em blocos de amostras sequenciais; (3) venezianas (*venetian blinds*), que separa amostras sistematicamente espaçadas; e (4) subconjuntos aleatórios, que separa, aleatoriamente, conjuntos de amostras [25].

A regressão PLS pode ser usada tanto para a calibração (PLS) quanto para classificação (PLS-DA) [26]. Para calibração os valores das variáveis dependentes são normalmente contínuos e para classificação são valores discretos, ou seja, o número um (1) indica que um conjunto pertence a uma determinada classe e o número zero (0) é atribuído ao conjunto que não pertence à classe em questão. Se há duas ou mais classes, uma matriz de classes precisa ser elaborada para análise PLS-DA. Para cada coluna o um (1) indica as observações da classe a ser analisada e o zero (0) indica as observações que não pertencem à classe definida com 1's. A classificação via PLS-DA é paramétrica e discriminante. Paramétrico porque é usado um método probabilístico para estimar a classe e discriminante porque as amostras podem ser previstas apenas em uma classe [27].

1.5. Seleção de Variáveis

A seleção de variáveis baseia-se na escolha de um número de variáveis do conjunto original que permitirá melhorar a exatidão da análise (melhorar a predição); facilitar (ou tornar possível) a interpretação das variáveis no modelo construído; identificar (bio)marcadores; projetar instrumentos portáteis; diminuir o tempo de processamento dos dados e fornecer modelos mais simples [28].

Modelos de calibração multivariada, como PLS foram desenvolvidos para análise quantitativa de dados espectrais devido à sua capacidade de reduzir o impacto de

problemas comuns, como colinearidade, sobreposições de sinais e dificuldade na resolução do sistema linear. Apesar de métodos baseados na compressão de dados, como o PLS, não ser limitado à dimensão da matriz original de dados, muitas variáveis podem não ser importantes e assim diminuir a qualidade do modelo de calibração construído [6,28–32]. Portanto, a seleção de variáveis contribui para melhorar o desempenho do modelo além de outras melhorias. Uma das principais aplicações da seleção de variáveis é diminuir o erro na predição tanto em calibração quanto na classificação [30,33].

Os métodos de seleção de variáveis mais utilizados para espectroscopia tem sido o algoritmo genético (GA) [34], o método de quadrados mínimos parciais por intervalo (iPLS) [35] e o método de seleção dos preditores ordenados (OPS) [36] . O OPS, abordado nesse trabalho, tem mostrado com alta capacidade de melhorar a previsão de modelos após a seleção de algumas variáveis importantes.

O OPS se baseia na obtenção de vetores informativos que darão a localização de determinadas regiões espectrais a fim do modelo de calibração minimizar os erros de predição. A seleção de variáveis pelo OPS pode ser realizada de três modos distintos, o primeiro deles, um método automático, *AutoOPS*, que desenvolve e executa seleção variáveis usando vários vetores informativos e suas combinações. Em segundo, o *FeedOPS,* apresenta uma nova estratégia as variáveis pré-selecionadas, retornando-as a uma nova seleção. Por último, foi estabelecido um método para aplicar OPS em subdivisões de matriz dos dados chamados OPS intervalado, *AutoiOPS* e *FeediOPS* [37].

Referências

[1] L.E. Agelet, C.R. Hurburgh, A Tutorial on Near Infrared Spectroscopy and Its Calibration, Crit. Rev. Anal. Chem. 40 (2010) 246–260. https://doi.org/10.1080/10408347.2010.515468.

[2] R.C. Skoog, D.A.; West, D. M.; Holler, F.J.; Stanley, Fundamentos de Química Analítica, Norte-amer, São Paulo-SP, (2014).

[3] L. Bokobza, Near Infrared Spectroscopy, J. Near Infrared Spectrosc. 6 (1998) 3–17. https://doi.org/10.1255/jnirs.116.

[4] K.H. Norris, Design and development of a new moisture meter, Agric. Eng. 45 (1964) 370–372.

[5] B.M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review, Postharvest Biol. Technol. 46 (2007) 99–118.

https://doi.org/10.1016/j.postharvbio.2007.06.024.

[6] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32. https://doi.org/https://doi.org/10.1016/j.aca.2010.03.048.

[7] C. Pasquini, Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications, J. Braz. Chem. Soc. 14 (2003) 198–219. https://doi.org/10.1590/S0103-50532003000200006.

[8] M.M.C. Ferreira, Quimiometria: conceitos, métodos e aplicações, Editora da Unicamp, (2015). https://doi.org/10.7476/9788526814714.

[9] R. Manne, Analysis of two partial-least-squares algorithms for multivariate calibration, Chemom. Intell. Lab. Syst. 2 (1987) 187–197. https://doi.org/10.1016/0169-7439(87)80096-5.

[10] K.R. Beebe, B.R. Kowalski, An Introduction to Multivariate Calibration and Analysis, Anal. Chem. 59 (1987) 1007A-1017A. https://doi.org/10.1021/ac00144a725.

[11] J. V. Roque, L.A.S. Dias, R.F. Teófilo, Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies, J. Braz. Chem. Soc. (2017). 1506-1516. https://doi.org/10.21577/0103-5053.20160332.

[12] M.-B. Gholivand, A.R. Jalalvand, H.C. Goicoechea, T. Skov, Chemometrics-assisted simultaneous voltammetric determination of ascorbic acid, uric acid, dopamine and nitrite: Application of non-bilinear voltammetric data for exploiting first-order advantage, Talanta. 119 (2014) 553–563. https://doi.org/10.1016/j.talanta.2013.11.028.

[13] R.G. Brereton, Introduction to multivariate calibration in analytical chemistry, Analyst. 125 (2000) 2125–2154. https://doi.org/10.1039/b003805i.

[14] R.G. Brereton, Pattern recognition in chemometrics, Chemom. Intell. Lab. Syst. 149 (2015) 90–96. https://doi.org/10.1016/j.chemolab.2015.06.012.

[15] P. Gemperline, Pratical Guide to Chemometrics, (2006). https://doi.org/10.1201/9781420018301.

[16] R.G. Brereton, Chemometrics: Data analysis for the laboratory and chemical plant, (2000) https://doi.org/10.1007/s10809-005-0223-6.

[17] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[18] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185 (1986) 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.

[19] Bhupinder. S. Dayal John F. MacGregor, Improved PLS algorithms, Hournal Chemom. (1998). https://doi.org/10.1002/(SICI)1099-128.

[20] F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for PLS, J. Chemom. 7 (1993) 45–59. https://doi.org/10.1002/cem.1180070104.

[21] S. de Jong, SIMPLS: An alternative approach to partial least squares regression, Chemom. Intell. Lab. Syst. 18 (1993) 251–263. https://doi.org/10.1016/0169-7439(93)85002-X.

[22] W. Wu, R. Manne, Fast regression methods in a Lanczos (or PLS-1) basis. Theory and applications, Chemom. Intell. Lab. Syst. 51 (2000) 145–161. https://doi.org/10.1016/S0169-7439(00)00063-0.

[23] J.P.A. Martins, R.F. Teófilo, M.M.C. Ferreira, Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets, J. Chemom. (2010) n/a-n/a. https://doi.org/10.1002/cem.1309.

[24] J.L. Barlow, N. Bosner, Z. Drmač, A new stable bidiagonal reduction algorithm, Linear Algebra Appl. 397 (2005) 35–84. https://doi.org/10.1016/j.laa.2004.09.019.

[25] R.G. Brereton, Applied Chemometrics for Scientists, John Wiley & Sons, Ltd, Chichester, UK, (2007). https://doi.org/10.1002/9780470057780.

[26] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[27] L.A. Berrueta, R.M. Alonso-Salces, K. Héberger, Supervised pattern recognition in food analysis, J. Chromatogr. A. 1158 (2007) 196–214. https://doi.org/10.1016/j.chroma.2007.05.024.

[28] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32. https://doi.org/10.1016/j.aca.2010.03.048.

[29] F.C.C. Oliveira, A.T.P.C. de Souza, J.A. Dias, S.C.L. Dias, J.C. Rubim, A escolha da faixa espectral no uso combinado de métodos espectroscópicos e quimiométricos, Quim. Nova. 27 (2004) 218–225. https://doi.org/10.1590/S0100-40422004000200009.

[30] R.M. Balabin, S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72. https://doi.org/10.1016/j.aca.2011.03.006.

[31] A. Höskuldsson, Variable and subset selection in PLS regression, Chemom. Intell. Lab. Syst. 55 (2001) 23–38. https://doi.org/10.1016/S0169-7439(00)00113-1.

[32] M. Goodarzi, Y. Vander Heyden, S. Funar-Timofei, Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models, TrAC Trends Anal. Chem. 42 (2013) 49–63. https://doi.org/10.1016/j.trac.2012.09.008.

[33] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemom. Intell. Lab. Syst. 118 (2012) 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010.

[34] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207. https://doi.org/10.1016/S0169-7439(98)00051-3.

[35] A. de Araújo Gomes, R.K.H. Galvão, M.C.U. de Araújo, G. Véras, E.C. da Silva, The successive projections algorithm for interval selection in PLS, Microchem. J. 110 (2013) 202–208. https://doi.org/10.1016/j.microc.2013.03.015.

[36] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. https://doi.org/10.1002/cem.1192.

[37] J. V. Roque, W. Cardoso, L.A. Peternelli, R.F. Teófilo, Comprehensive new approaches for variable selection using ordered predictors selection, Anal. Chim. Acta. 1075 (2019) 57–70. https://doi.org/10.1016/j.aca.2019.05.039.

CAPÍTULO 2

PREDICTING OIL CONTENT IN RIPE MACAW FRUITS (*Acrocomia aculeata*) FROM UNRIPE ONES BY NEAR INFRARED SPECTROSCOPY AND PLS REGRESSION

Abstract

A method for early quantification of oil content using near-infrared spectroscopy (NIR) of unripe macaw fruits and partial least squares (PLS) is presented. This method was compared with other methods using NIR spectra of ripe fruits. After harvest, the fruit takes about 30 days to reach its maximum oil accumulation. PLS models were built using NIR spectra of shell after five (Shell5) and thirty days (Shell30) and from mesocarp at thirty days (Pulp30) after harvest. Oil content was quantified after thirty days using Soxhlet extraction. Ordered predictors selection (OPS) was used to select the most informative variables. The best models presented root mean square error of prediction and correlation coefficient of prediction (R_P) of 4.87 % and 0.89 for Shell5; 5.83 % and 0.85 for Shell30*;* 4.76 % and 0.92 for Pulp30. Early prediction of oil content (Shell5) could reduce the time needed for decision-making related to quality control.

Keywords: macaw palm; multivariate regression; near-infrared spectroscopy; partial least squares; early prediction.

2.1. Introduction

The demand for vegetable oils has increased over the last decades. It may be due to food consumption and biofuel production related to the continue increase in the world's population and a demand for cleaner and renewable energy sources [1,2]. Nowadays, oil palm accounts for 35% of the edible vegetal oil market, followed by soybean with 27%, rapeseed with 16%, and sunflower with 9% [3]. In addition, Brazil's 2019 biodiesel production were of 6 million m^3, and from this total, soybean oil is accountable for 64%, other crops 10% and the rest of animal fat.[4]. Thus, there is a demand for new crops with food and bioenergetic potential [5].

The oleaginous palm, *Acrocomia aculeata*, commonly known as macaw palm, is found in savanna-like vegetation, semi-deciduous seasonal forests, and deforested areas, among other regions of Central and South America [6,7]. The estimated oil production from this crop is up to 6200 kg/ha, similar to the oil palm (*Elaeis guineensis*), the current major source of vegetal oil. However, it does not compete with rainforest and fertile land and presents a higher storage stability [2,8]. The macaw palm mesocarp oil contains about 60% oleic acid, 29% palmitic acid, and is rich in carotenoids, which plays an important role as antioxidant agent [9]. It has great potential to be used in food, pharmaceutical, cosmetics, and biodiesel production due to its higher oxidation stability and operation at low temperatures [10,11]. In addition, after oil extraction, the cake can be used to animal nutrition as it does not contain toxics compounds and the shell could be used to energy generation [12]. Therefore, the macaw palm has potential as an economic viable oil crop.

The macaw palm is still under process of domestication, and most studies involving the species aim at its breeding to increase its oil yield [13]. Even after harvest, the macaw fruits continue to increase its oil content [14,15]. This increase may be related to the availability of polysaccharide reserves since a reduction in starch and water content coincides with the increase in the oil content [16]. However, the oil acidity also increases, which decreases the oil's quality, being necessary a deacidification process, increasing, then, the process costs [12].

The current harvest practices based on collection of dropped fruits and extraction of its oil without adequate handling and storage has resulted in oil with low quality for food or biofuel purposes. Good practices can increase the fruit's oil content over 20% during a 30 days period of storage after harvest [2,8,12]. Therefore, good practices from

harvest to storage are necessary to maintain the oil's quality aiming its further use either for food or biofuel.

The oil content of ripe macaw fruits is often quantified using gravimetric analysis. The extraction is typically carried out using accelerated solvent extraction (ASE) [17,18] or Soxhlet extraction. ASE uses a combination of temperature, pressure, and solvent to extract the oil, and Soxhlet extraction uses solvent reflux in an intermittent extraction process, where the sample is in contact with the boiling solvent, avoiding lipid decomposition [17,18]. These methods are considered laborious, time-consuming, and destructive, and they use a considerable amount of sample and toxic solvents to perform the extraction [18].

Due to the continue increase in the oil content, long maturation time, and limitations of the conventional procedures for oil quantification, a method capable of early predicting the oil content of the ripe fruits would be advantageous for decision-making regarding the selection of high yield fruits, transportation, and storage. Considering the routine analyses carried out at breeding programs and quality assessments at industrial sites, a quick, inexpensive, nondestructive, and reliable analytical method is required in order to quantify the oil of numerous samples [8].

Noninvasive and nondestructive methods such as near-infrared spectroscopy (NIR) coupled with partial least squares (PLS) regression have been widely used to quantify agronomic properties and assess the quality of agricultural and food products [20,21]. This technique provides rapid information on physical and chemical properties with minimal or no sample preparation with a high predictive ability for unknown samples [22].

PLS regression stands out as one of the primary methods for building first-order regression models from chemical source data. This method does not require accurate knowledge of all the components present in the samples, and it can perform prediction of an analyte of interest even in the presence of interference, as long as the interferents are also present in the model built [23].

Several works have presented the use of NIR to quantify oil content of plants, such as avocado [24], olive [25,26], palm fruits [27], maize [28] and citrus [29]. A previous work reported the application of VIS-NIR region from 30770 to 9300 cm^{-1} to predict the oil content of the macaw palm in different maturation stages using the mesocarp [30]. However, studies involving early prediction from plant products are still scarce due to

their high complexity and possible changes in the matrix. Studies about early prediction of sugarcane features [31] and cassava productivity [32] can be found in the literature. Early predicting a fruit property is important to correctly designate the fruit before reaching its full maturity, therefore, improving its quality and saving time and costs with storage.

Studies involving the early prediction of oil content from plants have not been reported in the literature. Thus, the goal of this work is to establish for the first time a non-destructive, fast, inexpensive, and reliable method capable of early predicting the oil content of ripe macaw fruits using NIR spectra from 10000 to 4000 cm^{-1} of unripe fruits and PLS regression. Ordered predictors selection (OPS) [33,34] was used to improve the models' quality by selecting more informative and predictive variables. The method proposed was compared with other methods using the NIR spectra of ripe fruits.

2.2. Experimental

2.2.1. Macaw Palm Samples

The macaw fruits used in this study were part of the macaw palm germplasm active bank (MPGAB) of the Universidade Federal de Viçosa (UFV) located in the municipality of Araponga, Minas Gerais State (lat 20º 40'1'' S, long 42º 31'15'' W, alt ~ 980 m). The MPGAB is registered under the number 084/2013 – SECEX/CGEN and consists of 253 accessions, collected in several parts of Brazil. This bank is considered the largest MPGAB in the world.

The fruits were collected in the MPBAG, consisting of 44 accessions (half-sib families). Three plants were collected for each access, and five fruits were chosen from each plant for oil content analysis. On total, 630 fruits were collected. After harvest, the fruits were taken to the postharvest laboratory, where they remained stored at room temperature for ripening. After ripening, the fruits were frozen at -20 ºC refrigerator in order to stop the metabolism for conservation and further oil quantification.

2.2.2. Spectral Analysis

Due to logistics of transporting the macaw fruits, only after five days the near-infrared (NIR) spectra were collected from the shell of each unripe fruit (Shell5). Spectra were taken from shell to avoid damage the fruit before reaching its full maturity. After harvest, considering good practices of storage, the macaw palm takes thirty days to reach its maturity and, therefore, its maximum oil accumulation [12]. In order to evaluate the results obtained for the Shell5 model, NIR spectra were collected from the shell (Shell30) and mesocarp (Pulp30) of the fruits after 30 days of maturation. Fig. 2.1A shows the fruits before harvest, and Fig. 2.1B and 2.1C refer to how the NIR spectra were obtained for the macaw shell and mesocarp, respectively.

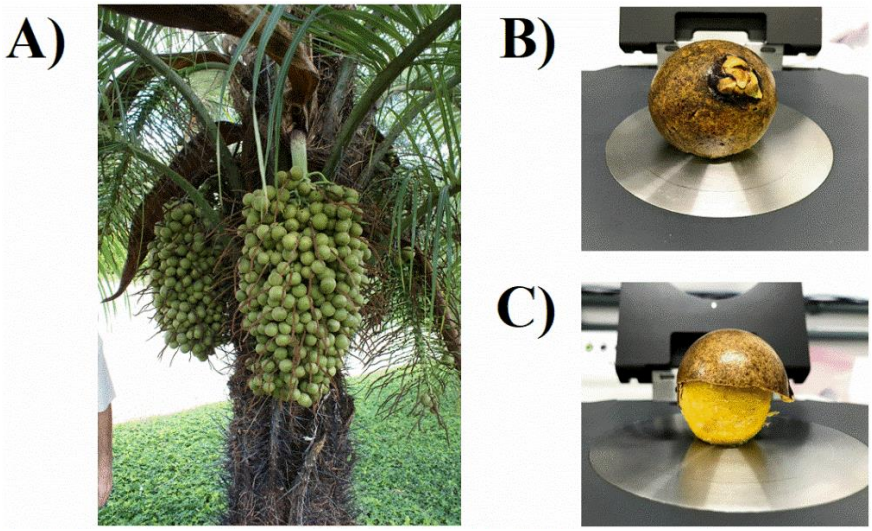

Figure 2.1. Macaw palm (*Acrocomia aculeata)* (A), NIR set up to acquire the macaw fruit's shell spectra (B), and NIR set up to acquire the macaw fruit's mesocarp spectra (C).

The NIR spectra were obtained using a Fourier transform NIR (FT-NIR) 660 spectrometer (Agilent Technologies) with a reflectance-integrating sphere accessory from PIKE Technologies. The range investigated was 10000 – 4000 cm^{-1} with an increment of 2 cm^{-1}. The spectra were obtained using the software Pro Resolutions Version 5.1 and stored as log (1/R), where R is the reflectance collected. For each sample, a total of 32 scans were performed, and the average was stored.

For Shell5 and Shell30, each fruit was directly placed on the instrument window without any sample preparation and scanned in two positions. Each side was measured three times, and the averaged spectrum of the two sides of each fruit was used for

modeling. For Pulp30, the fruit shell was removed using a knife to expose the mesocarp, and the spectrum was obtained placing the mesocarp directly on the instrument window. The spectrum was obtained from two different positions, and the averaged spectrum used for modeling. On total, 630 spectra were used for each model, named Shell5, Shell30, and Pulp30. The spectra were acquired in laboratory with controlled temperature at 21 ºC.

2.2.3. Oil Content Quantification

Oil extraction was performed using *n*-hexane solvent in a Soxhlet extractor according with the adapted procedure from Analytical Norms of Adolfo Lutz Institute [35]. The fruit's mesocarps were dried in a ventilated oven at 65 ºC for 72 h. After drying, 5 g of sample was placed in a filter paper cartridge and arranged in the Soxhlet extractor containing 150 mL of *n*-hexane; the extraction was carried out for 8 h. In sequence, the *n*-hexane extract was transferred to a beaker which was placed in an oven at 105 °C for 24 h for *n*-hexane evaporation. Finally, the remaining oil was cooled down to room temperature and weighed.

Oil extraction was performed for each fruit, which generated an oil content (OC) value for each sample (Equation 2.1).

$$OC = \left(\frac{M_0}{M_S} \right) 100 \tag{2.1}$$

where OC stands for oil content percent, M_0 stands for oil mass in grams, and M_s stands for the total sample mass in grams.

2.2.4. Multivariate Regression Models

The **y** vector (dependent variable) is associated with the oil content values. The **X** matrix of the NIR spectra (independent variables) was different for each condition (Shell5, Shell30, and Pulp30), but the **y** vector was the same. The Kennard and Stone algorithm [36] was used to split the samples into calibration and prediction sets. For all models, 504 and 126 samples were selected for the calibration and prediction sets, respectively.

The spectra were imported to Matlab 2019a environment (The Mathworks, Natick, USA). An inverse regression model (**y = Xb**) was built using the PLS and random cross-validation, where ten splits were used. Besides the raw data, two preprocessing methods, mean center and autoscale; eight transformations, smoothing, first derivative, second derivative, multiplicative scatter/signal correction (MSC), detrend, normalize, baseline, standard normal variate scaling (SNV), and their combinations were tested to find the best regression model.

Two OPS methods [33], *i.e.*, *AutoiOPS* and *FeediOPS*, were applied to select more predictive variables, aiming at improving the models. Firstly, the original response variables (**X** matrix columns) were subdivided into intervals, each one containing fifty variables. In each interval, informative vectors that contain information about the location of the best response variables for prediction were calculated. The original variables in each interval were differentiated according to the corresponding absolute values of the informative vector elements. The differentiated variables were sorted in descending order. An initial subset of two variables (window) for each interval was defined to build PLS models. The initial subset was extended by the addition of one variable (increment) over each window until all variables were taken. The variable subsets in each interval were compared using the cross-validation parameters calculated during validations, and the best variable subset was defined for each interval. These steps of variable selection were performed using several informative vectors (REG: regression coefficients; COR: correlation between each column of matrix X with y; SQR: residual information of the reconstructed matrix with h latent variables; NAS: net analyte signal; VIP: variable importance on projection; URXY: univariate regression between each column of matrix X with y; WGHT: weights; COV: covariance procedures; PRODALL: the product of all single vectors simultaneously.), and for each interval, the best vector was chosen. This method is called *AutoiOPS*.

Additionally, after performing the *AutoiOPS* selection, the selected variables returned to a new selection, and this procedure is called *FeediOPS*. This method was carried out until relative differences between two consecutive *RMSECV* values were less than 0.02% [33]. Then, a new matrix was created containing only the selected variables in each interval, and a new variable selection (*AutoiOPS* or *FeediOPS*) was performed considering the same parameters described previously.

Two optimum number of latent variables (*nlvs*) are employed in this work: one representing the component number for model building (*hMod*) and the other representing the component number employed to generate the best informative vector in the OPS method *(hOPS*) [33]. The OPS methods were applied using the algorithms available at www.deq.ufv.br/chemometrics. All of the calculations were performed in Matlab 2019a.

2.2.5. Figure of Merit

The quality of the models was evaluated using the parameters root mean square error (*RMSE*) and correlation coefficient (*R*), which were calculated by Equations (2.2) and (2.3), respectively.

$$RMSE = \sqrt{\sum_{i}^{N} (y_i - y_1)^2 / N} \tag{2.2}$$

$$R = \sum_{i=1}^{N} \left(\hat{y}_i - \overline{\hat{y}}\right)\left(y_i - \overline{y}\right) / \sqrt{\sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2} \sqrt{\sum_{i=1}^{N} \left(\hat{y}_i - \overline{\hat{y}}\right)^2} \tag{2.3}$$

where $\hat{y}_i$ $\overline{\hat{y}}$ and are the estimated value and mean of estimated values, respectively, *y* and $\overline{y}$ are the observed values and the mean of the observed values, respectively; and *N* represents the number of samples. When internal cross-validation (C_V) is applied, *N* represents the number of samples in the cross-validation set, and the error and correlation coefficient are called root mean square error of cross-validation (*RMSECV*) and correlation coefficient of cross-validation (R_{CV}), respectively. For the external validation, *N* represents the number of samples predicted (*P*) and, in this case, the error and correlation coefficients are named correlation root mean square error of prediction (*RMSEP*) and correlation coefficients of prediction (R_P), respectively.

The inverse of analytical sensitivity (γ^{-1}) is the minimum difference of concentration that can be determined by the model and is calculated by Equation 2.4. Selectivity (*SEL*) shows the models capacity to determine an analyte without interference of other compounds in the matrix and is calculated by Equation 2.5. Limit of detection (*LOD*) is the minimum detectable value of concentration and is calculated according to Equation 2.6.

$$\gamma^{-1} = \|\partial_x\| \times \|b\| \quad \textbf{(2.4)}$$

where $\|b\|$ is the Euclid.ean norm of regression coefficients of the PLS model and the $\|\partial_x\|$ standard deviation of the reference signal [37].

$$SEL = nas_i / \|x_i\| \quad \textbf{(2.5)}$$

where nas_i is the absolute scalar value of the net analytical signal for sample *i* and $\| x_i \|$ represents the Euclidean norm of the instrumental response vector for sample *i* [37,38].

$$LOD = \left(3.3 \times nas \times \|b\|\right) / \max(b) \quad \textbf{(2.6)}$$

where $\|b\|$ represents the Euclidean norm of the regression coefficients b.

Each model was verified for chance correlation. The **y** vector (dependent variables) was randomized 10,000 times, and models were built using these randomizations. The correlation coefficient between the reference and predicted values was evaluated. If the model's correlation parameter built using the authentic **y** was isolated from those of the models using the randomized **y**, then the model did not occur by chance.

2.3. Results and Discussion

2.3.1. Oil Content

The oil content no mesocarp of the samples ranged from 29.5 to 74.5% with an average of 55.7 % and a 10.7 standard deviation. The variability in oil content is related to the genetic variability of the plants [15]. This variability is also due to the different geographic regions where the fruits were harvested.

2.3.2. Spectral Interpretation

For all raw spectra shown in Fig. 2.2 (A, C and E), the band from 9000 to 7700 cm^{-1} (region 1) is related to C-H stretching second overtone, specific related oleic acid. The higher absorbance of this band is related to an oil rich in unsaturated fatty acids. The band from 7200 to 6500 cm^{-1} (region 2) is related to water and also combination bands. The bands in 5800 (region 3) and 5670 cm^{-1} (region 4) are related to saturated and unsaturated fatty acids, in this case, palmitic and oleic acids, respectively. The band around 5200 cm^{-1} (region 5) is related to water absorbance and combination bands. The region from 4500 to 4000 cm^{-1} (region 6) is related to combinations bands of edible oils [39,40].

The spectra of Shell5 and Shell30 are similar, the latter presenting a more scattered profile. It may be due to physical changes in the shell during the maturation. The Pulp30 spectra differ from Shell5 and Shell30 spectra, presenting a more accentuated absorbance in the regions from 8700 to 8000 cm^{-1}, 7200 to 6500 cm^{-1}, 6000 to 5500 cm^{-1}, 5300 to 5000 cm^{-1}, and 4500 to 4000 cm^{-1}. This may be because the Shell5 and Shell30 spectra were obtained from the shell, which is not rich in oil and water, while Pulp30 spectra were obtained directly on the pulp, which is rich in oil and water. For that reason, the spectra of Pulp30 presented a more accentuated profile in the regions related to fatty acids and water.

2.3.3. Modeling

The original NIR spectra and their respective transformations are shown in Fig. 2.2. Mathematical transformations in the matrix rows are essential to reduce systematic errors originated from undesired light scattering variations and to increase the signal-to-noise ratio. The best preprocessing and transformation combination that produced the smallest *RMSECV* and highest R_{CV} was chosen. The transformation that presented the best results for all data sets was the first derivative, followed by smoothing.

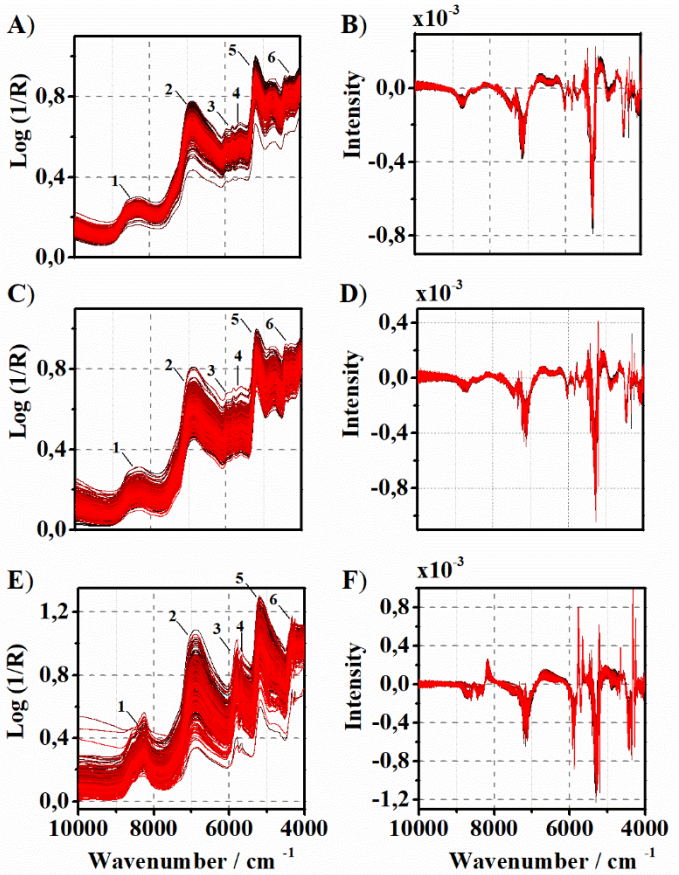

Figure 2.2. (A), (C), and (E) NIR spectra for Shell5 ($\mathbf{X}_{shell5}$), Shell30 ($\mathbf{X}_{shell30}$), and Pulp30 ($\mathbf{X}_{pulp30}$) respectively; (B), (F) and (G) represents the spectra transformed using first derivative and smoothing. Region **1**: 9000 to 7700 cm^{-1}; Region **2**: 7200 to 6500 cm^{-1}; Region **3**: 5800 cm^{-1}; Region **4**: 5670 cm^{-1}; Region **5**: 5200 cm^{-1}; Region **6**: 4500 to 4000 cm^{-1}.

Statistical parameters and figures of merit for the PLS models using all variables (*Full*) and selected variables (*AutoiOPS* and *FeediOPS*) are shown in Table 1. For Shell5, the model using the variables selected using *AutoiOPS* presented a lower *RMSECV* value than the *Full* model. However, a smaller *RMSEP* was achieved with the variables selected using *FeediOPS*, which means that it is a more predictive model than the model with all of the variables. For the Shell30 model, the variables selected improved both *RMSECV* and *RMSEP* values; the *AutoiOPS* was the best approach for this model. For the Pulp30 model, the variables selected improved the *RMSECV* values but not the *RMSEP* values, so the *Full* model was chosen.

Table 2.1. Statistical parameters and figures of merit for the PLS models with all variables (*Full*) and variables selected using *AutoiOPS* and *FeediOPS*.

	PLSmodel-Shell5			**PLSmodel-Shell30**			**PLSmodel-Pulp30**		
	Full	***Autoi*OPS**	***Feedi*OPS**	**Full**	***Autoi*OPS**	***Feedi*OPS**	**Full**	***Autoi*OPS**	***Feedi*OPS**
nlvs	6	6	6	7	7	7	6	6	6
*nVars**	3113	260	290	3113	200	155	3113	340	220
RMSEC (g/g)	3.81	3.33	3.31	3.65	3.49	3.99	3.14	2.52	2.19
Rc	0.93	0.95	0.96	0.94	0.94	0.93	0.95	0.97	0.98
RMSECV (g/g)	4.97	4.21	4.16	4.97	4.32	4.75	3.89	3.24	2.82
Rcv	0.88	0.91	0.92	0.88	0.91	0.89	0.93	0.95	0.96
RMSEP (g/g)	5.28	5.50	*4.87*	6.11	5.83	5.95	4.76	5.49	5.47
Rp	0.88	0.88	*0.89*	0.83	0.85	0.82	0.92	0.88	0.88
Υ^{-1} */ (g/g)*	$1.98\text{x}10^{-4}$	$5.30\text{x}10^{-3}$	$3.50\text{x}10^{-3}$	$4.37\text{X}10^{-4}$	$2.33\text{x}10^{-2}$	$3.12\text{x}10^{-2}$	$4.06\text{x}10^{-5}$	$3.30\text{x}10^{-3}$	$4.80\text{x}10^{-3}$
SEL	0.08	0.1	0.1	0.07	0.21	0.2	0.13	0.13	0.13
LOD (g/g)	26.19	11.43	18.32	25.89	13.23	7.01	13.66	8.83	8.66

nlvs: number of latent variables; *nVars*: number of variables; *Rc:* correlation coefficient of calibration; *Rcv:* correlation coefficient of cross-validation; *Rp:* correlation coefficient of prediction*; RMSEC*: root mean square error of calibration; *RMSECV*: root mean square error of cross-validation; *RMSEP*: root mean square error of prediction; *SEN*: sensitivity; *SEL*: selectivity; *LOD*: limit of detection..

Comparing the models Shell5-FeediOPS, Shell30-*Autoi*OPS and Pulp30-*Full* from the results shown in Table 2.1, it is possible to observe that the results involving Shell5-*Feedi*OPS were quite similar to the models built with spectra just before the determination of the oil content, *i.e.*, Shell30-*Autoi*OPS and Pulp30-*Full*. Furthermore, the Shell5-*Feedi*OPS model presented *RMSECV* and *RMSEP* values lower and *R* and *Rp* values higher than the Shell30-*Autoi*OPS model.

Matsimbe *et al.* determined the oil content in macaw palm using VIS-NIR spectra (30770 to 9300 cm^{-1}) of the mesocarp [30] and found *RMSEP*, *R,* and *nlvs* values equal to 7.08 %, 0.88, 9, respectively. Considering the Pulp30, the model presented better parameters, *RMSEP* of 4.76 % and *R* of 0.92*,* with 6 *nlvs*, a more parsimonious model. It may be due to the wavelength range used and a larger preprocessment screening, which may have resulted in a more corrected set of spectra, and therefore a better model.

For all models, low γ^{-1} values can be observed. A decrease in their values is noticed in the models using the selected variables, which may be due to the use of more interpretative and predictive variables. As expected, low *SEL* values were observed for all models as the spectra contain information about other components, which could be considered interferents. For both *AutoiOPS* and *FeediOPS* models, this parameter was higher as more selected regions of the NIR spectra related to the property were used. Except for the Pulp30 model where the selection did not improve the *SEL* values. The LOD values also decreased for the models using the selected variables

According to Table 2.1, *FeediOPS* models were chosen for Shell5, *AutoiOPS* for Shell30, and *Full* model for Pulp30, as variable selection did not improve the *RMSEP* value. Fig. 2.3 presents the measured versus predicted values of oil content (2.3A, 2.3D, and 2.3G) using the NIR spectra for the best model of each data set. For all datasets, a linear fit can be observed, indicating that the models can accurately predict the oil content in macaw fruit.

Fig. 2.3 also shows the relative errors of calibration (Fig. 2.3B, 2.3E and 2.3H) and prediction (Fig. 2.3C, 2.3F and 2.3I). Most of the relative errors were less than 10%, showing that by obtaining the spectra from the shell for model Shell5, it is possible to predict the amount of oil present in the fruit 25 days before its extraction.

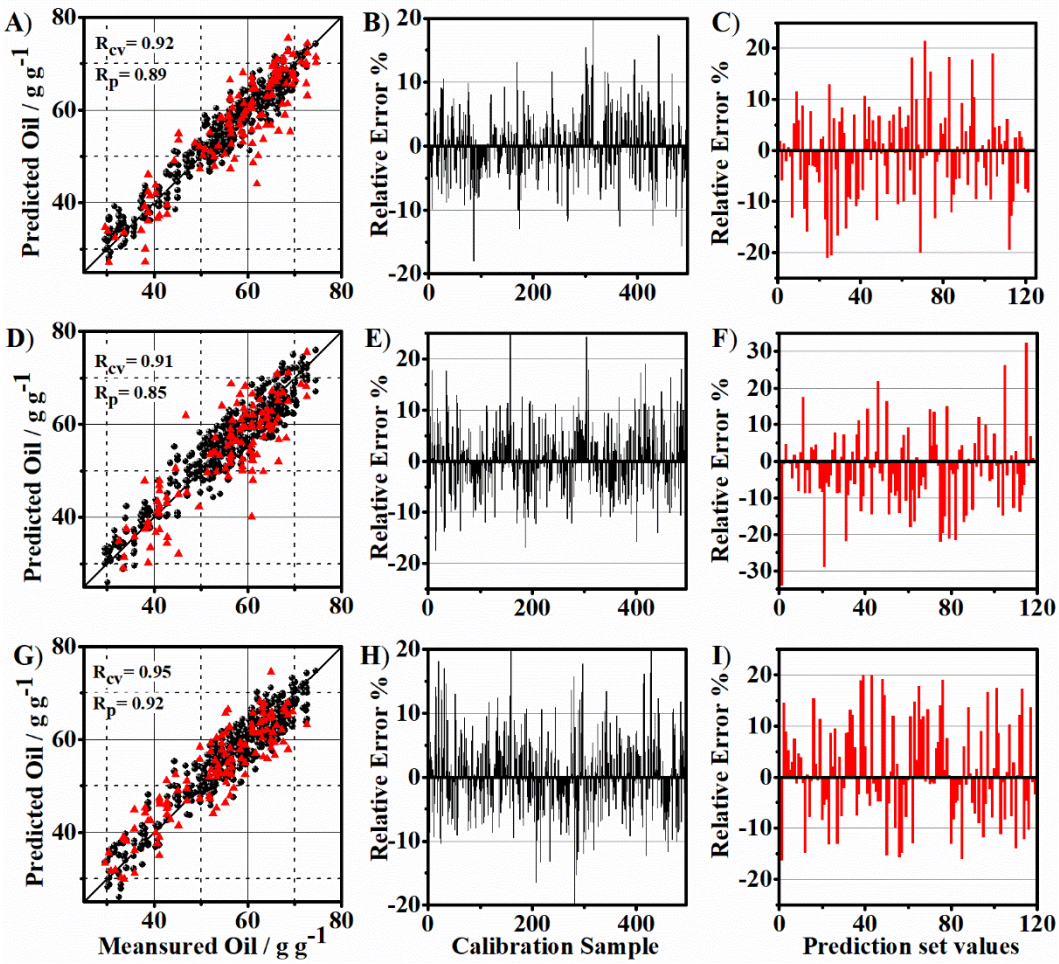

Figure 2.3. Measured and predicted oil content for (A) Shell5; (D) Shell30; (G) Pulp30 (● represents the regression calibration set and ● the prediction sample set). Relative error for calibration set. (B) Shell5; (E) Shell30; (H) Pulp30. Relative error for prediction set (C) Shell5; (F) Shell30; (I) Pulp30.

The above models were evaluated in order to verify if there was a correlation by chance. Fig. 2.4 shows which of the three models presented above (true models) are separated from the other models; so, the true model was not obtained by chance.

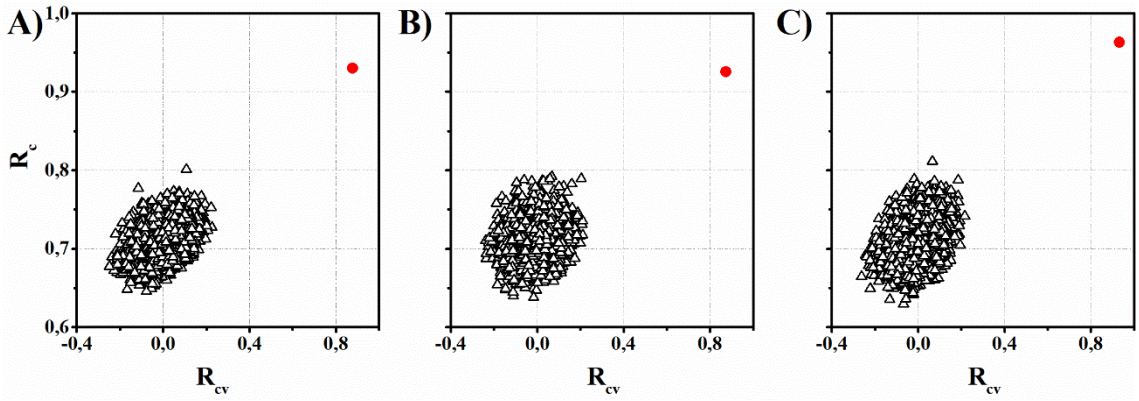

Figure 2.4. Correlation coefficient of cross-validation (R_{cv}) versus correlation coefficient of calibration (R_c) (chance correlation) for (A) Shell5; (B) Shell30; (C) Pulp30.

Fig. 2.5. presents the variables selected for Shell5 (Fig. 2.5A) and Shell30 (Fig. 2.5B). The variables selected for Shell5 and Shell30 were obtained by *FeediOPS* and *AutoiOPS*, respectively. The selected regions, 1000-9000, 8500-7500 cm^{-1} 6100-5500 cm^{-1}, were similar for both sets as the models were built using shell NIR spectra, differing only in the days after harvest. In addition, it can be observed that the selected regions are related to absorbance bands of fatty acids and the bands related to water absorbance were not selected.

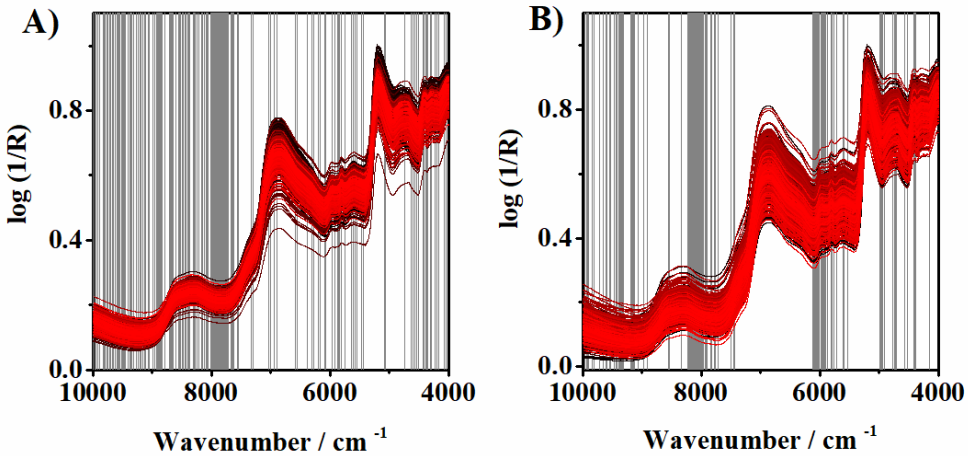

Figure 2.5. Variables selected for (A) Shell5 using *FeediOPS* and (B) Shell30 using *AutoiOPS*.

Thus, the prediction of the oil content in the macaw fruit was possible through NIR of the shell or mesocarp and PLS regression. Although all of the models presented reliable predictive capacity, the model Shell5 was the most advantageous, as the maximum oil content of the ripe fruit could be early predicted 25 days before its accumulation without damaging the fruit. This method has the potential to improve the harvest and storage practices aiming the oil's quality and reduce the time needed for decision-making related to selection of the best fruits for future extraction. In addition, this work opens the way to further studies relating to early prediction of fruits properties.

2.4. Conclusions

From this study, NIR and PLS regression proved to be helpful in the early quantification of oil content in macaw fruits using spectra from the shell of unripe fruits. A comparison of the results with data from Shell30 and Pulp30 showed that Shell5 obtained statistical parameters and errors similar to or even better. The OPS improved the models Shell5 and Shell30 by selecting more predictive and informative variables. Thus, NIR spectroscopy combined with PLS presented a high capability to early predict the oil content of ripe fruits from the spectra of unripe fruits. The presented method is faster and non-destructive, and it has a lower cost than the reference method.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (Project: CEX - APQ-02254-15).

References

[1] A. Muscat, E.M. de Olde, I.J.M. de Boer, R. Ripoll-Bosch, The battle for biomass: A systematic review of food-feed-fuel competition, Glob. Food Sec. (2019) 100330. https://doi.org/10.1016/j.gfs.2019.100330.

[2] P. Prates-Valério, J.M.F. Celayeta, E.C. Cren, Quality Parameters of Mechanically Extracted Edible Macauba Oils (Acrocomia aculeata) for Potential Food and Alternative Industrial Feedstock Application, Eur. J. Lipid Sci. Technol. 121 (2019) 1800329. https://doi.org/10.1002/ejlt.201800329.

[3] J. Poetsch, D. Haupenthal, I. Lewandowski, D. Oberlander, T. Hilger, Acrocomia aculeata – a sustainable oil crop, Rural 21. 03 (2012) 41–44.

[4] ANP, Evolução Anual de Produção de Biodiesel, (2020).

[5] L.R. V. Da Conceição, L.M. Carneiro, J.D. Rivaldi, H.F. de Castro, Solid acid as catalyst for biodiesel production via simultaneous esterification and transesterification of macaw palm oil, Ind. Crops Prod. 89 (2016) 416–424. https://doi.org/10.1016/j.indcrop.2016.05.044.

[6] É.C.M. Lanes, S.Y. Motoike, K.N. Kuki, C. Nick, R.D. Freitas, Molecular Characterization and Population Structure of the Macaw Palm, Acrocomia aculeata (Arecaceae), Ex Situ Germplasm Collection Using Microsatellites Markers, J. Hered. 106 (2015) 102–112. http://dx.doi.org/10.1093/jhered/esu073.

[7] W. Machado, A. Figueiredo, M.F. Guimarães, Initial development of seedlings of macauba palm (Acrocomia aculeata), Ind. Crops Prod. 87 (2016) 14–19.

https://doi.org/https://doi.org/10.1016/j.indcrop.2016.04.022.

[8] A.B. Evaristo, J.A.S. Grossi, L.D. Pimentel, S. de Melo Goulart, A.D. Martins, V.L. dos Santos, S. Motoike, Harvest and post-harvest conditions influencing macauba (Acrocomia aculeata) oil quality attributes, Ind. Crops Prod. 85 (2016) 63–73. https://doi.org/https://doi.org/10.1016/j.indcrop.2016.02.052.

[9] A.A. Nunes, S.P. Favaro, F. Galvani, C.H.B. Miranda, Good practices of harvest and processing provide high quality Macauba pulp oil, Eur. J. Lipid Sci. Technol. 117 (2015) 2036–2043. https://doi.org/10.1002/ejlt.201400577.

[10] S.G. Montoya, S.Y. Motoike, K.N. Kuki, A.D. Couto, Fruit development, growth, and stored reserves in macauba palm (Acrocomia aculeata), an alternative bioenergy crop, Planta. 244 (2016) 927–938. https://doi.org/10.1007/s00425-016-2558-7.

[11] G.F. Simiqueli, M.D.V. de Resende, S.Y. Motoike, E. Henriques, Inbreeding depression as a cause of fruit abortion in structured populations of macaw palm (Acrocomia aculeata): Implications for breeding programs, Ind. Crops Prod. 112 (2018) 652–659. https://doi.org/10.1016/j.indcrop.2017.12.068.

[12] W.W. Tilahun, J.A.S. Grossi, S.P. Favaro, C.S. Sediyama, S.D.M. Goulart, L.D. Pimentel, S.Y. Motoike, Increase in oil content and changes in quality of macauba mesocarp oil along storage, OCL. 26 (2019) 20. https://doi.org/10.1051/ocl/2019014.

[13] T.P. Pires, E. dos Santos Souza, K.N. Kuki, S.Y. Motoike, Ecophysiological traits of the macaw palm: A contribution towards the domestication of a novel oil crop, Ind. Crops Prod. 44 (2013) 200–210. https://doi.org/https://doi.org/10.1016/j.indcrop.2012.09.029.

[14] S. Monselise, Handbook of Fruit Set and Development, Boca Raton, 1986. https://doi.org/https://doi.org/10.1201/9781351073042.

[15] A.M. Costa, S.Y. Motoike, T.R. Corrêa, T.C. Silva, S.M. Coser, M.D.V. de Resende, R.F. Teófilo, Genetic parameters and selection of macaw palm (Acrocomia aculeata) accessions: an alternative crop for biofuels, Crop Breed. Appl. Biotechnol. 18 (2018) 259–266. https://doi.org/10.1590/1984-70332018v18n3a39.

[16] L. Ferreira de França, G. Reber, M.A.A. Meireles, N.T. Machado, G. Brunner, Supercritical extraction of carotenoids and lipids from buriti (Mauritia flexuosa), a fruit from the Amazon region, J. Supercrit. Fluids. 14 (1999) 247–256. https://doi.org/10.1016/S0896-8446(98)00122-3.

[17] A comparative study of various oil extraction techniques from plants , Rev. Chem. Eng. . 30 (2014) 605. https://doi.org/10.1515/revce-2013-0038.

[18] B.E. Richter, B.A. Jones, J.L. Ezzell, N.L. Porter, N. Avdalovic, C. Pohl, Accelerated Solvent Extraction: A Technique for Sample Preparation, Anal. Chem. 68 (1996) 1033–1039. https://doi.org/10.1021/ac9508199.

[19] J.A. Panford, J.M. deMan, Determination of oil content of seeds by NIR: Influence of fatty acid composition on wavelength selection, J. Am. Oil Chem. Soc. 67 (1990) 473–482. https://doi.org/10.1007/BF02540751.

[20] S. Tilahun, D.S. Park, M.H. Seo, I.G. Hwang, S.H. Kim, H.R. Choi, C.S. Jeong, Prediction of lycopene and β-carotene in tomatoes by portable chroma-meter and VIS/NIR spectra, Postharvest Biol. Technol. 136 (2018) 50–56. https://doi.org/10.1016/j.postharvbio.2017.10.007.

[21] D.S. Ferreira, O.F. Galão, J.A.L. Pallone, R.J. Poppi, Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for determination of quality

parameters in soybean samples, Food Control. 35 (2014) 227–232. https://doi.org/10.1016/j.foodcont.2013.07.010.

[22] C. Pasquini, Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications, J. Braz. Chem. Soc. 14 (2003) 198–219. https://doi.org/10.1590/S0103-50532003000200006.

[23] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS modeling, Chemom. Intell. Lab. Syst. 58 (2001) 131–150. https://doi.org/https://doi.org/10.1016/S0169-7439(01)00156-3.

[24] O.O. Olarewaju, I. Bertling, L.S. Magwaza, Non-destructive evaluation of avocado fruit maturity using near infrared spectroscopy and PLS regression models, Sci. Hortic. (Amsterdam). 199 (2016) 229–236. https://doi.org/10.1016/j.scienta.2015.12.047.

[25] U. Saha, D. Jackson, Analysis of moisture, oil, and fatty acid composition of olives by near-infrared spectroscopy: development and validation calibration models, J. Sci. Food Agric. 98 (2018) 1821–1831. https://doi.org/10.1002/jsfa.8658.

[26] G. Altieri, A. Matera, F. Genovese, G.C. Di Renzo, Models for the rapid assessment of water and oil content in olive pomace by near-infrared spectrometry, J. Sci. Food Agric. 100 (2020) 3236–3245. https://doi.org/10.1002/jsfa.10361.

[27] Sudarno, D.D. Silalahi, T. Risman, B.L. Widyastuti, F. Davrieux, Y.Y. Yuan, J.P. Caliman, Rapid determination of oil content in dried-ground oil palm mesocarp and kernel using near infrared spectroscopy, J. Near Infrared Spectrosc. 25 (2017) 338–347. https://doi.org/10.1177/0967033517732679.

[28] J.G. Tallada, N. Palacios-Rojas, P.R. Armstrong, Prediction of maize seed attributes using a rapid single kernel near infrared instrument, J. Cereal Sci. 50 (2009) 381–387. https://doi.org/10.1016/j.jcs.2009.08.003.

[29] B. Steuer, H. Schulz, E. Läger, Classification and analysis of citrus oils by NIR spectroscopy, Food Chem. 72 (2001) 113–117. https://doi.org/10.1016/S0308-8146(00)00209-0.

[30] S.F.S. Matsimbe, S.Y. Motoike, F.A. de C. Pinto, H.G. Leite, G.E. Marcatti, Prediction of oil content in the mesocarp of fruit from the macauba palm using spectrometry1, Rev. Ciência Agronômica. 46 (2015) 21–28. https://doi.org/10.1590/S1806-66902015000100003.

[31] N. de A. Porto, J. V. Roque, C.A. Wartha, W. Cardoso, L.A. Peternelli, M.H.P. Barbosa, R.F. Teófilo, Early prediction of sugarcane genotypes susceptible and resistant to Diatraea saccharalis using spectroscopies and classification techniques, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 218 (2019) 69–75. https://doi.org/10.1016/j.saa.2019.03.114.

[32] A.B. Vitor, R.P. Diniz, C.V. Morgante, R.P. Antônio, E.J. de Oliveira, Early prediction models for cassava root yield in different water regimes, F. Crop. Res. 239 (2019) 149–158. https://doi.org/10.1016/j.fcr.2019.05.017.

[33] J. V. Roque, W. Cardoso, L.A. Peternelli, R.F. Teófilo, Comprehensive new approaches for variable selection using ordered predictors selection, Anal. Chim. Acta. 1075 (2019) 57–70. https://doi.org/10.1016/j.aca.2019.05.039.

[34] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. https://doi.org/10.1002/cem.1192.

[35] IAL, Métodos químicos e físicos para análise de alimentos, in: O. Zenebon, N.S. Pascuet,

P. Tiglea (Eds.), Normas Analíticas Do Inst. Adolf Lutz, 4th ed., Instituto Adolfo Lutz, São Paulo, 2008: pp. 116–118.

[36] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148. https://doi.org/10.1080/00401706.1969.10490666.

[37] J. V. Roque, L.A.S. Dias, R.F. Teófilo, Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies, J. Braz. Chem. Soc. (2017). 1506-1516. https://doi.org/10.21577/0103-5053.20160332..

[38] M.K.D. Rambo, E.P. Amorim, M.M.C. Ferreira, Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues, Anal. Chim. Acta. 775 (2013) 41–49. https://doi.org/10.1016/j.aca.2013.03.015.

[39] P. Hourant, V. Baeten, M.T. Morales, M. Meurens, R. Aparicio, Oil and Fat Classification by Selected Bands of Near-Infrared Spectroscopy, Appl. Spectrosc. 54 (2000) 1168–1174. https://doi.org/10.1366/0003702001950733.

[40] T. Sato, S. Kawano, M. Iwamoto, Near infrared spectral patterns of fatty acid analysis from fats and oils, J. Am. Oil Chem. Soc. 68 (1991) 827–833. https://doi.org/10.1007/BF02660596.

CAPÍTULO 3

COMPARAÇÃO ENTRE INSTRUMENTOS NIR DE BANCADA E PORTÁTIL NA DETERMINAÇÃO DO TEOR DE AMILOSE EM FÉCULA DE MANDIOCA

Resumo

O objetivo deste trabalho foi comparar modelos de regressão e classificação multivariados construídos a partir de espectros NIR obtidos em dois diferentes instrumentos para determinação da amilose aparente (AM) em mandioca. A determinação da AM em mandioca foi realizada de acordo com o procedimento ISO-1987 modificado. Os valores variaram entre 8 e 43%. Espectros NIR de amostras de polpa de mandioca desidratada (*Manihot esculenta Crantz*) foram obtidos em dois instrumentos, um de bancada (NIRB) usando a região de 10000-4000 cm^{-1} com incremento de 1,93 cm^{-1} e outro portátil (NIRP) usando a região de 11111-5882 cm^{-1} com incremento de 8,64 cm^{-1}. Modelos quantitativos para estimar o teor de AM foram desenvolvidos utilizando a regressão PLS e a seleção de variáveis pelo método OPS. Modelos de classificação usando PLS-DA foram construídos a fim de classificar as amostras com valores de AM menores ou iguais a 20,0 % (classe 1) e com valores maiores que 20,0% (classe 2). Para os modelos de calibração após seleção de variáveis, os valores da *RMSEP* e R_P, foram, respectivamente, 3,37 % e 0,90 para o NIRB, e 5,05 % e 0,65 para o NIRP. Para os modelos de classificação (NIRB e NIRP), os resultados de sensibilidade, especificidade e erro de classificação para classe 1 foram iguais a 1,000; 0,952 e 0,000, respectivamente. Sendo tais resultados iguais para ambos os instrumentos. Os resultados deste trabalho mostraram a viabilidade e as limitações da utilização dos espectrômetros portáteis frente aos de bancadas para determinar e classificar o teor de AM em amostras de mandioca.

Palavras-Chave: mandioca, análise multivariada, quadrados mínimos parciais, análise discriminante, NIR portátil.

3.1. Introdução

A mandioca (*Manihot esculenta Crantz*) é atualmente a quarta mais importante cultura alimentar do mundo e a principal na região tropical, sendo consumida por mais de 800 milhões de pessoas [1–3]. A mandioca é consumida principalmente em lugares onde prevalecem a seca, a pobreza e a desnutrição. É predominantemente produzida por pequenos agricultores, tendo portanto, importância econômica e social [4,5].

Dentre os produtos derivados da mandioca, destacam-se os polvilhos doce e azedo, amplamente conhecidos no Brasil, sendo produzido por meio da fécula de mandioca, após extração, fermentação e secagem ao sol [6].

A fécula de mandioca é uma mistura heterogênea de duas macromoléculas, amilose e amilopectina (Fig. 3.1), que diferem no tamanho molecular e grau de ramificação [7]. A proporção dessas duas moléculas estão relacionadas com a atividade espessante, de retenção de água, estabilizador de emulsificação e agente gelificante, muito importantes na indústria de alimentos [8].

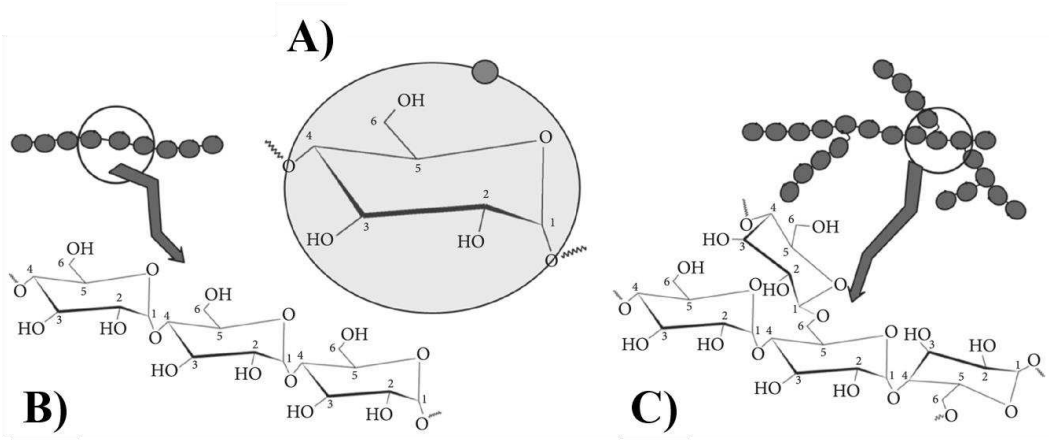

Figura 3.1. Estrutura básica (A) unidades de glicose, (B) amilose e (C) amilopectina, junto com a marcação dos átomos e ângulos de torção.

Sabe-se que há uma grande variabilidade genética para o conteúdo de nutrientes que podem ser armazenados na raiz da mandioca [3]. Assim a análise da amilose, e da amilopectina por diferença da amilose, de forma rápida e confiável pode possibilitar o isolamento e produção de genótipos de melhores qualidades nutricionais.

A técnica mais empregada na análise de amilose, também chamada de determinação da amilose aparente, é por espectrofotometria UV-Vis. Apesar da sua alta precisão e exatidão nas análises, essa técnica apresenta-se cara, demorada e faz-se o uso de reagentes tóxicos.

A espectroscopia NIR tem sido amplamente utilizada em diversas matrizes alimentícias [9,10], biológicas [11], ambientais [12] e vegetais [13–15]. Tal crescimento na utilização se deve às suas vantagens sobre outras técnicas analíticas, por ser não destrutiva e possui alta frequência analítica para amostras sólidas, líquidas e gasosas [16,17].

Uma inovação no campo da espectroscopia NIR está relacionada com os instrumentos portáteis. Alguns instrumentos com boa qualidade espectral podem ser encontrados a um custo inferior a US$ 1,000.00 [18]. Como exemplo, o utilizado nesse trabalho, embora muito compacto (100g, 10×10×5 cm), possui uma faixa de 900 – 1700 nm com incremento de 1,30 nm, podendo ser operado de forma autônoma usando baterias e possuir, entre outras, a conexão Bluetooth.

Os instrumentos portáteis tem se estabelecido com uma nova proposta em virtude da sua aplicação em campo, oferendo medições com alta velocidade, robustez, estabilidade, portabilidade e baixo consumo de energia, sendo usado com sucesso em várias aplicações [9,19,20].

A espectroscopia NIR aliada a métodos quimiométricos tem sido aplicada na quantificação e classificação de diversas propriedades da mandioca, como na previsão antecipada do crescimento das raízes [21], na previsão e seleção de características agronômicas e fisiológicas para tolerância ao déficit hídrico [21] e previsão do conteúdo de carotenoides [13,22]. Entretanto, a determinação de amilose em amostras de fécula de mandioca utilizando espectroscopia NIR e métodos quimiométricos ainda não foi estudada. Além disso, o estudo dessa propriedade pelos NIR portáteis possibilitará avanços na escolha de genótipos nas próprias áreas de plantações.

Os espectros obtidos tanto por instrumentos de bancada (NIRB) quanto por instrumentos portáteis (NIRP) possuem um grande número de variáveis. Devido à alta colinearidade das colunas da matriz de dados, torna-se necessário a utilização de métodos multivariados, como a calibração com PLS e a classificação via PLS-DA [16,23].

Neste contexto, este trabalho tem como objetivo comprar dois instrumentos NIR, bancada e portátil para predizer/classificar o teor de amilose aparente em vários genótipos de mandioca. Para alcançar o objetivo, análises multivariadas empregando os métodos PLS e PLS-DA foram executados e comparados para cada instrumento.

3.2. Materiais e Métodos

3.2.1. Origem das amostras

Os clones foram produzidos na área experimental da Embrapa Mandioca e Fruticultura no município de Cruz das Almas, Bahia, Brasil ((Latitude 12 ° 40'19" S, Longitude 39°06’22” O, altitude 226 m). O amido foi extraído de raízes colhidas com cerca de 12 meses de idade. Foram selecionadas 122 amostras de amidos de mandioca para a realização do trabalho.

3.2.2. Análise de amilose aparente

As raízes oriundas da Embrapa- Mandioca e Fruticultura foram limpas escovadas para a determinação do teor de amilose aparente segundo Fernandez *et al*. [24]. As amostras foram descascadas e cortadas em tamanhos menores e trituradas com água (1:1) em um liquidificador. Logo em seguida foram passadas em peneira com 150 mesh. A suspensão foi mantida em câmara fria (5°C) por 12h para a decantação do amido. Logo em seguida o amido foi recolhido por filtração e seco em temperatura ambiente por 24h.

A porcentagem de amilose presente no amido da mandioca foi determinada segundo a metodologia ISO 6647 [25]. Os grânulos de amido, 100 mg, foram transferidos para um balão volumétrico de 100 mL previamente identificados. Posteriormente foi adicionado 1 mL de álcool etílico 95% (v/v) e agitado cuidadosamente. A seguir, foi acrescento 9 mL de solução de NaOH 1 $molL^{-1}$. O balão foi tampado e deixado em repouso, por 12 horas, para gelatinização do amido. Posteriormente o volume do balão volumétrico foi completado com água destilada. Retirou-se uma alíquota de 5 mL da solução preparada e transferiu-se para outro balão volumétrico de 100 mL devidamente identificado. Então foram adicionados 1 mL de ácido acético 1 $molL^{-1}$ e 2 mL de solução de iodo-iodeto de potássio de I_2 1 g e 20 g de KI para 1 litro de solução.

O volume do balão volumétrico foi completado com água destilada e a solução foi deixada em repouso durante 30 minutos no escuro. Em seguida, foi feita a leitura de absorbância a 620 nm.

A curva analítica foi construída usando o mesmo procedimento acima, substituindo a amostras por concentrações conhecidas de amilose (0, 10, 20, 30, 40, 50 %) e amilopectina (Sigma Aldrich).

3.2.3. Obtenção dos espectros NIR

Os espectros NIR das amostras foram obtidos utilizando o espectrômetro de bancada Antaris II da Thermo Scientific (NIRB) e o espectrômetro portátil Nano-NIR, DLP NIRscan da Texas Instrument (NIRP).

O espectrômetro NIRB foi operado em um módulo de refletância difusa com esfera de integração. As varreduras das amostras foram o resultado médio de 32 varreduras medidas ao longo da faixa 10.000-4.000 cm^{-1} com incrementos de 1,92 cm^{-1}. Nesta configuração cada espectro foi obtido com 3112 valores de reflectância.

O espectrômetro NIRP foi também operado em um módulo de refletância difusa, realizando 50 varreduras para cada amostra ao longo de uma faixa de 11.111,1 – 5.882,4 cm^{-1}com incrementos de 8,64 cm^{-1}. Nesta configuração o espectro foi obtido com 605 valores de reflectância.

Para ambos os instrumentos o logaritmo na base 10 do inverso da refletância (log (1/R) foi coletado. Cada amostra de fécula foi colocada em um suporte de quartzo e foi centralizada diretamente na janela do instrumento (Fig. 3.1), sem qualquer preparação adicional. Foram obtidos dois espectros por amostra, e o espectro médio de cada amostra foi usado como variável independente.

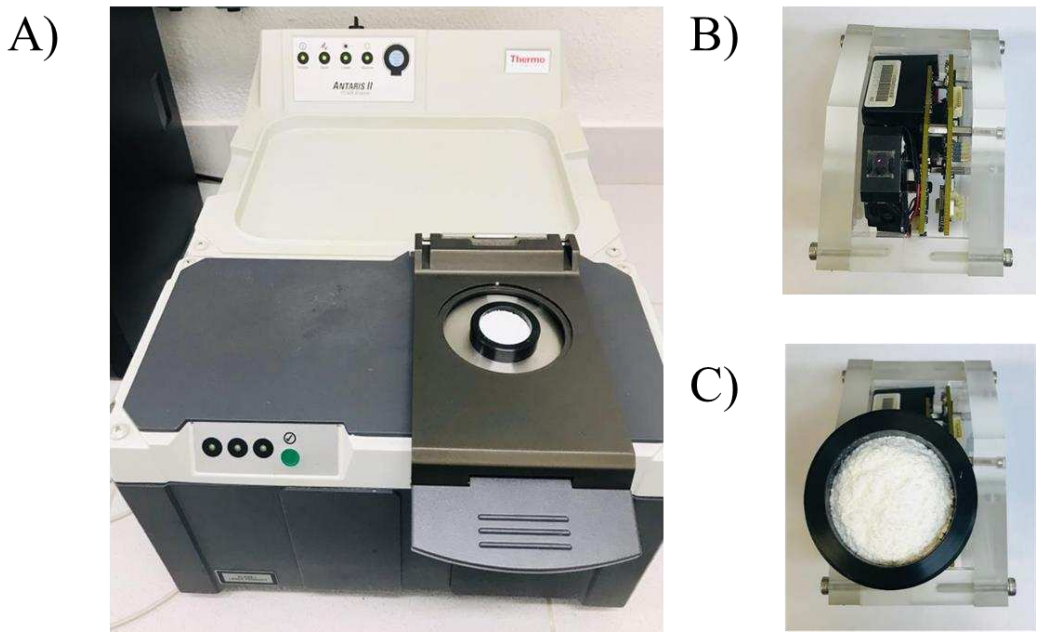

Figura 3.1. Espectrofotômetros Antaris II da Thermo Scientific-NIRB com suporte de quartzo mais amostra (A) e Nano-NIR, DLP NIRscan da Texas Instrument-NIRP sem (B) e com (C) suporte de quartzo mais amostra.

3.2.4. Modelos de calibração multivariada-PLS

O vetor **y** (variável dependente) e a matriz **X** dos espectros NIR (variáveis independentes) estão associados aos valores do teor de amilose aparente e os espectros NIR do amido de mandioca. O algoritmo Kennard-Stone [26] foi usado para dividir as amostras em conjuntos de calibração e previsão. Para todo os modelos, 98 e 24 amostras foram selecionadas para os conjuntos de calibração e predição, respectivamente.

Os espectros foram importados e organizados no *software* Matlab 2019a (Math Works, Natick, EUA). Um modelo de regressão inversa (**y = Xb**) foi construído usando o algoritmo bidiagonal PLS [27,28]. Os algoritmos para a importação de dados, a calibração e os modelos de validação foram escritos em nosso laboratório em função *.m* para o *software* Matlab. Diferentes métodos OPS [28] foram aplicadas utilizando os algoritmos disponíveis em www.deq.ufv.br/chemometrics. Todos os cálculos foram realizados no *software* Matlab.

3.2.5. Figuras de mérito PLS

A qualidade dos modelos foi avaliada pela raiz quadrada do erro quadrático médio *(RMSE*) e pelo coeficiente de correlação (*R*) calculados pelas Equações. (3.1) e (3.2), respectivamente.

$$RMSE = \sqrt{\frac{\sum_{i}^{N}(y_i - y_1)^2}{N}} \qquad \textbf{(3.1}$$

$$R = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \overline{\hat{y}}\right)\left(\hat{y}_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \overline{y}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \overline{\hat{y}}\right)^2}} \qquad \textbf{(3.2)}$$

onde $\hat{y}$ e $\overline{\hat{y}}$ são o valor estimado e a média dos valores estimados, respectivamente, sendo *y* os valores observados e a $\hat{y}$ média dos valores observados; e *N* representa o número de amostras. No caso em que a validação cruzada interna (*CV*) é usada, *N* representa o número de amostras no conjunto de validação cruzada, e o erro e o coeficiente de correlação são chamados de raiz quadrada do erro quadrático médio de validação cruzada (*RMSECV*) e coeficiente de correlação validação cruzada (R_{vc}). Para a

validação externa, representamos o número de amostras de predição (*P*) e, neste caso, os coeficientes de correlação de erros e são denominados raiz quadrada do erro quadrático médio de predição (*RMSEP*) e coeficientes de correlação de predição (R_P).

Sensibilidade analítica (γ^{-1}), seletividade (*SEL*) e limite de detecção (*LOD*) foram calculados de acordo com as Eqs. (3.3) a (3.5), respectivamente.

$$\gamma^{-1} = \left\|\partial_x\right\| \times \left\|b\right\| \tag{3.3}$$

em que $\| b \|$ é a norma euclidiana do vetor de coeficientes de regressão do modelo PLS e $\left\|\partial_x\right\|$ a norma do desvio padrão do sinal de referência [29].

$$SEL = \frac{nas_i}{\left\|\mathbf{x}_i\right\|} \tag{3.4}$$

em que nas_i é o valor escalar do sinal analítico da amostra *i*. A norma euclidiana do vector de resposta instrumental para a amostra *i* é representada por $\| \boldsymbol{x}_\mathrm{i} \|$ [30,31].

$$LOD = \frac{\Delta(\alpha,\beta)\, w_{y_0}\sigma}{\hat{a}} \tag{3.5}$$

em que o termo $w_{y_0}\sigma$ é uma estimativa do erro no intercepto e o fator $\Delta(\alpha,\beta)$ depende da probabilidade α (erro tipo I) e β (erro tipo II) e do número de graus de liberdade.

Cada modelo foi analisado frente a correlação por chance. O vetor **y** foi randomizado 10.000 vezes e os modelos foram construídos usando essas randomizações. A correlação dos modelos construídos foi avaliada. Se a correlação de **y** autêntico se distância dos valores de correlação dos valores aleatórios de **y**, é uma indicação de que o modelo não ocorreu por acaso.

3.2.6. Análise de dados PLS-DA

A organização dos dados se deu semelhantemente à do PLS, substituindo os valores de ***y*** por valores categóricos. Os algoritmos usados para construir e validar os modelos foram escritos em nosso laboratório usando o *software* Matlab R2019a (MathWorks Inc., Natick, EUA).

A validação cruzada aleatória foi aplicada com divisões ajustadas para 20% do conjunto de calibração.

O valor limite para a discriminação de classes foi determinado usando uma função normal de densidade de probabilidade [32]. Os valores previstos foram comparados com

a linha limite. As amostras com valores acima do limiar foram consideradas pertencentes à classe, e as amostras com um valor previsto abaixo do limiar foram consideradas como não pertencendo à classe.

O desempenho do modelo foi avaliado calculando a sensibilidade, especificidade e erro de classificação de calibração e previsão. A sensibilidade representa o número de amostras previsto como pertencente à classe dividido pelo número de amostras que pertencem a classe. A especificidade representa o número de amostras previstas como não pertencentes à classe dividido pelo o número real não está na classe. Sensibilidade, especificidade e erro foram calculados de acordo com as Equação. (3.6), (3.7) e (3.8), respectivamente.

$$\text{Sensibilidade} = \frac{VP}{VP + VN} \tag{3.6}$$

$$\text{Especificidade} = \frac{VN}{VN + VP} \tag{3.7}$$

$$\text{Erro} = \frac{FP + FN}{VP + VN + FP + FN} \tag{3.8}$$

onde *VP* é verdadeiro positivo e *VN* é verdadeiro negativo, *FN* é falso negativo e *FP* é falso positivo.

VP é o número de amostras que pertencem à classe *i* classificada como pertencente à classe *i*. *VN* é o número de amostras que não pertencem à classe *i* classificada como não pertencente à classe *i*. *FN* é o número de amostras que pertencem à classe *i* não classificadas como pertencentes à classe *i*. *FP* é o número de amostras que não pertencem à classe *i* classificada como pertencente à classe *i*.

3.2.7. Seleção dos preditores ordenados - OPS

Quatro métodos OPS [53], *AutoOPS, FeedOPS, AutoiOPS* e *FeediOPS*, foram aplicados para selecionar mais variáveis preditivas, visando aprimorar os modelos. Primeiramente, um método automático, *AutoOPS*, que desenvolve e executa seleção variáveis usando vários vetores informativos e suas combinações. Os Em segundo, o *FeedOPS* apresenta uma nova estratégia as variáveis pré-selecionadas, retornando a uma nova seleção. Por último, foi estabelecido um método para aplicar OPS em subdivisões de matriz dos dados chamados OPS intervalado, *AutoiOPS* e *FeediOPS.*

A seleção de variáveis foi aplicada para regressão via PLS e classificação por PLS-DA.

3.3. Resultados e Discussões

A equação do modelo e o coeficiente de determinação do teor de amilose aparente foram de *y = 0,1 +1,14x e 0,999*, respectivamente. A concentração de amilose na fécula de mandioca variou de 8 a 43 %.

A Fig. 3.2 apresenta os espectros NIR obtidos nos instrumentos NIRB e NIRP. Pode-se observar diferenças entre as regiões espectrais para dois instrumentos, onde o instrumento de bancada abrange uma faixa de 1000 a 2500 nm e o instrumento portátil de 900 a 1700 nm. Pode-se observar a alta semelhança nos espetros e a mesma variação da intensidade para a mesma região espectral.

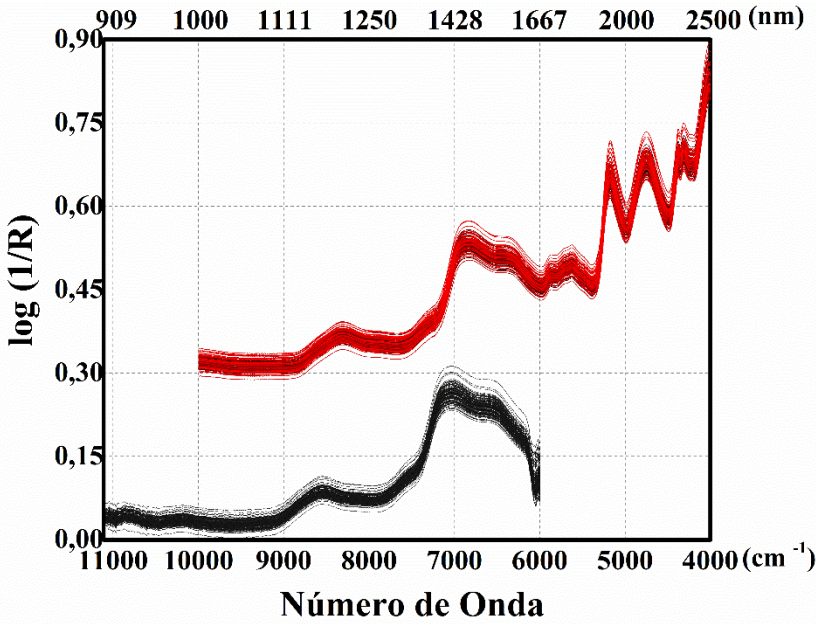

Figura 3.2. Espectros NIR do amido da mandioca para NIRB (vermelho) e NIRP (preto).

Os espectros de NIR originais com suas respectivas transformações são mostrados na Fig. 3.4A e 3.4B. O ruído e o deslocamento da linha de base eram características evidentes e indesejáveis [33]. Assim, transformações matemáticas nas linhas da matriz foram essenciais para reduzir erros sistemáticos e aleatórios e aumentar a relação sinal-ruído. As transformações aplicadas foram: (1) alisamento, (2) primeira derivada (D1), (3) segunda derivada (D2), (4) correção de sinal multiplicativo (MSC), (5) detrend (6) normalização, (7) variação normal padrão (SNV) e (8) combinações destas transformações foram testadas para encontrar o melhor modelo de regressão. Além das

transformações os pré-processamentos centrar na média e autoescalamento foram aplicados.

Para todos os conjuntos de dados apresentados, os parâmetros estatísticos *RMSECV* e R_{CV} foram comparados. A transformação que apresentou os melhores resultados para todos os conjuntos de dados NIRB foram o alisamento (21) seguida SNV e D1 com janela 9. O algoritmo Savitzky-Golay foi usado para obtenção de D1. Já para o conjunto de dados NIRP, as transformações utilizadas foram a norma (1) seguida por auto escalonamento. Os resultados dos pré-tratamentos são mostrados na Fig. 2.3A e 2.3D para os dados NIRB e NIRP, respectivamente.

Após escolher os melhores pré-tratamentos para ambos os conjuntos, os parâmetros estatísticos e figuras de mérito de modelos PLS usando todas as variáveis (*completo*), variáveis selecionadas pelo *Auto*OPS, *Feed*OPS, *Auto*iOPS e *Feed*iOPS para variáveis NIR são mostradas na Tabela 3.1.

Analisando primeiramente os resultados para o instrumento NIRB, temos cincos modelos construídos. A escolha do melhor modelo de calibração se baseou nos seus resultados de previsão. Assim sendo, o modelo que apresentou menores valores de *RMSEP* para os dados NIRB foi o *FeedOPS*, onde foram selecionadas apenas 145 variáveis.

Analisando os resultados dos modelos para o instrumento NIRP apresentados na Tabela 3.1, pode-se verificar que a seleção de variáveis *AutoOPS* conseguiu melhorar os resultados dos valores de previsão em relação ao modelo completo. Assim o modelo *AutoOPS* foi escolhido como melhor modelo para o instrumento NIRP.

A sensibilidade analítica (γ^{-1}) define a menor diferença de concentração entre amostras que podem ser distinguidas pelo método. Discutindo essa grandeza, o modelo *AutoiOPS* o menor valor de γ^{-1} para o conjunto NIRB. Já para os dados de NIRP o *Auto*OPS apresentou menor valor dessa grandeza.

Os valores SEL podem variar de 0 a 1, onde 0 significa que o sinal analítico contém informações de interferentes, enquanto o valor 1 indica que não possui interferentes. Assim, a baixa seletividade era esperada para ambos os conjuntos visto que os modelos foram construídos com base em espectros obtidos com interferentes.

Finalmente, o menor valor de LOD para o instrumento NIRB foi para o modelo *Auto*OPS, como esperado. Já para o instrumento NIRP seus valores não variaram, mas foram maiores que os para os conjuntos do instrumento NIRB.

Comparando o valor de *RMSEP,* do *FeedOPS* para os dados NIRB, como o menor valor dos dados de referência para o teor de amilose podemos perceber que seu valor é bem menor, aproximadamente 10% do seu valor. Isso indica que o erro está muito baixo perto do menor valor. O mesmo não acontece com o valor *RMSEP*, do *AutoOPS* para os dados NIRP, onde seu valor é certa de 50% em relação ao menor valor dos dados de referência.

Tabela 3.1. Parâmetros estatísticos do NIR-PLS com variáveis completas e variáveis selecionadas por OPS e figuras de mérito para os dados provenientes do NIRB e NIRP

	Modelos PLS – NIRB					**Modelos PLS – NIRP**				
	Completo	***AutoOPS***	***FeedOPS***	***AutoiOPS***	***FeediOPS***	***Completo***	***AutoOPS***	***FeedOPS***	***AutoiOPS***	***FeediOPS***
nVars	3112	475	145	205	120	605	505	385	325	340
RMSEC / (%)	0,978	0,641	0,484	0,747	0,823	2,350	2,323	2,619	2,355	2,435
R	0,988	0,995	0,997	0,993	0,993	0,927	0,928	0,929	0,826	0,921
RMSECV / (%)	2,787	1,403	0,988	1,679	1,572	4,686	4,824	4,449	4,424	4,354
Rcv	0,898	0,975	0,988	0,964	0,968	0,665	0,646	0,704	0,710	0,718
RMSEP / (%)	2,699	1,148	0,881	1,249	1,407	4,267	4,198	4,318	4,711	4,301
Rp	0,933	0,986	0,994	0,983	0,983	0,764	0,777	0,771	0,728	0,776
γ^{-1} / (%)	0,201	0,024	0,164	0,016	0,059	0,236	0,215	0,236	0,306	0,476
SEL	0,004	0,006	0,007	0,006	0,005	0,011	0,011	0,012	0,011	0,010
LOD / (%)	2,086	1,349	1,449	1,576	1,722	4,000	4,000	4,000	4,000	4,000

nlvs: número de variáveis latentes; *nVars*: número de variáveis; RMSEC: raiz quadrada do erro quadrático médio da calibração; RMSECV: raiz quadrada do erro quadrático médio da validação cruzada RMSEP: raiz quadrada do erro quadrático médio da previsão; γ^{-1}: sensibilidade analítica; SEL: seletividade; LOD: limite de detecção. O valor de *nlvs* foi igual a 10 para todos os modelos.

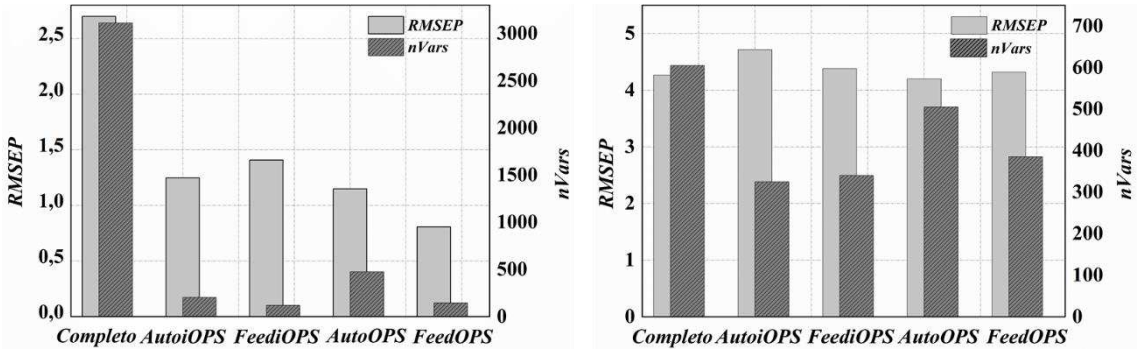

Figura 3.3. Gráfico com os valores *RMSEP* e *nVars* por modelos construidos para os dados NIRB (A) e NIRP (B).

A Fig. 3.3 representa os valores de *RMSEP* e *nVar* para os cinco modelos construídos. Na Fig. 3.3A e 3.3B podemos perceber a importância da seleção de variáveis, onde houve uma redução no valor de *RMSEP* com a seleção, selecionando-se as variáveis mais preditivas.

Os resultados do *FeedOPS* para o instrumento NIRB são apresentados na Fig. 3.5B. O melhor modelo para o modelo NIRP foi o *AutoOPS,* com valores ligeiramente melhores que o apresentado com todas as variáveis, sendo apresentado na Fig. 3.5E. Comparando visualmente os dois gráficos podemos perceber uma maior variação nos valores preditos e medidos para o conjunto NIRP. Isso é confirmado quando observamos as Fig. 3.5C e 3.5F, as quais se referem ao erro relativo de previsão para os conjuntos NIRB e NIRP, respectivamente, onde observamos dois erros altos NIRP, porém, a maioria das amostras possui erro menor que 20%.

Assim, o NIRP não apresentou, para este conjunto de dados, informações suficientes relacionadas à determinação de AM nas amostras de fécula de mandioca. Essa informação pode ser contatada analisando a Fig. 3.4 que apresenta as variáveis selecionadas pelo *FeedOPS* para o modelo NIRB. Pode-se observar por essa figura que existem um conjunto de variáveis preditivas para a construção do modelo na faixa 6000 a 4000 cm^{-1}, faixa essa não contemplada pelo NIRP. Essa faixa corresponde a região de combinação de bandas de pequena intensidade R-OH que estão presentes na estrutura da molécula [34].

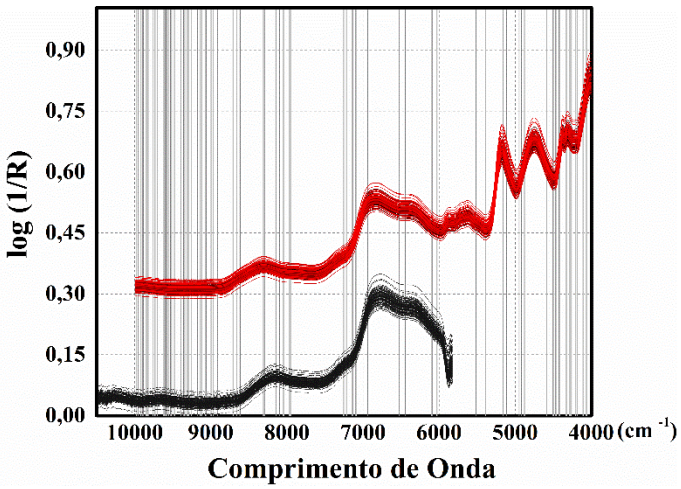

Figura 3.4. variáveis selecionadas pelo FeedOPS para conjunto de dados NIRB.

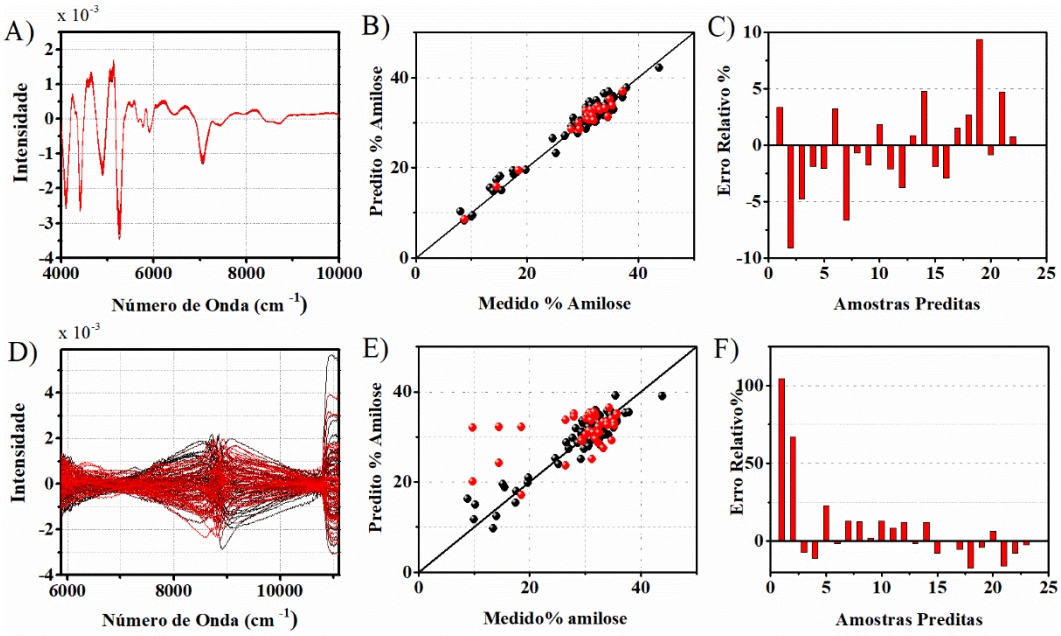

Figura 3.5. Espectros NIR pre-processados para os dados NIRB (A) e NIRP (D), Valores medidos *versus* preditos dos teores de AM para o conjunto de calibração (●) e de predição (●) NIRB 2B) e NIRP (E), Erro relativo da previsão para os dados NIRB (C) e NIRP (F).

Os modelos apresentados na Fig. 3.6 foram avaliados para verificar se houve correlação por acaso. Portanto, o vetor **y** foi randomizado e os modelos foram construídos com **y** aleatórios. Assim, as correlações desses modelos foram calculadas entre os valores de **y** aleatórios e seus valores estimados. Se a correlação de **y** medido com seus valores estimados se distanciam dos valores de correlação dos valores **y** aleatório com seus valores estimados, é uma indicação que o modelo não ocorreu por acaso. Como apresentados na Fig. 3.6, as correlações dos modelos verdadeiros estão separadas daquelas dos modelos construídos com **y** aleatorizado. Este resultado mostra que o modelo verdadeiro não foi obtido por acaso.

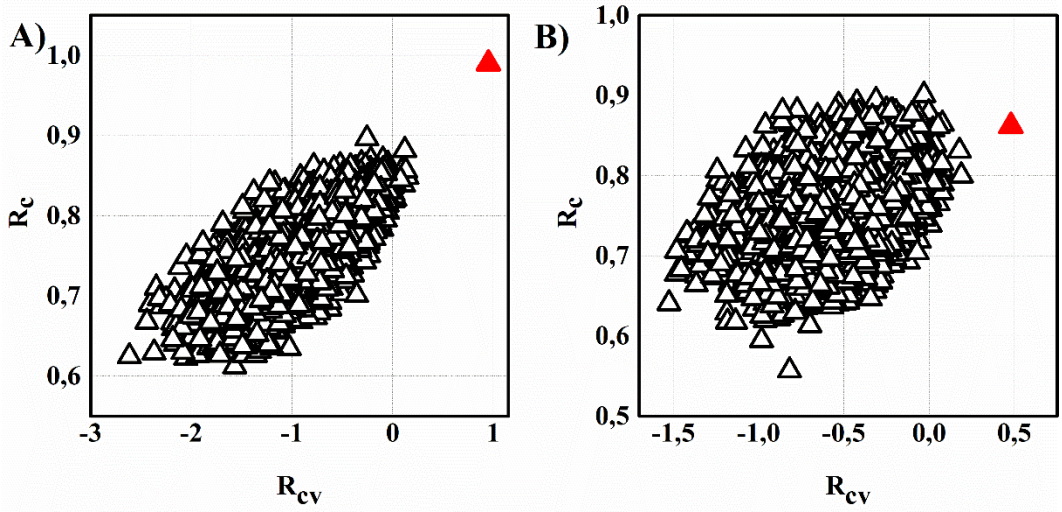

Figura 3.6. Gráfico de correlação por chance: modelo *FeedOPS* NIRB (A), modelo *AutoOPS* NIRP (B).

Pensando na possibilidade de se aplicar os dados do instrumento NIRP, modelos de classificação foram construídos para os mesmos conjuntos anteriormente citados.

Para diversos tipos de tubérculos, raízes e cereais, quantidades superiores a 20% de amilose são consideradas de intermediária a alta, como no caso do arroz, classificação como ceroso (1-2% amilose), muito baixo teor de amilose (2-12%), baixo teor de amilose (12-20 %), teor de amilose intermediária (20-25%) e alto teor de amilose (25-33%) [35]. Assim os modelos foram construídos considerando duas classes diferentes de teor de AM: teor menor que 20,0 % (Classe 1) e maior que 20,0 % (Classe 2).

Os resultados dos modelos PLS-DA para os dados obtidos nos instrumentos NIRB e NIRP são apresentados na Tabela 3.2. Foram construídos três modelos para cada conjunto de dados (NIRB e NIRP). Para todos os conjuntos de dados os modelos PLS-DA forneceram excelente discriminação para os conjuntos de calibração e previsão.

Analisando primeiramente os resultados para o NIRB, pode-se perceber que a seleção de variáveis melhorou a classificação para o conjunto de calibração, mas não os valores de previsão em relação ao modelo completo. Assim o modelo completo foi escolhido como melhor modelo para os dados NIRB. Os resultados para os dados NIRP mostraram que a seleção de variáveis melhorou a classificação em relação ao modelo com todas as variáveis. Os dois métodos de seleção, *AutoiOPS* e *FeediOPS*, tiveram os mesmos resultados.

Tabela 3.2. Parâmetros de classificação da amilose aparente nos conjuntos de dados NIRB e NIRP.

	MODELO PLS-DA – NIRB						**MODELO PLS-DA – NIRP**					
	Full		***Auto-OPSDA***		***Feed-OPSDA***		***Full***		Auto-OPSDA		***Feed-OPSDA***	
	Classe 1	**Classe 2**	**Classe 1**	**Classe 2**	**Classe 1**	**Classe 2**	**Classe 1**	**Classe 2**	**Classe 1**	**Classe 2**	**Classe 1**	**Classe 2**
nlvs	3		3		3		6		6		6	
nVars	3112	3112	307	307	307	307	531	531	70	70	55	55
Sensibilidade (Cal)	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,953	1,000	1,000	1,000	1,000
Especificidade (Cal)	1,000	1,000	1,000	1,000	1,000	1,000	0,953	1,000	1,000	1,000	1,000	1,000
Sensibilidade (Pred)	1,000	0,952	0,800	0,952	0,800	0,952	0,800	0,904	1,000	0,952	1,000	0,952
Especificidade (Pred)	0,952	1,000	0,952	0,800	0,952	0,800	0,904	0,800	0,952	1,000	0,952	1,000
Erro (Cal)	0,000	0,000	0,000	0,000	0,000	0,000	0,038	0,038	0,000	0,000	0,000	0,000
Erro (Pred)	0,038	0,038	0,076	0,076	0,076	0,076	0,115	0,115	0,038	0,038	0,038	0,038

nlvs: número de variáveis latentes; Cal: conjunto de calibração; Pred: conjunto de previsão.

A Fig. 3.7 mostra os resultados dos melhores modelos PLS-DA de treinamento e teste para os dados dos instrumentos NIRB e NIRP. Os quadrados pretos representam as amostras que pertencem à classe 1 e os círculos vermelhos as amostras que pertencem à classe 2. As formas preenchidas e vazias referem-se aos conjuntos de calibração e previsão, respectivamente. A linha preta tracejada é o limite estimado pelo PLS-DA. Idealmente, todos os quadrados pretos devem ser preditos acima do limite e os círculos vermelhos abaixo do limite. Pode-se perceber pela Fig. 3.7 que houve uma excelente discriminação entre as classes, com apenas um erro de previsão nas amostras da classe 1.

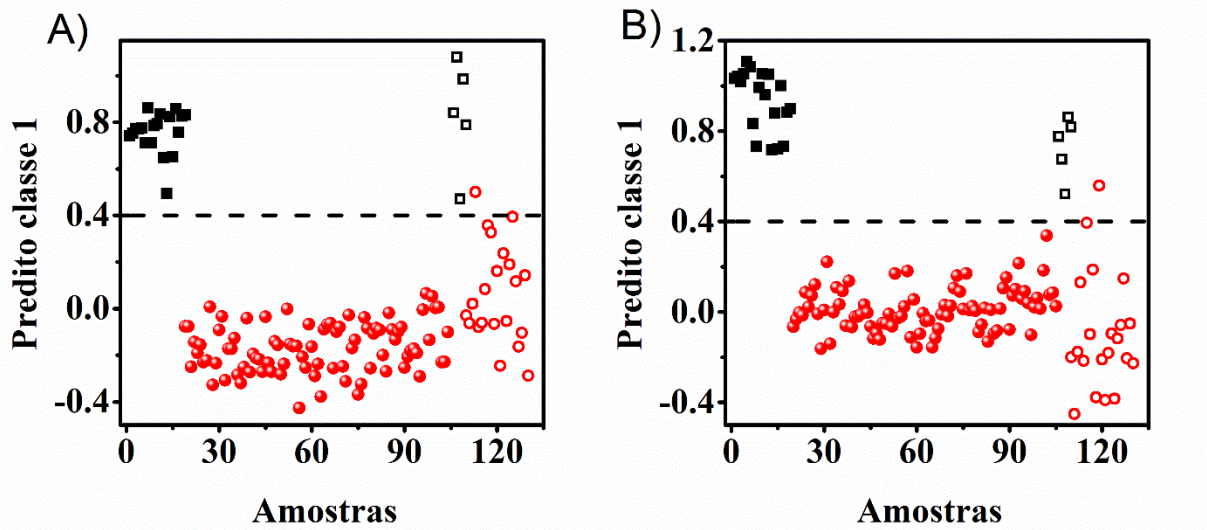

Figura 3.7. Classificação do teor de amilose para os conjuntos de dados (A) NIRB, (B) NIRP. Círculos vermelhos preenchidos (●) e círculos vermelhos vazios (○) são amostras da Classe 1. Quadrados pretos preenchidos (■) e quadrados pretos vazios (□) são amostras da Classe 2. Formas preenchidas e vazias referem-se ao conjunto de calibração e previsão, respectivamente . A linha preta tracejada (---) indica o limite estimado pelo algoritmo PLS-DA.

Assim podemos perceber que os resultados obtidos para os dados NIRP foram iguais para os dados NIRB. Assim pode-se constatar que os instrumento NIRP não pode ser aplicado para a quantificação de AM, mas pode ser aplicado com a mesma precisão para classificação que o NIRB.

3.4. Conclusão

O modelo PLS-*FeedOPS* construído a partir dos espectros de fécula desidratada de mandioca obtidos no instrumento NIRB apresentou capacidade de prever o teor de amilose aparente. O instrumento NIRP usado neste estudo não possui informações suficientes para construir modelos de predição para amilose aparente em fécula desidratada de mandioca.

Os modelos de classificação apresentaram boa capacidade preditiva usando espectros de ambos os instrumentos. Deste modo, para aplicações específicas de classificação de valores menores e maiores que 20% de amilose aparente em fécula desidratada de mandioca, o instrumento NIRP se torna uma boa opção.

Referências

[1] S.M. Chisenga, T.S. Workneh, G. Bultosa, B.A. Alimi, Progress in research and applications of cassava flour and starch: a review, J. Food Sci. Technol. 56 (2019) 2799–2813. https://doi.org/10.1007/s13197-019-03814-6.

[2] J. Liu, Q. Zheng, Q. Ma, K.K. Gadidasu, P. Zhang, Cassava Genetic Transformation and its Application in BreedingF, J. Integr. Plant Biol. 53 (2011) 552–569. https://doi.org/10.1111/j.1744-7909.2011.01048.x.

[3] H. Ceballos, C.A. Iglesias, J.C. Pérez, A.G.O. Dixon, Cassava breeding: opportunities and challenges, Plant Mol. Biol. 56 (2004) 503–516. https://doi.org/10.1007/s11103-004-5010-5.

[4] A.D. Uchechukwu-Agua, O.J. Caleb, U.L. Opara, Postharvest Handling and Storage of Fresh Cassava Root and Products: a Review, Food Bioprocess Technol. 8 (2015) 729–748. https://doi.org/10.1007/s11947-015-1478-z.

[5] S. Prochnik, P.R. Marri, B. Desany, P.D. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D.S. Rokhsar, S. Rounsley, The Cassava Genome: Current Progress, Future Directions, Trop. Plant Biol. 5 (2012) 88–94. https://doi.org/10.1007/s12042-011-9088-z.

[6] P.M. Alvarado, L. Grosmaire, D. Dufour, A.G. Toro, T. Sánchez, F. Calle, M.A.M. Santander, H. Ceballos, J.L. Delarbre, T. Tran, Combined effect of fermentation, sun-drying and genotype on breadmaking ability of sour cassava starch, Carbohydr. Polym. 98 (2013) 1137–1146. https://doi.org/https://doi.org/10.1016/j.carbpol.2013.07.012.

[7] X. Shen, W. Shang, P. Strappe, L. Chen, X. Li, Z. Zhou, C. Blanchard, Manipulation of the internal structure of high amylose maize starch by high pressure treatment and its diverse influence on digestion, Food Hydrocoll. 77 (2018) 40–48. https://doi.org/https://doi.org/10.1016/j.foodhyd.2017.09.015.

[8] N. Singh, J. Singh, L. Kaur, N. Singh Sodhi, B. Singh Gill, Morphological, thermal and rheological properties of starches from different botanical sources, Food Chem. 81 (2003) 219–231. https://doi.org/https://doi.org/10.1016/S0308-8146(02)00416-8.

[9] C. Malegori, E.J. Nascimento Marques, S.T. de Freitas, M.F. Pimentel, C. Pasquini, E. Casiraghi, Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms, Talanta. 165 (2017) 112–116. https://doi.org/https://doi.org/10.1016/j.talanta.2016.12.035.

[10] M.-T. Sanchez, M. la Haba, M. Benitez-Lopez, J. Fernandez-Novales, A. Garrido-Varo, D. Perez-Marin, Non-destructive characterization and quality control of intact strawberries based on NIR spectral data, J. Food Eng. 110 (2012) 102–108. https://doi.org/10.1016/j.jfoodeng.2011.12.003.

[11] I.O. Afara, I. Prasadam, R. Crawford, Y. Xiao, A. Oloyede, Near infrared (NIR) absorption spectra correlates with subchondral bone micro-CT parameters in osteoarthritic rat

models, Bone. 53 (2018) 350–357. https://doi.org/10.1016/j.bone.2012.12.042.

[12] J.-N. Tourvieille, F. Larachi, C. Duchesne, J. Chen, NIR hyperspectral investigation of extraction kinetics of ionic-liquid assisted bitumen extraction, Chem. Eng. J. 308 (2017) 1185–1199. https://doi.org/10.1016/j.cej.2016.10.010.

[13] H. Ceballos, F. Davrieux, E.F. Talsma, J. Belalcazar, P. Chavarriaga, M.S. Andersson, Carotenoids in Cassava Roots, in: Carotenoids, InTech, (2017). https://doi.org/10.5772/intechopen.68279.

[14] S. Lafhal, P. Vanloot, I. Bombarda, J. Kister, N. Dupuy, Chemometric analysis of French lavender and lavandin essential oils by near infrared spectroscopy, Ind. Crops Prod. 80 (2016) 156–164. https://doi.org/10.1016/j.indcrop.2015.11.017.

[15] R. Gislum, P. Nikneshan, S. Shrestha, A. Tadayyon, L. Deleuran, B. Boelt, Characterisation of Castor (Ricinus communis L.) Seed Quality Using Fourier Transform Near-Infrared Spectroscopy in Combination with Multivariate Data Analysis, Agriculture. 8 (2018) 59. https://doi.org/10.3390/agriculture8040059.

[16] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32. https://doi.org/https://doi.org/10.1016/j.aca.2010.03.048.

[17] R.M. Balabin, S. V Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72. https://doi.org/https://doi.org/10.1016/j.aca.2011.03.006.

[18] DLP® NIRscan™ Nano Evaluation Module, (n.d.). http://www.ti.com/tool/DLPNIRNANOEVM (accessed June 24, 2019).

[19] D.M. Friedrich, C.A. Hulse, M. von Gunten, E.P. Williamson, C.G. Pederson, N.A. O'Brien, Miniature near-infrared spectrometer for point-of-use chemical analysis, (2014) 899203–899211. https://doi.org/10.1117/12.2040669.

[20] C.G. Pederson, D.M. Friedrich, C. Hsiung, M. von Gunten, N.A. O'Brien, H.-J. Ramaker, E. van Sprang, M. Dreischor, Pocket-size near-infrared spectrometer for narcotic materials identification, (2014) 91010O-9101–11. https://doi.org/10.1117/12.2050019.

[21] A.B. Vitor, R.P. Diniz, C.V. Morgante, R.P. Antônio, E.J. de Oliveira, Early prediction models for cassava root yield in different water regimes, F. Crop. Res. 239 (2019) 149–158. https://doi.org/10.1016/j.fcr.2019.05.017.

[22] T. Sánchez, H. Ceballos, D. Dufour, D. Ortiz, N. Morante, F. Calle, T. Zum Felde, M. Domínguez, F. Davrieux, Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques, Food Chem. 151 (2014) 444–451. https://doi.org/10.1016/j.foodchem.2013.11.081.

[23] E.I.. Balabin, R.M.aEmail Author, Safieva, R.Z.a, Lomakina, Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction(, Chemom. Intell. Lab. Syst. 88 (2007) 183–188.

[24] A. Fernández, N. Zakhia, R. Ruiz, J. Trujillo, Desarrollo de un método sencillo para medir la calidad del almidón agrio de yuca : Impacto del método sobre la agroindustria rural en el departamento del Cauca (Colombia), (2002). http://hdl.handle.net/10568/72076 (accessed April 19, 2018).

[25] I.O.F. STANDARDIZATION, Norme Internacionale: Riz determination de la teneur en amylose, ISO 6647. (1987) 4.

[26] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148. https://doi.org/10.1080/00401706.1969.10490666.

[27] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. https://doi.org/10.1002/cem.1192.

[28] J.P.A. Martins, R.F. Teófilo, M.M.C. Ferreira, Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets, J. Chemom. (2010). https://doi.org/10.1002/cem.1309.

[29] J. V. Roque, W. Cardoso, L.A. Peternelli, R.F. Teófilo, Comprehensive new approaches for variable selection using ordered predictors selection, Anal. Chim. Acta. 1075 (2019) 57–70. https://doi.org/10.1016/j.aca.2019.05.039.

[30] J. V. Roque, L.A.S. Dias, R.F. Teófilo, Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies, J. Braz. Chem. Soc. (2017). 1506-1516. https://doi.org/10.21577/0103-5053.20160332..

[31] M.K.D. Rambo, E.P. Amorim, M.M.C. Ferreira, Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues, Anal. Chim. Acta. 775 (2013) 41–49. https://doi.org/10.1016/j.aca.2013.03.015.

[32] N.F. Pérez, J. Ferré, R. Boqué, Calculation of the reliability of classification in discriminant partial least-squares binary classification, Chemom. Intell. Lab. Syst. 95 (2009) 122–128. https://doi.org/10.1016/j.chemolab.2008.09.005.

[33] A.M.K. Pedro, M.M.C. Ferreira, Nondestructive Determination of Solids and Carotenoids in Tomato Products by Near-Infrared Spectroscopy and Multivariate Calibration, Anal. Chem. 77 (2005) 2505–2511. https://doi.org/10.1021/ac048651r.

[34] Z. Zhang, X. Yin, C. Ma, Development of simplified models for the nondestructive testing of rice with husk starch content using hyperspectral imaging technology, Anal. Methods. 11 (2019) 5910–5918. https://doi.org/10.1039/C9AY01926J.

[35] R.A. Olson, K.J. Frey, eds., Nutritional Quality of Cereal Grains: Genetic and Agronomic Improvement, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, USA, (1987). https://doi.org/10.2134/agronmonogr28.

CAPÍTULO 4

QUANTIFICAÇÃO DOS ÁCIDOS GRAXOS PRESENTES NO ÓLEO DE PINHÃO-MANSO (*Jatropha curcas L.)* POR ESPECTROSCOPIA NO INFRAVERMELHO E MÉTODOS QUIMIOMÉTRICOS

Resumo

Este trabalho teve como objetivo a construção de modelos de calibração multivariada usando as espectroscopias na região do infravermelho próximo (NIR) e médio (MIR) aliada à regressão por quadrados mínimos parciais (PLS), para a quantificação dos ácidos graxos (AG) presentes no óleo de pinhão-manso (*Jatropha curcas L.)*. A comparação de modelos provenientes de instrumentos NIR, portátil (NIRP) e de bancada (NIRB), e MIR de bancada com o acessório de Reflexão Total Atenuada (ATR) também foram estudados. Os AG presentes no óleo de pinhão-manso foram determinados por cromatografia gasosa com detector de ionização na chama (GC-FID). Os principais picos identificados e quantificados foram: AG palmítico (C16:00), esteárico (C18:00), oléico (C18:1) e linoléico (C18:2). Os resultados de predição dos modelos usando dados de NIRP e MIR foram inferiores ao do NIRB. Dentre os resultados do NIRB a seleção de variáveis teve papel importante na obtenção do melhor modelo. Para os modelos de regressão (NIRB), os valores da predição da raiz quadrada do erro quadrático médio de previsão (*RMSEP*) e coeficiente de correlação da previsão (R_P) foram, respectivamente, 3,85 mg mL^{-1} e 0,87 para AG palmítico; 2,20 mg mL^{-1} e 0,87 para AG esteárico; 8,61 mg mL^{-1} e 0,86 para AG oleico; 15,75 mg mL^{-1} e 0,85 mg mL^{-1} para AG linoleico. Portanto, este estudo apresenta um método rápido, barato e não-destrutivo para a determinação dos AG do óleo de pinhão-manso, possibilitando auxiliar a seleção de sementes de pinhão-manso em programas de melhoramento e outras aplicações.

Palavras-chaves: Pinhão-manso, regressão multivariada, espectroscopia no infravermelho próximo; quadrados mínimos parciais.

4.1. Introdução

O pinhão-manso (*Jatropha curcas* L.) é um arbusto pertencente à família das euforbiáceas, a mesma da mamona e da mandioca. Nativa da américa central, encontra-se difundida em diversas regiões do Brasil e do mundo. Possuindo um rápido crescimento, ela apresenta em média de dois a três metros, podendo alcançar até cinco metros em condições especiais [1,2].

De particular importância, os frutos do pinhão-manso contêm óleo viscoso, são ricos em ácidos graxos insaturados, possui boa estabilidade à oxidação e custo de extração relativamente baixo. Por esses fatores seu óleo tem sido amplamente usado para fabricação de sabões, produtos na indústria de cosméticos e na produção de biodiesel [3]. Esse último uso pode ser de particular importância ao examinar substitutos aos combustíveis a base de petróleo para combater a acumulação de gases de efeito estufa [1]. Tendo em vista os benefícios da utilização do óleo do pinhão-manso como fonte renovável de combustível, a descrição da composição química do óleo fornece informação que podem subsidiar a produção e exploração comercial desta espécie [4].

A determinação da qualidade do óleo é feita pela quantificação dos seus respectivos ácidos graxos (AG). Tal quantificação inicia-se pela reação de transesterificação do óleo para obtenção dos ésteres metílicos de ácidos graxos (FAME), e na maioria das vezes seguida da análise via cromatografia em fase gasosa (GC) [5,6]. Embora métodos convencionais, como GC ofereçam um alto nível de precisão e exatidão, eles parecem ser menos adequados, pois consomem tempo, são caros e exigem o manuseio de vários produtos químicos tóxicos no preparo e análise das amostras [7].

Métodos não invasivos e não destrutivos, como as espectroscopias vibracionais no infravermelho próximo (NIR) e médio (MIR), têm sido amplamente utilizados para quantificar propriedades em materiais agrícolas e avaliar a qualidade de produtos agrícolas e alimentares [8–12]. Essa técnica fornece informações rápidas sobre propriedades físicas e químicas, com preparação mínima ou inexistente de amostras [13]. Além disso, as espectroscopias vibracionais, juntamente com a regressão PLS, fornece modelos confiáveis e com alta capacidade preditiva para amostras desconhecidas.

A regressão PLS se destaca como um dos métodos principais para a construção de modelos de regressão de primeira ordem a partir de dados de fontes químicas. Esse método não requer conhecimento preciso de todos os componentes presentes nas amostras

e pode realizar a previsão de um analito de interesse mesmo na presença de interferência, desde que os interferentes também estejam presentes no modelo construído [14].

Vários trabalhos apresentaram o uso de NIR e PLS para quantificar o teor total de óleo em materiais vegetais [15–18]. A determinação dos FAME a partir da espectroscopia NIR e métodos quimiométricos tem sido aplicada em matrizes como o óleo de canola [12], azeite virgem [11,19] e até mesmo em amostras de sementes de pinhão-manso [20,21]. Apesar de haver trabalhos que usaram NIR e PLS para quantificação de AG em óleo de pinhão manso, os resultados de exatidão dos modelos não foram satisfatórios e assim, a capacidade preditiva é questionável [20,21].

Este trabalho tem como objetivo obter espectros diretamente do óleo de pinhão-manso usando espectroscopias vibracionais NIR e MIR em diferentes instrumentos, construir modelos de calibração PLS para prever os AG e comparar os resultados.

4.2. Materiais e Métodos

4.2.1. Coleta das amostras

As amostras utilizadas neste trabalho são pertencentes ao Banco Ativo de Germoplasma (GAB) de pinhão manso da Universidade Federal de Viçosa (UFV), constituído de 77 acessos de *J. curcas*, sendo 74 deles oriundos de diferentes regiões do Brasil e outros três do Camboja. Além destas, foram utilizadas amostras de um teste de progênies (PT) implantado em 2008 no Campo Experimental da UFV de Araponga, MG. A partir deste teste foram obtidas 121 progênies de polinização livre, assumidas meias-irmãs, de *J. curcas*. Essas amostras, obtidas em sementes, foram devidamente identificadas e armazenadas à temperatura ambiente até o momento das análises. Para realização das análises, foram selecionadas 99 amostras do conjunto total de amostras (GAB + PT).

4.2.2. Extração do óleo

O óleo foi extraído das sementes das amostras manualmente com o auxílio de uma prensa mecânica. Aproximadamente 20 mL de óleo foram coletados e colocados em frascos de vidro âmbar e armazenados sob refrigeração (−20 °C). Logo após as extrações, o óleo foi centrifugado para remoção de impurezas.

4.2.3. Reação de transesterificação

A transesterificação por catálise básica foi realizada de acordo com procedimento adaptado por Mir *et al.* [22], seguindo as seguintes etapas: a um tubo de polipropileno tipo falcon foi pipetado 50,00 µL do óleo, logo em seguida adicionou-se 2 mL de uma solução de hexano e 200,00 µL de uma solução a 0,500 $molL^{-1}$ de KOH em metanol. Esse sistema foi levado a agitação a 600 rpm por 15 min, em seguida, duas fases foram formadas. Foram extraídos 1,5 mL da fase contendo hexano e adicionado sulfato de sódio anidro para adsorver a água remanescente na fase orgânica. Um volume de aproximadamente 2 mL da fase orgânica foi coletado com seringa e este volume foi filtrado com filtro de seringa de teflon de 0,22 µm, transferido diretamente para vials de 2 mL. Os vials foram tampados e armazenados até a injeção.

4.2.4. Cromatografia gasosa

As separações foram realizadas em uma coluna de sílica, capilar SP-2560 da Sigma-Aldrich. O diâmetro interno e tamanho da coluna foram de respectivamente de 0,18 mm e 75 m. A coluna foi acondicionada em um cromatógrafo a gás (GC) Shimadzu GC-2010 com injetor automático e detector por ionização de chama (FID). A temperatura inicial da coluna foi 160 °C com tempo de espera de 3 min. Em seguida foi aquecida até 180 °C permanecendo nesta temperatura por 12 minutos. Por fim, a temperatura foi elevada até 220 °C permanecendo por 9 minutos, logo após o sistema foi resfriado para uma nova injeção de amostra. A temperatura do injetor foi 240 ºC, a temperatura do detector FID foi 240 °C. O volume injetado foi 0,5 µL e o modo de injeção foi o *split* de 1:50, o fluxo do gás de arraste (N_2) foi 0,41 mL/min e a pressão da coluna foi de 176,1 kPa.

A quantificação se deu pela curva analítica das concentrações e área dos picos dos padrões. Os compostos foram identificados pelo tempo de retenção usando padrões de referência certificados, utilizando o hexano como solvente.

A identificação e a curva analítica dos AG foram construídas utilizando o conjunto de padrões SULPECO Mix C14-C22 da Sigma-Aldrich. A identificação foi possível comparando com os tempos de retenção com o padrão. Já a construção da curva analítica, 100 mg do conjunto de FAMEs (massa da ampola) foi diluída com hexano nas seguintes concentrações totais, 30, 27, 20, 18, 15, 12 mg/ml.

4.2.5. Análise espectral das amostras

Os espectros NIR das amostras foram obtidos do óleo do pinhão-manso utilizando os espectrofotômetros da Bruker Optics modelo MPA (Billerica, MA, EUA) (NIRB) e Nano-NIR, DLP NIRscan da Texas Instrument (NIRP).

O espectrômetro NIRB, Fig. 4.1C, foi operado no modo transflectância. As varreduras das amostras foram o resultado médio de 32 varreduras medidas com uma resolução de 4 cm^{-1} ao longo da faixa de número de onda 12.000-3600 cm^{-1}. Nesta configuração o espectro foi obtido com 4200 valores de transflectância. A aquisição espectral, o controle instrumental e a manipulação preliminar de arquivos foram realizadas usando o software Bruker Optics OPUS (versão 6.5).

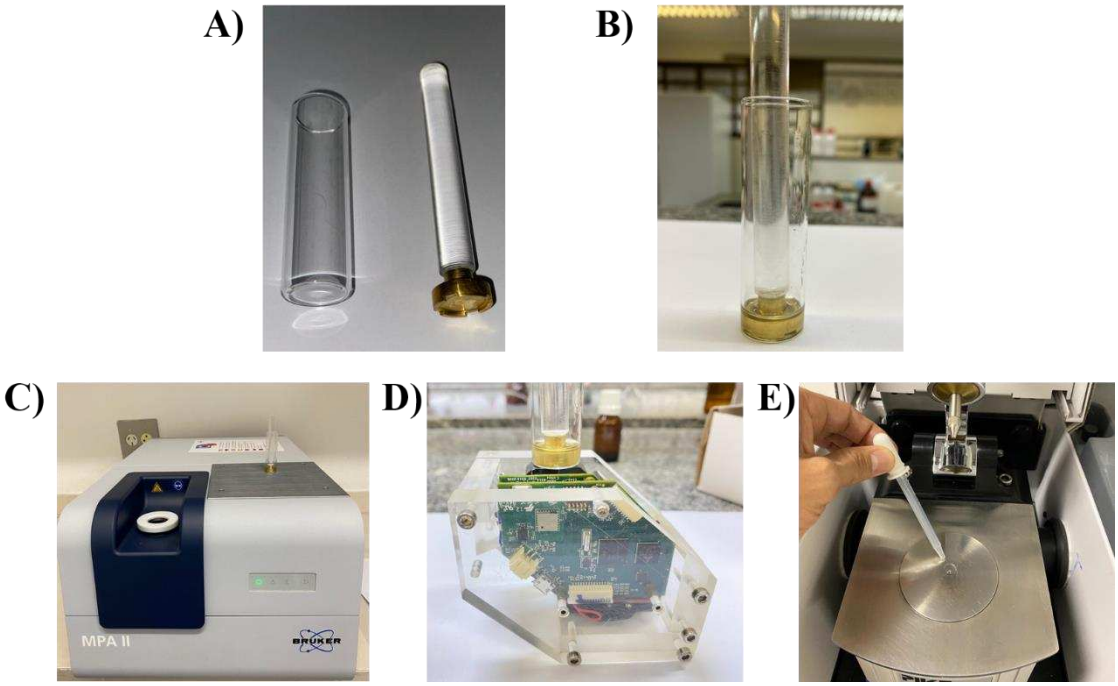

Figura 4.1. Tubo e pistão de cobre (A), Tubo e pistão de cobre com óleo (B), Modo de obtenção dos espectros no NIRB (C), NIRP (D) e no MIR (C).

O espectrômetro NIRP, Fig. 4.1D, foi também operado no modo transflectância, realizando 50 varreduras para cada amostra com resolução de 8 cm^{-1} ao longo de uma faixa de 11.111,1 – 5.882,4 cm^{-1}. Nesta configuração o espectro foi obtido com 605 valores de transflectância.

Para ambos os instrumentos o logaritmo na base 10 do inverso da transflectância foi coletado. As aquisições de cada amostra foram obtidas adicionando aproximadamente 300 µL em um tubo de quartzo e sobre esse um pistão de cobre (Fig. 4.1 A e 4.1B). Foram usados dois pistões com espessuras diferentes para aquisição dos espectros, 2mm e 1 mm.

Este foi centralizado diretamente na janela do instrumento, sem qualquer preparação adicional. Foram obtidos dois espectros por amostra, e o espectro médio de cada amostra foi usado como variável independente.

Os espectros coletados no instrumento FTIR – MIR, Fig. 4.1E, com acessório ATR foram obtidos usando o instrumento da VARIAN 660-IR equipado com acessório PIKE Gladi ATR. Os espectros foram obtidos à temperatura ambiente no modo absorbância com 16 varreduras coletadas com resolução de 4 cm^{-1} na faixa espectral de 4000 a 400 cm^{-1}.

4.2.6. Modelos de calibração multivariada

Os espectros dos instrumentos NIR e MIR foram importados pelo Matlab 2019a (Math Works, Natick, EUA), onde as análises foram realizadas. Os modelos de regressão inversa foram construídos usando o algoritmo bidiagonal PLS [23,24] e a validação cruzada foi do tipo randômica, onde foram utilizadas dez divisões [23,24]. Além dos dados brutos, dois métodos de pré-processamentos, *i.e*., (1) centrar na média e (2) autoescalar, e oito transformações, i.e., (1) alisamento, (2) primeira derivada, (3) segunda derivada, (4) correção multiplicativa de sinal (MSC), (5) detrend, (6) normalização, (7) correção de linha de base, (8) padronização normal de sinal (SNV) e suas combinações foram testadas para encontrar o melhor modelo de regressão.

Os algoritmos para a importação de dados, a calibração e os modelos de validação foram escritos em nosso laboratório em função *.m* para o Matlab. Os métodos OPS [25] foram aplicados utilizando os algoritmos disponíveis em www.deq.ufv.br/chemometrics. Todos os cálculos foram realizados no Matlab 2019a.

A qualidade dos modelos foi avaliada pela raiz quadrada do erro quadrático médio *(RMSE*) e pelo coeficiente de correlação (*R*) calculados pelas equações. (4.1) e (4.2), respectivamente.

$$RMSE = \sqrt{\frac{\sum_{i}^{N}(y_i - y_1)^2}{N}} \qquad \textbf{(4.1)}$$

$$R = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)\left(\hat{y}_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{y}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}} \qquad \textbf{(4.2)}$$

onde $\hat{y}$ e $\overline{\hat{y}}$ são o valor estimado e a média dos valores estimados, respectivamente, sendo y os valores observados e a $\hat{y}$ média dos valores observados respectivamente; e *N* representa o número de amostras. No caso em que a validação cruzada (*CV*) é usada, *N* representa o número de amostras no conjunto de validação cruzada, e o erro e o coeficiente de correlação são chamados de raiz quadrada do erro quadrático médio de validação cruzada (*RMSECV*) e coeficiente de correlação validação cruzada (R_{vc}). Para a validação externa, representamos o número de amostras de predição (*P*) e, neste caso, os coeficientes de correlação de erros e são denominados raiz quadrada do erro quadrático médio de predição (*RMSEP*) e coeficientes de correlação de predição (R_P). O algoritmo Kennard-Stone foi usado para selecionar amostras dos conjuntos de calibração e predição.

4.3. Seleção dos preditores ordenados - OPS

Quatro métodos OPS [53], *AutoOPS, FeedOPS, AutoiOPS* e *FeediOPS*, foram aplicados para selecionar mais variáveis preditivas, visando aprimorar os modelos. Primeiramente, um método automático, *AutoOPS*, que desenvolve e executa seleção variáveis usando vários vetores informativos e suas combinações. Os Em segundo, o *FeedOPS* apresenta uma nova estratégia as variáveis pré-selecionadas, retornando a uma nova seleção. Por último, foi estabelecido um método para aplicar OPS em subdivisões de matriz dos dados chamados OPS intervalado, *AutoiOPS* e *FeediOPS.*

4.4. Resultados e Discussão

A partir do padrão, os seguintes AG foram identificados: palmítico (C16:0), esteárico (C18:0), oleico (C18:1) e linoleico (C18:2). Esses AG somados representam em média mais de 97 % da quantidade total de óleo. A equação da curva e o coeficiente de determinação foram de *y=56932,44x - 34126 e 0,96*; *y=74668x – 46612,49 e 0,96*; *y=83733,26x – 219267,93 e 0,96*; *y=81282,80x – 283712,31 e 0,96* para os AG palmítico, esteárico, oleico e linoleico, respectivamente.

Os dois AG saturados palmítico e esteárico - variaram de 19,7 a 22,7% do óleo da semente e os dois AG insaturados oleico e linoleico variaram de 77,3 a 80,3% do óleo da semente. As concentrações dos AG palmítico, esteárico, oleico e linoleicos variaram de

97,2 a 162,5 mg mL^{-1}, 42,4 a 63,4 mg mlL^{-1}, 218,9 a 348,8 mg mL^{-1} e 290,9 a 498,2 mg mL^{-1}, respectivamente.

A Tabela 4.1 apresenta os valores máximo, mínimos e a média das determinações dos AG das amostras. Podemos observar uma boa variabilidade para a concentração os AG.

Tabela 4.1. Valores máximo, mínimos e a média das determinações dos AG.

AG	Min (mg mL^{-1})	Max (mg mL^{-1})	Med (mg mL^{-1})
Palmítico	97,2	162,5	115,2
Esteárico	42,4	63,4	49,1
Oleico	218,9	348,8	261,1
Linoleico	290,9	498,2	363,7

Foram realizadas 99 determinações para todos os AG determinados.

A Fig. 4.2 apresenta os espectros NIR provenientes do espectrofotômetro de bancada, NIRB, Fig. 4.2A, e portátil, NIRP, Fig. 4.2B e espectros MIR, Fig. 4.2C.

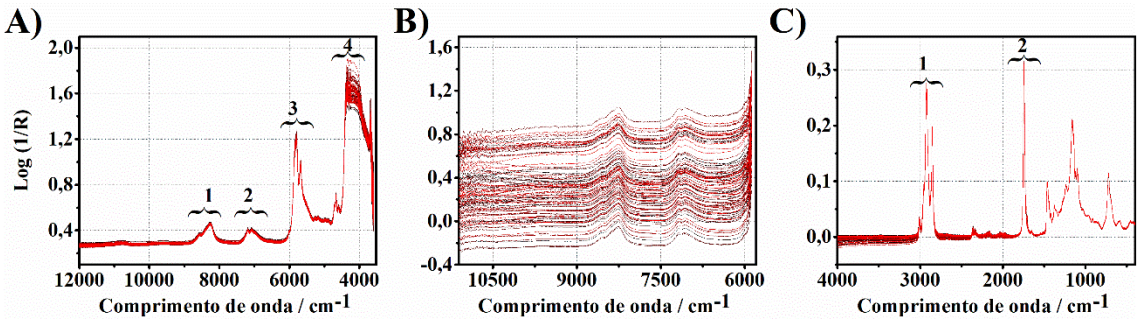

Figura 4.2. Espectros NIR para NIRB (A) e NIRP (B) e MIR (C) do óleo de pinhão-manso.

Em relação aos espectros adquiridos na região MIR (4000–400 cm^{-1}) Fig. 4.2C, cada pico identificado nos espectros corresponde a grupos funcionais e modos de vibração responsáveis pela absorção do infravermelho médio. A região (1) na Fig.4.2C, compreende a região de estiramento da ligação carbono com hidrogênio, sendo a banda em 3005 cm^{-1} atribuída ao estiramento da ligação = C-H (cis); as bandas 2959 e 2873 cm^{-1}, referem-se ao estiramento assimétrico e simétrico de -C-H de CH_3 alifático; as bandas em 2921 e 2851 cm^{-1} aos estiramentos assimétrico e simétrico de -C-H de CH_2 alifático. A região de estiramento da ligação dupla -C=O, em (2), Fig. 4.2C, apresenta uma banda a 1742 cm^{-1} , correlacionada com a vibração dos estiramentos dos grupos funcionais éster carbonil dos triglicerídeos [123–125]. A presença de um possível ombro fraco em 1700 cm^{-1} poderia estar correlacionada com o estado de oxidação do produto e a presença de ácidos graxos livres (estiramento da ligação do grupo carbonila de ácidos graxos livres),

o que não foi evidenciado, caracterizando a baixa presença de ácidos graxos livres no óleo [28].

Ao contrário da região MIR, os espectros NIR consistiam em bandas espectrais relativamente fracas e altamente sobrepostas. Essas bandas surgem como resultado de sobretons e combinações das bandas fundamentais na região do MIR. Portanto, análises dos espectros NIR requerer a utilização de ferramentas quimiométrica para extrair informações relevantes das informações espectrais sobrepostas.

A Fig. 4.2A e B apresentam, nas suas respectivas faixas, os espectros NIR. Na Fig. 4.2A temos em (1) o segundo sobretom do estiramento da ligação do CH (grupos metil, metileno e etileno), em (2) a combinação do estiramento da ligação do CH, em (3) o primeiro sobretom do estiramento da ligação do CH (grupos metil, metileno e etileno) e em (4) a combinação do estiramento da ligação CH com outros modos vibracionais [11].

Os espectros NIR obtidos pelo instrumento portátil (Fig. 4.2B), NIRP, apresenta um alto desvio aditivo na linha de base. Isso pode estar relacionado com a variação na fonte de radiação do instrumento. No entanto, essa variação pode ser minimizada com pré-tratamentos dos espectros [29].

O ruído e o deslocamento da linha de base são características evidentes e indesejáveis. Assim, transformações matemáticas nas linhas da matriz foram essenciais para reduzir erros sistemáticos e aleatórios de variações físicas indesejáveis e aumentar a relação sinal-ruído. Para todos os conjuntos de dados apresentados, os parâmetros estatísticos *RMSE* e *R* foram comparados.

As Tabelas 4.2, 4.3, 4.4 e 4.5, apresenta os parâmetros estatísticos dos modelos PLS usando todas as variáveis (*completo*), variáveis selecionadas pelas abordagens *Auto*OPS, *Feed*OPS, *Auto*iOPS e *Feed*iOPS para os espectros obtidos no NIRB, NIRP e MIR para os AG palmítico, esteárico, oleico e linoleico, respectivamente.

De um modo geral os resultados com o NIRP e MIR não foram satisfatórios. Este fato pode ser evidenciado comparando os resultados para cada determinação dos AG, Tabelas 4.2, 4.3, 4.4 e 4.5. Assim apenas os resultados do NIRB foram utilizados na construção dos modelos de calibração.

Dentre cada conjunto de dados, para cada instrumento, foram construídos cinco modelos, sendo um contendo todas as variáveis e os demais com seleção de variáveis. Para a construção desses modelos foram empregados pré-tratamentos.

Na construção do modelo para o AG palmítico aplicou-se a segunda derivada com janela sete (7) e logo em seguida os dados foram autoescalados. Os melhores modelos

foram escolhidos com base nos parâmetros de previsão, *RMSEP* e R_p. Sendo assim, o modelo do NIRB que apresentou menor valor de *RMSEP* e maior R_p para determinação do AG palmítico foi o com seleção de variáveis pelo *FeediOPS,* Tabela 4.2, com valores de 3,85 e 0,87, respectivamente. Pode-se observar também que os resultados para NIRP e MIR foram piores que para o NIRB, apresentando maiores valores de *RMSEP* e menor R_p.

Tabela 4.2. Parâmetros estatísticos dos modelos PLS com variáveis completas e variáveis selecionadas por OPS para os dados provenientes do NIRB, NIRP e MIR referentes ao AG Palmítico.

Ácido graxo			*nlvs*	*nVars*	*RMSEC (mg mL^{-1})*	*Rc*	*RMSECV (mg mL^{-1})*	*Rcv*	*RMSEP (mg mL^{-1})*	*Rp*
Palmítico C16:0	*NIRB*	***Completo***	*14*	*4200*	*0,00*	*1,00*	*6,38*	*0,06*	*7,40*	*0,01*
		AutoiOPS	*14*	*650*	*0,00*	*1,00*	*4,30*	*0,72*	*4,66*	*0,80*
		FeediOPS	***4***	***290***	***1,10***	***0,98***	***3,85***	***0,78***	***3,85***	***0,87***
		AutoOPS	*4*	*530*	*0,79*	*0,99*	*4,00*	*0,76*	*4,58*	*0,81*
		FeedOPS	*4*	*570*	*0,77*	*0,99*	*4,15*	*0,74*	*4,68*	*0,79*
	NIRP	***Completo***	*15*	*605*	*0,01*	*1,00*	*7,94*	*0,24*	*9,37*	*0,03*
		AutoiOPS	*15*	*230*	*0,05*	*1,00*	*8,17*	*0,26*	*8,50*	*0,29*
		FeediOPS	*4*	*210*	*2,43*	*0,94*	*7,28*	*0,38*	*8,02*	*0,28*
		AutoOPS	*4*	*130*	*2,49*	*0,94*	*7,56*	*0,39*	*8,84*	*0,05*
		FeedOPS	*4*	*270*	*2,65*	*0,93*	*7,13*	*0,38*	*7,55*	*0,34*
	MIR	***Completo***	*15*	*1869*	*0,05*	*1,00*	*7,16*	*0,27*	*8,47*	*0,21*
		AutoiOPS	*15*	*290*	*0,26*	*1,00*	*8,48*	*0,26*	*6,43*	*0,68*
		FeediOPS	*12*	*210*	*0,91*	*0,99*	*8,04*	*0,36*	*7,10*	*0,66*
		AutoOPS	*5*	*270*	*2,30*	*0,94*	*6,65*	*0,42*	*6,26*	*0,62*
		FeedOPS	*6*	*170*	*2,52*	*0,93*	*6,70*	*0,49*	*7,22*	*0,56*

nlvs: número de variáveis latentes; *nVars*: número de variáveis; *RMSEC*: raiz quadrada do erro quadrático médio da calibração; *RMSECV*: raiz quadrada do erro quadrático médio da validação cruzada; *RMSEP* :raiz quadrada do erro quadrático médio da previsão.

Para a determinação do AG esteárico utilizou-se da segunda derivada com janela sete (7), autoescalar e normalização como pré-tratamentos. Apesar dos resultados de *RMSEP* dos conjuntos NIRP e MIR terem ficado próximos aos do NIRB os resultados de R_p ficaram inferiores ao seu valor, Tabela 4.3. Dentro dos cinco modelos apresentados para o conjunto NIRB o modelo que trouxe os melhores resultados foi o pela seleção de variáveis *AutoOPS* com *RMSEP* e *Rp* de 2,20 e 0,87, respetivamente.

Tabela 4.3. Parâmetros estatísticos dos modelos PLS com variáveis completas e variáveis selecionadas por OPS para os dados provenientes do NIRB, NIRP e MIR referentes ao AG Esteárico.

Ácido graxo			*nlvs*	*nVars*	*RMSEC (mg mL^{-1})*	*Rc*	*RMSECV (mg mL^{-1})*	*Rcv*	*RMSEP (mg mL^{-1})*	*Rp*
Esteárico C18:0	***NIRB***	***Completo***	*10*	*4200*	*0,01*	*1,00*	*2,84*	*0,17*	*3,93*	*0,04*
		AutoiOPS	*10*	*410*	*0,02*	*1,00*	*2,04*	*0,63*	*2,33*	*0,86*
		FeediOPS	*8*	*190*	*0,14*	*1,00*	*2,09*	*0,62*	*2,09*	*0,86*
		AutoOPS	***5***	***410***	***0,29***	***0,99***	***1,71***	***0,79***	***2,20***	***0,87***
		FeedOPS	*10*	*410*	*0,30*	*1,00*	*1,93*	*0,68*	*2,27*	*0,87*
	NIRP	***Completo***	*2*	*605*	*2,37*	*0,30*	*2,51*	*0,08*	*2,38*	*0,31*
		AutoiOPS	*2*	*190*	*2,34*	*0,33*	*2,51*	*0,14*	*2,28*	*0,42*
		FeediOPS	*1*	*50*	*2,39*	*0,30*	*2,48*	*0,14*	*2,17*	*0,53*
		AutoOPS	*2*	*50*	*2,34*	*0,33*	*2,54*	*0,11*	*2,30*	*0,34*
		FeedOPS	*2*	*50*	*2,34*	*0,33*	*2,49*	*0,16*	*2,26*	*0,40*
	MIR	***Completo***	*15*	*1869*	*0,00*	*1,00*	*3,00*	*0,02*	*2,99*	*0,01*
		AutoiOPS	*15*	*50*	*1,32*	*0,87*	*4,02*	*0,24*	*4,27*	*0,37*
		FeediOPS	*1*	*110*	*2,06*	*0,65*	*2,30*	*0,54*	*2,07*	*0,57*
		AutoOPS	*1*	*90*	*2,08*	*0,64*	*2,31*	*0,53*	*2,09*	*0,56*
		FeedOPS	*1*	*50*	*2,06*	*0,65*	*2,21*	*0,58*	*2,01*	*0,61*

nlvs: número de variáveis latentes; *nVars*: número de variáveis; *RMSEC*: raiz quadrada do erro quadrático médio da calibração; *RMSECV*: raiz quadrada do erro quadrático médio da validação cruzada *RMSEP* :raiz quadrada do erro quadrático médio da previsão.

Na determinação do AG oleico utilizou-se da normalização seguida da segunda derivada com janela sete (7) e autoescalamento como pré-tratamentos. Dentre os modelos apresentados para NIRB o modelo pela seleção de variáveis *AutoOPS,* Tabela 4.4, obteve os melhores resultados, com *RMSEP* de 8,61 e R_p de 0,86. Os resultados de NIRP e MIR ficaram significativamente piores que o NIRB.

Tabela 4.4. Parâmetros estatísticos dos modelos PLS com variáveis completas e variáveis selecionadas por OPS para os dados provenientes do NIRB, NIRP e MIR referentes ao AG Oleico.

Ácido graxo			*nlvs*	*nVars*	*RMSEC (mg mL^{-1})*	*Rc*	*RMSECV (mg mL^{-1})*	*Rcv*	*RMSEP (mg mL^{-1})*	*Rp*
Oleico C18:1	***NIRB***	***Completo***	*15*	*4200*	*0,00*	*1,00*	*14,59*	*0,09*	*15,12*	*0,06*
		AutoiOPS	*15*	*590*	*0,00*	*1,00*	*8,47*	*0,81*	*8,95*	*0,82*
		FeediOPS	*4*	*370*	*1,51*	*0,99*	*7,47*	*0,85*	*9,01*	*0,82*
		AutoOPS	***6***	***690***	***0,69***	***1,00***	***8,26***	***0,84***	***8,61***	***0,86***
		FeedOPS	*4*	*430*	*1,83*	*0,99*	*7,54*	*0,86*	*9,28*	*0,79*
	NIRP	***Completo***	*2*	*605*	*13,69*	*0,33*	*14,15*	*0,23*	*14,88*	*0,16*
		AutoiOPS	*2*	*170*	*13,11*	*0,42*	*13,94*	*0,28*	*14,52*	*0,30*
		FeediOPS	*2*	*50*	*13,26*	*0,39*	*13,90*	*0,28*	*14,23*	*0,32*
		AutoOPS	*2*	*50*	*13,18*	*0,41*	*14,02*	*0,27*	*14,48*	*0,30*
		FeedOPS	*2*	*250*	*13,16*	*0,41*	*14,03*	*0,27*	*14,49*	*0,30*
	MIR	***Completo***	*15*	*1869*	*0,03*	*1,00*	*15,87*	*0,15*	*15,07*	*0,06*
		AutoiOPS	*15*	*190*	*0,60*	*1,00*	*16,70*	*0,24*	*20,80*	*0,10*
		FeediOPS	*1*	*190*	*10,62*	*0,67*	*12,02*	*0,54*	*11,87*	*0,47*
		AutoOPS	*1*	*110*	*10,46*	*0,68*	*11,81*	*0,57*	*12,60*	*0,39*
		FeedOPS	*1*	*110*	*10,46*	*0,68*	*11,79*	*0,56*	*12,60*	*0,39*

nlvs: número de variáveis latentes; *nVars*: número de variáveis; *RMSEC*: raiz quadrada do erro quadrático médio da calibração; *RMSECV*: raiz quadrada do erro quadrático médio da validação cruzada *RMSEP* :raiz quadrada do erro quadrático médio da previsão.

A determinação do AG linoleico seguiu a mesma tendência dos anteriores, onde os resultados de NIRP e MIR foram piores que os NIRB. Apesar disso, podemos observar na Tabela 4.5 que os resultados de *RMSEP* das variáveis selecionados pelo *AutoOPS* e *FeedOPS* do conjunto MIR foram melhores que o selecionado como melhor. No entanto, os demais parâmetros estatísticos dos modelos foram inferiores ao selecionado. Também nesse caso foi utilizado a segunda derivada com janela sete (7), logo em seguida autoescalados e normalizados. Assim, dentre os cinco modelos construídos para o conjunto de dados NIRB, o melhor modelo foi o obtido pela seleção de variáveis pelo *FeedOPS* com valores de *RMSEP* e R_p igual a 15,75 e 0,85, respectivamente.

Tabela 4.5. Parâmetros estatísticos dos modelos PLS com variáveis completas e variáveis selecionadas por OPS para os dados provenientes do NIRB, NIRP e MIR referentes ao AG Linoleico.

Ácido graxo			*nlvs*	*nVars*	*RMSEC (mg mL^{-1})*	*Rc*	*RMSECV (mg mL^{-1})*	*Rcv*	*RMSEP (mg mL^{-1})*	*Rp*
Linoleico C18:2	***NIRB***	***Completo***	*12*	*4200*	*0,10*	*1,00*	*20,49*	*0,06*	*25,41*	*0,18*
		AutoiOPS	*12*	*690*	*0,10*	*1,00*	*12,14*	*0,81*	*15,85*	*0,81*
		FeediOPS	*4*	*290*	*4,49*	*0,97*	*11,10*	*0,84*	*15,85*	*0,78*
		AutoOPS	*6*	*690*	*3,63*	*0,98*	*13,08*	*0,76*	*15,79*	*0,83*
		FeedOPS	***4***	***390***	***2,41***	***0,99***	***13,35***	***0,74***	***15,75***	***0,85***
	NIRP	***Completo***	*1*	*605*	*21,28*	*0,18*	*22,13*	*0,16*	*21,46*	*0,14*
		AutoiOPS	*1*	*470*	*21,26*	*0,18*	*21,73*	*0,01*	*21,45*	*0,14*
		FeediOPS	*1*	*70*	*20,40*	*0,6*	*21,00*	*0,37*	*19,73*	*0,48*
		AutoOPS	*1*	*70*	*21,30*	*0,18*	*21,73*	*0,02*	*21,44*	*0,15*
		FeedOPS	*1*	*50*	*21,38*	*0,58*	*21,02*	*0,36*	*20,30*	*0,41*
	MIR	***Completo***	*15*	*1869*	*0,01*	*1,00*	*21,86*	*0,11*	*21,19*	*0,16*
		AutoiOPS	*15*	*70*	*3,39*	*0,98*	*24,25*	*0,47*	*31,11*	*0,13*
		FeediOPS	*3*	*450*	*6,72*	*0,94*	*17,70*	*0,49*	*17,61*	*0,27*
		AutoOPS	*1*	*70*	*13,88*	*0,71*	*15,96*	*0,59*	*14,43*	*0,57*
		FeedOPS	*1*	*70*	*13,88*	*0,71*	*15,69*	*0,60*	*14,43*	*0,57*

nlvs: número de variáveis latentes; *nVars*: número de variáveis; *RMSEC*: raiz quadrada do erro quadrático médio da calibração; *RMSECV*: raiz quadrada do erro quadrático médio da validação cruzada *RMSEP* :raiz quadrada do erro quadrático médio da previsão.

Vaknin *et al.* [21] e Montes *et al.* [20] realizaram também a determinação dos AG do pinhão-manso. Seus trabalhos realizaram a determinação da porcentagem relativa dos AG e não sua quantificação. Mas comparando os resultados de *R* com nosso trabalho, podemos constatar que em nenhum dos trabalhos esse valor foi superior ao nosso trabalho. Para o AG palmítico temos um *R* de 0,87 contra 0,31 de Vaknin *et al.* e 0,40 Montes *et al.*, AG esteárico temos um *R* de 0,87 contra 0,39 de Vaknin *et al.* e 0, 0,77 Montes *et al.*, AG oleico temos um *R* de 0,86 contra 0,72 de Vaknin *et al.* e 0,78 Montes *et al.*, AG linoleico temos um *R* de 0,85 contra 0,85 de Vaknin *et al.* e 0,79 Montes *et al.*

Foram construídos também modelos de calibração com espetros NIRB do mesmo instrumento com caminho ótico maior transflectância, 2 mm, anteriormente aos demostrados acima. Os resultados não foram satisfatórios, Tabela 4.6. Assim foi feita a diminuição do caminho óptico para 1 mm possibilitando a modelagem dos AG. Este fato pode estar relacionado com o desvio na lei de Lambert-Beer.

Tabela 4.6. Parâmetros estatísticos dos modelos PLS com variáveis completas e variáveis selecionadas por OPS para os dados provenientes do NIRB com caminho ótico 2 mm.

Ácido graxo			nlvs	nVars	RMSEC (mg ml^{-1})	Rc	RMSECV (mg ml^{-1})	Rcv	RMSEP (mg ml^{-1})	Rp
NIRB	Palmítico C16:0	Completo	11	4200	0,12	1,00	8,48	0,14	10,48	0,19
		AutoiOPS	11	350	0,12	1,00	5,02	075	6,21	0,77
		FeediOPS	2	590	2,52	0,94	5,32	0,73	5,34	0,86
		AutoOPS	3	430	1,84	0,97	4,93	0,76	5,52	0,83
		FeedOPS	3	370	1,62	0,98	5,36	0,70	5,43	0,84
	Esteárico C18:0	Completo	13	4200	0,00	1,00	2,66	0,13	3,03	0,04
		AutoiOPS	13	490	0,00	1,00	1,62	0,76	1,69	0,79
		FeediOPS	3	290	0,52	0,98	1,79	0,67	1,35	0,87
		AutoOPS	3	210	0,70	0,96	1,68	0,72	1,52	0,87
		FeedOPS	5	310	0,20	1,00	1,83	0,65	1,35	0,87
	Oleico C18:1	Completo	15	4200	0,00	1,00	16,39	0,16	14,89	0,08
		AutoiOPS	15	510	0,00	1,00	8,80	0,81	8,86	0,74
		FeediOPS	7	310	0,67	1,00	8,53	0,82	9,41	0,72
		AutoOPS	4	350	2,96	0,98	9,25	0,78	8,74	0,74
		FeedOPS	8	810	0,11	1,00	9,62	0,77	8,96	0,71
	Linoleico C18:2	Completo	14	4200	0,00	1,00	18,52	0,29	30,55	0,18
		AutoiOPS	14	1070	0,04	1,00	15,85	0,62	22,41	0,65
		FeediOPS	6	550	2,52	0,99	14,33	0,65	21,56	0,68
		AutoOPS	5	1050	3,49	0,98	16,33	0,51	22,94	0,62
		FeedOPS	6	870	2,75	0,99	14,33	0,65	23,03	0,62

Nlv: número de variáveis latentes; *nVars*: número de variáveis; *RMSEC*: raiz quadrada do erro quadrático médio da calibração; *RMSECV*: raiz quadrada do erro quadrático médio da cross validation; *RMSEP* :raiz quadrada do erro quadrático médio da previsão.

Os resultados dos melhores modelos construídos, *FeediOPS* para AG palmítico, *AutoOPS* para AG esteárico e oleico e *FeedOPS* para AG linoleico, são apresentados na Fig. 4.3.

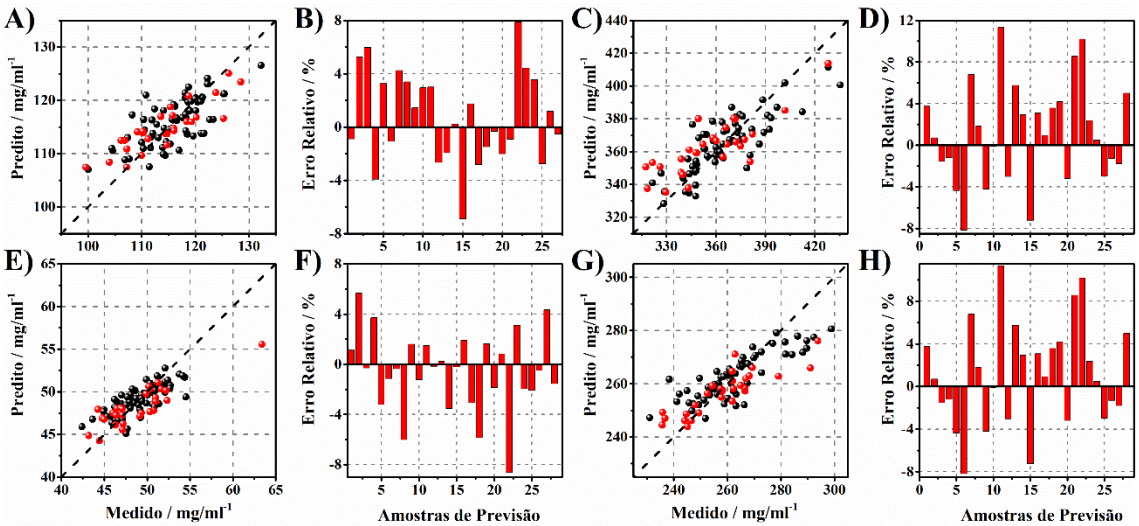

Figura 4.3. Valores medidos versus preditos para a determinação da concentração dos AG palmítico (A), esteárico (C), oleico (E) e linoleico (G) para o conjunto de calibração (●) e de predição (●). Erro relativo da previsão para AG palmítico (B), esteárico (D), oleico (F) e linoleico (H).

Como mostrado nas Tabela 4.2, 4.3, 4.4, e 4.5 e evidenciado na Fig. 4.3 os modelos foram bem parecidos para os quatro AG. Para todos os conjuntos de dados, uma

reta de 45º pode ser observada (Fig. 4.3A, 4.3C, 4.3E, 4.3G), indicando que os modelos podem prever com exatidão a concentração de AG no óleo de pinhão-manso. A Fig. 4.3 também mostra os erros relativos de previsão (Fig. 4.3B, 4.3D, 4.3F, 4.3H), onde a maioria dos erros relativos foi inferior a 5%.

Pode-se observar após a construção dos modelos que para todos os modelos a seleção de variáveis teve um papel muito importante, melhorando todos os parâmetros estatísticos.

Tendo em vista a importância da seleção de variáveis a Fig. 4.4 apresenta as variáveis selecionadas para os melhores conjuntos.

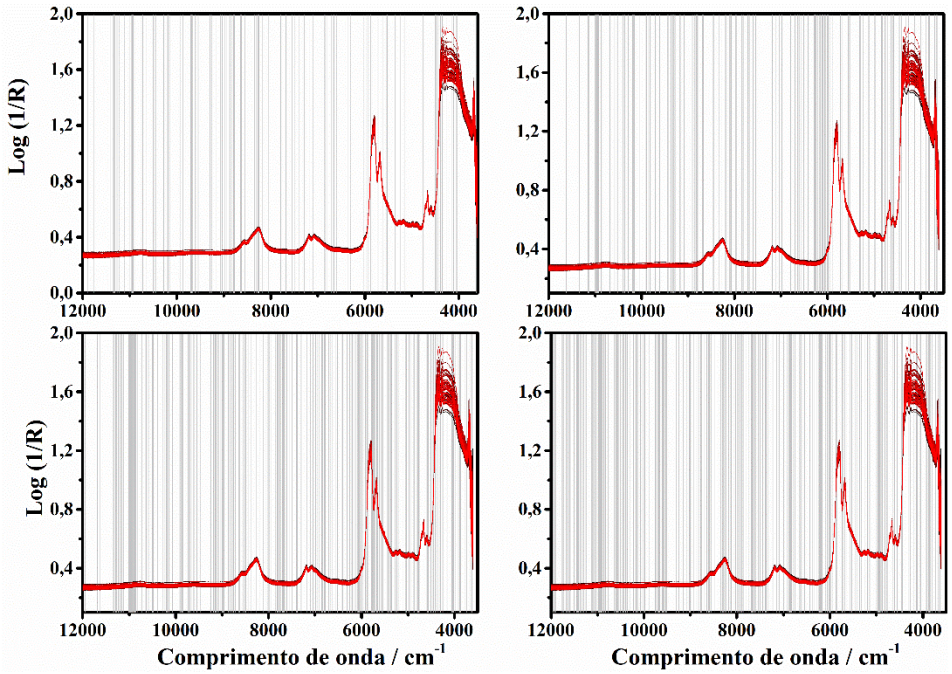

Figura 4.3. Variáveis selecionados para os AG palmítico (*FeediOPS*) (A), esteárico (*AutoOPS*) (B), oleico (*AutoOPS*) (C) e linoléico (*FeedOPS*) (D).

Como as estruturas dos quatros AG são muitos semelhantes, justifica a alta semelhança nas suas variáveis selecionadas. Também podemos observar que não ouve uma determinada região específica para um determinado AG, uma vez que todas as regiões estão relacionadas com os compostos determinados. Assim, pode-se constatar, que a seleção de variáveis pode retirar dentro de cada região as variáveis que realmente contribuíram para a determinação de cada AG.

4.5. Conclusão

Foi realizado a determinações da concentração dos AG palmítico, esteárico, oleico e linoleico provenientes do óleo do pinhão-manso. Foi possível também a comparação de modelos com dados provenientes de três instrumentos NIRB, NIRB e MIR, onde os melhores resultados foram para NIRB usando transflectância com caminho ótico de 1 mm. A seleção de variáveis teve importante papel na obtenção de melhores modelos, selecionando variáveis mais preditivas. Os parâmetros estatísticos dos modelos construídos neste trabalho mostraram-se melhores do que comparadas com resultado da literatura. Assim este trabalho apresenta um método rápido, barato e não-destrutivo para a determinação dos AG do óleo de pinhão-manso, possibilitando auxiliar a seleção de sementes de pinhão-manso em programas de melhoramento.

Referências

[1] K. Openshaw, A review of Jatropha curcas: an oil plant of unfulfilled promise, Biomass and Bioenergy. 19 (2000) 1–15. https://doi.org/10.1016/S0961-9534(00)00019-2.

[2] W.M.J. Achten, L. Verchot, Y.J. Franken, E. Mathijs, V.P. Singh, R. Aerts, B. Muys, Jatropha bio-diesel production and use, Biomass and Bioenergy. 32 (2008) 1063–1084. https://doi.org/10.1016/j.biombioe.2008.03.003.

[3] A. Suresh, N. Shah, M. Kotecha, P. Robin, Evaluation of biochemical and physiological changes in seeds of Jatropha curcas L. Under natural aging, accelerated aging and saturated salt accelerated aging, Sci. Hortic. (Amsterdam). 255 (2019) 21–29. https://doi.org/10.1016/j.scienta.2019.05.014.

[4] I.O. Virgens, R.D. de Castro, M.B. Loureiro, L.G. Fernandez, Revisão: Jatropha curcas L.: aspectos morfofisiológicos e químicos, Brazilian J. Food Technol. 20 (2017). https://doi.org/10.1590/1981-6723.3016.

[5] M.J. Salar-García, V.M. Ortiz-Martínez, P. Olivares-Carrillo, J. Quesada-Medina, A.P. de los Ríos, F.J. Hernández-Fernández, Analysis of optimal conditions for biodiesel production from Jatropha oil in supercritical methanol: Quantification of thermal decomposition degree and analysis of FAMEs, J. Supercrit. Fluids. 112 (2016) 1–6. https://doi.org/10.1016/j.supflu.2016.02.004.

[6] K. Openshaw, A review of Jatropha curcas: an oil plant of unfulfilled promise, Biomass and Bioenergy. 19 (2000) 1–15. https://doi.org/10.1016/S0961-9534(00)00019-2.

[7] W.W. Christie, Preparation of ester derivatives of fatty acids for Chromatographic analysis, Adv. Lipid Methodol. (1993) 69–111.

[8] S. Lafhal, P. Vanloot, I. Bombarda, J. Kister, N. Dupuy, Chemometric analysis of French lavender and lavandin essential oils by near infrared spectroscopy, Ind. Crops Prod. 80 (2016) 156–164. https://doi.org/10.1016/j.indcrop.2015.11.017.

[9] R. Gislum, P. Nikneshan, S. Shrestha, A. Tadayyon, L. Deleuran, B. Boelt, Characterisation of Castor (Ricinus communis L.) Seed Quality Using Fourier Transform

Near-Infrared Spectroscopy in Combination with Multivariate Data Analysis, Agriculture. 8 (2018) 59. https://doi.org/10.3390/agriculture8040059.

[10] A.J. Steidle Neto, D.C. Lopes, J. V. Toledo, S. Zolnier, T.G.F. Silva, Classification of sugarcane varieties using visible/near infrared spectral reflectance of stalks and multivariate methods, J. Agric. Sci. 156 (2018) 537–546. https://doi.org/10.1017/S0021859618000539.

[11] S. Laroussi-Mezghani, P. Vanloot, J. Molinet, N. Dupuy, M. Hammami, N. Grati-Kamoun, J. Artaud, Authentication of Tunisian virgin olive oils by chemometric analysis of fatty acid compositions and NIR spectra. Comparison with Maghrebian and French virgin olive oils, Food Chem. 173 (2015) 122–132. https://doi.org/10.1016/j.foodchem.2014.10.002.

[12] M. Tang, H. Miao, Q. Wu, N. Han, Y. Mu, J. Zhou, J. Ding, Y. Yang, Y. Feng, Z. Huang, Modeling of Fatty Acid Methyl Esters, Monoglycerides, Triglycerides and Diglycerides in Rapeseed Oil Biodiesel by Near Infrared Spectroscopy, J. Biobased Mater. Bioenergy. 13 (2019) 806–811. https://doi.org/10.1166/jbmb.2019.1913.

[13] C. Pasquini, Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications, J. Braz. Chem. Soc. 14 (2003) 198–219. https://doi.org/10.1590/S0103-50532003000200006.

[14] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[15] Q. Liu, Y. Sun, J. Chen, Z. Zhao, C. Li, G. Niu, L. Jiang, Determination of Fruit Oil Content and Fatty Acid Composition in Symplocos paniculata Using Near Infrared Reflectance Spectroscopy, J. Biobased Mater. Bioenergy. 10 (2016) 272–278. https://doi.org/10.1166/jbmb.2016.1607.

[16] J.A. Panford, J.M. deMan, Determination of oil content of seeds by NIR: Influence of fatty acid composition on wavelength selection, J. Am. Oil Chem. Soc. 67 (1990) 473–482. https://doi.org/10.1007/BF02540751.

[17] S. Lian-guang, L. Jun-hui, W. Yu-mei, L. Yu-hua, W. Dan, X. Min, H. Jin-ping, Establishment and Application of Model for Determining Oil Content of Cottonseed Using Near Infrared Spectroscopy, Spectrosc. Spectr. Anal. 35 (2015) 609–612. https://doi.org/10.3964/j.issn.1000-0593(2015)03-0609-04.

[18] K. Zhang, Z. Tan, C. Chen, X.S. Sun, D. Wang, Rapid Prediction of Camelina Seed Oil Content Using Near-Infrared Spectroscopy, Energy & Fuels. 31 (2017) 5629–5634. https://doi.org/10.1021/acs.energyfuels.6b02762.

[19] A.J. Marquez, A.M. Díaz, M.I.P. Reguera, Using optical NIR sensor for on-line virgin olive oils characterization, Sensors Actuators B Chem. 107 (2005) 64–68. https://doi.org/10.1016/j.snb.2004.11.103.

[20] J.M. Montes, F. Technow, B. Bohlinger, K. Becker, Grain quality determination by means of near infrared spectroscopy in Jatropha curcas L., Ind. Crops Prod. 43 (2013) 301–305. https://doi.org/10.1016/j.indcrop.2012.06.054.

[21] Y. Vaknin, M. Ghanim, S. Samra, L. Dvash, E. Hendelsman, D. Eisikowitch, Y. Samocha, Predicting Jatropha curcas seed-oil content, oil composition and protein content using near-infrared spectroscopy—A quick and non-destructive method, Ind. Crops Prod. 34 (2011) 1029–1034. https://doi.org/10.1016/j.indcrop.2011.03.011.

[22] M. Mir, S.M. Ghoreishi, Response Surface Optimization of Biodiesel Production via

Catalytic Transesterification of Fatty Acids, Chem. Eng. Technol. 38 (2015) 835–834. https://doi.org/10.1002/ceat.201300328.

[23] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. https://doi.org/10.1002/cem.1192.

[24] J.P.A. Martins, R.F. Teófilo, M.M.C. Ferreira, Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets, J. Chemom. (2010). https://doi.org/10.1002/cem.1309.

[25] J. V. Roque, W. Cardoso, L.A. Peternelli, R.F. Teófilo, Comprehensive new approaches for variable selection using ordered predictors selection, Anal. Chim. Acta. 1075 (2019) 57–70. https://doi.org/10.1016/j.aca.2019.05.039.

[26] O. Jović, T. Smolić, Z. Jurišić, Z. Meić, T. Hrenar, Chemometric Analysis of Croatian Extra Virgin Olive Oils from Central Dalmatia Region, Croat. Chem. Acta. 86 (2013) 335–344. https://doi.org/10.5562/cca2377.

[27] M.A. Moharam, L.M. Abbas, A study on the effect of microwave heating on the properties of edible oils using FTIR spectroscopy, AFRICAN J. Microbiol. Res. 4 (2010) 1921–1927.

[28] N. Vlachos, Y. Skopelitis, M. Psaroudaki, V. Konstantinidou, A. Chatzilazarou, E. Tegou, Applications of Fourier transform-infrared spectroscopy to edible oils, Anal. Chim. Acta. 573–574 (2006) 459–465. https://doi.org/10.1016/j.aca.2006.05.034.

[29] A.A.C. Yukihiro Ozaki, W. Fred McClure, Near-Infrared Spectroscopy in Food Science and Technology, (2006).

CONSIDERAÇÕES FINAIS

O trabalho apresentou aplicações envolvendo a espectroscopia NIR e métodos quimiométricos, PLS e PLS-DA, a diversas matrizes de origens vegetais. Os resultados mostraram a eficiência em se usar essas fermentas, pois, como se pode constatar, suas aplicações resultaram em determinações mais rápidas, simples, menor custo e com um mínimo preparo das amostras. O trabalho também se fez uso e comparou resultados proveniente de instrumentos NIRP, que são instrumentos mais baratos que os de bancada e podem realizar análises no próprio local da amostra. Portanto, os estudos realizados neste trabalho servem de base para seleção de melhores materiais, auxiliando programas melhoramento genético das espécies.