UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO CIÊNCIAS E TECNOLOGIA
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS GRADUAÇÃO EM QUÍMICA
CURSO DE MESTRADO EM QUÍMICA

MIRELLY ALEXANDRE GOMES

DETERMINAÇÃO DOS PARÂMETROS DE QUALIDADE DA POLPA DA *Spondias mombin* UTILIZANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO E ALGORITMO DE SELEÇÃO DE VARIÁVEIS

CAMPINA GRANDE – 2022

MIRELLY ALEXANDRE GOMES

DETERMINAÇÃO DOS PARÂMETROS DE QUALIDADE DA POLPA DA *Spondias mombin* UTILIZANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO E ALGORITMO DE SELEÇÃO DE VARIÁVEIS

Dissertação apresentada ao programa de pós-graduação em química como pré-requisito à obtenção do título de Mestre em Química.

ORIENTADOR: Prof. Dr. José Germano Véras Neto.

CAMPINA GRANDE

2022

G633d Gomes, Mirelly Alexandre.

Determinação dos parâmetros de qualidade da polpa da *Spondias mombin* utilizando espectroscopia de infravermelho próximo e algoritmo de seleção de variáveis [manuscrito] / Mirelly Alexandre Gomes. - 2022.

57 p.

Digitado.

Dissertação (Mestrado em Química - Mestrado) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2022.

"Orientação : Prof. Dr. José Germano Véras Neto , Coordenação do Curso de Licenciatura em Química - CCT."

1. Calibração multivariada. 2. Espectroscopia vibracional. 3. Algoritmo bioinspirado. 4. Spondias mombin L. I. Título

21. ed. CDD 540

Elaborada por Deivid A. M. Neves - CRB - 15/899 BC/UEPB

MIRELLY ALEXANDRE GOMES

DETERMINAÇÃO DOS PARÂMETROS DE QUALIDADE DA POLPA DA *Spondias mombin* UTILIZANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO E ALGORITMO DE SELEÇÃO DE VARIÁVEIS

Trabalho de Dissertação apresentado ao programa de pós-graduação em química da Universidade Estadual da Paraíba, como pré-requisito à obtenção do título de Mestre em Química.

Aprovada em 19 de Setembro de 2022.

BANCA EXAMINADORA

Prof.ª Dra. Maria Fernanda Pimentel Avelar
Universidade Federal de Pernambuco - UFPE

Prof.ª Dr. David Douglas de Sousa Fernandes
Universidade Federal da Paraíba - UFPB

A minha família, pela dedicação, companheirismo e amizade, **DEDICO**.

AGRADECIMENTOS

Agradeço a Deus a vida, a saúde, as oportunidades... As possibilidades de acertar e de errar, e de aprender em ambas as situações. Por ter tido lições que vão muito além da química...

A meu esposo, Francisco Furtado Neto, por ter renunciado de muitas coisas e momentos para que eu realizasse esta etapa tão importante para mim e para minha vida. Muito obrigada por simplesmente tudo que fez e faz por mim.

À minha família, em especial a minha mãe, irmã e irmão, pelo apoio e motivação. Sou grata por tudo, por ter me mostrado o caminho certo e o que eu sou hoje devo a ti, Mãe.

Aos meus "compadres", em especial, Olga, por ter contribuído de forma tão amiga, acolhendo-me em sua casa por todo período de estudo. A ti, meu mais sincero agradecimento.

Ao Dr. Germano Véras, pela orientação criteriosa, paciência, compreensão e por estar sempre disposto a discutir o trabalho, me fazendo pensar criticamente e incentivando meu crescimento profissional. Enfim, obrigada pela confiança no meu trabalho, obrigada por acreditar em mim e me "ensinar a pescar". Obrigada também pelo laço de amizade que se formou para além do profissional.

Aos professores membros da banca de qualificação, a Profª Drª. Simone Simões, por ter me acolhido como filha no seu grupo de pesquisa gMAQ (Grupo de Metodologias Analíticas e Quimiometria), pelos ensinamentos, paciência, colaboração dada e acima de tudo, pela amizade. Obrigada por mostrar-me que mesmo estando em "vagões diferentes, nada impedi que visitemos o vagão da outra". Ao Prof. Dr. Antonio José Ferreira Gadelha, pela participação na banca e considerações feitas, guiando-me à finalização da dissertação.

A banca examinadora Profª Dra. Maria Fernanda Pimentel Avelar, pelas brilhantes contribuições que trilharão ao trabalho final, e também ao pesquisador Dr. David Douglas de Sousa Fernandes cujas observações serão igualmente imprescindíveis.

Aos colegas do LQAQ, cuja amizade foi construída entre um café e uma brincadeira, pelos inesquecíveis momentos de descontração e conversas filosóficas, ou nem tão filosóficas assim...

A CAPES, pela bolsa concedida, que tornou possível a realização deste trabalho.

Aos que participaram de alguma forma e me ajudaram em mais esta etapa, obrigada.

"Corações gratos e felizes até atravessando obstáculos encontram razões para agradecer a Deus"

(Autor Desconhecido).

RESUMO

O Brasil ocupa o terceiro lugar na produção e comercialização de frutas tropicais consumidas, atrás apenas da China e da Índia. Além das frutas conhecidas há também uma grande variedade de espécies pouco conhecidas, mesmo pelos próprios brasileiros, mas que possuem grande potencial para comercialização *in natura* ou como polpas, por exemplo. Em termos de regiões brasileiras, o Nordeste possui frutas tropicais ainda pouco exploradas, dentre as quais se destaca o fruto de *Spondias mombin* L., conhecido como cajá ou taperebá no Norte. Além do sabor marcante da fruta, o cajá contém substâncias antioxidantes, flavonoides, ácidos fenólicos, vitaminas A e C, antocianinas e carotenoides em quantidades consideráveis. No entanto, quando a fruta é processada, essas substâncias podem ser perdidas, afetando a quantidade de sólidos solúveis totais, a acidez titulável total e o pH, por exemplo. Nesse sentido é fundamental garantir que o produto comercializado atenda às características que o consumidor deseja em termos de propriedades nutricionais e funcionais. No entanto, as análises recomendadas são caras e demoradas que dificultam o controle de qualidade dos produtos, tendo em vista que a maior parte da produção é realizada por pequenas empresas com comercialização local ou regional. Portanto, o objetivo deste trabalho foi avaliar os parâmetros de qualidade da polpa de *S. mombin* avaliando a acidez titulável total (ATT), pH e sólidos solúveis totais (SST), utilizando Espectroscopia NIR associada a algoritmos de Regressão de Mínimos Quadrados Parciais: iPLS, iSPA-PLS e FF-iPLSR em dados brutos e pré-processados. Os dados foram processados utilizando-se as transformações SNV, MSC e derivação de Savitzky-Golay (com combinações de derivadas de primeira e segunda, polinômio de primeiro e segundo grau e janela de 17, 21 e 25 pontos). Os espectros NIR foram obtidos utilizando 36 amostras na faixa espectral de 910 a 1408 nm. Os resultados foram avaliados utilizando a capacidade preditiva em termos de EJCR, RMSECV, RMSEP, Coeficiente de Determinação de calibração (R^2_{cal}) e de predição (R^2_{pred}), *bias e RPD*. As amostras dos conjuntos de calibração (27) e de predição (9) foram selecionadas pelo algoritmo SPXy. Em termos dos resultados obtidos, para a determinação de pH os três modelos obtidos foi com FF-iPLS. Este foi o único tipo de seleção de variáveis que funcionou para a determinação desta propriedade. Obtendo como valores de RMSEC =0,1330, R^2_{cal} = 0,7195, RMSEP = 0,2727, R^2_{pred} = 0,0242 (FF-iPLS dados brutos); RMSEC =0,0204, R^2_{cal} = 0,1748, RMSEP = 0,1412, R^2_{pred} = 0,00862 (FF-iPLS SG21_1_1) e RMSEC =0,1909, R^2_{cal} = 0,3298, RMSEP = 0,1617, R^2_{pred} = 0,0666 (SG21_2_2). Para SST, em termos de °brix, dois

modelos PLS foram adequados, com valores de RMSEC =0,5032, R^2_{cal} = 0,8159, RMSEP = 0,9322, R^2_{pred} = 0,9999 (modelo com SNV) e RMSEC =0,5043, R^2_{cal} = 0,8150, RMSEP = 0,9300, R^2_{pred} = 0,9999 (modelo com MSC). Para ATT, apenas um modelo funcionou, e este também foi com FF-iPLS, tendo valores de RMSEC =0,0072, R^2_{cal} = 0,3945, RMSEP = 0,0087, R^2_{pred} = 0,0167 (FF-iPLS com SG21_2_2). Desta forma, os modelos construídos com pré-processamento SNV e MSC e o com o algoritmo FF-iPLS apresentaram-se como ferramenta interessante para predição das propriedades físico-químicas de polpas de frutas, em especial da *Spondias mombin*.

Palavras-chave: calibração multivariada; espectroscopia vibracional, algoritmo bioinspirado.

ABSTRACT

Brazil occupies the third place in the production and commercialization of consumed tropical fruits, behind only China and India. In addition to the known fruits, there is also a wide variety of species that are little-known, even by Brazilians themselves, but which have great potential for commercialization in natura or pulp, for example. In terms of Brazilian regions, the Northeast has tropical fruits that are still little explored, among which the fruit of Spondias mombin L., known as cajá or taperebá in the North, stands out. In addition to the striking flavor of the fruit, cajá contains antioxidant substances, flavonoids, phenolic acids, vitamins A and C, anthocyanins and carotenoids in considerable amounts. However, when the fruit is processed, these substances can be lost, affecting total soluble solids, total acidity and pH, for example. In this sense, it is essential to ensure that the marketed product meets the characteristics that the consumer wants in terms of nutritional and functional properties. However, the recommended analyzes are expensive and time-consuming, which make it difficult to control the quality of the products, given that most of the production is carried out by small companies with local or regional marketing. Therefore, the objective of this work was to evaluate the quality parameters of the pulp of S. mombin by evaluating the titratable total acidity (TTA), pH and Total Soluble Solids (SST), using NIR Spectroscopy associated with Partial Least Squares Regression algorithms: iPLS , iSPA-PLS and FF-iPLSR in raw and preprocessed data. Data were preprocessed using SNV, MSC transformations and Savitzky-Golay derivation (with combinations of first and second derivatives, first and second degree polynomials and windows of 17, 21 and 25 points). NIR spectra were obtained using 36 samples in the spectral range from 910 to 1408 nm. The results were evaluated using the predictive capacity in terms of EJCR, RMSEC, RMSEP, Determination Coefficient for Calibration (R^2_{cal}) and Prediction (R^2_{pred}) and bias. The samples from the calibration (27) and prediction (9) sets were selected by the SPXy algorithm. In terms of the results obtained, for the determination of pH the three models obtained were with FF-iPLS. This was the only type of variable selection that worked for the determination of this property. Getting as values of RMSEC = 0.1330, R^2_{cal} = 0.7195, RMSEP = 0.2727, R^2_{pred} = 0.0242 (FF-iPLS raw data); RMSEC = 0.0204, R^2_{cal} = 0.1748, RMSEP = 0.1412, R^2_{pred} = 0.00862 (FF-iPLS SG21_1_1) and RMSEC =0.1909, R^2_{cal} = 0.3298, RMSEP = 0.1617, R^2_{pred} = 0, 0666 (SG21_2_2). For SST in terms of °brix, they had two suitable models and both used PLS, with values of RMSEC =0.5032, R^2_{cal} = 0.8159, RMSEP = 0.9322, R^2_{pred} = 0.9999 (model with SNV) and RMSEC =0.5043, R^2_{cal} = 0.8150, RMSEP = 0.9300, R^2_{pred} = 0.9999 (model with MSC). For

acidity, only one model worked, and this one was also with FF-iPLS, having values of RMSEC =0.0072, R^2_{cal} = 0.3945, RMSEP = 0.0087, R^2_{pred} = 0.0167 (FF-iPLS with SG21_2_2). In this way, the models built with SNV and MSC pre-processing and with the FF-iPLS algorithm presented themselves as an interesting tool for predicting the physical-chemical properties of theft pulp, especially *Spondias mombin*.

KEYWORD: multivariate calibration. vibrational spectroscopy. bioinspired algorthm.

SUMÁRIO

1. INTRODUÇÃO..... 12

2. OBJETIVOS..... 14

2.1. Objetivos Específicos..... 14

3. REFERENCIAL TEÓRICO..... 15

***3.1* Fruto..... 15**

***3.2* Frutos tropicais..... 15**

***3.3* Cajá..... 15**

3.3.1 Processamento do cajá..... **17**

3.3.2 Controle de qualidade de polpas..... **18**

3.4 Espectroscopia NIR no controle da qualidade de alimentos..... 19

3.5 Quimiometria..... 20

3.5.1 Regressão PLS..... **22**

3.5.2 Figuras de mérito em modelos de regressão..... **25**

3.5.2.1 Exatidão..... **26**

3.5.2.2 Bias..... **26**

3.5.2.3 EJCR **26**

3.5.2.4 RPD..... **26**

3.5.3 Seleção de variáveis..... **27**

3.5.3.1 Seleção de variáveis com iPLS..... **28**

3.5.3.2 Seleção de variáveis com iSPA-PLS..... **29**

3.5.3.3 Seleção de variáveis com FF-iPLS..... **30**

3.5.3.4 Aplicações de espectroscopia NIR e calibração PLS em análise de alimentos **31**

4 MATERIAL E MÉTODOS..... 35

4.1 Aquisição de amostras..... 35

4.2 Determinação dos parâmetros físico-químicos..... 35

4.2.1 Determinação do potencial hidrogeniônico (pH)..... **35**

4.2.2 Determinação de sólidos solúveis totais (SST)..... **35**

4.2.3 Determinação de acidez total titulável (ATT)..... **36**

4.3 Obtenção dos espectros NIR..... 36

4.4 Tratamento Quimiométrico..... 36

5 RESULTADOS E DISCUSSÃO..... 39

5.1 Métodos de Referência..... 39

5.2 Modelos quimiométricos para pH..... 42

5.3 Modelos quimiométricos para ATT..... 46

5.4 Modelos quimiométricos para SST..... 49

6 CONCLUSÃO..... 51

REFERÊNCIAS..... 52

1 INTRODUÇÃO

Segundo a Associação Brasileira dos produtores Exportadores de Frutas e Derivados (ABRAFRUTAS, 2019), a fruticultura brasileira ocupa uma área de mais de 2,5 milhões de hectares em todo País, com volume anual de produção de aproximadamente 44 milhões de toneladas. Assim, o Brasil é um grande produtor de frutos tropicais, ocupando o terceiro lugar mundial atrás apenas da China e da Índia, possuindo grande diversidade e sendo bastante apreciados em todo o mundo.

Ressalte-se que significativa parte dos frutos produzidos no Brasil é pouco conhecida ou até desconhecida inclusive dos próprios brasileiros e que representam grande potencial para o desenvolvimento de novos produtos (ANUÁRIO DA FRUTICULTURA, 2016 apud FREITAS et al., 2013). Na esteira do raciocínio, a região Nordeste do Brasil também possui uma grande variedade de frutas tropicais ainda pouco exploradas, dentre as quais se destaca o fruto da *Spondias mombin* L. *(Anacardiaceae*), com nomes populares de cajá, taperebá, cajazeira, cajá-mirim, cajá verdadeiro, dentre outros. O fruto suculento, azedo, apresentando uma quantidade significativa de nutrientes, minerais, fibras e vitaminas A e C. Também são ricos em carotenoides, compostos fenólicos e antioxidantes. Além do fruto bastante apreciado na cultura local, as folhas de cajazeira são amplamente utilizadas na medicina tradicional por suas propriedades anti-inflamatórias, antimicrobianas e antivirais.

O processamento do cajá em nível nacional está restrito às regiões de produção, em escala ainda pequena, sendo boa parte dos estabelecimentos de caráter doméstico, produzindo desde polpas até geleias e *licor*. Ressalte-se que o volume comercializado do fruto *in natura* e de seus produtos ainda é baixo devido a sua sazonalidade e perecibilidade. Assim sendo, buscam-se procedimentos de obtenção de produtos com maior integridade e meta de qualidade que permitirá expandir a produção com garantia da qualidade para mercados consumidores no Brasil e no mundo. Entretanto, não é apenas a limitação da obtenção de produtos que mantenham as propriedades nutricionais que afetam as empresas que processam os frutos. Outro gargalo dos Arranjos Produtivos Locais é a restrição na capacidade de avaliação da qualidade dos da matéria-prima e de seus derivados pelo alto custo das análises, repercutindo fortemente na rentabilidade comercial das pequenas indústrias que detém grande parte do processamento dos frutos.

Para contornar os inconvenientes supracitados, técnicas analíticas não destrutivas estão sendo estudadas para desenvolvimento de métodos de análise de polpas. Dentre estes se destacam os métodos espectroscópicos vibracionais, tal como a espectrometria de absorção

molecular na região do infravermelho próximo (NIRS, do inglês *"Near Infrared Spectroscopy"*), espectrofluorimetrias e voltametrias. Esses tipos de tecnologias permitem obter resultados com um tempo muito curto, com quantidade mínima de reagentes ou ausência destes, sendo em muitas aplicações procedimentos não destrutivos e não invasivos das amostras.

Entretanto, a espectroscopia NIR é uma técnica não-alvo (*untarget*), isto é, que não permite identificar ou quantificar uma molécula ou conjunto de moléculas específicas, e os produtos de origem biológica como as polpas possuem matrizes bastante complexas, com uma grande quantidade de substâncias em proporções muito distintas entre amostras. Assim sendo, ferramentas quimiométricas são necessárias para permitir a interpretação qualitativa ou quantitativa de analito(s) nas amostras.

Por outro lado, muito embora a quimiometria apresente ferramentas que processem grande número de variáveis advindas das amostras, isto demanda considerável desempenho computacional e justifica o uso de algoritmos que permitam escolher comprimentos de onda ou intervalos destes que sejam mais correlacionados com a(s) variável(is) de interesse (PAULA et al., 2014). Este processo de seleção de variáveis reduz informações desnecessárias, abstrai informações relevantes e assim reduz as zonas ruidosas. Neste sentido, um algoritmo bioinspirado foi desenvolvido pelo Laboratório de Química Analítica e Quimiometria da Universidade Estadual da Paraíba, em conjunto com A Universidade Federal do Rio Grande do Sul, denominado de *Firefly* (vagalume). Este algoritmo de seleção de variáveis é baseado no comportamento luminoso de vagalumes em busca de alimento e foi descrito para aplicação em Regressão por Mínimos Quadrados Parciais em intervalos (iPLS, do inglês *interval Partial Least Square*) (OLIVEIRA et al., 2021).

Deste modo, o objetivo do presente trabalho foi avaliar o método quimiométrico FF-iPLS para determinação dos parâmetros de qualidade de polpa de cajá utilizando espectroscopia NIR como técnica analítica e compará-lo com PLS, iSPA-PLS e iPLS.

2 OBJETIVOS

Avaliar os parâmetros de qualidade pH, sólidos solúveis totais (SST), e acidez total titulável (ATT) da polpa de fruta da *Spondias mombin* L. utilizando espectroscopia NIR com método quimiométrico FF-iPLS e compará-lo com PLS, iPLS e SPA-PLS.

2.1 Objetivos Específicos

- Determinar as propriedades físico-químicas, pH, sólidos solúveis totais (SST), e acidez total titulável (ATT) da polpa da *Spondias mombin* L através dos métodos clássicos de análise;
- Construir modelos quimiométricos das propriedades físico-químicas com o algoritmo FF-iPLS;
- Comparar os resultados dos modelos FF-iPLS com os algoritmos PLS, iSPA-PLS e iPLS;

3 REFERENCIAL TEÓRICO

3.1 Fruto

Em termos botânicos, o fruto é um órgão formado por um ou mais ovários desenvolvidos, aos quais podem se associar outras estruturas acessórias e ocorre exclusivamente nas Angiospermas.

Os frutos são alimentos benéficos para a saúde humana porque têm minerais, antioxidantes, vitaminas e ácidos graxos essenciais (KARASAKAL, 2020), além de uma grande quantidade de micronutrientes, como minerais, fibras e vitaminas.

Segundo a Vidal (2019), o Brasil é o terceiro maior produtor mundial de frutas com 42,2 milhões de toneladas, atrás apenas da China e da Índia, distribuindo-se em uma área de cerca de 2,5 milhões de hectares.

As frutas podem ser produzidas em diversos climas e suas características nutricionais estão associadas a estas condições edafoclimáticas. Em termos de origem de produção, os frutos podem ser classificados como frutos de clima temperado, tropical e subtropical. Visto que o Brasil se situa em uma zona climática tropical a maioria de seus frutos podem ser classificados como tropicais.

3.2 Frutos Tropicais

São consideradas espécies tropicais, aquelas cuja origem ou adaptação estejam relacionadas à zona climática Tropical. Uma característica marcante dos frutos tropicais é sua sazonalidade, isto é, produção restrita à determinadas épocas do ano (BARBOSA et a, 2014). Desta forma, pode-se dizer que frutas tropicais provêm de plantas dos mais diversos habitats, tendo como característica comum a intolerância às geadas. Os representantes dos frutos tropicais vêm de inúmeras famílias, incluindo Anacardiaceae (manga, cajá, imbu), Sapindaceae (rambutan, taun, lichia, longan), Passifloraceae (maracujá), Bromeliaceae (abacaxi) e Annonaceae (pinha, graviola, pinha) (UNDERHILL, 2003). Uma dos frutos mais apreciados no Nordeste e Norte brasileiros e que ainda não é um produto de circulação nacional é a *Spondias mombin* L., popularmente conhecido no Nordeste como cajá e no Norte como taperebá.

3.3 Cajá

Spondias L. é um gênero de árvores frutíferas que compreende a 18 espécies nativas da América tropical, Ásia e Madagascar. A família *Anacardiaceae* compreende setenta e seis

(76) gêneros e 600 espécies que são principalmente árvores e arbustos que crescem em zonas tropicais, subtropicais e temperadas (GUEDES, 2018).

Além desses climas, as *Spondias* são encontradas na região Norte e Nordeste. Tipicamente conhecidas como “cajazeiras”, seus frutos recebem nomes distintos. Nas regiões acima citadas, as espécies deste gênero com maior potencial de exploração e uso agroindustrial são: mombim amarelo (*Spondias mombin* L.), conhecido em certas regiões brasileiras, como cajá; cajá-mirim ou taperebá; mamãe vermelha ou ciriguela (S. *purpurea* L.); umbu ou imbu (*S. tuberosa Arruda* Câmara); ambarella; maçã dourada; cajarana ou cajá-manga (*S. dulcis Parkinson*); e duas espécies não definidas taxonomicamente (Santos 1996 apud Júnior et al 2004), mas considerados híbridos naturais cajá- umbu (*S. mombin x S. tuberosa*) e o umbuguela (*S. tuberosa x S. purpurea*). Muito embora grandes designações do gênero, as Spondias de modo geral apresentam características semelhantes à do umbuzeiro, produzindo deliciosos e suculentos frutos.

A *S. mombin* pode atingir 30 metros de comprimento e seus frutos possuem uma coloração amarela brilhante, contendo uma pequena camada de polpa ao redor de um caroço volumoso (PEREIRA, 2017). Os frutos são saborosos, nutritivos, apresentam odor agradável e a cor amarelada indica fidedignamente a presença de carotenoides.

Os frutos de *S. mombin* são ricos em carotenoides, vitaminas A e C, sais minerais, fibras, além de compostos fenólicos (NUNES, 2019). Além disso, apresenta como principal carotenoide a β-criptoxantina, seguido da luteína. Entre os compostos fenólicos mais abundantes em *S. mombin* estão a rutina, o ácido elágico e a quercetina, com sua capacidade antiviral intrínseca e capacidade de gerenciar distúrbios gastrointestinais crônicos (BRITO et al., 2018).

O cajá apresenta efeitos farmacológicos importantes a saúde humana, atuando como cicatrizante e estimulante para produção de suco gástrico. Brito et al (2018) avaliaram que o extrato etanólico do cajá possui atividade antiulcerogênica mediada por atividade antioxidante, além de ser antisecretor e anti-Helicobacter pylori.

Além do fruto, as cascas e folhas do cajá também possuem propriedades medicinais. Estudos fotoquímicos demonstraram que a casca possui propriedades antidiarreica e antidisentérica. As folhas apresentam taninos, saponinas, esteróis, flavonoides e alcaloides (GUEDES, 2018), sendo usadas para dores de estômago, complicações do parto e enfermidades dos olhos e laringe (PEREIRA, 2017). As folhas também são utilizadas como anti-inflamatório, antitumoral, antidepressivo e ansiolítico (AYOKA et al., 2005; CABRAL et

al., 2016; ALBUQUERQUE et al., 2007; ELUFIOYE et al., 2017; SAMPAIO et al., 2018; GUEDES, 2018).

Ainda sobre as propriedades medicinais da espécie, estudo mostrou que as folhas da cajazeira apresentaram atividade antibacteriana contra *Pseudomonas aeruginosa* e *Shigella dysenteriae* e os extratos da casca do caule inibiram o crescimento das bactérias *Escherichia coli* e *Klebsiella pneumoniae* (GUEDES, 2018). Oladunmoye (2007) mostrou que a folha também impede o crescimento das bactérias *Bacillus cereus* e *Clostridium sporogenes*

As folhas da cajazeira também são utilizadas na medicina tradicional africana para tratar desordens neurológicas. Alguns trabalhos desenvolvidos comprovaram os efeitos ansiolítico, sedativo, antiepiléptico e antipsicótico (SAMPAIO, 2018). Apesar dessa abordagem terapêutica, os estudos químicos ainda são escassos, e alguns ácidos, flavonoides, taninos e triterpenos são "isolados" da *S. mombin* (CORTHOUT et al., 1991; CABRAL, 2016).

Com base nos estudos sobre os efeitos farmacológicos das cascas e das folhas da *S. mombin*, além das propriedades nutricionais do fruto, o aumento da área plantada e da produção deve ser estimulada. Entretanto, o fruto possui alta perecibilidade, podendo ser só consumida *in natura* na época de safra. Neste sentido, técnicas de processamento do fruto podem permitir o fornecimento dos derivados do cajá, garantindo a ingesta de compostos fenólicos, que possuem ação antioxidante, além de protetora contra o surgimento e/ou desenvolvimento de processos degenerativos que conduzem a doenças crônicas não transmissíveis (VIEIRA et al, 2011).

3.3.1 Processamento do cajá

Por se tratar de um produto orgânico, o cajá é altamente perecível, devido a fatores físico-quimicos e biológicos, resultando em curta vida de prateleira. Além disso, a alta perecibilidade aumenta à medida que alguns cuidados são negligenciados, como transporte, lavagem, armazenamento e colheita inadequados. Isto faz com que sejam muito elevadas as perdas pós-colheita ou que sejam reduzidas suas propriedades nutricionais.

Assim, o processamento do cajá visa atender vários segmentos do setor de alimentos. Dentre os procedimentos mais simples está a produção de polpas, que permite a utilização posterior para preparo de sucos, sorvetes, produtos de confeitaria e lácteos (MACIEL, 2020), além da indústria farmacêutica e de cosméticos. Além disso, devido às mudanças ocorridas no perfil dos consumidores também está ocorrendo um crescimento no mercado de polpas para consumidores familiares (MACHADO et al., 2007).

Modo geral, países como Brasil, Colômbia, Chile, México e Reino Unido entre outros, possuem legislação que definem e padronizam a qualidade para a produção de polpas de frutas (como valores mínimos de acidez, sólidos solúveis ou requisitos microbiológicos), com finalidade de garantir a qualidade do produto final (ALAMAR et al., 2016).

Atualmente, muitos estudos têm demonstrado a inadequação tecnológica do processo de produção das polpas de frutas brasileiras para consumo humano, que refletem a falta de: mão de obra qualificada, falta de métodos de processamento padronizados e falta de boas práticas de fabricação (BPF) em pequenas e médias empresas.

3.3.2 Controle de qualidade de polpas

Quando se refere a alimentos, é essencial um monitoramento adequado e tenaz das propriedades nutricionais que esse alimento apresenta. Considera-se de grande importância a determinação e avaliação de componentes que sejam específicos do alimento. E para isto, são utilizadas técnicas analíticas capazes de monitorar os parâmetros de qualidade, de forma a conduzi-los ao permissível para consumo humano, seguindo os parâmetros dos órgãos de controle e fiscalização, que no caso do Brasil é a Agência Nacional de Vigilância Sanitária (ANVISA).

Para Freitas (2017), "é necessária a determinação da composição centesimal, onde são realizados procedimentos que têm por finalidade de fornecer informações em relação a composição química, físico-química e, ou, física do alimento".

Tais aspectos são considerados importantes, uma vez que garantem o adequado controle de qualidade das matérias-primas e alimentos processados. Tendo em vista que no processamento do fruto ocorre uma possível degradação dos nutrientes presentes por oxidação (SILVA et al, 2017). Assim sendo, o setor industrial vem se mostrando mais interessado em garantir uma qualidade maior a seus produtos. A grande preocupação em melhorar seus processos industriais para que possa ser possível diminuir os custos na produção, sem diminuir o faturamento e a qualidade se tornou importante. Neste sentido, técnicas que consigam predizer as propriedades das frutas processadas permite que o produtor do derivado do fruto, o consumidor final e os órgãos de fiscalização possam acompanhar a qualidade do que está sendo produzido, comercializado e consumido.

As principais ferramentas utilizadas no controle de risco são as Boas Práticas de Fabricação (BPF) e a Análise de Perigos e Pontos Críticos de Controle (APPCC). As BPF são um conjunto de procedimentos higiênico-sanitários instituídos pela Agência Nacional de Vigilância Sanitária do Ministério da Saúde (MACIEL, 2020).

Em termos do controle de qualidade de bebidas à base de frutas, o qual a polpa está inserida, é realizado no Brasil a partir do que determina os Padrões de Identidade e Qualidade (PIQ), definidos pelo Ministério da Agricultura Pecuária e Abastecimento (MAPA), que indicam as características físicas, químicas e organolépticas, estabelecendo valores de limites mínimos e máximos específicos para cada tipo de produto (NOGUEIRA et al., 2020). Porém, o que a indústria privada busca para as análises é um custo e benefício viável. E desta forma, a busca por aparatos mais ágeis, economicamente viáveis e que apresentem desempenho analítico de um método oficial de rotina quanto a sensibilidade, exatidão e precisão estão cada vez maior.

Diante do exposto, Morgan e Haley (2019) apontam que o controle de processo industrial alimentício pode incluir o uso de sensores, atuadores, controladores, softwares e redes totalmente compatíveis e integrados com os demais. Neste sentido, a espectroscopia no infravermelho próximo (NIR, do inglês *Near Infared*) é uma alternativa viável para análises industriais de monitoramento dos padrões de qualidade das polpas de frutas, por englobar técnicas não destrutivas, com obtenção de resultados rápidos, reprodutíveis e com custo financeiro baixo.

3.4 Espectroscopia NIR no controle da qualidade de alimentos

As espectroscopias são um conjunto de métodos analíticos baseados na interação da radiação eletromagnética com a matéria, podendo esta radiação ser absorvida, emitida, refletida, dentro outras. Uma das técnicas espectroscópicas que permitem a análises de alimentos, com custo baixo e com medidas *on-line*, durante o processo de produção, com consequente ajuste as condições operacionais automaticamente com base nos resultados das análises obtidos, é a espectroscopia NIR. Além disso, é uma técnica que permite o desenvolvimento de métodos não invasivos e não destrutivos das amostras.

A espectroscopia NIR é denominada de espectroscopia vibracional, pois consiste na exposição de várias amostras a radiação eletromagnética. A combinação da energia da radiação com a diferença de energia entre dois níveis vibracionais provoca uma resposta seletiva do sistema molecular à radiação incidente. Isso significa que em determinado comprimento de onda, uma frequência correspondente é absorvida. Este fenômeno corresponde a interação da matéria com fótons que possuem energia na faixa de $2{,}65 \times 10^{-19}$ a $7{,}96 \times 10^{-20}$ J, o que corresponde ao intervalo de comprimentos de onda de 750 a 2.500 nm. O resultado do fenômeno pode ser observado pela visualização do espectro de absorção de uma amostra que é dado pela representação gráfica da combinação da intensidade de absorção

versus o comprimento de onda. Nestes espectros estão caracterizadas bandas relacionadas aos sobretons e bandas de combinação de vibrações fundamentais que estão presentes na região do infravermelho médio, bandas de vibração fundamentais (CHANDRASEKARAN et al., 2019).

A versatilidade do uso da espectroscopia NIR abrange áreas muito amplas, tais como: aplicações de ciências básicas à engenharia de alimentos e agrícola e, ciências ambientais. Em termos de alimentos, o resultado da interação da radiação eletromagnética com as espécies químicas, pode fornecer a identificação e a quantidade da espécie analisada ou de determinada propriedade física ou físico-química, tal como acidez titulável total, pH, sólidos solúveis totais, dentre outros. Além disso, é possível separar frutos defeituosos de sadios, detectar adulterações, quantificar diversos parâmetros nas mais variadas matrizes, como leite e derivados, frutas e vegetais, azeite, mel e cervejas, dentre outros.

A capacidade da espectroscopia NIR de avaliar matéria orgânica é devido ao intervalo de comprimentos de onda estar situado nas vibrações das ligações C – H, O – H e N – H (PRIETO, 2017), componentes principais de compostos orgânicos presentes em espécies vegetais.

Embora a espectroscopia NIR seja referida como uma nova tecnologia, ela foi descoberta em 1800, quando Herschel identificou a dispersão de ondas eletromagnéticas além do visível usando uma série de termômetros com lâmpadas escurecidas (PRIETO, 2017). Entretanto, foi a partir da década de 1960 que os sinais na região do NIR passaram a ser utilizados analiticamente com o advento de computadores com capacidade de processamento adequado e ferramentas matemático-estatísticas que permitissem a interpretação dos espectros. Isto foi necessário visto que os espectros gerados são demasiadamente complexos para uma interpretação direta visual por parte do analista. A área da química que tem como objeto de estudo o uso de ferramentas matemático-estatísticas de interpretação de grande quantidade de dados químicos é denominada de Quimiometria.

3.5 Quimiometria

O avanço tecnológico ocorrido nos últimos anos repercutiu em todas as áreas da ciência, incluindo a área da química analítica. A busca por ferramentas matemáticas e estatísticas que possam tratar grandes conjuntos de dados analíticos foi a principal causa do surgimento e rápido desenvolvimento de uma nova temática que é nomeada de Quimiometria (FERREIRA, 2015). Como consequência deste avanço tecnológico, e maior volume de dados

gerados a partir de medidas analíticas, há necessidade de incrementar metodologias multivariadas para sua interpretação.

A Quimiometria pode ser dividida em diversas áreas, compreendendo desde a análise exploratória de dados, a calibração multivariada para a predição quantitativa das propriedades de uma amostra e a classificação de amostras de acordo com os critérios de agrupamento, dentre outras (HASHIMOTO, 2015).

As ferramentas quimiométricas permitem, entre muitas coisas, observar relação entre a estrutura de moléculas e sua atividade biológica, denominada de relação quantidade-estrutura (QSAR, do inglês *Quantitative structure–activity relationship*); reconhecer e discriminar padrões de efeitos de dadas moléculas em termos, por exemplo, farmacológico ou nutricional; e modelar e monitorar processos químicos industriais, observando alterações no perfil espectral do produto durante a transformação química, física ou físico-química.

Concomitantemente, diversas técnicas foram bastante difundidas para o tratamento de dados químicos: métodos de reconhecimento de padrões tais como: Análise Discriminante Multivariada, como, Análise de Componente Principal (PCA, do inglês *Principal Component Analysis*), Modelagem Independente Flexível por Analogia de Classe (SIMCA, do inglês *Soft Independent Modelling of Class Analogy)*, K- vizinhos mais próximos (KNN, do inglês *k-Nearest Neighbour),* Mínimos Quadrados Parciais para Análise Discriminante (PLS-DA, *Partial Least Squares - Discriminant Analysis),* Análise Hierárquica por Agrupamentos (HCA, do inglês *Hierarchical Cluster Analysis*).

Calibração Multivariada é o procedimento que consiste em desenvolver modelos para quantificar uma propriedade de interesse da amostra (FERREIRA, 2015). Esta foi rapidamente difundida a partir do final dos anos 1980, uma vez que sua utilização possibilitou a realização de análises quantitativas sem a necessidade de resolução do sinal analítico, permitindo a modelagem na presença de interferentes.

Na maioria das vezes, ao realizar-se análise em uma amostra, objetiva-se determinar a concentração das espécies em questão. Porém, como supracitado, a concentração não é uma grandeza gerada diretamente pelo instrumento, o que a faz ser prevista a partir de suas propriedades físicas. Para isto, utiliza-se uma técnica de calibração, capaz de correlacionar a medida de uma propriedade física com o parâmetro analisado. Além disso, deve ser selecionado um conjunto de calibração que contenha amostras com qualquer espécie interferente, ou melhor, as espécies interferentes devem estar no conjunto de calibração em quantidades variáveis para que seja possível construir um modelo de calibração que compense o efeito das espécies interferentes com o método multivariado (GEMPERLINE, 2006).

A utilização dos métodos de calibração multivariada é válida para extrair informações químicas dos espectros, no qual pode ser aplicada a vários métodos instrumentais, tais como, UV-Visível, Raman, Infravermelho Médio, Fluorescência, RMN e Espectros de massa, mas sua importância tem se destacado na espectroscopia NIR, devido à dificuldade de utilizar métodos univariados do espectro gerado.

No que tange as ferramentas de calibração multivariada, as principais são: Regressão Linear Múltipla (MLR, do inglês *Multiple Linear Regression*), Regressão por Componente Principal (PCR, *Principal Component Regression*) e Regressão por Mínimos Quadrados Parciais (PLS, do inglês *Partial Least Squares Regression*), dentre outros. É digno de nota que o método PLS é o mais difundido entre os três e, portanto, será descrito na sequência.

Como citado anteriormente, o método mais comumente empregado de calibração multivariada para análise química é a Regressão PLS (FEUDALE et al, 2002). Este método consiste em obter uma sequência de modelos parciais ajustados por quadrados mínimos a matrizes **X** e **Y**.

3.5.1 Regressão PLS

A regressão PLS foi desenvolvida por Herman Wold em 1966 para a modelagem de conjunto de dados complexos (ZIMMER; ANZANELO, 2014). No decorrer da década de 70, os grupos liderados por S. Wold e H. Martens popularizaram o uso deste método para aplicações químicas (DINIZ, 2010).

A regressão PLS busca relacionar a matriz **X** (composta por informações, como os espectros NIR) à matriz **Y** (composta por propriedades das respectivas amostras, como propriedades químicas, reatividade, atividade biológica), permitindo analisar dados com forte correlação, com presença de ruídos (WOLD; SJÖSTRÖM; ERIKSSON, 2001).

De modo geral, o método PLS consta com duas etapas: Construção do modelo de calibração ou treinamento e a construção do conjunto de validação. A primeira etapa consiste em uma matriz de espectros de absorbância obtidos, variáveis independentes (matriz **X**) e de uma matriz concentração contendo valores determinados por um método de referência, variáveis dependentes (matriz **Y**). A segunda fase, composta pelas amostras teste, contem medidas de amostras independentes das usadas no conjunto de treinamento, onde são utilizadas para avaliar o desempenho da calibração (GELADI; KOWALSKI, 1986).

A regressão PLS transforma as variáveis de resposta e de processo em um número reduzido de combinações lineares (ZIMMER; ANZANELO, 2014). Onde as matrizes **X** e **Y**

são decompostas por PCA em escores e pesos respectivamente, denominadas variáveis latentes. Podendo ser explicadas matematicamente pelas equações a seguir:

$$X = T.P^t + E \qquad \textbf{Equação 1}$$

$$Y = U.Q^t + F \qquad \textbf{Equação 2}$$

Onde **T** e **U** são matrizes de escores (*scores*) das matrizes **X** e **Y**, respectivamente; **P** e **Q** são as matrizes de pesos (*loadings*) ou fatores das respectivas matrizes **X** e **Y**; e **E** e **F** são os resíduos, a parte não modelada das matrizes (DINIZ, 2010).

Após decomposições das respectivas matrizes, é feita uma relação linear entre os "blocos" das amostras (**T** e **U**) das matrizes **X** e **Y**, onde se é obtido o coeficiente de regressão entre os *scores* de **X** e os *scores* de **Y** para cada VL. Representada pelas **Equações 3** e **4** abaixo:

$$u_h = u_{h.}t_h \qquad \textbf{Equação 3}$$

$$T = X.W \qquad \textbf{Equação 4}$$

O "h" representa as componentes principais na **Equação 3**, e a **Equação 4** está expresso o coeficiente de regressão obtido da relação linear dos blocos **U** e **T**, em que o **W** é designado como a matriz de fatores-peso. A melhor relação linear possível entre os "scores" desses dois blocos é obtida a partir de pequenas rotações das variáveis latentes dos blocos de **X** e **Y** (POPPI et al, 2000). Essa rotação é feita com intuito de captar a maior informação possível de **Y**, e assim aumentar a correlação (covariância) de **X** com **Y**. Posto isto, a componente **Y** passa agora a ser chamada de variável latente.

Para que os vetores **t** e **y** sejam correlacionados é preciso fazer com que cada fator-peso **w** seja proporcional a covariância **X** e **y** (FERREIRA, 2015). Há duas variantes do método PLS, conhecidas como PLS1 e PLS2. No PLS1, o **y** é um vetor, ou seja, vai determinar apenas uma variável de interesse. Já no PLS2, **Y** é uma "matriz", onde se é calculado um conjunto de escores e variáveis latentes para todas as propriedades de interesse.

O algoritmo NIPALS (Wold) e sua versão não ortogonal (Martens & Naes), foram os primeiros algoritmos PLS usados para construir modelos de regressão (FERREIRA, 2015). O algoritmo inicia-se estimando os elementos fator-peso, **w**, para primeira variável latente, e em

seguida faz o mesmo "*loop*" para o número "h" de variáveis latentes. Expresso matematicamente pela expressão abaixo:

$$\widehat{w}_{1=}\, \mathrm{X^T y(y^T y)^{-1}} \qquad \textbf{Equação 5}$$

A **Equação 5** expressa o coeficiente de regressão de um ajuste de mínimos quadrados das colunas **X** em **y**. Em seguida o vetor peso é determinado, como mostra a **Equação 6**.

$$\mathrm{w_1} = \frac{\mathrm{X^T y}}{\|\mathrm{X^T y}\|} \qquad \textbf{Equação 6}$$

Em sequência, o *score* é determinado a partir da multiplicação (combinação linear das variáveis originais) da matriz **X** com o peso **w** padronizado, obtido como segue a **Equação 7.**

$$\mathrm{t_1} = \mathrm{X}.\,\widehat{\mathrm{w}}_1 \qquad \textbf{Equação 7}$$

Segundo Ferreira 2015, para uma matriz de dados com três variáveis $\mathbf{x_1}$, $\mathbf{x_2}$, e $\mathbf{x_3}$, cada uma delas é projetada em um vetor **y** quando estimados os coeficientes de regressão $\widehat{w}_1$, $\widehat{w}_2$ e $\widehat{w}_3$. Desta forma, sua projeção gera um vetor resultante definido por esses elementos, obtido pela **Equação 8.**

$$\widehat{\mathrm{w}}_1^{\mathrm{T}} = \widehat{[\mathrm{w}_1\, \widehat{\mathrm{w}}_2\, \widehat{\mathrm{w}}_3]} \qquad \textbf{Equação 8}$$

Posto isto, o vetor $\mathbf{t_1}$ colocado na **Equação 7** é utilizado para determinar os *loadings* (vetor peso $\mathbf{l_1}$) e sucessivamente o valor do coeficiente q_1. Através da determinação de **tt** e rearranjos de equações iniciais é possível fazer a estimativa tanto do vetor peso $\mathbf{l_1}$ quanto do coeficiente q_1 para a estrutura **X,** como mostra a **Equação 9** e **Equação 10**.

$$\mathrm{X^T} = \mathrm{l_1 t_1^T} \rightarrow \hat{\mathrm{l}}_1 = \mathrm{X^T t_1 (t_1^T t_1)^{-1}} \qquad \textbf{Equação 9}$$

$$\hat{\mathrm{q}}_1 = \mathrm{(t_1^T t_1)^{-1} t_1^T y} \qquad \textbf{Equação 10}$$

A determinação do vetor $\mathbf{l_1}$ é essencial para a determinação da matriz de resíduo **E,** ou seja, a parte não modelada**.** A interpretação segue a mesma lógica da **Equação 8**, e desta forma o vetor $\mathbf{l_1}$ agora passa a ser o novo coeficiente de regressão. E assim como o vetor $\widehat{w}_1$, este também passa a ser definido no espaço original das variáveis. Geometricamente falando esta etapa consiste na projeção de **y** em $\mathbf{t_1}$.

Em seguida, repete-se o "*loop*" até completar o número de variáveis latentes escolhido para o modelo. E por fim, ocorre a determinação do vetor de regressão $\hat{b}$, posto na **Equação 11**.

$$\hat{\mathrm{b}} = \mathrm{W_A}(\mathrm{L_A^T}\,\mathrm{W_A})^{-1}\hat{\mathrm{q}} \qquad \textbf{Equação 11}$$

O vetor de regressão obtido é resultado de rearranjos dos vetores w da matriz de peso, com a matriz de *loadings* e o valor da diagonal também da matriz de *loadings* da estrutura **y**.

3.5.2 Figuras de mérito em modelos de regressão

Na quimiometria, existem alguns parâmetros analíticos de desempenho que devem ser seguidos com a finalidade de garantir a confiabilidade dos modelos construídos, conhecido por figuras de mérito. A partir dos modelos gerados, são obtidas as figuras de mérito.

3.5.2.1 Exatidão

Essa figura de mérito expressa o grau de concordância entre o valor estimado ou medido e o valor de referência (VALDERRAMA, 2009). O erro de predição é uma métrica utilizada para selecionar o número ideal de fatores a serem incluídos no modelo. Em calibração multivariada, normalmente, é expressa através da raiz quadrada do erro médio quadrático de previsão, conforme a **Equação 12**:

$$RMSEP = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \hat{y})^2}}{n_v} \qquad \textbf{Equação 12}$$

Onde:

RMSEP = erro quadrático médio de predição;

y_i = concentração da i-ésima amostra de previsão;

$\hat{y}$ = concentração prevista para esta amostra;

n_v = número de amostras do conjunto de validação.

3.5.2.2 Bias

É uma constante que verifica o erro sistemático (VALDERRAMA et al., 2009). Logo, um valor de BIAS muito próximo de zero não garante que o modelo possui capacidade preditiva adequada, apenas indica que, em termos absolutos, há uma equidade entre pontos subestimados pelo modelo (resíduos negativos) e pontos sobrestimados (resíduos positivos), o que é necessário, mas não suficiente (BARBOSA, 2019). O cálculo de bias é dado conforme ilustra **Equação 13**:

$$\text{bias} = \frac{\sum_{i=1}^{n_v}(y_1 - \hat{y}_1)}{n_v} \qquad \textbf{Equação 13}$$

Onde:

bias= erro sistemático;

ŷ e yi = valores de referência pela propriedade predita de interesse;

N = número de amostras do conjunto de validação.

3.5.2.3 Elipse de confiança (EJCR)

Na comparação de resultados entre os dois métodos distintos, a exatidão pode ser acessada por meio da comparação dos valores obtidos para a inclinação e o intercepto de uma reta ajustada entre os valores de referência e os estimados pelo modelo (VALDERRAMA et al., 2009). Porém, isto é melhor observado quando se é posto uma elipse de confiança, onde os valores esperados 1 e 0, para inclinação e intercepto devam estar contidos nela. Desta forma pode-se inferir estatisticamente que os modelos são equivalentes no nível de confiança esperado. Do contrário, quando tais valores se encontram fora da elipse, ou seja, fora de seus intervalos de confiança indicam que há presença de erros sistemáticos proporcionais e constantes.

3.5.2.4 RPD

A relação de desempenho do desvio (do inglês *Residual Prediction Deviation*) é uma figura de mérito mais recente utilizada para estimar a capacidade preditiva dos modelos multivariados (CORDEIRO, 2020). Essa figura de mérito foi estimada para os conjuntos de calibração e validação.

$$RPD_{val} = \frac{DP_{val}}{RMSEP} \qquad \textbf{Equação 14}$$

Onde:

RPD_{val} = Desempenho do desvio para o conjunto de validação;

DP_{val} = Desvio padrão dos valores de referência para as amostras no conjunto de validação;

RMSEP = Erro quadrático médio de predição.

3.5.3 Seleção de variáveis

A razão da calibração multivariada acoplada à espectroscopia NIR ser considerado um importante método analítico é por ser uma técnica rápida, barata e categórica para solucionar os problemas reais.

Por outro lado, apesar de obter uma grande quantidade de dados para interpretação, nem toda a região do espectro NIR é informativa para descrever as amostras, seja em termos de calibração multivariada ou de técnicas de reconhecimento de padrões. No sentido de permitir que variáveis não informativas sejam excluídas da análise dos dados e reduzir o tempo computacional, diversos trabalhos utilizam técnicas de seleção de variáveis.

A seleção de variáveis bem como a redução de dimensionalidade vem sendo investigada na construção de modelos eficientes tanto para classificação quanto para regressão, auxiliando ao analista na tomada de decisão (FISTER et al., 2013; GOMES et al., 2013; GOODARZI et al., 2014; SHI et al., 2016; ATTIA et al., 2017; ZHANG et al., 2018). Desta forma, a seleção de variáveis pode ser compreendida como um problema de otimização combinatorial com restrições, no qual o objetivo é encontrar um subconjunto de preditores, capaz de produzir modelos de calibração mais exatos e robustos quando análogos ao modelo com todas as variáveis.

Vale ressaltar que apesar da calibração PLS permitir um ranqueamento de variáveis mais informativas, ainda sim os modelos podem ser afetados por aquelas não informativas (XIAOBO et al., 2010). Assim sendo, a seleção de variáveis constitui maneira de aperfeiçoar e otimizar os modelos quimiométricos de forma parcimoniosa (ZHANG et al., 2018). Desta forma, procedimentos de seleção de variáveis são de particular interesse quando se trata de dados espectroscópicos.

Existem muitos tipos de técnicas de seleção de variáveis. A escolha da melhor abordagem de seleção de variáveis geralmente vai depender do problema. Durante a avaliação do método de seleção de variáveis, deve-se levar em consideração a forma como os algoritmos de seleção de variáveis são executados em relação ao elevado número de variáveis.

Modo geral, as técnicas de seleção de variáveis podem ser classificadas em determinísticos e estocásticos.

Algoritmos Determinísticos são aqueles cujo não estão associados a nenhuma probabilidade a priori, apresentando como solução uma única resposta (subconjunto) de variáveis. Ou seja, expressam ao final do cálculo uma solução única. Muito embora apresentando bom desempenho, esses apresentam limitações, uma vez que carece ao passo que o número de variáveis é aumentado. São exemplos de algoritmos determinísticos para regressão por mínimos quadrados parciais o iPLS e o iSPA-PLS.

3.5.3.1 Seleção de variáveis com iPLS

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

O iPLS é um algoritmo desenvolvido para utilização interativa em modelos PLS. O mesmo desenvolve modelos analisando os espectros em intervalos (regiões) equidistantes do espectro completo, onde a quantidade de intervalos é definida pelo usuário, a fim de construir uma sequência para o algoritmo iPLS sem interferência de outras regiões e rica em informações importantes.

A escolha de intervalos iPLS deve ser criteriosa, uma vez que quantidades de intervalos pequenos acarretará faixas maiores, implicando em possíveis informações desnecessárias, tendo como consequência um sobreajuste. Em contrapartida, intervalos muitos estreitos podem fragmentar as informações úteis dos dados, gerando possível subajuste (FERNANDES, 2016).

Posteriormente a divisão dos intervalos, o algoritmo iPLS calcula um modelo a partir da validação cruzada, o que possibilita ao usuário decidir quantos fatores serão necessários para descrever o sistema. Feito isto, cada intervalo é calculado gerando um modelo PLS, e o que apresentar menor Raiz Quadrada do Erro Médio Quadrático para Validação Cruzada (RMSECV, do inglês *Root Mean Square Error of Croos Validation*) é o intervalo selecionado (NØRGAARD et al. 2000). Porém, outros parâmetros também são avaliados, tais como, R^2 (coeficiente de correlação quadrado), *slope* (inclinação) e *offset* (deslocamento) com intuito de garantir uma visão mais abrangente do modelo criado.

É crucial que a aplicação do algoritmo seja feita após remoção de anormalidade de amostra e/ou medição (*outlier*). Essas amostras são muito diferentes do restante do conjunto de dados, e sua detecção é crucial no desenvolvimento de modelos multivariados.

3.5.3.2 Seleção de variáveis com iSPA-PLS

O Algoritmo das Projeções Sucessivas para seleção de intervalos em PLS (iSPA-PLS, do inglês *Successive Projections Algorithm for Interval Selection in PLS*), proposto por Gomes et al. (2013), é uma extensão do Algoritmo das Projeções Sucessivas em MLR (SPA-MLR, do inglês *Successive Projections Algorithm in MLR*).

Os modelos SPA-MLR apresentam vantagem com relação à simplicidade e facilidade de interpretação, em comparação aos modelos PLSR. Entretanto, a vantagem do iSPA-PLS frente ao SPA-MLR é ser menos sensível a ruídos instrumentais, bem como, apresentar modelos mais robustos (GOMES et al., 2013).

No trabalho desenvolvido por Krepper et al 2018 outra vantagem do iSPA-PLS foi apresentar resultados mais parcimoniosos e com maior capacidade preditiva, quando confrontado a modelo full PLS, e iPLS em amostras de hamburguer de frango.

No iSPA-PLS a validação cruzada é empregada para determinar um número apropriado de variáveis latentes em cada modelo PLS. A melhor combinação de intervalos é então escolhida com base no menor RMSECV.

Primeiramente, a matriz de repostas instrumentais é centrada na média das colunas, dividindo em **w** intervalos não sobrepostos. A condição estabelecida é que o número de variáveis **k** incluído no intervalo seja maior que o número ótimo de fatores. Posteriormente, os intervalos são submetidos à etapa de projeção via SPA, contendo as colunas de matriz **Xcal**. Tais projeções são utilizadas para formar cadeias K de M variáveis cada, onde M = min {Ncal - 1, K} é o número máximo de variáveis que podem ser incluídas em um modelo MLR com termo de interceptação (MARTINS et al., 2010). A notação {**SEL(1, k), SEL (2, k), …, SEL(M, k)**} incluem os índices das cadeias dos intervalos e é usado para denotar o conjunto índice de variáveis pertencentes à cadeia inicializada com **xk**.

No segundo momento, o conjunto de intervalos selecionados é utilizado na construção dos modelos PLS, empregando validação cruzada, onde será selecionado como “melhores” modelos delimitados pelo menor RMSECV (FERNANDES, 2016).

Em contraste, os algoritmos Estocásticos são aqueles associados a certo grau de probabilidade, possuindo alguma variável de entrada aleatória. Desta forma, apresentam subconjuntos de variáveis distintos ao passo que se realizam as replicatas. Uma classe de algoritmos estocásticos com particular interesse da comunidade científica são os denominados de algoritmos meta-heurísticos. Eles são baseados em uma procura não exaustiva, considerando que há informação insuficiente para descrever a relação entre variáveis de

entrada e saída, envolvendo, portanto, aprendizagem de máquina com convergência de respostas ou limite de número de ciclos executados para encontrar resposta ótima (OLIVERIA et al, 2021).

3.5.3.3 Seleção de variáveis com FF-iPLS

Alguns dos algoritmos meta-heurísticos utilizam métrica não-linear empregando regras de busca para simular aspectos do comportamento de seres vivos, e são denominados, portanto, como algoritmos bioinspirados. A inteligência de bando ou enxame é inspirada no comportamento coletivo de bandos de animais sociais, como formigas, cupins, abelhas, vermes, leões, morcegos, peixes, dentre outros, que apesar de serem indivíduos relativamente pouco sofisticados, apresentam comportamento coordenado que direciona os enxames para os objetivos desejados.

Alguns exemplos de algoritmos estocásticos, meta-heurísticos bioinspirados: otimização por colônias de formigas (*ACO*, do inglês *ant colony optimization*), otimização por enxame de partículas (*PSO*, do inglês *particle swarm optimization*), colônia de abelhas (inglês *bee colony*), algoritmo morcego (do inglês *bat algorithm*), algoritmo leão para otimização (do inglês *lion optimization algorithm*), além do algoritmo vagalume (FF, do inglês *firefly algorithm*) (FISTER et al., 2013).

O algoritmo *firefly* foi desenvolvido por Yang (2008) pode ser aplicado para solucionar problemas mais difíceis de otimização, não havendo garantia que a solução ideal seja encontrada em um período de tempo razoável. Em contrapartida, a randomização permite que o processo de pesquisa evite que a solução fique presa em ótimos locais (FISTER et al., 2013).

Poucos são os artigos publicados em química analítica utilizando o algoritmo firefly. O que implica na abertura de grande lacuna em torno de pesquisas que consolidem a eficácia deste algoritmo para seleção de variáveis.

Dos artigos vistos, os algoritmos que envolvem o conceito Firefly atrelado a calibração multivariada em PLS apenas selecionam variáveis individuais (LI et al., 2015; ATTIA et al., 2017; XU et al.,2018).

No entanto, o algoritmo firefly (FF-iPLS), proposto por Oliveira et al., 2021, seleciona intervalos a partir do comportamento inspirado heuristicamente pelas características dos vaga-lumes, por meio do modelo de regressão PLS, ao qual se emprega o cálculo a todo espectro e posteriormente nos intervalos selecionados.

A matriz inicial de calibração (**Xcal**) é particionada em **n** intervalos não sobrepostos e definidos pelo usuário, onde são posteriormente otimizados através do método (OLIVEIRA et al., 2021).

Os vaga-lumes criados artificialmente tem brilho proporcional à sua capacidade de produzir uma boa aptidão dada pelo RMSECV (*leave-one-out*), onde posteriormente é apresentado graficamente.

O Parâmetro de entrada ffpop corresponde à quantidade de vaga-lumes na população inicial gerada aleatoriamente. Ciclos é número de iteração do processo de otimização; ω é a atratividade em r (distância) = 0; α é a porcentagem aleatória do movimento do vaga-lume; e γ é o brilho. Em cada ciclo do otimizador, o comportamento do enxame será guiado por um componente aleatório α, e pelo brilho (γ). Os valores dos parâmetros selecionados foram: 100 (ffpop), 100 (ciclos). Tais parâmetros são

Cada vaga-lume recebe uma atratividade que decai rapidamente com a distância. Inicialmente, todos os vaga-lumes têm atratividade unitária (i = 1). Ao final dos ciclos de otimização, o vaga-lume de maior brilho e atratividade corresponde à solução otimizada de problema de seleção de variáveis. Por fim, o modelo PLS final é construído, considerando apenas os intervalos armazenados no vaga-lume mais brilhante.

3.5.3.4 Aplicações de espectroscopia NIR e calibração PLS em análise de alimentos

A determinação de parâmetros de qualidade em alimentos utilizando espectroscopia NIR com técnicas de calibração multivariada PLS é bastante utilizada e alguns exemplos são apresentados abaixo.

Galdino et al (2020) desenvolveram um método para determinação da composição aproximada de queijo de cabra (umidade, sólidos totais, cinzas, proteína, gordura e carboidrato por diferença). Neste trabalho foram adequados os modelos quimiométricos PLSR para determinação da quantidade de gordura e de proteína nos dados sem pré-processamento com erros relativos em relação aos métodos de referência de 0,97 e 0,81%, respectivamente para os dois parâmetros.

O modelo desenvolvido para determinação de gordura obteve R^2 de calibração de 0,99%, Raiz Quadrada do Erro Médio Quadrático para Calibração (RMSEC, do inglês *Root Mean Square Error of Calibration*) de 0,01% e para Predição (RMSEP, do inglês *Root Mean Square Error of Prediction*) de 0,07%. Para proteína os resultados foram R^2 de calibração de 0,63%, RMSEC de 0,11% e RMSEP de 0,15%. Em suma, não foi observada diferença significativa em todos os parâmetros analisados ($p > 0,05$). Com o valor do erro relativo,

pode-se inferir que o modelo construído pode prever, com erros reduzidos, os teores de gordura, em que uma excelente correlação entre os dois métodos pode ser observada.

Zhu et al (2021) propuseram um método para determinação quantidade de lipídio e proteína em grãos verdes de café de diferentes origens usando modelos *full*-PLS e redução do número critico de ondas utilizando o cálculo do valor das respostas das medições X do coeficiente de regressão (β). Os modelos quimiométricos obtiveram as seguintes figuras de mérito: R^2 de predição de 0,758% e RMSEP de 0,381% para lipídios e R^2 de predição de 0,892% e RMSEP de 0,209% para lipídios.

Haruna et al (2022) desenvolveram uma aplicação para quantificação rápida de índice de acidez e de peróxido em óleo bruto de amendoim e comparou o modelo full-PLS construído com três métodos de seleção de variáveis, bootstrapping soft shrinkage-PLS (BOSS-PLS); eliminação de variável não informativa-PLS (UVE-PLS), e Amostragem competitiva-adaptativa reponderada-PLS (CARS-PLS). Apesar de modelos menos parcimoniosos, com maior número de variáveis latentes, os resultados com a totalidade dos dados, os resultados foram comparáveis aos demais, com R^2 de calibração de 0,9377% e 0,9559%, RMSEP de 0,1020 mg/g e 0,6820 mEq/kg, respectivamente para índice de acidez e índice de peróxido.

Wang et al (2022) obtiveram um método para quantificação no nível de adulteração de farinha de quinoa com cinco outros tipos de farinhas. Os modelos *full*-PLS foram comparados com dois métodos de seleção de variáveis e os valores das figuras de mérito foram similares entre si, com R^2 de calibração de 0,99%, RMSEC de 1,80%, R^2 de predição de 0,94% e RMSEP de 3,04% para determinação de farinha de trigo como adulterante.

Huan et al (2021) determinaram a quantidade de proteína em grãos de trigo, comparando *full*-PLS com quatro algoritmos com seleção de variáveis, Amostragem reponderada adaptativa competitiva (CARS), combinação variável análise populacional (VCPA), análise populacional de combinação variável de Monte Carlo (MCVCPA) e análise populacional de combinação de variáveis de ponderação automática (AWVCPA). Neste artigo, mesmo com resultados piores do que os demais, os modelos quimiométricos com full-PLS obtiveram figuras de mérito aceitáveis, com R^2 de predição de 0,8876% e RMSEP de 0,04936%.

Assis et al (2018) desenvolveram modelos quimioméricos *full*-PLS e iPLS para determinação de sólidos solúveis, pH e acidez titulável em frutos de *Dovyalis abyssinica* Warb. Os resultados dos modelos para sólidos solúveis foram R^2 de calibração de 0,99 e 0,98%, RMSEC de 0,08 e 0,27 °brix, R^2 de predição de 0,74 e 0,56% e RMSEP de 0,71 e

0,96 °brix para determinação com *full*-PLS e iPLS, respectivamente. Para acidez titulável os resultados foram R^2 de calibração de 0,96 e 0,84%, RMSEC de 2,0 e 3,8 g/kg, R^2 de predição de 0,14 e 0,30% e RMSEP de 7,4 e 10,1 g/kg, para determinação com *full*-PLS e iPLS, respectivamente. Os resultados para pH foram R^2 de calibração de 0,96 e 0,99%, RMSEC de 0,03 e 0,06, R^2 de predição de 0,69 e 0,67% e RMSEP de 0,09 e 0,06, para determinação com *full*-PLS e iPLS, respectivamente.

Moraes et al (2019) estimaram ácido ascórbico em acerola intacta e compararam os modelos *full*-PLS com iPLS. Os melhores modelos obtiveram as seguintes métricas das figuras de mérito: R^2 de calibração de 0,93 e 0,98%, R^2 de predição de 0,19 e 0,12% e RMSEP de 358,8 e 365,3 mg/100g para determinação com *full*-PLS e iPLS, respectivamente.

Guo et al (2020) quantificaram sólidos solúveis totais e grau de pingo-de-mel, doença degenerativa em maçã. Os autores compararam modelos full-PLS com PLS usando três métodos de seleção de variáveis, incluindo SPA-PLS. Os resultados das figuras de mérito dos modelos foram: R^2 de calibração de 0,9610 e 0,9316%, RMSEC de 0,461 e 0,615 °brix, R^2 de predição de 0,9514 e 0,9208% e RMSEP de 0,521 e 0,660 °brix para sólidos solúveis totais nos modelos *full*-PLS e SPA-PLS respectivamente; R^2 de calibração de 0,8557 e 0,8288%, RMSEC de 2,571 e 2,780%, R^2 de predição de 0,8465 e 0,8242% e RMSEP de 2,590 e 2,740% para grua de pingo-de-mel nos modelos *full*-PLS e SPA-PLS respectivamente.

Li et al (2019) detectaram adulteração de Aspergillus flavus em amendoim a partir da contagem total de bolores. Os modelos obtiveram as seguintes figuras de mérito: R^2 de calibração de 0,9678 e 0,9374%, RMSEC de 0,2053 e 0,2887 Log CFU/g, R^2 de predição de 0,9500 e 0,9029% e RMSEP de 0,2545 e 0,3537 Log CFU/g para os modelos *full*-PLS e SPA-PLS, respectivamente.

Krepper et al (2018) determinaram a quantidade de gordura em hambúrguer de frango utilizando espectroscopia NIR associado aos modelos quimiométricos *full*-PLS, iPLS e iSPA-PLS. Os valores obtidos para as métricas de desempenho foram: R^2 de calibração (0,74, 0,73 e 0,76%), R^2 de predição (0,90, 0,88 e 0,94%) e RMSEP (2,33, 2,34 e 1,59 mg/kg) para os modelos *full*-PLS iPLS e iSPA-PLS, respectivamente. Os autores concluíram que a metodologia proposta representa uma excelente alternativa ao método de referência, com resultados concordantes àquelas com nível de confiança estatística de 95, sendo, portanto, promissora em análise de rotina.

Pereira et al (2020) desenvolveram um método para determinar adulteração de leite de vaca em leite de cabra e a quantificação de gordura e de proteína em leite de cabra, leite de vaca e suas misturas usando *full*-PLS, iPLS e iSPA-PLS como ferramentas quimiométricas. O

melhor desempenho para determinação de adulteração de leite de vaca em cabra foi para o modelo iSPA-PLS com pré-processamento de média móvel de 13 pontos e correção de linha de base por offset obtendo 0,9996 para R^2 de calibração, 0,9955 para R^2 de predição e 3,6597 g/100g para RMSEP. Em termos da quantidade de gordura, o melhor desempenho também foi para o modelo iSPA-PLS com pré-processamento de média móvel de 13 pontos e correção linear da linha de base obtendo 0,98 para R^2 de calibração, 0,96 para R^2 de predição e 0,20 g/100g para RMSEP. Por fim, para o conteúdo de proteína, o melhor desempenho também foi para o modelo *full*-PLS com pré-processamento de média móvel de 13 pontos, correção linear e por offset da linha de base obtendo 0,989 para R^2 de calibração, 0,960 para R^2 de predição e 0,047 g/100g para RMSEP. Os autores concluíram, portanto, que os modelos propostos obtiveram boas performances de qualidade.

Dentre todos os trabalhos aqui ressaltados, saliento ainda a inexistência de um que reporte a utilização da espectroscopia NIR e os métodos de seleção de variáveis para determinação dos parâmetros de qualidade da polpa de cajá. Diante do exposto, este trabalho propõe avaliar os parâmetros de qualidade pH, sólidos solúveis totais (SST), e acidez total titulável (ATT) da polpa de fruta da *Spondias mombin* L. utilizando espectroscopia NIR com método quimiométrico FF-iPLS e compará-lo com PLS, iPLS e SPA-PLS.

4 MATERIAL E MÉTODOS

4.1 Aquisição das amostras

Polpas congeladas de cajá, 36 lotes, foram adquiridas em sítios das cidades de Vieirópolis/PB, Sousa/PB, Lavras da Mangabeira/CE e Rio Grande do Norte. Foram coletados os frutos entre os meses de abril a agosto, nos anos de 2020 e de 2021, período da safra de *S. mombin*. Imediatamente após a colheita, as amostras foram processadas para obter a polpa em Indústria de Processamento na cidade de Vieirópolis.

As amostras foram processadas para obtenção de polpas no processamento industrial na cidade de Vieirópolis – Brasil. Quando os frutos chegam à indústria, são pesados e submetidos a seleção manual, descartando-se os estragados e os em estágio avançado de maturação. Em seguida, ocorre a etapa de limpeza em três estágios sequenciais, sendo o primeiro em tanque de imersão para remover sujeiras mais grosseiras; a segunda, com banho de aspersão auxiliado por escovas para remover a sujeira que ficou; e, por fim, aplicação de água clorada, com teor de cloro residual livre entre 5 e 10 ppm por 15 minutos para desinfecção do material a ser processado. Em seguida, há a etapa de separação da polpa das demais partes do fruto (material fibroso, casca e sementes). As polpas obtidas são embaladas a vácuo em plástico e congeladas a -23°C.

4.2 Determinação dos parâmetros físico-químicos

Para realização das análises físico-químicas, as polpas são descongeladas até a temperatura ambiente, sem que haja necessidade de processamento para homogeneização. Foram feitas análises de pH, SST e ATT, todas realizadas em triplicatas.

4.2.1 Determinação do potencial hidrogeniônico (pH)

A avaliação do pH é feita de forma direta através de um pHmetro digital de bancada, modelo LUCA-210, marca Tecnopon, previamente calibrado em solução com pH 4,0 e 7,0. As amostras não são diluídas, porém homogeneizadas com auxílio de uma espátula de metal.

4.2.2 Determinação de sólidos solúveis totais (SST)

A determinação de sólidos solúveis totais é medida através de um refratômetro, e em seguida os valores obtidos pelo índice de refração são comparados a valores tabelados. São colocadas três gotas da amostra homogeneizada no refratômetro de mesa, espera-se um minuto para realização da leitura, e sucessivamente o resultado é dado em escala de

graus °brix. Para realização das demais réplicas, o refratômetro é sempre limpo com água destilada.

4.2.3 Determinação de acidez titulável total (ATT)

A ATT é realizada por meio de titulação, com solução de hidróxido de sódio (NaOH) a 0,1 mol/L, e a acidez da polpa é estimada pelo volume gasto até que o ponto de viragem seja atingindo. A fenolftaleína é utilizada como indicador para essa reação. Os resultados são expressos em termos de ácido cítrico %.

Para realização da análise, cinco gramas (5,0 g) da amostra homogeneizada são pesados em um Erlenmeyer, em seguida adicionadas 50 mL de água destilada e duas gotas do indicador fenolftaleína. Em seguida a solução foi titulada sob agitação constante com hidróxido de sódio a 0,1 mol/L até que atingisse a cor rósea (ponto de viragem). Para realização do cálculo da ATT, usa-se a **Equação 14:**

$$\mathrm{ATT}_{\left(\frac{\mathrm{g}}{100\mathrm{mL}}\right)}=\frac{\mathrm{C_{NaOH}}.\mathrm{V_{NaOH}}.\mathrm{MM_{C6H8O7}}}{\mathrm{V_{amostra}}}.100 \qquad \textbf{Equação 15}$$

Onde:

C = Concentração molar de NaOH;

v_{NaOH} = volume de NaOH (L)

MM = massa molar de $C_6H_8O_7$

v = volume da alíquota em mL.

4.3 Obtenção dos espectros NIR

Neste estudo, utilizou-se um espectrofotômetro UV – Vis - NIR PerkinElmer Lambda 750. Os espectros de reflectância difusa (Prying Mantis modelo, Harrick) das amostras de polpas foram obtidos de 310 a 1430 nm, com incremento de 1 nm e resolução de 0,20 nm. As medidas de cada amostra foram obtidas em triplicatas. O tratamento dos dados foi baseado na média de todos os espectros em cada amostra.

4.4 Tratamento Quimiométrico

Após a obtenção do banco de dados dos espectros NIR médios para cada uma das amostras foi avaliada a necessidade de retirada de variáveis explicitamente apresentadas com sinal espúrio, aplicando seleção de variáveis *a priori*.

Em seguida, os dados foram pré-processados usando Padronização Normal de Sinal (SNV, do inglês *Standard Normal Variate*), Correção Múltipla de Espalhamento (MSC, *Multiplicative Scatter Correlation*) e por derivação Savitzky-Golay, com combinações de derivadas de primeira e segunda, polinômio de primeiro e segundo grau e janela de 17, 21 e 25 pontos.

A transformação MSC, assim como a SNV são utilizadas para corrigir os efeitos de espalhamentos causados muitas vezes por fenômenos físicos (mudanças no caminho ótico, variações de temperatura e pressão), assim como diferenças no tamanho e forma das partículas, emulsões, dispersões e etc. Tais efeitos de espalhamentos, portanto, devem ser removidos, uma vez que acrescentam informações irrelevantes para o problema.

O alisamento por Savitzky-Golay faz uso da média ponderada entre os pontos da janela, onde os pesos são obtidos a partir de um ajuste polinomial. Desta forma, basta ajustar um polinômio de determinado grau as respostas da janela. E então as repostas são substituídas pelo valor do polinômio ajustado fazendo com que as janelas se movam e o valor do polinômio é calculado no novo centro. Este ciclo se repete até que todas as janelas tenham sido de fato ajustadas.

Em seguida, foram construídos os conjuntos de calibração, com 27 amostras, e de predição, com nove amostras, usando o algoritmo SPXY.

Os modelos de calibração iPLS, iSPA-PLS e FF-iPLS foram construídos particionando o conjunto de variáveis em 20 janelas.

Os modelos construídos com o algoritmo FF-iPLS usaram as condições padrão do algoritmo, com os seguintes parâmetros de entrada: 100 indivíduos na população randomicamente gerada, 100 ciclos de iteração do algoritmo, wo igual 0,97, α igual a 0,2 e γ igual a 1.

Os modelos quimiométricos construídos usando os algoritmos de calibração PLS em cada um dos bancos de dados (dados brutos e pré-processados) foram avaliados utilizando a capacidade preditiva em termos da Região de Confiança Elíptica (EJCR, do inglês *Elliptical Joint Confidence Region*), Raiz Quadrada Média do Erro de Calibração (RMSEC) e Predição (RMSEP), Coeficiente de Determinação de calibração (R^2_{cal}) e predição (R^2_{pred}) e bias.

A validação dos modelos foi obtida a partir da validação cruzada completa (do inglês *full cross-validation*), também denominada de "*leave-one-out*". Tal método consiste em dividir o conjunto de calibração em dois subconjuntos, calibração e validação. Neste, constrói-se o conjunto de calibração e assim usa-o para prever o de validação. Este processo é repetido até que todas as amostras de calibração e validação tenham sido incluídas no

subconjunto de validação. Os resultados obtidos na predição são utilizados para validar o modelo criado.

Todos os pré-processamentos e tratamentos quimiométricos foram executados em plataforma MATLAB, versão 2011b.

5 RESULTADOS E DISCUSSÃO

5.1. Métodos de Referência

Os resultados das análises químicas das amostras de polpas de cajá estão listados na **Tabela 1.** Os resultados em sequência mostram as medias de três medições. As faixas de teor de sólidos solúveis totais (SST) os valores de acidez total titulável (ATT) e os valores de pH. Os resultados de SST e ATT nas amostras foram considerados semelhantes aos obtidos por Silva et al (2018), adequados para o desenvolvimento de modelos de calibração NIR confiáveis.

Tabela 1. Distribuição de conteúdo de Sólidos Solúveis Totais (SST), acidez total titulável (ATT) e potencial hidrogênionico (pH) determinados pelos métodos de referência.

Parâmetros	Mínimo	Média	Máximo	Desvio Padrão
TTA(%ácido cítrico)	0.039	0.056	0.086	0.004
TSS (°Brix)	8.30	10.70	13.00	0.72
pH	3.10	3.57	3.94	0.211

Fonte: Elaborado pelo próprio autor, 2022.

O primeiro passo do tratamento dos dados foi uma seleção de variáveis *a priori*. A eliminação das medidas de reflectância nos intervalos de 310 a 909 nm e de 1409 a 1430 nm está associada à presença de sinais apenas na linha de base e muito ruidosos. Assim, a região espectral de trabalho mantida para sequência do trabalho foi de 910 a 1408 nm.

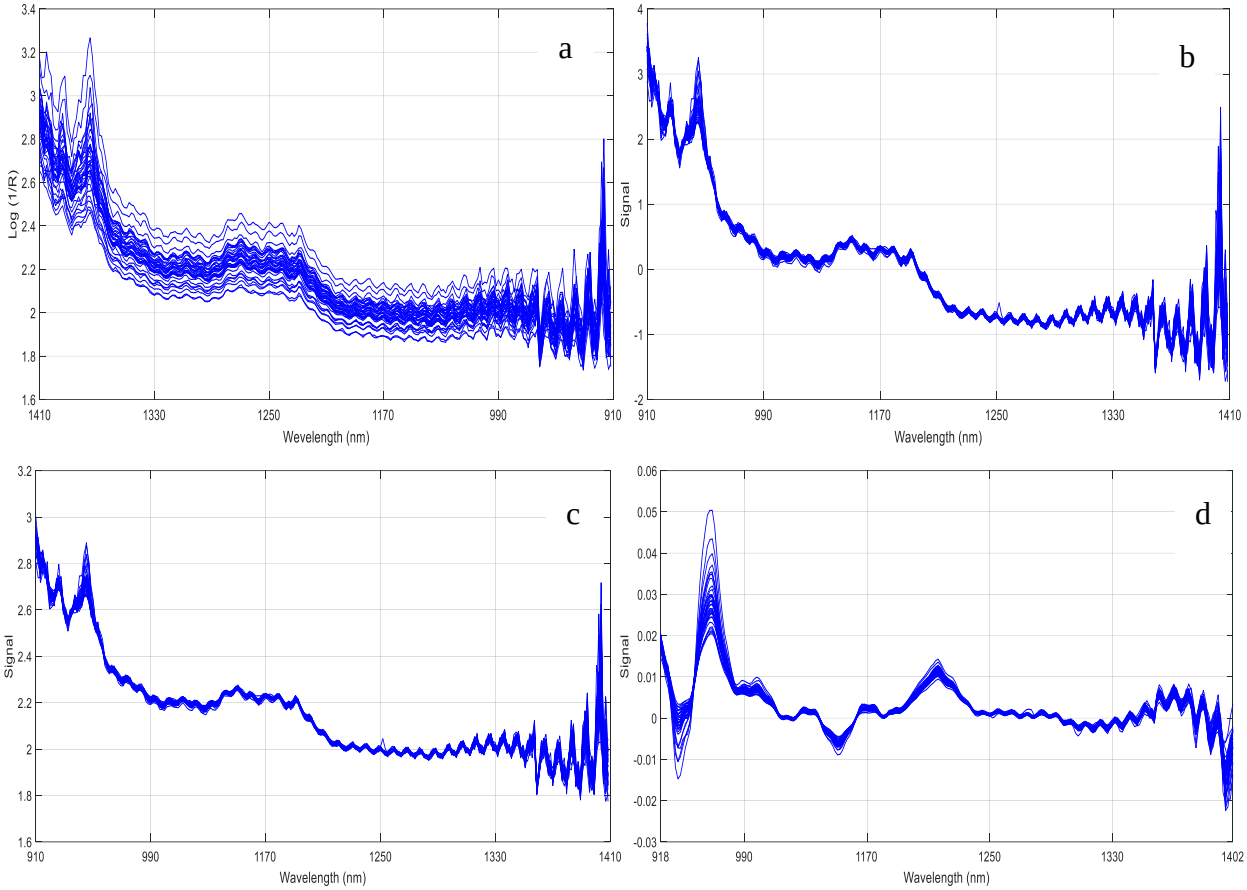
É importante ressaltar que na região espectral do NIR os picos existentes de 1200 a 1400 nm são correspondentes ao segundo sobretom de estiramento das ligações C-H e primeiro sobretom de combinação de C-H, respectivamente. A água apresenta bandas de absorção centradas em aproximadamente 970 nm dentre outras (WILLIAMS; NORRIS, 2001). O amido e os açúcares encontrados nas frutas, tais como sacarose, glicose e frutose, presentes nas polpas de frutos, apresentam bandas de absorção de difícil visualização, pois estão muito próximas das regiões onde a água apresenta forte absorção (DELWICHE et al., 2008). Em geral, estas bandas são provenientes do segundo (920 nm) sobreton de estiramento de ligações O-H e terceiro (910 nm) sobreton de estiramento de ligações C-H (GOLIC et al., 2003; SUBEDI et al., 2007).

Após a etapa de seleção de variáveis *a priori*, os dados passaram por pré-processamentos para eliminar efeitos deletérios na interpretação dos espectros e ressaltar sinais correlacionados com as propriedades de interesse. Os dados foram processados usando SNV, MSC e derivação Savitzky-Golay, com combinações de derivadas de primeira e segunda, polinômio de primeiro e segundo grau e janela de 17, 21 e 25 pontos. Na **Figura 1a-**

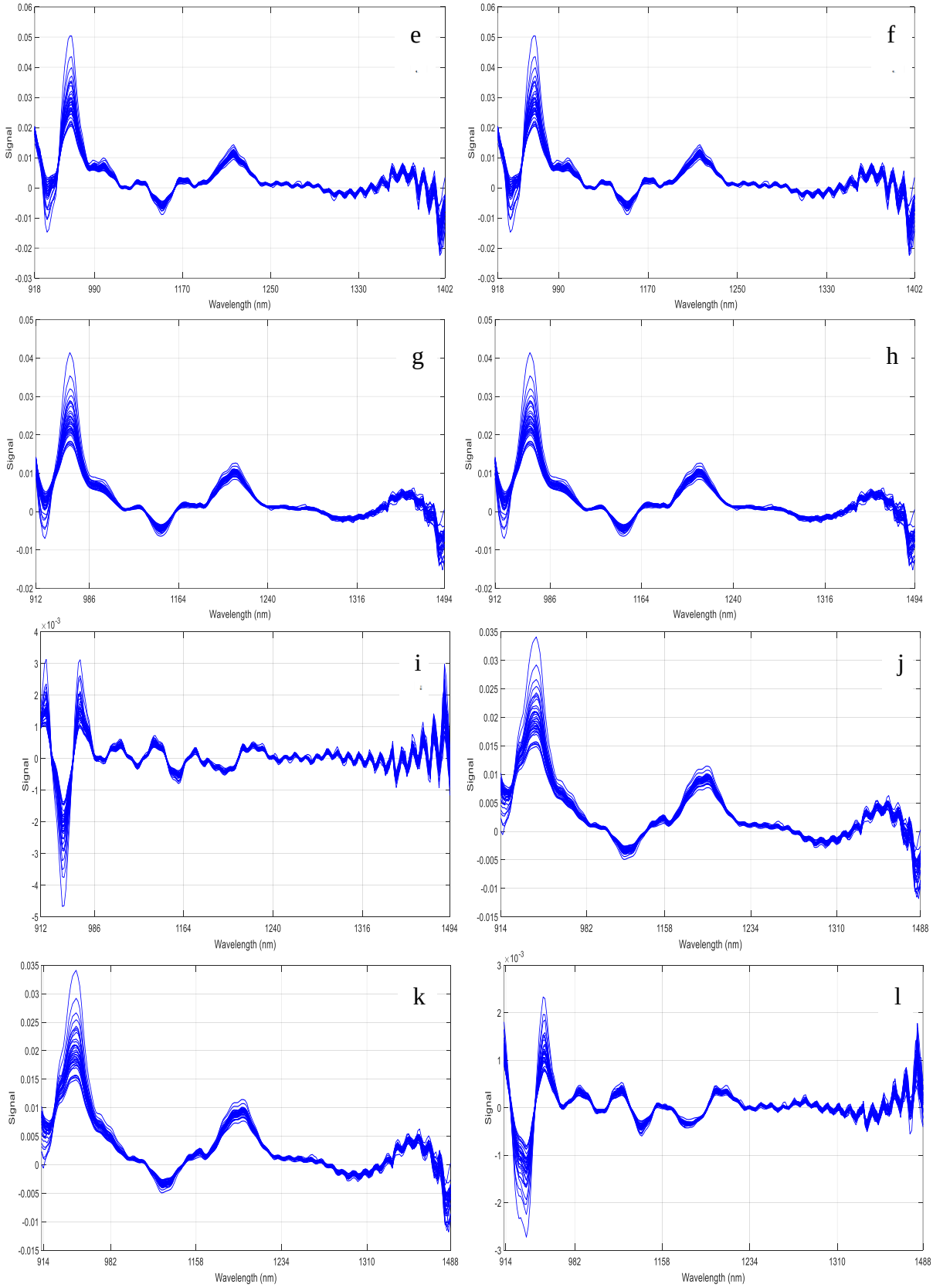
l é possível observar os espectros brutos e nos demais onze pré-processamentos indicados acima.

Em seguida, as amostras para cada um dos bancos de dados, brutos e pré-processados, foram divididos em termos de conjunto de calibração (75%, 27 amostras) e de predição (25%, nove amostras). As amostras do conjunto de calibração foram usadas para construção dos modelos quimiométricos para cada uma das propriedades em estudo (SST, ATT e pH) com PLS, iPLS, iSPA-PLS e FF- iPLS. As amostras do conjunto de predição foram usadas para validar os modelos quimiométricos.

Figura1. Representação dos espectros (a) brutos e pré-processados por: (b) SNV; (c) MSC; (d)SG17_1_1*; (e) SG17_2_1; (f) SG17_2_2; (g) SG21_1_1; (h) SG21_2_1; (i) SG21_2_2; (j) SG25_1_1; (k) SG25_2_1 e (l) SG25_2_2.



Continuação.



Fonte: Elaborado pelo próprio autor, 2022.

*SG17_1_1 indica que o pré-processamento foi obtido por uma derivação de Savitzky-Golay usando janela de 17 pontos, derivada de primeira e polinômio de primeira ordem. Os demais seguem o mesmo processo.

Os resultados dos modelos foram avaliados utilizando a capacidade preditiva em termos de EJCR, RMSEC e RMSEP, R^2_{cal}, R^2_{pred}, bias e RPD. Na sequência serão apresentados os melhores modelos de regressão (PLS, iSPA-PLS, iPLS e FF-iPLS) construídos baseados nos espectros brutos e sem pré-processamento, na faixa espectral de 910 a 1408 nm. Inicialmente, os modelos quimiométricos foram avaliados com base no EJCR, região de confiança para a inclinação da reta e intercepto incluindo os valores esperados de 1 e 0, respectivamente. Se a elipse de confiança contiver o par de dados indicado anteriormente indica que não há evidência de erro sistemático nos modelos. O Bias foi usado, também, para confirmar a ausência de erro sistemático. Em seguida, e com os valores dentro da elipse de confiança e com bias não-significativo, foram avaliadas as figuras de mérito: menor número de variáveis latentes (LVs), maiores valores de R^2_{cal}, R^2_{pred}, e menores valores de RMSECV e RMSEP, além do RPD.

5.2 Modelos quimiométricos para pH

Inicialmente os modelos foram construídos utilizando toda faixa espectral, 910 a 1408 nm, com o conjunto de calibração de 27 amostras e de predição com nove amostras. *A posteriori* foi utilizado o algoritmo de seleção de variáveis FF-iPLS e então criado o modelo quimiométrico para a propriedade pH. O número de variáveis latentes foi escolhido utilizando validação por *full cross validation* tendo como base o menor RMSECV.

Os modelos quimiométricos foram avaliados com base no EJCR, região de confiança para a inclinação da reta e intercepto incluindo valores esperados de 1 e 0, respectivamente. Quando a elipse de confiança mostra o par de dados contido nela, indica a ausência de erro sistemático nos modelos. O bias também foi usado para confirmar a ausência de erro sistemático. RMSECV, RMSEP, RPD e R^2_{cal} e R^2_{pred} foram usados para determinar a qualidade dos modelos. Em seguida, serão apresentados todos os modelos de quimiométrico de regressão, FF-iPLS, iPLS, iSPA-PLS.

Dentre todos os modelos, apenas os em destaque apresentaram os pontos de intercepto dentro da elipse de confiança. Estes apresentaram valores baixos RMSECV, RMSEP e bias, muito embora dando valores de RPD > 3,0. Quando observados com mais atenção, percebe-se que os valores previstos não estão tão distantes dos valores de referência, com RMSEP baixos de 0,178, 0,141 e 0,161, respectivamente. Nestes modelos foram utilizadas apenas duas ou três Variáveis Latentes (LVs) indicando resultados parcimoniosos.

Tabela 2: Resultados quimiométricos para modelos de calibração PLS do teor de pH para polpa de cajá com pré-processamento adequado.

Pré-processamento	Modelo Quimiométrico	LVs	RMSECV	R^2_{cal}
BRUTO	FF-iPLS	3	0.210	0.727
BRUTO	iPLS	1	0.231	0.001
BRUTO	iSPA-PLS	1	0.231	0.001
BRUTO	PLS	1	0.243	0.019
SNV	FF-iPLS	2	0.209	0.365
SNV	iPLS	1	0.226	0.323
SNV	iSPA-PLS	1	0.226	0.323
SNV	PLS	1	0.233	0.255
MSC	FF-iPLS	2	0.205	0.365
MSC	iPLS	1	0.226	0.328
MSC	iSPA-PLS	1	0.226	0.328
MSC	PLS	1	0.233	0.254
SG17_1_1	FF-iPLS	1	0.232	0.016
SG17_1_1	iPLS	2	0.230	0.228
SG17_1_1	iSPA-PLS	2	0.228	0.218
SG17_1_1	PLS	2	0.241	0.226
SG17_2_1	FF-iPLS	1	0.232	0.016
SG17_2_1	iPLS	2	0.230	0.228
SG17_2_1	iSPA-PLS	2	0.228	0.218
SG17_2_1	PLS	1	0.241	0.226
SG17_2_2	FF-iPLS	1	0.233	0.007
SG17_2_2	iPLS	1	0.226	0.269
SG17_2_2	iSPA-PLS	1	0.226	0.269
SG17_2_2	PLS	1	0.262	0.090
SG21_1_1	FF-iPLS	2	0.237	0.174
SG21_1_1	iPLS	2	0.237	0.174
SG21_1_1	iSPA-PLS	2	0.237	0.174
SG21_1_1	PLS	2	0.255	0.185
SG21_2_1	FF-iPLS	1	0.240	0.162
SG21_2_1	iPLS	2	0.237	0.174
SG21_2_1	iSPA-PLS	2	0.237	0.174
SG21_2_1	PLS	2	0.255	0.185
SG21_2_2	FF-iPLS	2	0.228	0.329
SG21_2_2	iPLS	2	0.228	0.240
SG21_2_2	iSPA-PLS	2	0.224	0.321
SG21_2_2	PLS	2	0.236	0.305
SG25_1_1	FF-iPLS	2	0.236	0.195
SG25_1_1	iPLS	2	0.236	0.195
SG25_1_1	iSPA-PLS	2	0.236	0.195
SG25_1_1	PLS	2	0.252	0.227
SG25_2_1	FF-iPLS	2	0.236	0.195
SG25_2_1	iPLS	2	0.236	0.195
SG25_2_1	iSPA-PLS	2	0.236	0.195
SG25_2_1	PLS	2	0.252	0.227
SG25_2_2	FF-iPLS	1	0.240	0.008
SG25_2_2	iPLS	1	0.239	0.105
SG25_2_2	iSPA-PLS	1	0.239	0.105
SG25_2_2	PLS	1	0.255	0.077

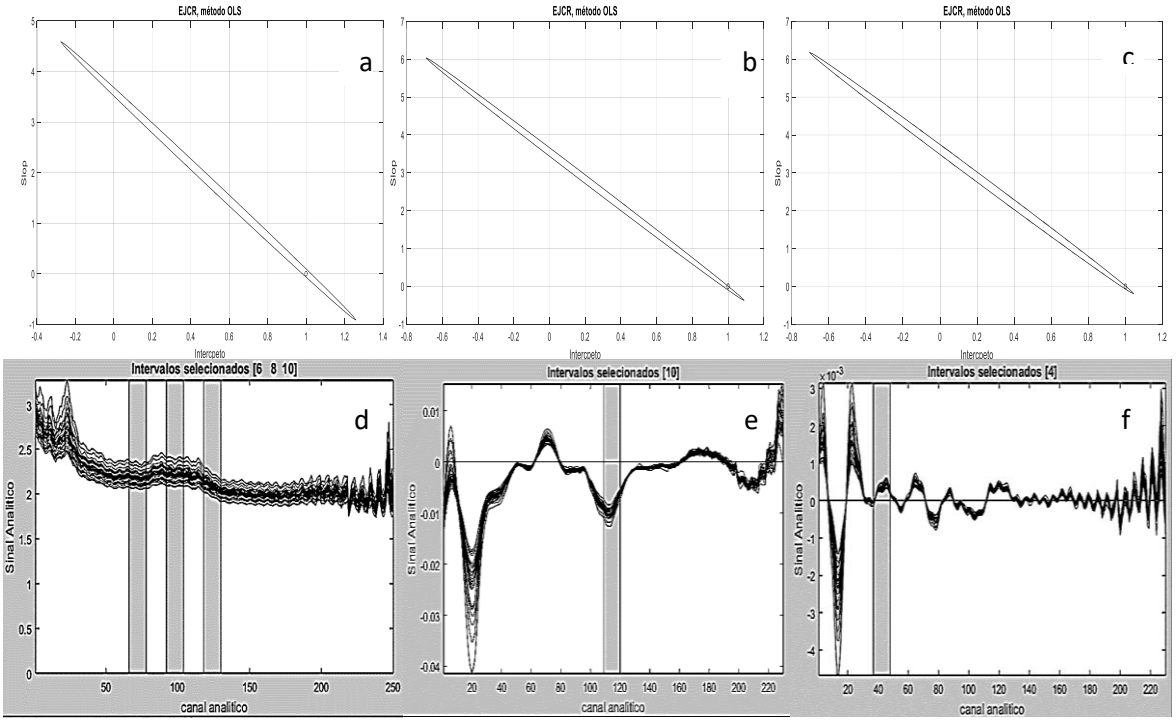
Fonte: Elaborado pelo próprio autor, 2022.

Tabela 3: Resultados quimiométricos para modelos de predição PLS do teor de pH para polpa de cajá com pré-processamento adequado.

Pré-processamento	RMSEP	$Bias_{pred}$	RPD
BRUTO	0.178	0.045	13,581
SG21_1_1	0.141	0.045	13,444
SG21_2_2	0.161	0.022	12,950

Fonte: Elaborado pelo próprio autor, 2022.

Figura 2: a) EJCR para dados brutos; b) EJCR para dados pré-processados com derivação SG21_1_1; c) EJCR pré-processados com derivação SG21_2_2; d) intervalo selecionado para construção do modelo FF-iPLS com dados brutos; e) intervalo selecionado para construção do modelo FF-iPLS com pré-processamento SG21_1_1; f) intervalo selecionado para construção do modelo FF-iPLS com pré-processamento SG21_2_2.



Fonte: Elaborado pelo próprio autor, 2022.

A **Figura 2a-f** apresenta os resultados do modelo. Os gráficos de EJCR são apresentados na **Figura 2a-c**, em que é possível verificar que o ponto ideal de cada um dos modelos se encontra dentro da elipse de confiança. Assim, infere-se que os modelos quimiométricos não possuem erro sistemático para os modelos construídos, sendo eficiente para determinação da propriedade pH. Em seguida, na **Figura 2d-f** é possível observar que os modelos FF-iPLS selecionaram os intervalos 6, 8, 10 e 4(faixa espectral de 1266 a 1288, de 1218 a 1240 e de 1170 a 1192, 1314 a 1336 nm respectivamente).

Tabela 4: Resultados quimiométricos para modelos de calibração PLS do teor de pH para polpa de cajá com pré-processamento adequado.

Pré-processamento	Modelo Quimiométrico	LVs	RMSECV	R^2_{cal}
BRUTO	FF-iPLS	1	0.009	0.002
BRUTO	iPLS	1	0.009	0.037
BRUTO	iSPA-PLS	1	0.009	0.037
BRUTO	PLS	1	0.009	0.018
SNV	FF-iPLS	1	0.009	0.060
SNV	iPLS	1	0.008	0.261
SNV	iSPA-PLS	1	0.008	0.261
SNV	PLS	1	0.009	0.184
MSC	FF-iPLS	1	0.009	0.050
MSC	iPLS	1	0.008	0.259
MSC	iSPA-PLS	1	0.008	0.259
MSC	PLS	1	0.009	0.184
SG17_1_1	FF-iPLS	3	0.008	0.425
SG17_1_1	iPLS	3	0.008	0.477
SG17_1_1	iSPA-PLS	3	0.008	0.477
SG17_1_1	PLS	3	0.010	0.492
SG17_2_1	FF-iPLS	3	0.008	0.425
SG17_2_1	iPLS	3	0.008	0.477
SG17_2_1	iSPA-PLS	3	0.008	0.477
SG17_2_1	PLS	3	0.010	0.492
SG17_2_2	FF-iPLS	1	0.009	0.066
SG17_2_2	iPLS	1	0.009	0.192
SG17_2_2	iSPA-PLS	1	0.009	0.192
SG17_2_2	PLS	1	0.011	0.179
SG21_1_1	FF-iPLS	1	0.008	0.028
SG21_1_1	iPLS	1	0.008	0.058
SG21_1_1	iSPA-PLS	1	0.008	0.058
SG21_1_1	PLS	1	0.009	0.037
SG21_2_1	FF-iPLS	1	0.008	0.028
SG21_2_1	iPLS	1	0.008	0.058
SG21_2_1	iSPA-PLS	1	0.008	0.058
SG21_2_1	PLS	1	0.009	0.037
SG21_2_2	FF-iPLS	2	0.008	0.497
SG21_2_2	iPLS	1	0.008	0.177
SG21_2_2	iSPA-PLS	1	0.008	0.177
SG21_2_2	PLS	1	0.010	0.259
SG25_1_1	FF-iPLS	1	0.008	0.058
SG25_1_1	iPLS	2	0.236	0.195
SG25_1_1	iSPA-PLS	2	0.236	0.195
SG25_1_1	PLS	2	0.252	0.227
SG25_2_1	FF-iPLS	2	0.236	0.195
SG25_2_1	iPLS	2	0.236	0.195
SG25_2_1	iSPA-PLS	2	0.236	0.195
SG25_2_1	PLS	2	0.252	0.227
SG25_2_2	FF-iPLS	1	0.240	0.008
SG25_2_2	iPLS	1	0.239	0.105
SG25_2_2	iSPA-PLS	1	0.239	0.105
SG25_2_2	PLS	1	0.255	0.077

Fonte: Elaborado pelo próprio autor, 2022.

Estes intervalos espectrais escolhidos pelo algoritmo FF-iPLS para construção do modelo quimiométrico podem estar relacionados à combinação de primeiro sobretom de estiramento OH (1220 nm) e combinações e flexões em trechos COH (1240 nm) presentes geralmente em ácidos carboxílicos. Isto é familiar com as bandas presentes do ácido elágico, composto fenólico pertencente ao grupo dos taninos, que se encontra no cajá. O intervalo de 1170 a 1192 nm pode estar associado ao estiramento C-H de anéis aromáticos presentes na rutina, quercetina e ácido elágico. O intervalo 4 (faixa espectral de 1314 a 1336 nm) esta banda pode está associada a sobreposição de bandas de primeiro sobretom de C-O presente em ácidos carboxílicos, uma função existente no ácido elágico.

5.3 Modelo quimiométrico para ATT

O modelo quimiométrico adequado para ATT foi baseado em dados pré-processados pelo algoritmo Savitzky-Golay com janela de 21 pontos, segunda derivada e polinômio de segunda ordem (**Tabela 4**). O modelo apresentara região de confiança incluindo 1 como inclinação da linha reta e zero como intercepto.

Tabela 5. Resultados Quimiométricos para modelos de predição PLS para ATT de polpa de cajá com processamento adequado.

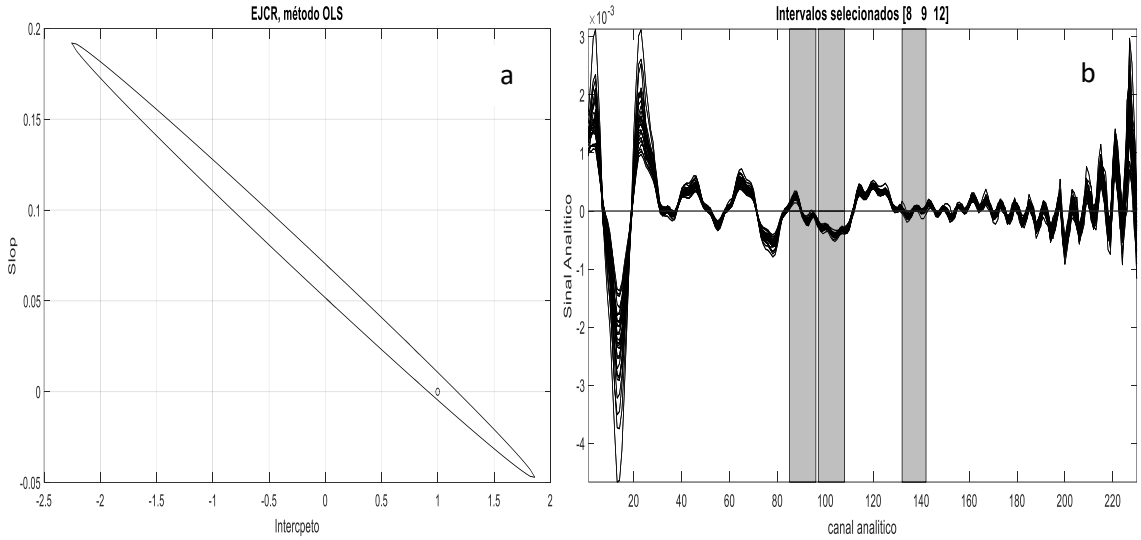
Pré-processamento	RMSEP	$Bias_{pred}$	RPD
SG21_2_2	0.008	0.001	14,535

Fonte: Elaborado pelo próprio autor, 2022.

Os parâmetros estudados dos melhores modelos foram comparáveis entre si. Nesse sentido, apenas o modelo quimiométrico para ATT baseado nos dados pré-processados de SG21_2_2 foi destacado nesta tabela.

Inicialmente o modelo foi construído utilizando toda a faixa espectral, 910 a 1408 nm, com um conjunto de calibração de 27 amostras e um conjunto de previsão de nove amostras. Em seguida, foi utilizado o algoritmo de seleção de variáveis FF-iPLS e, em seguida, foi criado o modelo quimiométrico para a propriedade ATT. O número de variáveis latentes foi escolhido usando validação cruzada completa com base no RMSECV mais baixo. A **Figura 3** apresenta os resultados do EJCR para todos os modelos quimiométricos para determinar o ATT. O gráfico do EJCR é apresentado na **Figura 3a**, onde é possível verificar que o ponto ideal do modelo está dentro da elipse de confiança. Assim, infere-se que o modelo quimiométrico não apresenta erro sistemático para o modelo construído, sendo eficiente para determinação da propriedade ATT.

Figura 3. EJCR para FF-iPLS modelo quimiométrico para ATT usando pré-processamento: a) SG21_2_2. b) intervalo selecionado para construção do modelo FF-iPLS com pré-processamento SG21_2_2.



Fonte: Elaborado pelo próprio autor, 2022.

Os resultados quimiométricos dos modelos quimiométricos de ATT desenvolvidos são comparados na **Tabela 4**. Os modelos apresentaram valores elevados de RPD (> 3,0) e baixos RMSECV, RMSEP e bias. Quando observados com mais atenção, percebe-se que os valores previstos não estão tão distantes dos valores de referência, com um RMSEP baixo de 0,008. Neste modelo foi utilizado apenas duas Variáveis Latentes (LVs) indicando resultados parcimoniosos.

É possível observar na **Figura 3b** que foram selecionados os intervalos de dados 8, 9 e 12 (faixas espectrais de 1218 a 1240, de 1194 a 1216 e de 1122 a 1144 nm) para a construção do modelo quimiométrico. A faixa espectral 1218 a 1240 nm pode estar associada à flexão vibracional em C-OH, característico de ácidos orgânicos. Já a região de 1194 a 1216 nm pode estar associada às bandas fracas de combinações OH, relacionados a presença de água. Por fim, o intervalo de comprimentos de onda de 1122 a 1144 nm está associado a segundo sobretom de C-H, característico de presença de anéis aromáticos. Desta forma, pode estar associado aos compostos fenólicos de maior abundância na polpa de fruta estudada, sendo eles rutina, quercentina e ácido elágico.

Tabela 6. Resultados quimiométricos para modelos de calibração PLS para SST para conteúdo de polpa de cajá com adequado pré-processamento.

Pré-processamento	Modelo Quimiométrico	LVs	RMSECV	R^2_{cal}
BRUTO	FF-iPLS	1	1.117	0.016
BRUTO	iPLS	1	1.117	0.016
BRUTO	iSPA-PLS	1	1.108	0.036
BRUTO	PLS	1	1.114	0.0239
SNV	FF-iPLS	2	1.050	0.309
SNV	iPLS	2	1.050	0.309
SNV	iSPA-PLS	2	1.047	0.373
SNV	PLS	4	1.318	0.815
MSC	FF-iPLS	2	1.048	0.311
MSC	iPLS	2	1.048	0.311
MSC	iSPA-PLS	2	1.045	0.375
MSC	PLS	4	1.318	0.815
SG17_1_1	FF-iPLS	1	1.116	0.037
SG17_1_1	iPLS	1	1.113	0.045
SG17_1_1	iSPA-PLS	1	1.113	0.045
SG17_1_1	PLS	1	1.132	0.030
SG17_2_1	FF-iPLS	1	1.116	0.036
SG17_2_1	iPLS	1	1.113	0.045
SG17_2_1	iSPA-PLS	1	1.113	0.045
SG17_2_1	PLS	1	1.132	0.030
SG17_2_2	FF-iPLS	1	1.076	0.189
SG17_2_2	iPLS	1	1.045	0.272
SG17_2_2	iSPA-PLS	1	1.045	0.272
SG17_2_2	PLS	1	1.141	0.051
SG21_1_1	FF-iPLS	1	1.116	0.037
SG21_1_1	iPLS	1	1.092	0.170
SG21_1_1	iSPA-PLS	1	1.092	0.170
SG21_1_1	PLS	1	1.131	0.028
SG21_2_1	FF-iPLS	1	1.118	0.031
SG21_2_1	iPLS	1	1.092	0.170
SG21_2_1	iSPA-PLS	1	1.092	0.170
SG21_2_1	PLS	1	1.131	0.028
SG21_2_2	FF-iPLS	1	1.086	0.257
SG21_2_2	iPLS	1	1.086	0.205
SG21_2_2	iSPA-PLS	1	1.086	0.205
SG21_2_2	PLS	1	1.145	0.070
SG25_1_1	FF-iPLS	1	1.130	0.067
SG25_1_1	iPLS	1	1.130	0.067
SG25_1_1	iSPA-PLS	2	0.236	0.195
SG25_1_1	PLS	1	1.142	0.039
SG25_2_1	FF-iPLS	1	1.130	0.067
SG25_2_1	iPLS	1	1.130	0.067
SG25_2_1	iSPA-PLS	1	1.130	0.067
SG25_2_1	PLS	1	1.142	0.039
SG25_2_2	FF-iPLS	1	1.080	0.087
SG25_2_2	iPLS	1	0.991	0.290
SG25_2_2	iSPA-PLS	1	0.991	0.290
SG25_2_2	PLS	1	1.114	0.049

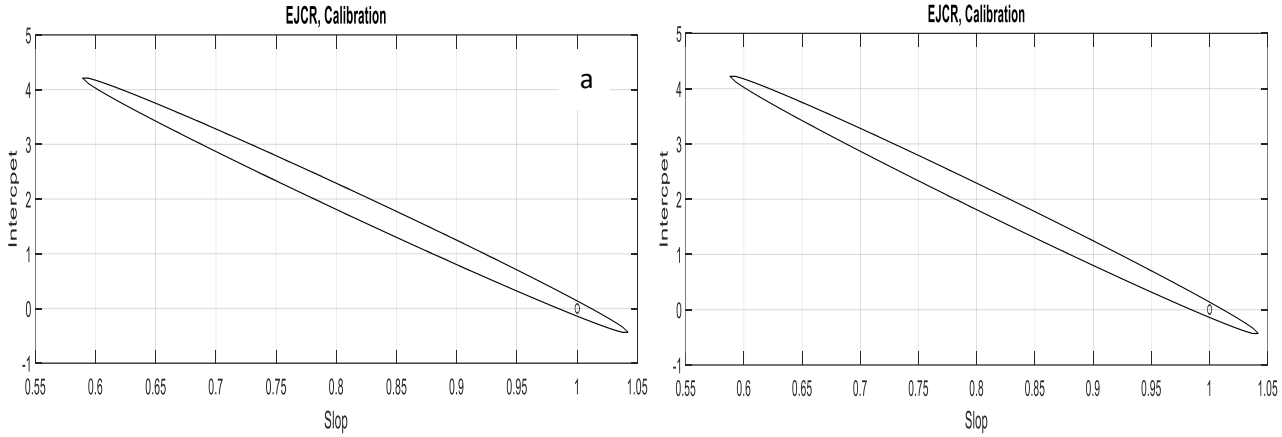
Fonte: Próprio autor, 2022.

5.4 Modelos quimiométricos para SST

Os melhores modelos PLS para pH foram obtidos utilizando faixa espectral 910 a 1408 nm e com pré-processamento SNV e MSC para as 36 amostras. Os pré-processamentos SNV e MSC utilizados foram aplicados com intuito de corrigir possível espalhamento de luz e deslocamento de linha de base.

Como o ponto ideal (1,0) se encontra dentro da EJCR calculada, **Figura 5a** e **b**, pode-se dizer que, estatisticamente, ele é equivalente ao encontrado nas regressões, e que não há diferença significante entre os valores reais e previstos

Figura 5. EJCR para FF-iPLS modelo quimiométrico para SST usando pré-processamento: a) SNV. b) MSC.



Fonte: Elaborado pelo próprio autor, 2022.

Tabela 7: Resultados Quimiométricos para modelos de predição PLS para SST de polpa de cajá com processamento adequado**.**

Pré-processamento	RMSEP	$Bias_{pred}$	RPD
SNV	0.932	0.108	9,053
MSC	0.930	0.105	13,647

Fonte: Elaborado pelo próprio autor, 2022.

Os modelos apresentaram valores elevados de RPD (> 3,0) , baixos RMSECV e bias, porém altos de RMSEP. Estas discrepâncias entre as respostas podem ser explicadas pela baixa amplitude de valores de SST das amostras de predição.

Porém, de modo geral, o algoritmo FF-iPLS utilizado e o modelo PLSR podem ser utilizados para construir modelos quimiométricos eficazes e parcimoniosos que podem associar as respostas de cada variável e obter as melhores respostas para o problema proposto e de forma muito menos dispendiosa em termos de custo e tempo.

Para SST em termos de °brix, tiveram dois modelos adequados e os dois utilizaram PLS. E neste ponto não é interessante a seleção de variáveis, uma vez que no PLS utiliza-se toda faixa. O que explica o fato do FF-iPLS ou iPLS não terem funcionado bem, uma vez que o *Full PLS* distribui a importância de cada uma variável no modelo, logo então, todas as variáveis são importantes de modo geral. Para ATT, teve-se que apenas um modelo funcionou, e este também foi com FF-iPLS.

Desta forma, os modelos construídos com pré-processamento SNV e MSC e o com o algoritmo FF-iPLS apresentaram-se como ferramenta interessante para predição das propriedades pH, SST e ATT. Sendo assim, o FF-iPLS um algoritmo eficaz na determinação destas propriedades em polpas de frutas.

Os resultados mostrados nas tabelas acima mostram que para pH os três modelos obtidos foram com FF-iPLS. Este foi o único tipo de seleção de variáveis que funcionou para a determinação da propriedade pH.

6 CONCLUSÃO

O presente estudo mostrou que as técnicas espectroscópicas podem ser uma técnica alternativa rápida e aplicável para analisar os parâmetros de qualidade de polpas de frutas. No entanto, vale ressaltar que a técnica NIR aplicada apresentou resultados diferentes para cada um dos parâmetros de qualidade analisados.

Para fins de comparação, o algoritmo de seleção de variáveis FF-iPLS foi usado para construir modelos de PLSR com espectroscopia NIR em polpas de cajá *Spondias* para as propriedades físico-químicas pH, ATT e SST. Os modelos PLSR foram construídos com e sem pré-processamento em uma faixa específica de espectros (910 - 1408 nm). Eles foram avaliados em termos de EJCR, validação cruzada, predição e viés.

Com relação às propriedades analisadas, os modelos SST (SNV e MSC), ATT (SG21_2_2 com FF-iPLS) e pH para dados brutos (FF-IPLS) apresentaram melhores desempenhos, com valores permitidos de RMSECV e RMSEP, resultando em modelos aceitáveis. Todos os modelos foram aceitáveis em termos de EJCR, o que concretiza a ideia de um ajuste no modelo quimiométrico criado.

Isso aponta que o método de seleção de variáveis FF-iPLS é um método robusto e adequado para realizar a determinação das propriedades físico-químicas e químicas de polpas de frutas, mas especificamente de cajá *Spondias*. E que pode ser usado para construir modelos quimiométricos harmônicos e parcimoniosos capazes de associar a variável a uma resposta.

REFERÊNCIAS

ABRAFRUTAS. abrafrutas.org. ABRAFRUTAS, 2019. Disponivel em: https://abrafrutas.org/dados-estatisticos/. Acesso em: 10 agosto 2021.

ALAMAR , et al. Quality evaluation of frozen guava and yellow passion fruit pulps by NIR spectroscopy and chemometrics. Food Research International, Campinas, São Paulo, v. 85, p. 209-214, February 2016.

ALVES, V. D. Quantificação de ivermectina em medicamentos veterinários por espectrofluorimetria associada a quimiometria. p. 39. TCC (Bacharelado em Química industrial). Universidade Estadual da Paraíba, Campina Grande, 2019.

ASSIS, C. Aplicação de técnicas espectroscópicas, métodos quimiométricos, fusão de dados e seleção de variáveis no controle de qualidade de blends das espécies de café arabica e robusta. Tese (Doutorado em Ciências -Química), Universidade Federal de Minas Gerais, Belo Horizonte, 2018.

ASSIS, M. W.; FUSCO, D. O.; COSTA, R. C.; LIMA, K. MG.; JUNIOR, L. C. C..; TEXEIRA, G.H.A. PLS, iPLS, GA-PLS models for soluble solids content, pH and acidity determination in intact dovyalis fruit using near-infrared spectroscopy. J Sci Food Agric, Dec; 98 (15), 2018.

BARBOSA, L.S., MACEDO, J. L., SANTOS, C. M., MACHADO, A. V. Estudo da secagem de frutos tropicais do Nordeste. Revista Verde (Mossoró –RN), v. 9, n.1, p.186-190, jan-mar, 2014.

BARBOSA, B.V. Uma nova abordagem na seleção de variáveis para analisadores virtuais via Regressão por Mínimos Quadrados Parciais. Dissertação (Mestrado em Engenharia Industrial), Universidade Federal da Bahia, Salvador, 2019.

BEHLING, E.B., SENDÃO, M.C., FRANCESCA, H.D.C., ANTUNES, L.M.G., BIANCHI, M.L.P. Flavonoide quercetina: aspectos gerais e ações biológicas. Alim. Nutr., Araraquara, v.15,n.3,p.285-292,2004.

BRASIL. Agencia Nacional de Vigilância Sanitária. Resolução RE n°13, de 29 de maio de 2003. Aprova o Guia para validação de métodos analíticos e bioanalíticos. Diário Oficial da União. Brasilia, DF, 2003.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento - MAPA. Regulamento técnico para fixação de qualidade e identidade do pólen apícola. Instrução Normativa n.3, de 19 de janeiro de 2001.

BRITO, S.A., BARBOSA, I.S., ALMEIDA, C.L.F., MEDEIROS, J.W., NETO, J.C.S, ROLIM, L.R., SILVA, **T**.G., XIMENES, R.M., MENEZES, I.R.A, CALDAS, G.F.R, WANDERLEY, A.G. Evaluation of gastroprotective and ulcer healing activities of yellow mombin juice from Spondias mombin L. PLoS ONE 13(11), November, 2018.

CABRAL, B., SIQUEIRA, E.M.S., BITERNCOURT, M.A.O., LIMA, M.C.J.S., LIMA, A.K., ORTMANND, C.F., CHAVES, V.C., PEDROSA, M.F.F., ROCHA, H, A.O., SCORTECCI, K.C., REGINATTO, F.H., GIORDANI, R.B., ZUCOLOTTO, S.M..

Phytochemical study and anti-inflammatory and antioxidant potential of *Spondias mombin* leaves. Rev. bras. farmacogn. 26 (3), May-Jun 2016.

CHANDRASEKARAN, I., PANIGRAHI, S.S., RAVIKANTH, L., SINGH, C.B. Potential of Near-Infrared (NIR) Spectroscopy and Hyperspectral Imaging for Quality and Safety Assessment of Fruits: an Overview Indurani. Food Analytical Methods. May 2019.

CORDEIRO, L.D. Calibração Multivariada e Imagens Digitais no controle de qualidade de farinha de trigo. Dissertação (Mestrado em Inovações Tecnológicas) Universidade Tecnológica Federal do Paraná, Campo Mourão, 2020.

DELWICHE, S. R.; MEKWATANAKARN, W.; WANG, C. Y. Soluble solids and simple sugars measurement in intact mango using near infrared spectroscopy. HortTechnology, v.18 (3), p.325-544, 2008.

DINIZ, P.H.G. Determinação do teor de NaCl, Glicose e KCl em medicamentos injetáveis por fotometria usando exploração do efeito Schlieren em sistemas FIA e PLS1. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa. p. 89, 2010.

FERREIRA, M.M.; ANTUNES, A.M.; MELGO, M.S.; VOLPE, P.L.O. Quimiometria I: Calibração multivariada, um tutorial. Química Nova, v. 22, n. 5, 1999.

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

FERREIRA, M. M. C. Quimiometria - conceitos, métodos e aplicações. Campinas, SP: Editora da Unicamp, 2015.

FISTER, I.; FISTER, I.; YANG, X.-S.; BREST, J. A comprehensive review of firefly algorithms. Swarm and Evolutionary Computation, v. 13, p. 34–46, 2013.

FREITAS, B.S.M. Estudo da caracterização e qualidade físicas e químicas do fruto de cajá (Spondias Mombin L.), e aproveitamento da polpa. Dissertação (Mestrado em Tecnologia em Alimentos) Instituto Federal de Ciência e Tecnologia Goiano, Rio Verde- Goiás, 2017.

FEUDALE, R.N.; WOODY, N.A.; TAN, H.; MYLES, A.J.; BROWN, S.D.; FERRÉ, J. Transfer of multivariate calibration models: a review. Chemometrics and Intelligent Laboratory Systems, v. 64, p. 181– 192, 2002.

GALDINO, I. K. C. P. O.; SALLES, H. O. S.; SANTOS, K. M. O.; VERAS, G.; BURITI, F. C. A. Proximate composition determination in goat cheese whey by near infrared spectroscopy (NIRS). PeerJ, DOI 10.7717/peerj.8619. 2020.

GELADI, P.; KOWALSKI, B.R. Partial Least-Squares regression – A Tutorial. Analytica Chimica Acta, v. 185, p. 1 – 17, 1986.

GEMPERLINE, p. O Practical Guide To Chemometrics. Boca Raton: Taylor & Francis, 2006. Disponível em: Practical Guide To Chemometrics.

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

GUO, Z.; WANG, M.; AGYEKUM, A. A.; WU, J.; CHEN, Q.; ZUO, M.; EL-SEEDI, H. R.; TAO, F.; SHI, J.; OUVANG, Q.; ZOU, X. Quantitative detection of apple watercore and soluble solids content by near infrared transmittance spectroscopy. Journal of Food Engineering, v.279, 2020.

GUEDES, J.A.C., Estudo do perfil metabolbrômico de folhas de cajazeira, umbuzeiro e abacaxizeiro e sua correlação com potencial atividade anticâncer por meio de análise multivariada. Tese (Doutorado em Química) Universidade Federal do Ceará, Fortaleza, 2018.

HARUNA, S. A.; L, H.; ZAREEF, M.; HASSAN, M. M.; ARSLAN, M.; GENG, W.; WEI, W.; DANDAGO, M. A.; ADADE, Y. S. S.; CHEN, Q. Application of NIR spectroscopy for rapid quantification of acid and peroxide in crude peanut oil coupled multivariate analysis. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy v. 267, 2022.

HASHIMOTO, J. C. Caracterização de amêndoas de cacau produzidas em diferentes estados brasileiros e aplicação de espectroscopia no infravermelho próximo e quimiometria como alternativa para o controle de qualidade. 2015. Tese (Doutorado em Ciência de Alimentos) – Universidade Estadual de Campinas, Campinas, 2015.

HUAN, K.; CHEN, X.; SONG, X.; DONG, W. Variable selection in near-infrared spectra: Application to quantitative non-destructive determination of protein content in wheat. Infrared Physics & Technology, v. 119, 2021.

KARASAKAL, A. Determination of Major, Minor, and Toxic Elements in Tropical Fruits by ICP-OES After Different Microwave Acid Digestion Methods. Food Analytical Methods, October 2020.

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

LÓPEZ-LÓPEZ, J.A., CORCHADO, C.B., MÁNUEL, M.P., BELLIDO, E.E. A simple and economical spectrofluorimetric alternative for Al routine analysis in seafood. Talanta, v. 182, p.210–217, 2018.

MACHADO, H.; NAGEM, T. J.; PETERS, V. M.; FONSECA, C. S.; OLIVEIRA, T. T. Flavonóides e seu potencial terapêutico. Boletim do Centro de Biologia da Reprodução, Juiz de Fora, v. 27, n. 1/2, p. 33-39, 2008.

MACIEL, G.D. Sistematização do controle de qualidade dos sucos tropicais via software computacional. Dissertação (Mestrado em Engenharia de Produção) Universidade Federal do Amazonas, Manaus, 2020.

MARTINS, M.N.; GALVÃO, R. K.H.; PIMENTEL, M.F. Multivariate Calibration Transfer Employing Variable Selection and Subagging. J. Braz. Chem. Soc., Vol. 21, n. 1, p.127-134, 2010.

MORAIS, F. P.; COSTA, R. C.; MORAIS, C. L. M.; MEDEIROS, F. G. M.; FERNADES, T. R. N.; HOSKIN, R. T.; LIMA, K. M. G. Estimation of Ascorbic Acid in Intact Acerola (Malpighia emarginata DC) Fruit by NIRS and Chemometric Analysis Horticulturae, v. 5, 2019.

NOGUEIRA, A.M.P.; FIGUEIRA, R.; IMAIZUMI, V.M.; FILHO, W.G.V. Avaliação físico-química e legislação brasileira de polpas, sucos tropicais e néctares de goiaba comerciais. Energia na Agricultura, v.35, n. 1, p. 136-142, Botucatu,2020.

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

OLIVEIRA, V.P., ESPECSHIT, A.C.R., PELUZIO, M.C.G. Flavonoides e doenças cardiovasculares: ação antioxidante. Rev Med Minas Gerais, v.16, p.234-8, 2006.

OLIVEIRA, A.C.S. Determinação de polifenóis totais em vinhos por espectrofluorimetria. TCC (Bacharelado em Química Industrial) – Universidade Federal Fluminense, Niterói, p.49. 2016.

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

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

PEREIRA, D.V.C., Composição fenólica e atividade antioxidante de cajá (Spondias mombin) nativo no Distrito Federal. Trabalho de conclusão de curso (Bacharel em Farmácia) Faculdade de Ceilândia, Brasília, 2017.

PEREIRA, E.V.S.; FERNANDES, D.D.S.; ARAÚJO, M. C. U.; DINIZA, P. H. G.; MACIEL, M. I.S. Simultaneous determination of goat milk adulteration with cow milk and their fat and protein contents using NIR spectroscopy and PLS algorithms. LWT - Food Science and Technology, p. 127, 2020.

POPPI, R.J.; SENA, M.M.; FRIGHETTO, R.T.S.; VALARINI, P.J. Avaliação do uso de métodos quimiométricos em análise de solos. Química Nova, v. 23, n. 4, 2000.

PRIETO, N., PAWLUCZYK, O., DUGAN, M. E. R., AALHUS, J.L. A Review of the Principles and Applications of Near-Infrared Spectroscopy to Characterize Meat, Fat, and Meat Products. Journals Applied Spectroscopy, p. 1- 24, 2017.

RAPHAEL, B.; SMITH, I. F. C. A direct stochastic algorithm for global search. Applied Mathematics and computation, v. 146, n. 2-3, p. 729-758, 2003.

RIBEIRO, F.A.D.L., FERREIRA, M.M.C., MORANO, S.C., DA SILVA, L.R. SCHNEIDER, R.P. Validation Spreadsheet: A New Tool For Estimating The Analytical Figures Of Merit For The Validation Of Univariate Methods [planilha De Validação: Uma Nova Ferramenta Para Estimar Figuras De Mérito Na Validação De Métodos Analíticos Univariados]. Quimica Nova, v. 31, n. 1, p. 164 - 171, 2008.

RAPHAEL, B.; SMITH, I. F. C. A direct stochastic algorithm for global search. Applied Mathematics and computation, v. 146, n. 2-3, p. 729-758, 2003.

SILVA, C. E. F., ABUD, A. K. S. Tropical Fruit Pulps: Processing, Product Standardization and Main Control Parameters for Quality Assurance. Braz. Arch. Biol. Technol. v.60, Jan/Dec 2017.

Silva, T.L.L.; Silva, E.P.; Asquieri, E.R.; Vieira, E.C.S.; Silva, J.S.; Silva, F.A.; Damiani, C. Physicochemical characterization and behavior of biocompounds of caja-manga fruit (*Spondias mombin* L.). Food Sci. Technol., v. 38, p. 399-406, 2018.

SUBEDI, P. P.; WALSH, K. B.; & OWENS, G. Prediction of mango eating quality at harvest using short-wave near infrared spectrometry. Postharvest Biology and Technology, 43, 326–334, 2007.

UNDERHILL, S. J. R. FRUITS OF TROPICAL CLIMATES | Commercial and Dietary Importance. Encyclopedia of Food Sciences and Nutrition, 2780–2785, 2003.

VALDERRAMA, P., BRAGA, J; W. B. e POPPI, R. J.Estado da arte de figuras de mérito em calibração multivariada. Química Nova, v. 32, n. 5, pp. 1278-1287. 2009.

VIDAL, A. P., Fruticultura na área de atuação do BNB: produção e mercado. Caderno setorial ETENE. Ano 4, n. 84, Junho, 2019.

VIEIRA, V. M., SOUSA, M. S. B., FILHO, J.M., LIMA, A. Fenólicos totais e capacidade antioxidante in vitro de polpas de frutos tropicais. Rev. Bras. Frutic., Jaboticabal - SP, v. 33, n. 3, p. 888-897, setembro, 2011.

WANG, Z.; WU, Q.; KAMRUZZAMAN, M. Portable NIR spectroscopy and PLS based variable selection for adulteration detection in quinoa flour. Food Control v. 138, 2022.

WILLIAMS, P., & NORRIS, K. H. Variable affecting near infrared spectroscopic analysis. In: P. Williams & K. H. Norris (Eds.). Near infrared technology in the agriculture and food industries. 2nd ed., p.171–185. St Paul: The American Association of Cereal Chemists, 2001.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, v. 58, p. 109 – 130, 2001.

XIAOBO Z., JIEWEN, Z, POVEY M.J. W, HOLMES M., HANPIN M. Variables selection methods in near-infrared spectroscopy.Analytica Chimica Acta. v. 667, p. 14-32, 2010.

YANG, X.F., GUO, X.Q., ZHAO, Y.B. Novel spectrofluorimetric method for the determination of sulfite with rhodamine B hydrazide in a micellar médium. Analytica Chimica Acta v.456, p.121–128, 2002.

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

ZIMMER, J.; ANZANELLO, M.J. Um novo método para seleção de variáveis preditivas com base em índices de importância. Production, v. 24, n. 1, p. 84 – 93, 2014.

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

ZHU, M.; LONG, Y.; CHEN, Y.; HUANG, Y.; TANG, L.; GAN, B.; YU, Q.; XIE, J. Fast determination of lipid and protein content in green coffee beans from different origins using NIR spectroscopy and chemometrics. Journal of Food Composition and Analysis, v.102, 2021.