UNIVERSIDADE ESTADUAL PAULISTA – UNESP
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CÂMPUS DE JABOTICABAL

CLASSIFICAÇÃO DE MAMOEIROS QUANTO AO TIPO SEXUAL E CULTIVARES USANDO A ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO, ANÁLISES MULTIVARIADAS E MACHINE LEARNING

Thiago Feliph Silva Fernandes
Engenheiro Agrônomo

2021

UNIVERSIDADE ESTADUAL PAULISTA – UNESP
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CÂMPUS DE JABOTICABAL

CLASSIFICAÇÃO DE MAMOEIROS QUANTO AO TIPO SEXUAL E CULTIVARES USANDO A ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO, ANÁLISES MULTIVARIADAS E MACHINE LEARNING

Thiago Feliph Silva Fernandes

Orientador: Prof. Dr. Gustavo Henrique de Almeida Teixeira

Dissertação apresentada à Faculdade de Ciências Agrarias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Mestre em Agronomia (Produção Vegetal).

2021

Fernandes, Thiago Feliph Silva
F363c Classificação de mamoeiros quanto ao tipo sexual e cultivares usando a espectroscopia no infravermelho próximo, análises multivariadas e machine learning / Thiago Feliph Silva Fernandes. --Jaboticabal, 2021 83 p.: il., tabs., fotos

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Gustavo Henrique de Almeida Teixeira

1. Carica papaya L.. 2. Espectroscopia NIR. 3. Sexagem.
4. Quimiometria. 5. Flores.. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: CLASSIFICAÇÃO DE MAMOEIROS QUANTO AO TIPO SEXUAL E CULTIVARES USANDO A ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO, ANÁLISES MULTIVARIADAS E MACHINE LEARNING

AUTOR: THIAGO FELIPH SILVA FERNANDES

ORIENTADOR: GUSTAVO HENRIQUE DE ALMEIDA TEIXEIRA

Aprovado como parte das exigências para obtenção do Título de Mestre em AGRONOMIA (PRODUÇÃO VEGETAL), pela Comissão Examinadora:

Prof. Dr. GUSTAVO HENRIQUE DE ALMEIDA TEIXEIRA (Participação Virtual)
Departamento de Ciências da Produção Agrícola (Produção Vegetal) / FCAV / UNESP - Jaboticabal

Prof. Dr. LUIS CARLOS CUNHA JÚNIOR (Participação Virtual)
Universidade Federal de Goiás/UFG-Campus Samambaia / Goiânia/GO

Prof. Dr. GLAUCO DE SOUZA ROLIM (Participação Virtual)
Departamento de Engenharia e Ciências Exatas (DCEx) / FCAV / UNESP - Jaboticabal

Jaboticabal, 28 de julho de 2021

DADOS CURRICULARES DO AUTOR

THIAGO FELIPH SILVA FERNANDES, nasceu em Castanhal, Pará (PA). Possui graduação em Agronomia, habilitação Engenheiro Agrônomo, pela Universidade Federal Rural da Amazônia (UFRA), Campus Capitão Poço - PA. Durante a graduação foi estagiário voluntário na disciplina de fruticultura no período de 2014 a 2015, realizando pesquisas nas áreas de tecnologia e produção de sementes e mudas de frutíferas nativas e exóticas. Além disso, participou como estagiário voluntário na Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) Amazônia Oriental, no período de 2015 a 2018, auxiliando nas pesquisas de seleção de genótipos de *Citrus* sp. para a região citrícola do Nordeste Paraense. Foi ainda membro grupo de pesquisa Interação Solo-Planta-Atmosfera na Amazônia (ISPAAm) no ano de 2017. Foi bolsista de Iniciação Cientifica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) em parceria com EMBRAPA Amazônia Oriental (cota 2018-2019), auxiliando nas avaliações morfológicas e agronômicas de 12 progênies de *Citrus* sp. para fins de melhoramento genético, bem como na qualidade físico-química dos frutos das mesmas progênies. Em 2019 ingressou no curso de Mestrado acadêmico do Programa de Pós-Graduação em Agronomia (Produção Vegetal), da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), Câmpus de Jaboticabal, sob orientação do Prof. Dr. Gustavo Henrique de Almeida Teixeira. Entre 2019 e 2021 desenvolveu pesquisas com o uso da espectroscopia no infravermelho próximo (NIR) visando a classificação de mamoeiros quanto ao tipo sexual e diferentes cultivares. Recentemente foi aprovado no curso de Doutorado do referido Programa de Pós-graduação.

“Eu sou parte de uma equipe.
Então, quando venço, **não sou eu apenas quem vence**.
De certa forma termino o trabalho de um **grupo enorme de pessoas**.”
Ayrton Senna

A Deus, pelo dom da vida.

À minha família, pelo auxílio e esforço.

Aos meus amigos e professores (principalmente aos da UNESP – Jaboticabal) que tanto contribuíram para minha trajetória acadêmica em especial, ao meu orientador e amigo, Professor Dr. Gustavo Teixeira, peça fundamental na minha formação acadêmica e intelectual.

Dedico.

AGRADECIMENTOS

A Deus, Senhor de todas as coisas e destinos, que guia minha trajetória de vida, inspira meu trabalho e não me deixa fraquejar. A ele toda a honra e glória.

À minha mãe Ana Maria N. Silva, por ser minha base e não medir esforços para que eu pudesse realizar meus sonhos, em tantas dificuldades que surgiram no caminho. Aos demais familiares, pois sem o esforço deles, não estaria concluindo mais uma etapa na minha vida.

A todos os professores que cooperaram para minha trajetória acadêmica, contribuindo com a minha formação intelectual e pessoal que, apesar de ser uma profissão importantíssima para a construção de uma sociedade melhor, é tão desvalorizada.

Agradeço ao meu orientador Prof. Dr. Gustavo Henrique de Almeida Teixeira pelos ensinamentos, oportunidades, paciência e preocupação com minha formação

À instituição de ensino Universidade Estadual Paulista (UNESP), Campus Jaboticabal, pela oportunidade de contribuir para minha formação profissional, concedida por meio da realização deste curso.

A todo o corpo docente da instituição UNESP, que repassou um pouco dos seus conhecimentos para agregação da minha formação acadêmica, em especial aos professores que tive durante o mestrado, como Prof. Arthur Bernardes, Prof. Pedro Luiz, Profa. Mara Cristina, Prof. Cristiano Zerbato e Prof. Rouverson Silva que compartilharam seus conhecimentos em disciplinas cursadas durante o mestrado

Aos amigos (a), funcionários do Ripado de Fruticultura, em especial aos funcionários João Brito, Claudemir dos Santos e ao técnico agrícola Adenilson Aparecido Servidone, pois sem eles o desenvolvimento da pesquisa teria encontrado maiores dificuldades.

À Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro - Código Financeiro 148354/2019-0.

Aos amigos: Alex Sanches, Maryelle Barros, Kolima Peña, Jonatan, Antonio Maricélio, Elbys Bastos, Lucas Soares e Valquíria Soares pelos momentos compartilhados com muita alegria.

A todas as pessoas que direta ou indiretamente contribuíram para a realização desta etapa na minha vida acadêmica e pessoal.

Meus sinceros agradecimentos,

Thiago Feliph S. Fernandes.

SUMÁRIO

CLASSIFICAÇÃO DE MAMOEIROS QUANTO AO TIPO SEXUAL E CULTIVARES USANDO A ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO, ANÁLISES MULTIVARIADAS E MARCHINE LEARNING

RESUMO

A sexagem do mamoeiro (*Carica papaya* L.) é uma das práticas hortícolas mais importantes para o sucesso desta cultura, pois esta espécie apresentar flores femininas, hermafroditas e masculinas. Todavia, apenas as plantas hermafroditas produzem frutos com o formato exigido pelo mercado, ou seja, frutos mais alongados e com polpa carnosa. Desta forma, é necessário o plantio de pelo menos três mudas por cova para se garantir a presença de plantas hermafroditas, o que acarreta um maior custo para o produtor. Assim, o objetivo geral deste projeto foi verificar a possibilidade de se utilizar a espectroscopia no infravermelho próximo (NIR) como um método não destrutivo para a sexagem e identificação de cultivares de mamoeiros. Foram utilizados as cultivares 'T2', 'Formosa' e 'Calimosa', do grupo Formosa, e 'THB' e 'Ouro', do grupo Solo. Os espectros NIR foram coletados nas sementes e nas folhas dos respectivos *seedlings*. Os resultados para a classificação do tipo sexual dos mamoeiros utilizando as sementes resultou em um valor de F-score de 0,81 para o grupo de validação externa ao se utilizar a análise de componentes principais e análise discriminante quadrática (PCA-QDA). Para as folhas dos *seedlings* o F-score foi 0,79 usando a PCA e análise discriminante linear (PCA-LDA). Em relação aos resultados para a classificação das cultivares, ao se utilizar as sementes foi possível obter valores de F-score apenas para as cultivares 'THB' (0,85) e 'Ouro' (0,77) no grupo de validação externa com o emprego da regressão por mínimos quadrados parciais e análise discriminante (PLS-DA). Para as demais cultivares não foi possível obter modelos de classificação robustos utilizando tanto as sementes quanto as folhas. Conclui-se que é possível utilizar a espectroscopia NIR associada às técnicas multivariadas como método não destrutivo para a sexagem e classificação de cultivares de mamoeiro utilizando tanto as sementes quanto as folhas dos *seedlings.*

Palavras-chave: *Carica papaya* L.; espectroscopia NIR; sexagem; quimiometria; flores.

CLASSIFICATION OF PAPAYA PLANTS BY SEX AND CULTIVAR USIND NEAR INFRARED SPECTROSCOPY, MULTIVARIATE ANALYSIS AND MACHINE LEARNING

ABSTRACT

Papaya plants can have typical pistillate or female flowers, hermaphrodite flowers and typical staminate or male flowers, but hermaphrodite plants produce fruit with the shape required by the market, that is, more elongated fruit with fleshy pulp. Therefore, it is necessary to plant at least three seedlings per hole to ensure the presence of hermaphrodite plants, which entails a higher cost for the producer. Thus, the general objective of this study was to verify the possibility of using near infrared spectroscopy (NIR) as a non-destructive method for sexing and identification of papaya cultivars. The cultivars ‘T2’, ‘Formosa’ and ‘Calimosa’, from the Formosa group, and ‘THB’ and ‘Ouro’, from the Solo group, were used. NIR spectra were collected from the seeds and leaves of the respective seedlings. The results for the classification of the sexual type of papaya trees using the seeds resulted in a F-score value of 0.81 for the external validation group applying principal component analysis and quadratic discriminant analysis (PCA-QDA). For the leaves of the seedling, this value 0.79 using PCA and linear discriminant analysis (PCA-LDA). Regarding the results for the cultivar classification, when seeds were used it was possible to obtain F-score values just for the cultivars 'THB' (0.85) and 'Ouro' (0.77) in the external validation group applying partial least squares regression and discriminant analysis (PLS-DA). For the other cultivars it was not possible to obtain good classification models using both seeds and leaves. It is possible to use NIR spectroscopy associated with multivariate techniques as a non-destructive method for sexing and classification of papaya cultivars using both seeds and leaves of seedlings. However, more studies are needed to improve the performance of the developed models.

Keywords: *Carica papaya* L.; NIR spectroscopy, sexing; fraud; flowers.

LISTA DE TABELAS

CAPÍTULO 2 - Determinação sexual em sementes e folhas de mamoeiros utilizando espectroscopia no infravermelho próximo combinadas a técnicas multivariadas e machine learning

CAPÍTULO 3 - Classificação de cultivares de mamoeiros por espectroscopia no infravermelho próximo combinada à PLS-DA e machine learning

LISTA DE FIGURAS

CAPÍTULO 2 - Determinação sexual em sementes e folhas de mamoeiros utilizando espectroscopia no infravermelho próximo combinadas a técnicas multivariadas e machine learning

CAPÍTULO 3 - Classificação de cultivares de mamoeiros por espectroscopia no infravermelho próximo combinada à PLS-DA e machine learning

CAPÍTULO 1 - Considerações gerais

1. Introdução

O Brasil é um dos principais produtores mundiais de frutas tropicais. Dentre as espécies de frutíferas tropicais cultivadas no país pode-se destacar o mamoeiro (*Carica papaya* L.), que apresenta grande importância social e econômica para o agronegócio brasileiro. De acordo com a *Food and Agriculture Organization* (Faostat, 2021a), o Brasil é o terceiro maior produtor mundial de mamão (1.400.00 toneladas métricas), com 10,72% na produção mundial. Esta produção ocorreu em uma área cultivada de 26.526 hectares (Embrapa, 2019). Em 2018, foram exportadas 42.669,06 toneladas, o que gerou uma renda em torno de US$ 50.061.420,00 (Mapa, 2019).

Em função da grande incidência de doenças, o cultivo do mamoeiro requer renovações frequentes dos pomares (2,5 a 4 anos), demandando constantes investimentos, em especial na aquisição de sementes e mudas (Silva et al., 2016; Oliveira et al., 1994). Estes materiais propagativos são geralmente obtidos via propagação sexuada em função de sua maior eficiência econômica (Boas, 2019; Farias et al., 2009; Simão, 1998). Todavia, o mamoeiro apresenta três classes de plantas, ou seja, monóicas, dióicas e monóclinas (Ming et al., 2007a; Storey, 1953; Hofmeyr, 1941) e o cruzamento entre estes tipos gera sementes com diferentes tipos sexuais (Marin e Gomes, 1986).

As plantas dos diferentes tipos sexuais produzirão frutos com características distintas, ou seja, as plantas hermafroditas irão produzir frutos de formato alongado, piriforme, com pequena cavidade interna e maior valor comercial (Dantas e Castro Neto, 2000). As plantas femininas irão produzir frutos de formato arredondado ou ligeiramente ovalado, cuja cavidade interna é grande, em relação à espessura da polpa, e necessitam da fecundação por pólen de plantas masculinas ou hermafroditas para produzirem frutos (Costa e Pacova, 2003). Assim, as plantas de mamoeiro podem apresentar flores pistiladas ou femininas típicas, flores hermafroditas e flores estaminadas ou masculinas típicas (Marin et al., 1995; Oliveira et al., 1994). Desta forma, visando garantir o plantio de plantas hermafroditas que

produzem frutos com o formato dos frutos exigidos pelo mercado consumidor, ou seja, frutos mais alongados e com polpa carnosa, é necessário o plantio de pelo menos três mudas por cova, o que acarreta um maior custo de produção (Costa e Pacova, 2003; Simão, 1998).

Em função da importância dos tipos sexuais para a cultura do mamoeiro, vários métodos foram desenvolvidos para identificar os tipos sexuais precocemente, na tentativa de diminuir o número de sementes e/ou mudas a serem utilizados. Desta forma, pode-se citar os métodos citológicos (Datta, 1971), por biomarcadores (Jindal e Singh, 1976), de determinação dos compostos fenólicos (Parasnis et al., 1999), e as ferramentas biomoleculares (Parasnis et al., 2000). Apesar da precisão, estes métodos são dispendiosos e requerem laboratórios especializados, dificultando a sua aplicação em condições de cultivo.

Similarmente, as diferentes cultivares de mamoeiro são comumente caracterizadas por meio de descritores morfológicos (Xu et al., 2009), porém este método é considerado limitado para a cultura do mamoeiro em função da baixa diversidade genética desta espécie, o que dificulta a identificação fenotípica das plantas (Santana et al., 2004). Desta forma, métodos mais acurados de diferenciação das cultivares de mamoeiros vem sendo empregados. Por exemplo, o uso de marcadores moleculares, tais como *Random Amplification of Polymorphic DNA* (RAPD) e microssatélites (*Inter Simple Sequence Repeats* – ISSR e *Simple Sequence Repeats* - SSR) foram relatados por Degel Barbosa et al. (2011) e Ming et al. (2007a), respectivamente. Contudo, estes métodos são destrutivos, geram resíduos químicos, demandam tempo e necessitam de laboratórios especializados para sua execução.

Neste sentido, métodos mais simples e eficazes de seleção de mamoeiros de acordo com o tipo sexual e cultivares, nas sementes e/ou mudas, poderiam minimizar os problemas decorrentes da sexagem nos plantios comerciais. A espectroscopia no infravermelho próximo (NIR) por ser uma técnica analítica rápida e não destrutiva poderia ser utilizada para estas finalidades, uma vez que os espectros NIR são fontes de informações para a identificação qualitativa de amostras (Pasquini, 2003).

Desta forma, a espectroscopia NIR foi utilizada com sucesso para a sexagem de diferentes materiais biológicos, tais como larvas do bicho-da-seda (Tao et al., 2019), bezerros e vacas (O'Neill et al., 2017), em outros. Similarmente, a espectroscopia NIR tem sido utilizada para a diferenciação de cultivares de soja, arroz e feijão (Singh et al., 2018), de cultivares de milho doce (Qiu et al., 2019) e nozes de macadâmia (Rahman et al., 2021), porém não foram encontrados trabalhos relacionados à sexagem e à classificação de cultivares de mamoeiros.

Portanto, o objetivo geral deste trabalho foi verificar a possibilidade de se utilizar a espectroscopia NIR como um método não destrutivo para a sexagem de mamoeiros e à classificação de cultivares, por objetivos específicos: *i*. verificar a possibilidade de se utilizar as sementes e folhas dos *seedlings* para sexagem de mamoeiros e *ii*. desenvolver modelos matemáticos visando a predição do tipo sexual dos mamoeiros e à classificação de cultivares.

2. Revisão de literatura

2.1 O Mamoeiro

O mamoeiro (*Carica papaya* L.) é uma angiosperma que pertencente à classe Cicotyledoneae, subclasse Archiclamydeage, ordem Violales, família Caricaceae e gênero *Carica*, sendo a única espécie de interesse comercial deste gênero (Marin et al., 1995). Além do gênero *Carica*, há mais quatro gêneros na família Caricaceae, três originários da América tropical (*Jarilla*, *Jacaratia* e *Vasconcella*) e um da África equatorial (*Cylicomorpha*), Paull e Duarte (2011).

Originário do continente americano, o mamoeiro foi descoberto pelos espanhóis no Panamá e sua provável introdução no Brasil ocorreu em 1587 (Serrano e Cattaneo, 2010). Segundo Paull e Duarte (2011), a dispersão inicial desta espécie na América Central e do Sul foi ajudada pela abundância de sementes nos frutos e a longa viabilidade das sementes. No século 18 os navegadores e botânicos levaram sementes do Caribe para Malaca na Malásia e depois para a Índia (Storey, 1941). De Malaca o mamoeiro continuou a ser introduzido em diversas localidades da Ásia e do Pacífico Sul. A introdução do mamoeiro no estado do Havaí no Estados Unidos da América (EUA) foi creditada ao explorador e horticultor espanhol Don

Francisco Marín a partir de sementes oriundas das Ilhas Marquesas no início de 1800 (Paull e Duarte, 2011).

Atualmente o mamoeiro é cultivado em todos os países tropicais e subtropicais do globo e seus frutos são apreciados tanto para o consumo *in natura* quanto para a elaboração de variados produtos como doces, geléias, iogurtes e demais subprodutos da industrialização (Salomão et al., 2007).

Em função de sua origem, o mamoeiro se desenvolve melhor em locais cujas temperaturas médias anuais se situam na faixa de 21 ºC a 33ºC, com precipitações pluviométricas de 1.500 mm anuais bem distribuídos e com no mínimo 66% de umidade relativa (Paull e Duarte, 2011). A produção de fruto no Brasil ocorre quase o ano todo, sendo a menor safra nos meses de junho a agosto.

O mamoeiro é uma das espécies frutíferas tropicais de maior interesse econômico e nutricional (Orrillo et al., 2019). Em 2019, foram produzidas 13,73 milhões de toneladas (t) deste fruto em todo o mundo (Faostat, 2021b). Apesar de ter se originado na América tropical, atualmente a Índia é o maior produtor mundial de mamões (6,05 milhões de t), seguido da República Dominicana (1,17 milhões de t) e do Brasil (1,16 milhões de t), (Faostat, 2021b). Dentre os estados brasileiros produtores de mamão, destaca-se a Bahia, Espírito Santo, Ceará, Rio Grande do Norte, Minas Gerais e Paraíba (Embrapa, 2019).

Em relação ao histórico da cultura do mamoeiro no Brasil, principalmente no tocante à qualidade genética, esta foi marcada pela ocorrência da doença denominada mosaico do mamoeiro que é causada pelo vírus *Papaya ringspot virus* (PRSV-p), cujos primeiros relatos ocorreram na região de Monte Alto – SP, cidade que era considerada a capital do mamão (Ruggiero et al., 2011).

No processo de controle da doença, a cultura do mamoeiro migrou para outras regiões agrícolas de São Paulo até o seu desaparecimento do estado em meados de 1970. No entanto a cultura migrou para outras regiões do país, como o Nordeste do Pará, o extremo Sul da Bahia e Norte do Espírito Santo nos anos de 1975/1976. Nessas regiões foram desenvolvidas técnicas de prevenção contra o mosaico do mamoeiro o que permitiu a estabilização da doença. A partir da década de 1980 a migração da cultura foi impulsionada por motivos mais comerciais do que fitossanitários. Outros locais de produção como Inhumas – GO, estado do Ceará,

Janaúba – MG e estado do Rio Grande do Norte, surgiram assim como a expansão da cultura destinada ao mercado internacional no final da década de 90 (Durigan, 2014).

A base genética da cultura do mamoeiro no Brasil é restrita, sendo limitada o número de cultivares plantadas pelos produtores. Reis et al., (2015), destacaram que as cultivares dos grupos Solo e Formosa como sendo as mais conhecidas no país. Estes ressaltaram ainda que dentre os híbridos do grupo Solo a cultivar 'Sunrise Solo' e 'improved Sunrise Solo cv 72/12' são popularmente conhecidas como mamão Papaia e Havaí. No tocante ao grupo Formosa, as cultivares 'Tainung n.1' e 'Tainung n.2' são híbridos reconhecidos pela alta qualidade de seus frutos e alta produtividade, que pode chegar a 100 toneladas por hectare (Serrano e Cattaneo, 2010).

2.2 Tipos de flores

O mamoeiro é uma espécie dióica cujas flores se desenvolvem em inflorescências do tipo cimosa ou simpodial (Paull e Duarte, 2011), porém estas são classificadas de maneiras diferenciadas devido à grande quantidade de formas de flores observadas no campo. Em função das flores se desenvolverem em três formas bem distintas, ou seja, flores pistiladas ou femininas, flores estaminadas ou masculinas e flores hermafroditas (Simão, 1998; Storey, 1969), os mamoeiros apresentam diferentes tipos sexuais, a saber:

Plantas femininas ou pistiladas. Estas produzem flores globosas, medindo de 3 a 4 cm de comprimento por 2 a 2,5 cm de diâmetro. As flores femininas apresentavam cinco pétalas carnosas livres ou ligeiramente unidas na base e cinco sépalas rudimentares. Possuem ovário grande, globoso ou cilíndrico, afunilando-se para o ápice, onde se inserem cinco estigmas sésseis em forma de leque. Estas flores não apresentavam estames, constituindo o tipo dióico verdadeiro (Oliveira et al., 2007; Storey, 1941).

Plantas masculinas ou estaminadas. Estas produzem flores dispostas em rácimos longos e em pêndulos atingindo mais de um metro de comprimento, sendo originadas das axilas das folhas superiores. As flores masculinas são uniformes e

esbranquiçadas, medindo de 2 a 2,5 cm de comprimento por 0,5 cm de diâmetro. No tubo da corola, na região próxima à base das pétalas, encontram-se 10 estames dispostos em duas séries de cinco. As pétalas são mais delicadas que as dos outros tipos. Estas flores apresentam pistilo rudimentar ou ausente e, quando presente, atinge metade do tubo da corola. As anteras são amarelas, rica em grãos de pólen (Dantas e Castro Neto, 2000; Storey, 1941)

Plantas hermafroditas. Estas plantas produzem flores em inflorescências com 6 a 12 cm de comprimento e cada rácimo constituí de duas a seis flores. As flores femininas são alongadas, com 4 a 5 cm de comprimento e 2 cm de diâmetro, de base tubular, que se abre em forma de taça e se alarga em cinco pétalas grossas, de cor creme-claro. As flores apresentam de cinco a 10 estames sésseis, que se encontravam na região tubular, próximos à base das pétalas do ovário. Estas apresentam ovário desenvolvido, porém com volume inferior ao da flor pistilada(Deputy et al., 2002; Storey, 1941).

As plantas hermafroditas podem apresentar muitas formas, tais como pentandra, intermediárias, enlongata e estéril, que podem resultar em frutos sem valor comercial, como é o caso dos frutos carpelóides e pentândricos (Storey, 2020, 1941).

2.3 Reprodução e tipos sexuais

Segundo Storey (2020), a variação do sexo no mamoeiro é um sistema intrigante, pois esta é uma espécie trióica por apresenta três classes flores, ou seja, monóicas, dióicas e monóclinas, sendo que o cruzamento entre estes tipos irá gerar sementes com diferentes tipos sexuais. A variação sexual tem despertado investigações por parte dos melhoristas devido aos problemas econômicos ocasionados pela segregação dos tipos sexuais (Parasnis et al., 1999).

Desta forma, por se tratar de uma espécie trióica, a variação do sexo no mamoeiro é controlada por fatores sexuais XY com dois cromossomos Y ligeiramente diferente, um MSY e um Yh hermafrodita, sendo que a combinação desse genes são transmitidas para seus parentais em proporções inesperadas (Liu et al., 2004). Além dessas proporções incertas, a recombinação em torno desses

genes com o tempo pode acumular mutações recessivas deletérias que ocasiona perdas de características de interesse agronômicos e econômicos desejáveis agronômica (Storey, 1938).

Acredita-se que esse processo de segregação sexual possa ser uma possível resposta das plantas de mamoeiros às mudanças fisiológicas, ocasionadas por fatores ambientes, em resposta a determinadas estações do ano (Hofmeyr, 1939). Contudo, é necessário se conhecer o gene responsável por expressar o tipo sexual quando a planta está sob estresse.

Embora, os recentes estudos no mapeamento genômico do mamoeiro indique que as sutis variações estruturais do fragmento de DNA masculino provocados pelo estresse ambiental tenha relação com a expressão sexual da planta (Liao et al., 2021).

Com base no processo de reversão de pequenos genes localizados na região cromossômica de um gene de três alelos, Storey (1953) propôs que o esse processo é o fator determinante para a expressão sexual, visto que é uma região dominada por genótipos homozigotos masculinos.

Em progênies com dois cromossomos X, Hamilton e Izuno (1967) relataram que plantas tinham pêndulos longos apresentavam algumas flores masculinas ou hermafroditas, apesar da maioria das flores caracterizarem femininas.

2.4 Herança do sexo

O caráter de herança sexual do mamoeiro é simples e controlado por fatores genéticos, sendo essa característica monogênica, com três alelos (Ming et al., 2007b). O mamoeiro possui três tipos de inflorescência que são tipicamente são dióicas, com flores unissexuais, ou seja, plantas masculinas e femininas, ou ginodióica, com flores bissexuais e unissexuais e hermafrodita e plantas femininas (Storey, 1953).

Assim como as formas sexuais são distintas, a herança sexual é repassada em proporções inesperadas para seus desentendes, visto que o alelo dominante masculino configura-se como fator letal (Liu et al., 2004). Dessa forma, a forte

influência que esse alelo exerce na supressão do *crossing-over*, pode afetar característica secundários como quantidade de flores (Ming et al., 2007b).

Baseado em cruzamento entre espécies de mamoeiro, Storey (1953) propôs que a determinação do sexo seria do tipo feminino (mm), masculino (M_1m) e hermafrodita (M_2m), sendo que as plantas ferminas são homozigotas para o alelo recessivo m e as plantas masculinas e hermafroditas heterozigotas para os alelos M_1 e M_2, respectivamente.

Dessa forma, a partir das combinações de cruzamentos entre os três tipos sexuais conclui-se que as plantas hermafroditas autofecundadas sempre se segregarão em hermafroditas e fêmeas na proporção 2:1, ou seja, 66% de plantas hermafroditas para 33% de femininas. Plantas femininas segregarão na proporção de 1:1 para plantas hermafroditas e femininas, ou seja, 50% hermafroditas e 50% femininas se forem fertilizadas com pólen de plantas hermafroditas ou na proporção de 1:1 de plantas masculinas e femininas quando fertilizadas com pólen de uma planta masculina (Storey, 1953).

2.5 Métodos de sexagem

Os mamoeiros são tradicionalmente sexados em função do fenótipo das plantas quando estas atingem a fase reprodutiva e emitem as primeiras flores, que ocorre de três a quatro meses após o plantio das mudas (Paull e Duarte, 2011). Para isso, é necessário o plantio de pelo menos três mudas por cova visando aumentar as chances de pelo menos uma seja hermafrodita (Costa e Pacova, 2003; Simão, 1998). Esta prática aumenta os custos de produção em função do número de sementes e/ou mudas a serem adquiridos, além da manutenção das plantas no campo até que a sexagem seja realizada (Dantas et al., 2013).

Neste sentido, muitos métodos foram desenvolvidos visando a reduzir os custos de produção da sexagem nos plantios comercias (Oliveira et al., 2007).

Dentre estes métodos pode-se citar o citogenético, em que os cromossomos do mamoeiro são analisados viando identificar as alterações cromossômicas relacionadas às características moleculares. Contudo, estudos citogenéticos com a

cultura do mamoeiro são limitados, pois os cromossomos apresentam muita semelhança e exigem laboratórios especializados (Bajpai e Singh, 2006).

Embora estudos citológicos sejam limitados, Datta (1971) relatou que com o uso da microscopia de campo claro em amostras de cinco linhagens diferentes de mamoeiro foi possível observar diferenças no comprimento nas contrições primarias dos cromossomos. Comportamento semelhante foi relatado por Araújo (2008).

O método do teste colorimétricos foi utilizado para diferenciar o tipo sexual em mamoeiros por Jindal e Singh (1976) ao extraírem os compostos fenólicos de folhas de plantas masculinas e femininas e também das folhas de mudas em fase de viveiro. Como isto foi possível discriminar em 80% as plantas femininas e em 60% as plantas masculinas com o teste Azul da Prússia. Enquanto no teste usando os compostos fenólicos totais a precisão foi de 86% e 77% para as plantas femininas e masculinas, respectivamente.

Em relação aos estudos empregando marcadores genéticos para identificar a região no DNA ligada a expressão sexual, Chaves-Bedoya et al. (2009) relataram que o uso de marcadores moleculares em mudas de mamoeiro para determinar precocemente o tipo sexual resultou na identificação de três marcadores polimórficos de DNA (*Random Amplified Polymorphic DNA* – RAPD*, Inter Simple Sequence Repeats* – ISSR e *Simple Sequence Repeats* - SSR). Desta forma, estes autores conseguiram diferenciar os tipos sexuais do mamoeiro com uso de dois marcadores específicos para plantas masculinas e hermafroditas e um terceiro para plantas femininas.

Apesar de todos os estudos direcionados à obtenção de métodos de determinação precoce em sementes e mudas de mamoeiro, estas metodologias têm aplicabilidade limitada para os produtores, pois estes são procedimentos demorados, onerosos e exigem laboratórios especializados. Desta forma, a sexagem realizada baseada no fenótipo dos mamoeiros no campo ainda é o métodos mais utilizado pelos produtores, principalmente em função dos custo serem menores quando comparados aos métodos anteriormente descritos (Jiménez, 2002).

2.6 Espectroscopia no infravermelho próximo (NIR)

A espectroscopia no infravermelho próximo com transformada de Fourier (FT-NIR) tem sido amplamente utilizada como uma ferramenta rápida, precisa e não destrutiva em vários campos de aplicação (Haq et al., 2018; Einax, 2007). Estudos recentes têm demonstrados a possibilidade dessa ferramenta detectar e identificação de qualidade de sementes soja (Ferrazza et al., 2020), identificação de cultivares de milho (Qiu et al., 2019), diferenciação sexual de plantas (Matias et al., 2020; Tormena et al., 2020) e animais (Gebru et al., 2018; O'Neill et al., 2017).

A espectroscopia NIR é um método analítico voltado para a interação da matéria com a energia eletromagnética (Vogel, 1981). Essa interação ocorre pelo estado de excitação que a energia provoca nas moléculas, como a vibracional, rotacional e translacional (Beć et al., 2021). Neste processo a energia oriunda de ondas eletromagnéticas ou fótons na faixa que compreende os comprimentos de onda de 750 nm a 2.500 nm ou equivalente número de ondas de 13.300 a 4.000 cm^{-1} e energia entre 160 kJ mol^{-1} e 48 kJ mol^{-1} é absorvida pelas moléculas de uma amostra, sendo convertida em energia vibracional e rotacional podendo apresentar movimentos harmônicos e combinações de nível energético para dois ou mais níveis de maior energia. A intensidade da absorção da radiação é dominadas por sobretons e bandas de combinação das transições vibracionais de moléculas orgânicas, principalmente em grupos funcionais -CH, -NH, -OH e -SH (Pasquini, 2018).

Esta interação foi descrita primeiramente em 1800 pelo cientista anglo-alemão, Frederick William Herschel, ao estudar o efeito da luz solar, destacadamente do efeito das diferentes cores do espectro visível, no aumento da temperatura. Em seus experimentos utilizando um prisma de vidro e termômetros com bulbos de cor pretas, Herschel constatou que a temperatura aumentava significativamente quando se movia os termômetros além da luz visível, fato esse intitulado por ele como região calorífica, e posteriormente, chamado de infravermelho (William Herschel, 1800). Todavia, as primeiras aplicações analíticas usando a espectroscopia NIR foram desenvolvidas após um século e meio deste descobrimento. Isto ocorreu em função de fatores limitantes como a falta de equipamentos analíticos adequados, da qualidade espectral na região NIR e a

dificuldade da se interpretar os espectros, principalmente, por inúmeras sobreposições nas bandas de absorção (Ciurczak et al., 2021).

Assim sendo, os trabalhos mais relevantes com a utilização da espectroscopia NIR foram obtidos a partir da década de 50, devido principalmente ao desenvolvimento da instrumentação eletrônica e ferramentas matemáticas para contornar a falta de seletividade da informação analítica espectral (Williams, 2019). Além disso, estudos pioneiros desenvolvidos e disseminados no meio acadêmico pelo pesquisador Karl Norris impulsionou a técnica (Ben-Gera and Norris, 1968).

Neste contexto, a espectroscopia NIR, por apresentar promissores resultados de forma rápida, não destrutiva, não invasiva e precisos, sendo aplicado em métodos analíticos de caracterização físicas-químicas, concentrações de compostos e qualidade de amostras de alimentos (Véstia et al., 2019; Sirisomboon, 2018), fármacos (Mishra et al., 2021; Pedersen et al., 2021), solos (Bao et al., 2021; Barthès e Chotte, 2021) e plantas (Eshkabilov et al., 2021; Lee et al., 2018; MingMei; et al., 2017).

Além destas aplicações, a espectroscopia NIR surge como uma possibilidade para a sexagem do mamoeiro, pois vários estudos tem demostrado a eficiência da espectroscopia NIR para a identificação sexual em diferentes materiais biológicos, tais como, larvas do bicho-da-seda (Tao et al., 2019), bezerros e vacas (O'Neill et al., 2017). No tocante à espécies vegetais, a espectroscopia NIR foi utilizada para a diferenciação sexual de *Phoenix dactylifera* L. (Khan et al., 2021), seleção de genitores sexuais e híbridos de *Urochloa decumbens* (Matias et al., 2020) e dimorfismo sexual em plantas de erva-mate (Tormena et al., 2020).

Quimiometria

A quimiometria se relaciona ao uso de ferramentas e métodos matemáticos, estatísticos e computacionais para a análise de dados químicos de natureza multivariada (Martens e Næs, 1984), sendo bastante utilizada em dados coletados no infravermelho próximo (Slutsky, 1998). Esta pode ser dividida em quatro principais vertentes: pré-processamento dos dados, análise exploratória dos dados, classificação supervisionada e calibração multivariada (Neto et al., 2006).

Neste sentido, o reconhecimento de tendências dos dados por meio de ferramentas computacionais (*Softwares*) permite uma melhor resolução dos dados. Desse modo, esse reconhecimento acontece a partir de um primeiro conjunto de dados que de forma categorizada é utilizado para a construção de modelos de classificação, com base em aferições de parâmetros relativos à cada espécie, com a finalidade de se verificar níveis de similaridade entre as amostras exploradas (Williams, 2019; Otto, 2016).

A calibração multivariada consiste nas propriedades matemáticas matriciais, sendo essa vertente como uma das mais importantes da quimiometria, pois com isso é possível estabelecer relações matemáticas baseadas em observações indiretas e/ou instrumentais para se estimar propriedades físicas e químicas de amostras transformadas em parâmetros numéricos (Souza e Poppi, 2012). Vale ressaltar que a calibração é dividida em três etapas, ou seja, a análise treinamento ou calibração e validação ou teste.

Ao se utilizar os espectros NIR, antes de se realizar a calibração, é necessário empregar o pré-processamento dos dados. Essa etapa consiste na eliminação ou redução de fontes de variação aleatória ou sistemáticas não desejáveis no sentido de maximizar a extração das informações das amostras que de fato interessam ao pesquisador (Pasquini, 2003).

Pré-processamento dos dados

Os procedimentos de pré-processamentos aplicados à espectroscopia NIR se relacionam com frequência à correção do espalhamento da luz. Esses procedimentos visam a redução ou eliminação de fontes de variação sistemáticas indesejáveis no conjunto de calibração (Pasquini, 2003). Dessa maneira, pode-se maximizar a extração de informações das amostras com qualidade e de acordo com o objetivo do estudo.

Para a correção do espalhamento da luz são utilizados com frequência a correção do espalhamento multiplicativo de luz (*multiplicative scatter correction* - MSC) e a variação normal padrão (*standard normal variate* - SNV). O métodos MSC é baseado na aplicação matemática de regressão em razão de um espectro de

referência que agrupa os dados espectrais brutos em torno da sua média espectral (Martens e Næs, 1984). Já o SNV não utiliza um espectro médio de referência, e sim um vetor de normalização por meio nas linhas das matrizes, no qual é subtraído um valor de todos os dados médios e, em seguida, divididos pelos respectivos desvios-padrões (Dhanoa et al., 1994).

Além destes, os procedimentos de alinhamento da linha de base são realizados com objetivo de reduzir os ruídos provocados pelo deslocamento da linha da base por meio da relação sinal/ruído espectrais, sendo o método de Savitzky-Golay o mais utilizado (Véras et al., 2012).

O alisamento por Savitzky-Golay acentua os picos espectrais pela correção do deslocamento da linha da base por sucessíveis ajustes polinomiais de primeira e segunda ordem, denominada de parte/janela do espectro. Portanto, a execução de uma regressão polinomial dos pontos (primeira e segunda ordem) para determinar um valor do ponto central (x) da janela, sendo este a ser o valor de referência para subavaliação (Savitzky and Golay, 1964).

Após do pré-processamento, a próxima etapa é a construção do modelo ou calibração multivariada. Essa etapa, corresponde as várias técnicas matemáticas utilizadas de forma simultaneamente ou não para análise do conjunto de dados, sendo esse um processo complexos, e por isso realizado por meio de aplicações de programas computacionais (Martens e Næs, 1984).

Por esse motivo, as ferramentas matemáticas multivariadas são selecionadas de acordo com os objetivos e hipóteses da pesquisa (Hair et al., 2009). Dentre essas ferramentas, as amplamente utilizada são: Análise de Componente Principal - Análise Discriminante Linear (PCA-LDA); Análise de Componente Principal - Análise Discriminante Quadrática (PCA-QDA); Análise de componentes principais - Máquina de vetores de suporte (PCA – SVM); Algoritmo de projeções sucessivas - Análise Discriminante Linear (SPA-LDA); Algoritmo de projeções sucessivas - Análise Discriminante Quadrática (SPA-QDA); Algoritmo de projeções sucessivas - Máquinas de vetor de suporte (SPA-SVM); Algoritmos genéticos - Análise Discriminante Linear (GA-LDA); Algoritmos genéticos - Análise Discriminante Quadrática (GA-QDA); Algoritmos genéticos - Máquinas de vetor de suporte (GA-SVM); Modelagem Soft Independente da Analogia de Classes (SIMCA); mínimos

quadrados parciais para análise discriminante (PLS-DA); Máquinas de vetor de suporte (SVM); Programa de modelagem Independente para analogia de classes (SIMCA) (Qiu et al., 2019; Balabin e Lomakina, 2011; Hair et al., 2009; Sjöström et al., 1986).

Análise de componentes principais (PCA)

A análise de componentes principais (PCA) é uma técnica de projeção dos dados multivariados com o objetivo de visualizar a estrutura, encontrar similaridades entre amostras, detectar amostras anômalas (*outliers*) e reduzir a dimensionalidade do conjunto de dados (Souza e Poppi, 2012). É baseada no princípio algébrico linear, entre dois grupos de dados (**X** e **Y**), em que **X** é a variável resposta e **Y** a variável depende, sendo que suas possíveis correlações for transformadas em um terceiro conjunto de dados mais simples (Neto et al., 2006).

Regressão por mínimos quadrados parciais (PLS)

A regressão por mínimos quadrados parciais (PLS) tem como objetivo criar uma relação linear entre os valores originais das variáveis medidas (**X**) e as variáveis de interesse (**Y**) para encontrar uma solução. Assim, a regressão de PLS pode encontrar os componentes principais (denominados de variáveis latentes) pela covariância entre os escores de **X** e **Y** determinados a partir da maximização da matriz **X** (Sjöström et al., 1986). Vale ressaltar que para se obter uma boa relação de explicação da variância do eixo **X** e a predição dependente **Y** é imprescindível que as variáveis latentes sejam harmônicas.

Dessa forma, a construção do modelo de regressão é baseada no número de variáveis latentes e calculando-se um limiar para que seja proporcionado um menor erro possível de predição (Zorzetti e Harynuk, 2011). Esses valores obtidos entre os valores de referência e previstos devem ser testados, mediante um conjunto de dados externos para avaliar a real capacidade preditiva do modelo (Souza et al., 2013).

Por fim, neste método todos os componentes presentes nas amostras podem ser desconhecidos e apresentar variáveis anômalas que ainda pode ser realizado, porém deste que os mesmos parâmetros e amostras sejam similares aos utilizados no modelo de calibração e validação (Lorber e Kowalski, 1988)

Análise discriminante (AD)

A análise discriminante é um método de para classificar novos parâmetros a partir de grupos existentes, sendo assim, é necessário o entendimento das características mais importantes das amostras. Como o número de observações em cada grupo é diferente, a probabilidade de ocorrência de um grupo pode ser maior do que a do outro grupo, portanto as regras de classificação levam em consideração a probabilidade prévia dessas ocorrências, que pode ser estimada em cada classe pelas proporções de observações (Hair et al., 2009).

Os métodos de análise discriminante podem ser divididos em: análise discriminante linear (LDA) e análise discriminante quadrática (QDA). Ambas partem do pressuposto que o conjunto de dados possui uma distribuição normal em cada classe especifica, sendo que na LDA a matriz de covariância de cada grupo é considera igual, enquanto na QDA essas matrizes podem ou não ser diferentes (Hair et al., 2009; Johnson e Wichern, 2007).

Neste sentindo, caso a quantidade de classes seja maior de dois, é necessário realizar um ajuste na extensão natural do discriminante linear de Fisher mediante a análise discriminante múltipla. A partir do modelo construído combinado com seus ajustes, ele pode ser utilizado para classificar amostras desconhecidas (Morais et al., 2019; Ballabio e Consonni, 2013).

Máquina de vetores de suporte (SVM)

A máquina de suporte de vetores (SVM) é um algoritmo computacional com grande aplicabilidade (Haykin, 2009). Criado incialmente para resolução de problemas de aprendizagem de maquinas, o seu uso foi ampliado para classificação

de padrões e calibração em métodos quimiométricas, principalmente pela sua robustez (Balabin e Lomakina, 2011).

A técnica SVM é uma ferramenta estatística baseada na classificação binária para resolver problemas não lineares. Sua construção é por meio de um hiperplano como superfície de decisão para maximização de suas margens de separação entre classes positivas e negativas, sendo assim possível caracterizar os mesmos dados de formar linear (Noble, 2006).

A sua utilização para análise de dados espectrais é principalmente pela flexibilidade de minimizar o risco de erros. Com a introdução de um coeficiente de determinação é possível limitar a classificação pela soma da taxa do erro de treinamento e padrões de separação previamente determinados (Otto, 2016).

Avaliação dos modelos

A avaliação da real eficácia do desempenho dos modelos desenvolvidos é a última etapa para então ser aplicado e recomendado para pesquisa. Assim sendo, os parâmetros que fornece a descrição da eficácia dos modelos é obtido, por exemplo, por meio da tabela de confusão (Hair et al., 2009).

Os parâmetros utilizados para a construção desta tabela são as taxas dos falsos positivos (FP), falso negativos (FN), verdadeiro positivos (VP), verdadeiros negativos (VN), sensibilidade, especificidade, precisão, eficiência, acurácia e F-escore (Benos et al., 2021)

Dessa formar, esses parâmetros medem a capacidade de predição de um dado modelo através de equações, onde seus parâmetros são obtidos através dos resultados encontrados para um dado conjunto de teste. Assim sendo, temos: a acurácia representa o número de amostras classificadas corretamente, para ambas as classes. A sensibilidade e especificidade mensuram a proporção de amostras positivas e negativas, aqui sendo consideradas para fins de classificação como pertencentes ao grupo de amostras das classes definidas. E por fim, o F-score que é o produto da média harmônica entre sensibilidade e especificidade que representando a acurácia real do modelo (Hair et al., 2009).

3. Referências

Araújo, F.S., 2008. Estudos citogenéticos e citométricos em mamoeiro (Carica papaya L.). Universidade Federal de Viçosa.

Bajpai, A., Singh, A.K., 2006. Meiotic behavior of Carica papaya L.: Spontaneous chromosome instability and elimination in important cvs. in North Indian conditions. Cytologia (Tokyo). 71, 131–136. https://doi.org/10.1508/cytologia.71.131

Balabin, R.M., Lomakina, E.I., 2011. Support vector machine regression (SVR/LS-SVM) - An alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst 136, 1703–1712. https://doi.org/10.1039/c0an00387e

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: Linear models. PLS-DA. Anal. Methods. https://doi.org/10.1039/c3ay40582f

Bao, N., Liu, S., Yang, T., Cao, Y., 2021. Characterization and prediction of soil organic matter content in reclaimed mine soil using visible and near-infrared diffuse spectroscopy. Arid L. Res. Manag. https://doi.org/10.1080/15324982.2020.1867935

Barthès, B.G., Chotte, J., 2021. Infrared spectroscopy approaches support soil organic carbon estimations to evaluate land degradation. L. Degrad. Dev. 32, 310–322. https://doi.org/10.1002/ldr.3718

Beć, K.B., Grabska, J., Huck, C.W., 2021. Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers. Chem. - A Eur. J. https://doi.org/10.1002/chem.202002838

Ben-Gera, I., Norris, K.H., 1968. Direct Spectrophotometric Determination of Fat and Moisture in Meat Products. J. Food Sci. 33, 64–67. https://doi.org/10.1111/j.1365-2621.1968.tb00885.x

Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D., Bochtis, D., 2021. Machine Learning in Agriculture: A Comprehensive Updated Review. Sensors 21, 3758. https://doi.org/10.3390/s21113758

Boas, R.V., 2019. Influência da seleção de sementes no desenvolvimento de plantas e na biologia floral do mamoeiro (Carica papaya L). TCC - ICA - Agron. 50.

Chaves-Bedoya, G., Pulido, M., Sánchez-Betancourt, E., Núñez, V., 2009. RAPD markers for sex identification in papaya (Carica papaya L.) in Colombia. Agron. Colomb. 27, 145–149.

Ciurczak, E., Igne, B., Jr, J.W., Burns, D., 2021. Handbook of near-infrared analysis.

Costa, A.F.S., Pacova, B.E. V., 2003. Caracterizaçãoo de cultivares, estratégias e

perspectivas do melhoramento genético do mamoeiro., in: A Cultura Do Mamão: Tecnologia e Produçãor. pp. 59–102.

Dantas, J.L.L., Castro Neto, M.D., 2000. Aspectos botânicos e fisiológicos. Mamão, produção Asp. técnicos. Brasília Embrapa Comun. para Transf. Tecnol. 11–14.

Dantas, J.L.L., Junghans, D.T., Lima, J.F. de, 2013. Mamão: o produtor pergunta, a Embrapa responde., 2ª rev. e atual. ed. Embrapa, Brasília, DF .

Datta, P.C., 1971. Chromosomal Biotypes of Carica Papaya Linn. Cytologia (Tokyo). 36, 555–562. https://doi.org/10.1508/cytologia.36.555

Degel Barbosa, C., Pio Viana, A., Silva, S., Quintal, R., Pereira, M.G., 2011. CD Barbosa et al. Brazilian Society of Plant Breeding. Printed in Brazil, Crop Breeding and Applied Biotechnology.

Deputy, J.C., Ming, R., Ma, H., Liu, Z., Fitch, M.M.M., Wang, M., Manshardt, R., Stiles, J.I., 2002. Molecular markers for sex determination in papaya (Carica papaya L.). Theor. Appl. Genet. 106, 107–111. https://doi.org/10.1007/s00122-002-0995-0

Dhanoa, M.S., Sanderson, R., Barnes, R.J., Lister, S.J., 1994. The Link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) Transformations of NIR Spectra. J. Near Infrared Spectrosc. Vol. 2, Issue 1, pp. 43-47 2, 43–47.

Durigan, M.F.B., 2014. Tecnologia Pós-colheita e Processamento de Mamão: Qualidade e Renda aos Produtos Roraimenses. Embrapa Roraima, Boa Vista, RR, p. 27.

Einax, J.W., 2007. Paul Gemperline (Ed.): Practical guide to chemometrics, 2nd Ed. Anal. Bioanal. Chem. 388, 511–512. https://doi.org/10.1007/s00216-007-1201-7

Embrapa, E.B. de P.A., 2019. Produção Brasileira de mamão em 2019.

Eshkabilov, S., Lee, A., Sun, X., Lee, C.W., Simsek, H., 2021. Hyperspectral imaging techniques for rapid detection of nutrient content of hydroponically grown lettuce cultivars. Comput. Electron. Agric. 181, 105968. https://doi.org/10.1016/j.compag.2020.105968

Faostat, F. and A.O. of the U.N.S., 2021a. Harvested area, yield and production in the main producing countries of Carica papaya L., in: Statistics Division.

Faostat, F. and A.O. of the U.N.S., 2021b. Major tropical fruits - Statistical compendium 2021., in: Statistics Division. Rome, pp. 20–31.

Farias, A.R.N., Oliveira, A.M.G., Santos Filho, H.P., Oliveira, J.R.P., Dantas, J.L.L., Oliveira, M.A., Sanches, N.F., Medina, V.M., Cordeiro, Z.J.M., 2009. A cultura do mamão, Coleção Plantar, 65. ed. Brasília, DF.

Ferrazza, F.L.F., Jacoboski, D.T.K., Wyrepkowski, A., Rodrigues, L., Figueiró, A.G., Paraginski, R.T., 2020. Qualidade de sementes e parâmetros produtivos de sementes de soja submetidas a diferentes tratamentos de sementes antes da semeadura. Res. Soc. Dev. 9, e47996232. https://doi.org/10.33448/rsd-v9i9.6232

Gebru, A., Jansson, S., Ignell, R., Kirkeby, C., Prangsma, J.C., Brydegaard, M., 2018. Multiband modulation spectroscopy for the determination of sex and species of mosquitoes in flight. J. Biophotonics 11, e201800014. https://doi.org/10.1002/jbio.201800014

Hair, J., Black, W., Babin, B., Anderson, R., Tatham, R., 2009. Análise multivariada de dados.

Hamilton, R.A., Izuno, T., 1967. A revised concept of sex inheritance in Carica papaya. Trop Agron 17, 2–401.

Haq, Q.M.I., Mabood, F., Naureen, Z., Al-Harrasi, A., Gilani, S.A., Hussain, J., Jabeen, F., Khan, A., Al-Sabari, R.S.M., Al-khanbashi, F.H.S., Al-Fahdi, A.A.M., Al-Zaabi, A.K.A., Al-Shuraiqi, F.A.M., Al-Bahaisi, I.M., 2018. Application of reflectance spectroscopies (FTIR-ATR & FT-NIR) coupled with multivariate methods for robust in vivo detection of begomovirus infection in papaya leaves. Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 198, 27–32. https://doi.org/10.1016/j.saa.2018.02.065

Haykin, S.S., 2009. Neural Networks and Learning Machines/Simon Haykin.

Hofmeyr, J., 1941. Genetics of Carica papaya L. Chron. Bot. 6, 245–247.

Hofmeyr, J.D.J., 1939. SEX REVERSAL IN CARICA PAPAY A L, AFRICAN JOURNAL OF SCIENCE. https://doi.org/10.10520/AJA00382353_7503

Jiménez, J.A., 2002. Manual práctico para el cultivo de la papaya hawaina.

Jindal, K.K., Singh, R.N., 1976. Sex determination in vegetative seedlings of Carica papaya by phenolic tests. Sci. Hortic. (Amsterdam). 4, 33–39. https://doi.org/10.1016/0304-4238(76)90062-5

Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis.: Pearson Prentice Hall. Pearson Prentice Hall 15–30.

Khan, A.L., Al-Harrasi, A., Numan, M., AbdulKareem, N.M., Mabood, F., Al-Rawahi, A., 2021. Spectroscopic and Molecular Methods to Differentiate Gender in Immature Date Palm (Phoenix dactylifera L.). Plants 10, 536. https://doi.org/10.3390/plants10030536

Lee, L.C., Liong, C.Y., Jemain, A.A., 2018. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. Analyst. https://doi.org/10.1039/c8an00599k

Liao, Z., Zhang, Xunxiao, Zhang, S., Lin, Z., Zhang, Xingtan, Ming, R., 2021. Structural variations in papaya genomes. BMC Genomics 22, 1–13. https://doi.org/10.1186/s12864-021-07665-4

Liu, Z., Moore, P.H., Ma, H., Ackerman, C.M., Ragiba, M., Yu, Q., Pearl, H.M., Kim, M.S., Charlton, J.W., Stiles, J.I., Zee, F.T., Paterson, A.H., Ming, R., 2004. A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature 427, 348–352. https://doi.org/10.1038/nature02228

Lorber, A., Kowalski, B.R., 1988. Estimation of prediction error for multivariate calibration. J. Chemom. 2, 93–109. https://doi.org/10.1002/cem.1180020203

Mapa, M. da A.P. e A., 2019. AGROSTAT – Estatísticas de Comércio Exterior do Agronegócio Brasileiro. [WWW Document]. MAPA - Ministério da Agric. Pecuária e Abast. URL http://indicadores.agricultura.gov.br/agrostat/index.htm (accessed 5.5.21).

Marin, S.L.D., Gomes, J.A., 1986. Morfologia e biologia floral do mamoeiro. Inf. Agropecuário 12, 10–14.

Marin, S.L.D., Gomes, J.A., Salgado, J.S., Martins, D.S., Fullin, E.A., 1995. Recomendacoes para a cultura do mamoeiro dos grupos solo e formosa no Estado do Espirito Santo, 4th ed. Circular Técnica - Empresa Capixaba de Pesquisa Agropecuária (Brazil)., Vitoria, ES (Brazil).

Martens, H., Næs, T., 1984. Multivariate Calibration, in: Kowalski, B.R. (Ed.), Chemometrics: Mathematics and Statistics in Chemistry. Springer Netherlands, Dordrecht, pp. 147–156. https://doi.org/10.1007/978-94-017-1026-8_5

Matias, F.I., Do Valle, C.B., Gouveia, B.T., Moro, G.V., Barrios, S.C.L., 2020. Using additive indices and principal components to select sexual genitors and hybrids of urochloa decumbens. Crop Breed. Appl. Biotechnol. 20, 2020. https://doi.org/10.1590/1984-70332020v20n2a18

Ming, R., Yu, Q., Moore, P.H., 2007a. Sex determination in papaya. Semin. Cell Dev. Biol. https://doi.org/10.1016/j.semcdb.2006.11.013

Ming, R., Yu, Q., Moore, P.H., 2007b. Sex determination in papaya. Semin. Cell Dev. Biol. https://doi.org/10.1016/j.semcdb.2006.11.013

MingMei;, W., ChengBin;, L., YuYu, L., 2017. Analysis method of similarity match of NIR spectral in differences of chemical components of tobacco leaves under different aging conditions. J. Yunnan Agric. Univ., 2 32, 269–275.

Mishra, P., Nordon, A., Roger, J.M., 2021. Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques. J. Pharm. Biomed. Anal. 192, 113684. https://doi.org/10.1016/j.jpba.2020.113684

Morais, C.L.M., Santos, M.C.D., Lima, K.M.G., Martin, F.L., 2019. Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach. Bioinformatics 35, 5257–5263. https://doi.org/10.1093/bioinformatics/btz421

Neto, B.D.B., Scarminio, I.S., Bruns, R.E., 2006. 25 Years of chemometrics in Brazil. Quim. Nova. https://doi.org/10.1590/s0100-40422006000600042

Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. https://doi.org/10.1038/nbt1206-1565

O'Neill, C.J., Roberts, J.J., Cozzolino, D., 2017. Identification of beef cattle categories (cows and calves) and sex based on the near infrared reflectance spectroscopy of their tail hair. Biosyst. Eng. 162, 140–146. https://doi.org/10.1016/j.biosystemseng.2017.07.007

Oliveira, A.M.G., Farias, A.R.N., Filho, H.P.S., Oliveira, J.R.P., Dantas, J.L.L., Santos, I. B., Oliveira, M.A., Souza Junior, M.T.S., Silva, M.J., Almeida, O.A., Nickel, O., Medina, V.M., Cordeiro, Z.J.M., 1994. Propagação e plantio: mamão para exportação: aspectos técnicos da produção.

Oliveira, E.J. de, Dantas, J.L.L., Castellen, M.D.S., De Lima, D.S., Barbosa, H.D.S., Motta, T.B.N., 2007. Marcadores moleculares na predição do sexo em plantas de mamoeiro. Pesqui. Agropecu. Bras. 42, 1747–1754. https://doi.org/10.1590/S0100-204X2007001200012

Orrillo, I., Cruz-Tirado, J.P., Cardenas, A., Oruna, M., Carnero, A., Barbin, D.F., Siche, R., 2019. Hyperspectral imaging as a powerful tool for identification of papaya seeds in black pepper. Food Control 101, 45–52. https://doi.org/10.1016/j.foodcont.2019.02.036

Otto, M., 2016. Quimiometria: estatística e aplicação computacional em química analítica, Analista.

Parasnis, A.S., Gupta, V.S., Tamhankar, S.A., Ranjekar, P.K., 2000. A highly reliable sex diagnostic PCR assay for mass screening of papaya seedlings. Mol. Breed. 6, 337–344. https://doi.org/10.1023/A:1009678807507

Parasnis, A.S., Ramakrishna, W., Chowdari, K. V., Gupta, V.S., Ranjekar, P.K., 1999. Microsatellite (GATA)(n) reveals sex-specific differences in papaya. Theor. Appl. Genet. 99, 1047–1052. https://doi.org/10.1007/s001220051413

Pasquini, C., 2018. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. Anal. Chim. Acta. https://doi.org/10.1016/j.aca.2018.04.004

Pasquini, C., 2003. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc. https://doi.org/10.1590/S0103-

50532003000200006

Paull, R.E., Duarte, O., 2011. Tropical fruits. CABI.

Pedersen, T., Karttunen, A.P., Korhonen, O., Wu, J.X., Naelapää, K., Skibsted, E., Rantanen, J., 2021. Determination of Residence Time Distribution in a Continuous Powder Mixing Process With Supervised and Unsupervised Modeling of In-line Near Infrared (NIR) Spectroscopic Data. J. Pharm. Sci. 110, 1259–1269. https://doi.org/10.1016/j.xphs.2020.10.067

Qiu, G., Lü, E., Wang, N., Lu, H., Wang, F., Zeng, F., 2019. Cultivar classification of single sweet corn seed using fourier transform near-infrared spectroscopy combined with discriminant analysis. Appl. Sci. 9, 1530. https://doi.org/10.3390/app9081530

Rahman, A., Wang, S., Yan, J., Xu, H., 2021. Intact macadamia nut quality assessment using near-infrared spectroscopy and multivariate analysis. J. Food Compos. Anal. 102, 104033. https://doi.org/10.1016/j.jfca.2021.104033

Reis, R.C., Viana, E. de S., de Jesus, J.L., Dantas, J.L.L., Lucena, R.S., 2015. Caracterização físico-química de frutos de novos híbridos e linhagens de mamoeiro. Pesqui. Agropecu. Bras. 50, 210–217. https://doi.org/10.1590/S0100-204X2015000300004

Ruggiero, C., Marin, S.L.D., Durigan, J.F., 2011. Mamão, Uma História de Sucesso. Rev. Bras. Frutic. 33, 076–082. https://doi.org/10.1590/s0100-29452011000500011

Salomão, L.C.C., Siqueira, D.L. de, Santos, D. dos S., Borba, A.N., 2007. Cultivo do Mamoeiro, 1st ed. Editora UFV, Viçosa, MG.

Santana, L.R.R., Matsuura, F.C.A.U., Cardoso, R.L., 2004. Genótipos melhorados de mamão (Carica papaya L.): avaliação sensorial e físico-química dos frutos. Ciência e Tecnol. Aliment. 24, 217–222. https://doi.org/10.1590/s0101-20612004000200010

Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047

Serrano, L.A.L., Cattaneo, L.F., 2010. O cultivo do mamoeiro no Brasil. Rev. Bras. Frutic. 32. https://doi.org/10.1590/s0100-29452010000300001

Silva, M.R.R. da, Vanzela, L.S., Pinheiro, L.C., Souza, J.F. dos S., 2016. Eect of Differen Compound in Producttion of Papaya Seedlings. Nucleus 13, 63–70. https://doi.org/10.3738/1982.2278.1044

Simão, S., 1998. Tratado de fruticultura, II. FEALQ, Piracicaba, SP.

Singh, S., Patel, S., Litoria, N., Gandhi, K., Faldu, P., Patel, K.G., 2018. Comparative Efficiency of Conventional and NIR Based Technique for Proximate Composition of

Pigeon Pea, Soybean and Rice Cultivars. Int. J. Curr. Microbiol. Appl. Sci. 7, 773–782. https://doi.org/10.20546/ijcmas.2018.701.094

Sirisomboon, P., 2018. NIR Spectroscopy for Quality Evaluation of Fruits and Vegetables, in: Materials Today: Proceedings. Elsevier Ltd, pp. 22481–22486. https://doi.org/10.1016/j.matpr.2018.06.619

Sjöström, M., Wold, S., Söderström, B., 1986. PLS DISCRIMINANT PLOTS, in: Pattern Recognition in Practice. Elsevier, pp. 461–470. https://doi.org/10.1016/b978-0-444-87877-9.50042-x

Slutsky, B., 1998. Handbook of Chemometrics and Qualimetrics: Part A By D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, and J. Smeyers-Verbeke. Data Handling in Science and Technology Volume 20A. Elsevier: Amsterdam. 1997. Xvii + 867 pp. ISBN 0-444-89724-0. $293.25. J. Chem. Inf. Comput. Sci. 38, 1254–1254. https://doi.org/10.1021/ci980427d

Souza, A.M. de, Breitkreitz, M.C., Filgueiras, P.R., Rohwedder, J.J.R., Poppi, R.J., 2013. Experimento didático de quimiometria para calibração multivariada na determinação de paracetamol em comprimidos comerciais utilizando espectroscopia no infravermelho próximo: Um tutorial, parte II. Quim. Nova 36, 1057–1065. https://doi.org/10.1590/S0100-40422013000700022

Souza, A.M. de, Poppi, R.J., 2012. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte i., Quim. Nova.

Storey, B.W., 1938. Segregation sex types in Solo papaya and their application to selection of seed. Proc. Am. Soc. Hortic. Sci. 35, 83–85.

Storey, W.B., 2020. Carica Papaya, in: CRC Handbook of Flowering. CRC Press, pp. 147–157. https://doi.org/10.1201/9781351072540-23

Storey, W.B., 1969. Papaya (Carica papaya L)., Papaya (Carica papaya L). Wageningen.

Storey, W.B., 1953. Genetics of the papaya. J. Hered. 44, 70–78. https://doi.org/10.1093/oxfordjournals.jhered.a106358

Storey, W.B., 1941. The botany and sex relationships of the papaya. Hawaii Agric. Exp. Bull. 87, 5–23.

Tao, D., Wang, Z., Li, G., Xie, L., 2019. Sex determination of silkworm pupae using VIS-NIR hyperspectral imaging combined with chemometrics. Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 208, 7–12. https://doi.org/10.1016/j.saa.2018.09.049

Tormena, C.D., Pauli, E.D., Marcheafave, G.G., Scheel, G.L., Rakocevic, M., Bruns,

R.E., Scarminio, I.S., 2020. FT-IR biomarkers of sexual dimorphism in yerba-mate plants: Seasonal and light accessibility effects. Microchem. J. 158, 105329. https://doi.org/10.1016/j.microc.2020.105329

Véras, G., De Brito, A.L.B., Da Silva, A.C., Da Silva, P., Da Costa, G.B., Félix, L.C.N., De Sousa Fernandes, D.D., De Fontes, M.M., 2012. Classificação de biodiesel na região do visível. Quim. Nova 35, 315–318. https://doi.org/10.1590/S0100-40422012000200015

Véstia, J., Barroso, J.M., Ferreira, H., Gaspar, L., Rato, A.E., 2019. Predicting calcium in grape must and base wine by FT-NIR spectroscopy. Food Chem. 276, 71–76. https://doi.org/10.1016/j.foodchem.2018.09.116

Vogel, A.I., 1981. Química analítica qualitativa/Arthur I. Vogel; tradução por Antonio Gimeno 5, in: Rev. Por G. Svehla–São Paulo: Mestre Jou. São Paulo.

William Herschel, S.B., 1800. XIV. Experiments on the refrangibility of the invisible rays of the sun. Philos. Trans. R. Soc. London 90, 284–292. https://doi.org/10.1098/rstl.1800.0015

Williams, P., 2019. Karl H. Norris, the Father of Near-Infrared Spectroscopy. NIR news 30, 25–27. https://doi.org/10.1177/0960336019875883

Xu, H.R., Yu, P., Fu, X.P., Ying, Y. Bin, 2009. On-site variety discrimination of tomato plant using visible-near infrared reflectance spectroscopy. J. Zhejiang Univ. Sci. B 10, 126–132. https://doi.org/10.1631/jzus.B0820200

Zorzetti, B.M., Harynuk, J.J., 2011. Using GC × GC-FID profiles to estimate the age of weathered gasoline samples. Anal. Bioanal. Chem. 401, 2423–2431. https://doi.org/10.1007/s00216-011-5130-0

CAPÍTULO 2 – Sexagem de sementes e folhas de mamoeiros utilizando espectroscopia no infravermelho próximo combinadas a técnicas multivariadas e machine learning

Resumo - As plantas de mamoeiro (*Carica papaya* L.) podem apresentar flores femininas, hermafroditas e masculinas. Todavia, somente as hermafroditas produzem frutos com o formato alongado que é exigido e valorizado pelo mercado consumidor. A seleção destas plantas é realizada baseada no fenótipo das mesmas e ocorre após o plantio das mudas, aumentando os custos de produção. Desta forma, este trabalho teve por objetivo verificar a possibilidade de se utilizar a espectroscopia no infravermelho próximo (NIR) como método não destrutivo para a sexagem de mamoeiros. Foram utilizados as cultivares 'T2', 'Formosa' e 'Calimosa', do grupo Formosa, e 'THB' e 'Ouro', do grupo Solo. Os espectros NIR foram coletados nas sementes e nas folhas dos respectivos *seedlings*. Ao se utilizar as sementes, foi possível obter um valor de F-score de 0,81 para o grupo de validação externa usando a análise de componentes principais e análise discriminante quadrática (PCA-QDA). Para as folhas, este valor foi um pouco menor, ou seja, 0,79 usando PCA e análise discriminante linear (PCA-LDA). Conclui-se que é possível utilizar a espectroscopia NIR associada às técnicas multivariadas como método não destrutivo para a sexagem de mamoeiros utilizando tanto as sementes quanto as folhas dos *seedlings*.

Palavras-chaves: *Carica papaya* L, sexagem, quimiometria, análise multivariada.

1. Introdução

O mamoeiro (*Carica papaya* L.) é comercialmente propagado por via sexuada (sementes) em função da maior eficiência econômica (Boas, 2019). No entanto, este método de propagação apresenta algumas limitações devido à grande variabilidade genética e fenotípica (Jiménez, 2002).

Por ser uma espécie polígama, o mamoeiro pode apresentar três classes de plantas, ou seja, monóicas, dióicas e monóclinas (Hofmeyr, 1941; Storey, 1953). Assim, o cruzamento entre estes tipos, mesmo de forma controlada, irá gerar sementes com diferentes tipos sexuais. Desta forma, as plantas de mamoeiro podem apresentar flores pistiladas ou femininas típicas, flores hermafroditas e flores estaminadas ou masculinas típicas (Marin e Gomes, 1986; Oliveira et al., 1996).

As plantas que apresentam flores pistiladas são denominadas femininas e os frutos oriundos destas flores são mais arredondados. Por outro lado, as plantas com flores hermafroditas produzem frutos com formato piriforme ou alongados e polpa carnosa, sendo estes preferidos pelo mercado consumidor (Paull and Duarte, 2011). Desta forma, os mamoeiros são tradicionalmente sexados em função do fenótipo das plantas quando estas atingem a fase reprodutiva e emitem as primeiras flores. Para isso, é necessário o plantio de pelo menos três mudas por cova visando aumentar as chances de pelo menos uma ser hermafrodita (Simão, 1998; Costa e Pacova, 2003). Esta prática aumenta os custos de produção em função do número de sementes e/ou mudas a serem adquiridos, além da manutenção das plantas no campo até que a sexagem seja realizada (Dantas et al., 2013)

Neste sentido, vários trabalhos foram desenvolvidos visando a identificação precoce do tipo sexual dos mamoeiros no sentido de se diminuir o número de sementes e/ou mudas a serem adquiridas. Dentre os métodos desenvolvidos, pode-se citar os citológicos que utilizam a determinação da dimensão dos estômatos e a contagem dos cloroplastos presente nas células-guarda (Datta, 1971), os citogenéticos que utilizam o número e análise do comportamento cromossômicos das plantas masculinas, femininas e hermafroditas (Araújo, 2008), os teores de compostos fenólicos pela extração destes compostos presente nas folhas, pecíolos e tecidos das plantas masculinas e femininas (Jindal e Singh, 1976) e ferramentas

biomoleculares, tais como marcadores do tipo *Random Amplified Polymorphic* DNA (RAPD) que mapeiam os padrões aleatórios da banda de DNA polimórfico amplificado ligados ao tipo sexual das plantas de mamoeiro (Chaves-Bedoya et al., 2009; Honoré et al., 2020).

Apesar da precisão, estes métodos são demorados, dispendiosos e requerem laboratórios especializados, dificultando a sua aplicação. Dessa forma, métodos mais rápidos e eficientes para a seleção dos tipos sexuais do mamoeiro usando as sementes ou nas mudas, poderiam minimizar os problemas decorrentes da sexagem desta espécie. Nesse sentido, espectroscopia no infravermelho próximo (NIR) surge como uma possibilidade para a sexagem do mamoeiro, pois vários estudos tem demostrado a eficiência da espectroscopia NIR para a identificação sexual em diferentes materiais biológicos, tais como, larvas do bicho-da-seda (Tao et al., 2019), bezerros e vacas (O'Neill et al., 2017).

No tocante à espécies vegetais, a espectroscopia NIR foi utilizada para a diferenciação sexual de *Phoenix dactylifera* L. (Khan et al., 2021), seleção de genitores sexuais e híbridos de *Urochloa decumbens* (Matias et al., 2020) e dimorfismo sexual em plantas de erva-mate (Tormena et al., 2020). No entanto, não foram encontrados trabalhos relacionados à sexagem de mais espécies vegetais, especialmente em sementes e folhas de mamoeiros. Portanto, o objetivo geral deste trabalho foi verificar a possibilidade de se utilizar a espectroscopia NIR como um método não destrutivo para a sexagem de mamoeiros e, por objetivos específicos: *i*. verificar a possibilidade de se utilizar as sementes e folhas dos *seedlings* para sexagem de mamoeiros e *ii*. desenvolver modelos matemáticos visando a predição do tipo sexual dos mamoeiros.

2. Material e Métodos

2.1 Material Vegetal

Foram utilizados dois lotes de sementes de cinco cultivares de C. papaya L. fornecidas pela empresa Feltrin® Sementes (Farroupilha, Brasil). As amostras foram constituídas pelas cultivares 'T2', 'Formosa' e 'Calimosa', do grupo Formosa, e 'THB' e 'Ouro', do grupo Solo. Esses lotes, foram adquiridos em agosto de 2019 (primeiro

lote) e em abril de 2020 (segundo lote), constituindo assim o conjunto de calibração e de validação externa (Tabela 1).

Além das sementes, foram utilizados também as folhas recém maduras que apresentavam três folíolos dos seedlings de cada semente.

2.2 Aquisição dos espectros NIR

A aquisição dos espectros NIR foi realizada em duas etapas, ou seja, a primeira relativa à aquisição dos espectros das sementes e folhas dos seus respectivos *seedlings* (conjunto de treinamento). A segunda etapa, se relacionou ao segundo lote de sementes e das folhas dos respectivos *seedlings* (conjunto de validação externa). Os espectros foram coletados da seguinte forma:

2.2.1 Sementes

Para a coleta dos espectros NIR foram utilizadas 150 sementes de cada uma das cultivares anteriormente descritas, o que totalizou 750 sementes para o conjunto de treinamento. Para o conjunto de validação externa foram coletados 350 espectros das sementes, 70 de cada cultivar. Estas foram agrupadas por cultivares e armazenadas em becker de vidro, sendo transferidos para um dessecador de vidro por 24 horas à temperatura ambiente (~25 °C) visando a uniformização do teor de umidade das sementes.

Em seguida, cada semente, previamente identificada com numeração única, foi transferida para um suporte metálico acoplado ao acessório de reflectância difusa *Near Infrared Reflectance Accessory* (NIRA) do espectrofotômetro FT-IR Spectrum 100N (PerkinElmer, Shelton, Estados Unidos) previamente calibrado, e os espectros NIR foram obtidos duas vezes, sendo as sementes reviradas aleatoriamente após cada leitura. O suporte foi colocado diretamente na saída do feixe de luz do NIRA no intuito de se evitar a entrada de luz externa.

2.2.2 Folhas

Após a obtenção dos espectros NIR das sementes, estas foram individualmente semeadas a dois centímetros de profundidade em sacos plásticos pretos medindo 15 cm de diâmetro x 20 cm de altura, contendo o substrato comercial composto de cascas de pinus e eucaliptos (Multiplant® Citrus), e estes foram identificados visando associar cada semente ao seu respectivo *seedlings*.

À medida que os *seedlings* apresentavam as folhas totalmente expandidas contendo três folíolos, foram coletados dois espectros NIR em folhas alternadas, nas posições da segunda lâmina foliar e próximo da nervura central, utilizando o acessório de fibra óptica do espectrofotômetro FT-IR Spectrum 100N (PerkinElmer, Shelton, Estados Unidos), colocando o mais próximo possível da superfície das folhas.

Os espectros NIR das sementes e folhas foram obtidos no modo de reflectância difusa na região do infravermelho próximo (NIR) com transformada de Fourier (FT). Previamente à coleta de espectros, o equipamento foi calibrado utilizando o padrão Spectralon® e após 150 leituras espectrais, realizava-se uma nova calibração do equipamento. Os dados foram obtidos na faixa espectral entre 4.000 a 10.000 cm^{-1}, sendo realizados 64 *scans*, com resolução espectral de 16 cm^{-1} e um intervalo de 2 cm^{-1}. Os espectros foram coletados como log(1/R), onde R é a refletância relativa (Miralbés, 2008).

2.3 Método de referência: identificação do tipo sexual

Quando as mudas atingiram uma altura média de 16 cm, estas foram transplantados para o solo utilizando o espaçamento de 50 cm x 50 cm. No momento em que as plantas iniciaram a emissão dos botões florais, procedeu-se o processo de identificação visual do tipo sexual de acordo com as características morfológicas descrita por Simão (1998) e Storey (1969), sendo identificadas as plantas femininas, hermafroditas e masculinas (Figura 1).

2.4 Quimiometria

Todos os dados espectrais foram importados e processados em ambiente MATLAB® versão 8.4 (R2014b) (The MathWorks Inc., Natick, MA, EUA) juntamente com a plataforma PLS Toolbox versão 7.9.3 (Eigenvector Research, Inc, Manson, WA, EUA) e algoritmos feitos no laboratório de pesquisa.

Aos dados espectrais referentes às sementes e às folhas foram aplicados diversos pré-processamentos, sendo estes avaliados pelos resultados para o conjunto de validação cruzada pelo método venetian blinds (CV VB) com 5 splits. Para as sementes, o melhor conjunto de validação interna foi obtido através da suavização espectral do tipo Savitzky e Golay - SG (1964) com 21 pontos de janela e ordem polinomial de segunda ordem, seguido de uma correção de linha de base do tipo automatic weighted least aquares (AWLS) com normalização vetorial, sendo este o pré-processamento selecionado. Já para os dados obtidos a partir das folhas, foi utilizada a primeira derivada SG (31 pontos de janela, segunda ordem polinomial), seguido de AWLS correção de linha de base com normalização vetorial.

Para os dados de sementes e folhas, todo o espectro na região de 10.000 – 4.000 cm^{-1} foi utilizado. Para todos os casos, a média espectral foi utilizado como espectro representativo para cada amostra. Similarmente, foram utilizados um conjunto de treinamento, composto de 70 % dos dados iniciais e o conjunto de validação interna, composto de 30 % dos dados (Pasquini, 2018). O primeiro subconjunto incluiu a maioria das amostras e foi o conjunto utilizado para a construção e otimização dos modelos supervisionados. O segundo subconjunto, foi composto pelas demais amostras e serviu para a avaliação parcial da performance dos modelos.

Os dados foram organizados de forma a se obter uma matriz final com a presença das cinco cultivares e os diferentes tipos sexuais (feminino e hermafrodita), tanto para o conjunto espectral das sementes e folhas. A seleção das amostras foi realizada a partir do algoritmo de seleção de Kennard e Stone (1969). Este foi aplicado, com o objetivo de reduzir o viés na divisão entre os conjuntos de treinamento e validação interna dos modelos. A divisão foi realizada de forma a se obter porcentagens semelhantes, para as duas classes, em cada uma das cultivares, construindo assim um modelo que continha informação proporcional relacionada tanto às diferentes cultivares quanto aos tipos sexuais (Tabela 1).

Para testar a real eficácia dos modelos de sementes e folhas, foi utilizado integralmente o segundo lote amostral, no processo de validação externa, sendo os resultados estatísticos para este conjunto os mais importantes e representativos da real eficácia dos modelos.

2.5 Modelos supervisionados

A análise discriminante linear (LDA) e quadrática (QDA), bem como máquina de vetores suporte (SVM) foram utilizadas tendo como os *scores* obtidos a partir da PCA como dados de entrada (Fisher, 1936; Bedin et al., 2021).

A PCA-LDA é uma técnica multivariada supervisionada de classificação (Wold, 1966; Moo-Young, 2019;) e os *scores* de classificação da LDA, na forma não-Bayesiana, podem ser obtidos para uma i-ésima amostra para uma k-ésima classe através da seguinte equação (Tibshirani, 1996):

$$\mathrm{L_{ik}} = (\mathbf{x}_\mathrm{i} - \bar{\mathbf{x}}_\mathrm{k})^\mathrm{T} \sum\nolimits_{\mathrm{pooled}}^{-1} (\mathbf{x}_\mathrm{i} - \bar{\mathbf{x}}_\mathrm{k}) \quad (1)$$

Em que x_i representa uma medida desconhecida para uma i-ésima amostra, $\bar{x}_k$ representa uma medição média para a k-ésima classe k, e finalmente, $\sum_{pooled}$ como a matriz de covariância combinada.

A PCA foi utilizada pois apresenta a vantagem de ser simples e rápida, embora não seja recomendada para conjuntos de dados de poucas classes, além de que o conjunto amostral ter que atender o pressuposto de normalidade.

Para PCA-QDA, de forma análoga e não-Bayesiana, pode ser definida como (Tibshirani, 1996):

$$\mathrm{Q_{ik}} = (\mathbf{x}_\mathrm{i} - \bar{\mathbf{x}}_\mathrm{k})^T \Sigma_k^{-1} (\mathbf{x}_\mathrm{i} - \bar{\mathbf{x}}_\mathrm{k}) \quad (2)$$

Como representado nas equações, como principal diferença, LDA e QDA assume diferentes estruturas de dados: LDA considera que as classes envolvidas têm uma matriz de variância-covariância semelhantes, contrariamente a QDA, que assume a presença de diferenças significantes entre os grupos e suas matrizes, calculando, para cada classe, uma matriz de variância-covariância. Ambos assumem que o conjunto de dados segue uma distribuição normal em cada classe especificada. Na LDA, a matriz de covariância de cada grupo é considerada igual, enquanto em QDA essas matrizes podem ser diferentes.

Já para a classificação através de PCA-SVM, há uma clara diferenciação: este modelo assume uma relação não-linear dos dados, efetivando uma mudança de espaço através de uma função de núcleo, assim modificando a estrutura dos dados, classificando as amostras através de hiperplanos que têm o objetivo de maximizar as diferenças entre os grupos. Neste estudo, apenas a função de base radial (RBF) foi utilizada por ser costumeiramente utilizada em problemas de classificação binária, apresentando resultados satisfatórios em diversos casos. Para esta função, temos a seguinte equação (Mazumder et al., 2011):

$$\mathrm{k}(\mathbf{x}_\mathrm{i}, \mathbf{z}_\mathrm{j}) = \exp(-\gamma \left\| \mathbf{x}_\mathrm{i} - \mathbf{z}_\mathrm{j}^2 \right\|) \quad (3)$$

Na qual, $\mathbf{x}_\mathrm{i}$ e $\mathbf{z}_\mathrm{j}$ são vetores de medidas para as amostras e γ é um parâmetro para ajuste para espaçamento da função RBF.

O classificador SVM é obtido através da seguinte função:

$$\mathrm{f(x)} = \mathrm{sign}\left(\sum_{\mathrm{i}=1}^{\mathrm{N_{SV}}} \alpha_\mathrm{i} \mathrm{y_i} \mathrm{k}(\mathbf{x}_\mathrm{i}, \mathbf{z}_\mathrm{j}) + \mathrm{b}\right) \quad (4)$$

Em que, N_{SV} é o número de vetores de suporte; α_i representa o multiplicador de Lagrange, y_i é a classe, podendo assumir o valor de ± 1, $\mathrm{k}(\mathbf{x}_\mathrm{i}, \mathbf{z}_\mathrm{j})$ sendo uma função de núcleo, definido pela equação (x) e b é a tendência (bias) do parâmetro.

O SVM é recomendado para a determinação de classes a partir de muitas variáveis, por ser rápido e eficiente para separações não lineares. Contudo, a desvantagem desse método é a demora no processamento quando se tem um grande conjunto de repetições, não se adequando bem a variáveis com muitos ruídos.

2.2.3 Avaliação dos modelos

Os modelos foram avaliados quanto a sua eficácia mediante a parâmetros estatísticos. Tais parâmetros medem a capacidade de predição de um dado modelo por meio de equações, onde seus parâmetros são obtidos através dos resultados encontrados para um dado conjunto de teste, que, neste caso, será considerado como conjunto de teste (validação interna) e validação externa. A acurácia representa o número de amostras classificadas corretamente, para ambas as classes. A sensibilidade e especificidade mensuram a proporção de amostras positivas e negativas, aqui sendo consideradas para fins de classificação como

pertencentes ao grupo de amostras de femininas e hermafroditas, respectivamente. Por fim, foi avaliado a média harmônica entre sensibilidade e especificidade, conhecida como F-score, representando a acurácia real do modelo, quando é levada em conta os diferentes tamanhos das classes, calculada a partir dos valores de sensibilidade e especificidade. Os parâmetros aqui citados podem ser calculados a partir das seguintes equações:

$$Acurácia\ (\%) = \left(\frac{TP+TN}{TP+FP+TN+FN}\right)\ x\ 100 \quad (5)$$

$$Sensibilidade\ (\%) = \left(\frac{TP}{TP+FN}\right)\ x\ 100 \quad (6)$$

$$Especificidade\ (\%) = \left(\frac{TN}{TN+FP}\right)\ x\ 100 \quad (7)$$

$$F-score\ (\%) = \frac{2\ x\ Sensitivity\ x\ Specificity}{Sensitivity+Specificity} \quad (8)$$

Em que os valores de TP e TN representam o número de verdadeiros positivos (TP) e negativos (TN), enquanto FP e FN equivalem ao número de falsos positivos (FP) e negativos (FN), respectivamente. Todos estes valores foram obtidos a partir do conjunto de treinamento e de validação externa.

3. Resultados

3.1 Sementes

3.1.1 Espectros NIR

Os espectros NIR sem pré-processamento (brutos) das sementes de todas as cultivares combinadas em função de cada tipo sexual (feminino e hermafrodita) podem ser observados na Figura 2. Vale destacar que foram identificadas apenas plantas femininas e hermafroditas, não sendo observadas plantas masculinas no conjunto amostral.

De maneira geral os espectros NIR foram semelhantes entre as cultivares e entre as sementes hermafroditas e femininas, não sendo possível identificar diferenças espectrais entre os tipos sexuais (Figura 2). Os espectros NIR foram dominados por dois picos, um na faixa de 5.158 cm^{-1} e outro na de 6.984 cm^{-1} (Figura 2). Estes estão associados às bandas de ligações e de vibrações assimétricas do OH e de estiramento e ao primeiro sobretom da ligação OH, relacionados à presença de água nas sementes (Purcell et al., 2009). Outra região

que mostrou alguma diferença entre os tipos sexuais foi na faixa de 4.720 cm^{-1} que corresponde ao primeiro sobretom de CH. Esta região se relaciona a presença de lipídeos que são abundantes nas sementes de mamoeiro (Chielle, 2014).

3.1.2 Desenvolvimento dos modelos de classificação

No tocante ao valor de F-score do conjunto de validação externa (0,81), o modelo PCA-QDA foi o considerado o mais adequado visando a classificação das sementes femininas e hermafroditas. De maneira geral, os resultados encontrados no conjunto de validação externa foram muito bons, ou seja, valores superiores a 0,80 para acurácia, 0,84 para sensitividade e 0,78 para especificidade (Tabela 2).

Para distinguir e classificar o comportamento das amostras em hermafroditas e femininas, foram construídas as matrizes de confusão. Usando o modelo PCA-QDA para os conjuntos de validação cruzada (teste) foram observados 12 erros e 58 acertos em um total de 70 amostras consideradas hermafroditas e 17 erros e 60 acertos em um total de 77 amostras consideradas femininas (Tabela 3). Para o conjunto de validação externa, foram observados 21 erros e 74 acertos para as amostras consideradas hermafroditas e 10 erros e 53 acertos para as amostras consideradas femininas (Tabela 3).

3.2 Folhas

3.2.1 Espectros NIR

Similarmente ao observado para as sementes, os espectros NIR das folhas não diferiram quanto as plantas femininas e hermafroditas, exibindo um padrão espectral semelhante. Todavia, vale ressaltar que os espectros médios mostraram uma pequena diferenciação entre os tipos sexuais (Figura 3).

Os espectros NIR das folhas foram dominados pelos picos nas regiões de 4.720 e 6.984 cm^{-1} que se relacionam à presença de água, ao primeiro sobretom da C-H e aos primeiros sobretons das combinações N-H e O-H (Williams, 2001).

3.2.2 *Desenvolvimento dos modelos de classificação*

Foram testados diferentes modelos matemáticos visando obter um com o melhor desempenho a partir das folhas dos mamoeiros (Tabela 4). Em função dos resultados de F-score, sensibilidade e especificidade, principalmente em relação aos valores de F-score obtidos no conjunto de validação externa (0,79), o modelo PCA-LDA se mostrou o mais adequado visando a classificação das folhas oriundas dos *seedlings* femininos e hermafroditas (Tabela 4). Para o conjunto de validação externa das folhas foram observados valores superiores de 0,79 de acurácia, 0,83 de sensitividade e 0,76 de especificidade (Tabela 4).

O mesmo procedimento utilizado para o desenvolvimento dos modelos de classificação das sementes foi aplicado às folhas. Desta forma, construiu-se matrizes de confusão para os conjuntos de validação cruzada e validação externa (Tabela 5). Aplicando-se o modelo PCA-LDA para o conjunto de validação cruzada (teste) foram observados 4 erros e 21 acertos em um total de 70 amostras consideradas hermafroditas e 3 erros e 23 acertos em um total de 77 amostras consideradas femininas (Tabela 5). Para o conjunto de validação externa, foram observados 19 erros e 60 acertos para as amostras consideradas hermafroditas e 9 erros e 44 acertos para as amostras consideradas femininas (Tabela 5).

4. Discussão

4.1 Sementes

De um total de 750 espectros NIR coletados a partir das sementes, apenas 495 foram utilizados, pois muitas sementes da cultivar ‘Ouro’ não germinaram e muitas plantas das demais cultivares não sobreviveram até a emissão das primeiras flores (Tabela 1). Do total de plantas usadas para a identificação do tipo sexual, 235 foram classificadas como femininas (47,5%) e 260 como hermafroditas (52,5%). Estes resultados ficaram próximos aos obtidos pelas empresas produtoras de sementes que utilizam apenas progenitores hermafroditas (M_2 m) para a produção de sementes de mamoeiros, ou seja 33% de plantas femininas e 67% de plantas hermafroditas (Costa, 2003).

Apesar das diferenças entre cultivares e tipos sexuais, os espectros NIR das sementes pouco diferiram em relação à estas variáveis e se caracterizaram pela presença de água (5.158 e 6.984 cm^{-1}), ou seja, 47±0,9% de umidade, mesmos tendo sido submetidas ao processo de desidratação em dessecador por 24 h. Por outro lado, foram observadas variações na região na faixa de 4.720 cm^{-1} (Figura 2) que corresponde a presença de lipídeos que são abundantes nas sementes de mamoeiro (Chielle, 2014). Desta forma, pode haver diferenças na composição de ácidos graxos entre as sementes de mamoeiros femininos e hermafroditas que levaram as diferenças espectrais na região do primeiro sobretom de CH.

O maior número de amostras e as diferenças espectrais podem ter contribuído para a melhor performance dos modelos de classificação desenvolvidos com os espectros NIR das sementes (Tabela 2) em relação à das folhas (Tabela 5). Foram testados 12 tipos diferentes de modelos matemáticos e, apesar dos modelos mais sofisticados terem apresentados ótimos resultados (Tabela 2 e 5), destacando o uso de algoritmos genéticos e análise discriminante quadrática (GA-QDA), modelos mais simples como o PCA-LDA, PCA-QDA e PCA–SVM também apresentaram resultados satisfatórios quanto às estatísticas para a discriminação das sementes quanto ao tipo sexual (Tabela 2).

A escolha do modelo desenvolvido usando PCA-QDA com dois PCs se relacionou à simplicidade e aos bons valores de acurácia, sensibilidade, especificidade e, principalmente, de F-score, tanto para as amostras do conjunto de validação interna (treinamento) quanto para o de validação externa (Tabela 2), o que demonstra a robustez deste modelo. Nicolaï et al. (2007), ressaltaram a importância de se utilizar conjuntos de validação externa no desenvolvimento de modelos de calibração multivariada, pois grande parte das calibrações são feitas usando os mesmos conjuntos amostrais. Desta forma, a performance obtida com o lote de sementes do conjunto de validação externa demonstra o ótimo desempenho do modelo escolhido.

4.1 Folhas

Em relação às folhas, um número ainda menor de plantas foi utilizado para a identificação do tipo sexual, ou seja, 64 plantas foram classificadas como femininas (42,1%) e 88 como hermafroditas (57,9%), Tabela 1. Estes resultados se aproximaram mais da proporção teórico obtida da autofecundação de plantas hermafroditas (M_2m), 33% de plantas femininas e 67% de plantas hermafroditas (Costa, 2003). Destaca-se que houve uma incidência severa de ácaros nos *seedlings* de todas as cultivares, o que comprometeu a qualidade das folhas a serem utilizadas para a obtenção dos espectros NIR, sendo necessária a aplicação de enxofre para controlar esta praga (Vieira et al., 2004). Apesar do número de amostras ter sido menor, o número total de plantas (n=152) foi superior à faixa de 50 a 100 amostras recomendada por Pasquini (2003) para estudo envolvendo a espectroscopia NIR.

Os espectros NIR das folhas refletiram o alto teor de água neste tecido, sem grandes diferenças entre as cultivares e tipos sexuais (Figura 3). As folhas de mamoeiro apresentam em média 96,9 - 98,6% de umidade como relatado por Cruz et al., (2004), o que se relacionou aos picos associados às bandas de ligações e de vibrações assimétricas do OH (5.158 cm^{-1}) e de estiramento e ao primeiro sobretom da ligação OH (6.984 cm^{-1}). O mesmo foi observado por Haq et al. (2018), que utilizou espectros NIR de folhas de mamoeiro para discriminar plantas infectadas com begomovírus.

Apesar do menor número de amostras e da ausência de diferenças espectrais, a performance dos modelos de classificação desenvolvidos com os espectros NIR das folhas podem ser considerados como muito bons (Tabela 5). Também foram testados 12 modelos matemáticos e, ao contrário do observado para as sementes, os modelos mais sofisticados não apresentaram melhores resultados em relação as mais simples (Tabela 5). Desta forma, a aplicação do PCA-LDA resultou em melhor classificação das folhas oriundas dos *seedlings* femininos e hermafroditas, sendo observados valores superiores de acurácia, sensitividade, especificidade, e F-score no conjunto de validação externa (Tabela 5).

O uso de folhas imaturas também foi eficiente para a diferenciação do tipo sexual de tamareiras (*Phoenix dactylifera* L.) e o uso do PCA-LDA possibilitou a discriminação tanto no conjunto de calibração (R^2=0,90) quanto no de validação

interno (R^2=0,88) (Khan et al., 2021). Desta forma, as folhas, tanto quanto as sementes, se caracterizam com tecido que pode ser utilizado visando a classificação das plantas quanto aos diferentes tipos sexuais.

5. Conclusão

É possível a sexagem de mamoeiros utilizando tanto as sementes quanto as folhas dos *seedlings* utilizando a espectroscopia no infravermelho próximo (NIR) associada às técnicas multivariadas de calibração.

Os valores de F-escore obtidos com o uso do PCA-QDA para as sementes (0,81) e com o PCA-LDA para as folhas (0,79), no conjunto de validação externa, demonstram a acurácia dos resultados encontrados.

Os modelos desenvolvidos podem ser utilizados como um método rápido e precoce de determinação do tipo sexual dos mamoeiros, o que pode contribuir para a redução do custo de produção desta cultura.

6. Referências

Araújo, F.S., 2008. Estudos citogenéticos e citométricos em mamoeiro (Carica papaya L.). Universidade Federal de Viçosa.

Bedin, F.C.B., Faust, M.V., Guarneri, G.A., Assmann, T.S., Lafay, C.B.B., Soares, L.F., de Oliveira, P.A.V., dos Santos-Tonial, L.M., 2021. NIR associated to PLS and SVM for fast and non-destructive determination of C, N, P, and K contents in poultry litter. Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 245, 118834. https://doi.org/10.1016/j.saa.2020.118834

Boas, R.V., 2019. Influência da seleção de sementes no desenvolvimento de plantas e na biologia floral do mamoeiro (Carica papaya L). TCC - ICA - Agron. 50.

Chaves-Bedoya, G., Pulido, M., Sánchez-Betancourt, E., Núñez, V., 2009. RAPD markers for sex identification in papaya (Carica papaya L.) in Colombia. Agron. Colomb. 27, 145–149.

Chielle, D.P., 2014. Estudo da secagem de sementes de mamão papaya (carica papaya) em secador convectivo horizontal e leito de jorro e a influência na extração de óleo. Universidade Federal de Santa Maria, Santa Maria, RS.

Costa, A. de F.S. da, 2003. Aspectos Gerais do Melhoramento do Mamoeiro, in:

Papaya Brasil - 2003. Vitória : Incaper, 2003., pp. 157–170.

Cruz, J.L., Ferreira Coelho, E., Pelacani, C.R., Coelho Filho, M.A., Tosta Dias, A., Taluana Dos Santos, M., 2004. Growth and dry matter and carbon partition in papaya plants in response to nitrogen nutrition. Bragantia 63, 351–361. https://doi.org/10.1590/s0006-87052004000300005

Costa, A.F.S., Pacova, B.E. V., 2003. Caracterização de cultivares, estratégias e perspectivas do melhoramento genético do mamoeiro., in: A Cultura Do Mamão: Tecnologia e Produção. pp. 59–102.

Dantas, J.L.L., Junghans, D.T., Lima, J.F. de, 2013. Mamão: o produtor pergunta, a Embrapa responde., 2ª rev. e atual. ed. Embrapa, Brasília, DF .

Datta, P.C., 1971. Chromosomal Biotypes of Carica Papaya Linn. Cytologia (Tokyo). 36, 555–562. https://doi.org/10.1508/cytologia.36.555

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Haq, Q.M.I., Mabood, F., Naureen, Z., Al-Harrasi, A., Gilani, S.A., Hussain, J., Jabeen, F., Khan, A., Al-Sabari, R.S.M., Al-khanbashi, F.H.S., Al-Fahdi, A.A.M., Al-Zaabi, A.K.A., Al-Shuraiqi, F.A.M., Al-Bahaisi, I.M., 2018. Application of reflectance spectroscopies (FTIR-ATR & FT-NIR) coupled with multivariate methods for robust in vivo detection of begomovirus infection in papaya leaves. Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 198, 27–32. https://doi.org/10.1016/j.saa.2018.02.065

Hofmeyr, J., 1941. Genetics of Carica papaya L. Chron. Bot. 6, 245–247.

Honoré, M.N., Belmonte-Ureña, L.J., Navarro-Velasco, A., Camacho-Ferre, F., 2020. Effects of the size of papaya (Carica papaya L.) seedling with early determination of sex on the yield and the quality in a greenhouse cultivation in continental Europe. Sci. Hortic. (Amsterdam). 265, 109218. https://doi.org/10.1016/j.scienta.2020.109218

Jiménez, J.A., 2002. Manual práctico para el cultivo de la papaya hawaina.

Jindal, K.K., Singh, R.N., 1976. Sex determination in vegetative seedlings of Carica papaya by phenolic tests. Sci. Hortic. (Amsterdam). 4, 33–39. https://doi.org/10.1016/0304-4238(76)90062-5

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

Khan, A.L., Al-Harrasi, A., Numan, M., AbdulKareem, N.M., Mabood, F., Al-Rawahi, A., 2021. Spectroscopic and Molecular Methods to Differentiate Gender in Immature Date Palm (Phoenix dactylifera L.). Plants 10, 536. https://doi.org/10.3390/plants10030536

Marin, S.L.D., Gomes, J.A., 1986. Morfologia e biologia floral do mamoeiro. Inf. Agropecuário 12, 10–14.

Matias, F.I., Do Valle, C.B., Gouveia, B.T., Moro, G.V., Barrios, S.C.L., 2020. Using additive indices and principal components to select sexual genitors and hybrids of urochloa decumbens. Crop Breed. Appl. Biotechnol. 20, 2020. https://doi.org/10.1590/1984-70332020v20n2a18

Mazumder, R., Friedman, J.H., Hastie, T., 2011. SparseNet: Coordinate descent with nonconvex penalties. J. Am. Stat. Assoc. 106, 1125–1138. https://doi.org/10.1198/jasa.2011.tm09738

Miralbés, C., 2008. Discrimination of European wheat varieties using near infrared reflectance spectroscopy. Food Chem. 106, 386–389. https://doi.org/10.1016/j.foodchem.2007.05.090

Moo-Young, M., 2019. Comprehensive Biotechnology. Universidade de Waterloo, Waterloo, ON, Canadá.

Nicolaï, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., Lammertyn, J., 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biol. Technol. https://doi.org/10.1016/j.postharvbio.2007.06.024

O'Neill, C.J., Roberts, J.J., Cozzolino, D., 2017. Identification of beef cattle categories (cows and calves) and sex based on the near infrared reflectance spectroscopy of their tail hair. Biosyst. Eng. 162, 140–146. https://doi.org/10.1016/j.biosystemseng.2017.07.007

Oliveira, R.D., Dantas, J., Almeida, E.D., NIckel, O., Vilarinhos, A.D., Morales, C., 1996. Uso da biotecnologia no melhoramento genético e propagação do mamoeiro. Mamão no Bras.

Pasquini, C., 2018. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. Anal. Chim. Acta. https://doi.org/10.1016/j.aca.2018.04.004

Pasquini, C., 2003. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc. https://doi.org/10.1590/S0103-50532003000200006

Paull, R.E., Duarte, O., 2011. Tropical fruits. CABI.

Purcell, D.E., O'shea, M.G., Johnson, R.A., Kokot, S., 2009. Near-infrared spectroscopy for the prediction of disease ratings for fiji leaf gall in sugarcane clones. Appl. Spectrosc. 63, 450–457. https://doi.org/10.1366/000370209787944370

Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified

Least Squares Procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047

Simão, S., 1998. Tratado de fruticultura, II. FEALQ, Piracicaba, SP.

Storey, W.B., 1969. Papaya (Carica papaya L)., Papaya (Carica papaya L). Wageningen.

Storey, W.B., 1953. Genetics of the papaya. J. Hered. 44, 70–78. https://doi.org/10.1093/oxfordjournals.jhered.a106358

Tao, D., Wang, Z., Li, G., Xie, L., 2019. Sex determination of silkworm pupae using VIS-NIR hyperspectral imaging combined with chemometrics. Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 208, 7–12. https://doi.org/10.1016/j.saa.2018.09.049

Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tormena, C.D., Pauli, E.D., Marcheafave, G.G., Scheel, G.L., Rakocevic, M., Bruns, R.E., Scarminio, I.S., 2020. FT-IR biomarkers of sexual dimorphism in yerba-mate plants: Seasonal and light accessibility effects. Microchem. J. 158, 105329. https://doi.org/10.1016/j.microc.2020.105329

Vieira, M.R., Correa, L. de S., Castro, T.M.M.G. de, Silva, L.F.S. da, Monteverde, M. de S., 2004. Efeito do cultivo do mamoeiro (Carica papaya L.) em ambiente protegido sobre a ocorrência de ácaros fitófagos e moscas-brancas. Rev. Bras. Frutic. 26, 441–445. https://doi.org/10.1590/s0100-29452004000300017

Williams, P., 2001. Implementation of Near-infrared Technology, Near-infrared Technology in the Agricultural and Food Industries, and Edition. American Assoc. of Cereal Chemists Inc.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. Multivar. Anal. 391–420.

Tabelas

Tabela 1. Número de sementes e folhas utilizadas no conjunto de calibração (treinamento), validação cruzada (teste) das cultivares de mamoeiro ‘Calimosa’, ‘Formosa’, ‘T2’, ‘Ouro’ e ‘THB’.

Cultivares	Sementes				Folhas			
	número de amostras							
	Treinamento		Teste		Treinamento		Teste	
	Hermafoditas	Femininas	Hermafroditas	Femininas	Hermafroditas	Femininas	Hermafroditas	Femininas
Calimosa	47	48	20	20	17	16	7	7
Formosa	50	55	21	24	11	19	5	8
T2	50	66	21	28	23	23	10	10
Ouro	5	4	2	1	0	0	0	0
THB	13	10	6	4	8	4	3	1
Total	165	183	70	77	59	62	25	26

Tabela 2. Desempenho do conjunto de treinamento, teste e validação externa para o tipo sexual de sementes de mamoeiros.

Modelo	Conjunto	Acurácia	Sensitividade	Especificidade	F-score
[1]PCA-LDA (1 PC)	Treinamento	0.72	0.73	0.71	0.72
	Teste	0.76	0.78	0.74	0.76
	Validação Externa	0.78	0.87	0.72	0.78
[2]PCA-QDA (2 PC)	**Treinamento**	**0.85**	**0.86**	**0.85**	**0.85**
	Teste	**0.80**	**0.78**	**0.83**	**0.80**
	Validação Externa	**0.80**	**0.78**	**0.83**	**0.81**
[3]PCA-SVM (10 PC)	Treinamento	0.98	0.97	0.98	0.97
	Teste	0.59	0.62	0.54	0.58
	Validação Externa	0.43	0.86	0.15	0.25
[4]SPA-LDA	Treinamento	0.56	0.59	0.54	0.56
	Teste	0.61	0.62	0.59	0.60
[5]SPA-QDA	Treinamento	0.66	0.61	0.70	0.65
	Teste	0.64	0.52	0.77	0.62
[6]SPA-SVM	Treinamento	0.57	0.63	0.50	0.56
	Teste	0.51	0.58	0.43	0.49
[7]GA-LDA	Treinamento	0.57	0.59	0.54	0.56
	Teste	0.51	0.53	0.49	0.51
[8]GA-QDA	Treinamento	0.76	0.78	0.75	0.76
	Teste	0.73	0.73	0.73	0.73
[9]GA-SVM	Treinamento	0.67	0.60	0.75	0.67
	Teste	0.55	0.55	0.56	0.55
[10]SIMCA (1 PCs)	Treinamento	0.58	0.62	0.54	0.58
	Teste	0.52	0.82	0.20	0.32
[11]PLS-DA (3 LVs)	Treinamento	0.54	0.50	0.58	0.54
	Teste	0.62	0.58	0.66	0.62
[12]SVM (k = 0.14; c=1000)	Treinamento	0.85	0.89	0.81	0.85
	Teste	0.54	0.64	0.44	0.52

[1]PCA-LDA: Análise de Componente Principal - Análise Discriminante Linear; [2] PCA-QDA: Análise de Componente Principal - Análise Discriminante Quadrática; [3]PCA – SVM: Análise de componentes principais - Máquina de vetores de suporte; [4]SPA-LDA: Algoritmo de projeções sucessivas - Análise Discriminante Linear; [5]SPA-QDA: Algoritmo de projeções sucessivas - Análise Discriminante Quadrática; [6]SPA-SVM: Algoritmo de projeções sucessivas - Máquinas de vetor de suporte; [7]GA-LDA: Algoritmos genéticos - Análise Discriminante Linear; [8]GA-QDA: Algoritmos genéticos - Análise Discriminante Quadrática; [9]GA-SVM: Algoritmos genéticos - Máquinas de vetor de suporte; [10]SIMCA: Modelagem Soft Independente da Analogia de Classes; [11]PLS-DA: mínimos quadrados parciais para análise discriminante; [12]SVM: Máquinas de vetor de suporte.

Tabela 3. Matriz de confusão para os conjuntos de validação cruzada (teste) e validação externa do desempenho do modelo PCA-QDA das sementes de todas as cultivares combinadas, separadas de acordo com o tipo sexual do mamoeiro, femininas e hermafroditas.

		Teste		Validação externa	
Out Class	Hermafrodita	TP = 39,5% (n=58)	FP = 11,6% (n=17)	TP = 46,8% (n=74)	FP = 6,3% (n=10)
	Feminina	FN = 8,2% (n=12)	TN = 40,8% (n=60)	FN = 13,3% (n=21)	TN = 33,5% (n=53)
		Hermafrodita	Feminina	Hermafrodita	Feminina
		Target Class		Target Class	

TP = True Positive; FP=False Positive; FN=False Negative; TN=True Negative

Tabela 4. Desempenho do conjunto de treinamento, teste e validação externa para o tipo sexual de folhas de seedlings de mamoeiros.

Modelo	Conjunto	Acurácia	Sensitividade	Especificidade	F-score
[1]PCA-LDA (1 PC)	**Treinamento**	**0.74**	**0.74**	**0.75**	**0.74**
	Teste	**0.86**	**0.88**	**0.84**	**0.86**
	Validação Externa	**0.79**	**0.83**	**0.76**	**0.79**
[2]PCA-QDA (10 PC)	Treinamento	0.88	0.90	0.85	0.87
	Teste	0.84	0.92	0.76	0.83
	Validação Externa	0.70	0.87	0.58	0.70
[3]PCA-SVM (6 PC)	Treinamento	0.88	0.87	0.88	0.87
	Teste	0.59	0.62	0.56	0.59
	Validação Externa	0.43	0.49	0.39	0.43
[4]SPA-LDA	Treinamento	0.56	0.57	0.55	0.56
	Teste	0.55	0.58	0.53	0.55
[5]SPA-QDA	Treinamento	0.79	0.81	0.76	0.78
	Teste	0.76	0.81	0.72	0.76
[6]SPA-SVM	Treinamento	0.79	0.65	0.49	0.56
	Teste	0.55	0.73	0.36	0.48
[7]GA-LDA	Treinamento	0.69	0.70	0.68	0.69
	Teste	0.55	0.56	0.54	0.55
[8]GA-QDA	Treinamento	0.92	0.90	0.93	0.91
	Teste	0.76	0.69	0.84	0.76
[9]GA-SVM	Treinamento	0.86	0.95	0.76	0.84
	Teste	0.71	0.69	0.72	0.70
[10]SIMCA (3 PCs)	Treinamento	0.52	0.53	0.51	0.52
	Teste	0.55	0.85	0.24	0.37
[11]PLS-DA (3 LVs)	Treinamento	0.58	0.53	0.63	0.58
	Teste	0.61	0.58	0.64	0.61
[12]SVM (k = 0.57; c=1000)	Treinamento	0.65	0.69	0.61	0.65
	Teste	0.63	0.77	0.48	0.59

[1]PCA-LDA: Análise de Componente Principal - Análise Discriminante Linear; [2] PCA-QDA: Análise de Componente Principal - Análise Discriminante Quadrática; [3]PCA – SVM: Análise de componentes principais - Máquina de vetores de suporte; [4]SPA-LDA: Algoritmo de projeções sucessivas - Análise Discriminante Linear; [5]SPA-QDA: Algoritmo de projeções sucessivas - Análise Discriminante Quadrática; [6]SPA-SVM: Algoritmo de projeções sucessivas - Máquinas de vetor de suporte; [7]GA-LDA: Algoritmos genéticos - Análise Discriminante Linear; [8]GA-QDA: Algoritmos genéticos - Análise Discriminante Quadrática; [9]GA-SVM: Algoritmos genéticos - Máquinas de vetor de suporte; [10]SIMCA: Modelagem Soft Independente da Analogia de Classes; [11]PLS-DA: mínimos quadrados parciais para análise discriminante; [12]SVM: Máquinas de vetor de suporte

.

Tabela 5. Matriz de confusão para os conjuntos de validação (teste) e validação externa do desempenho do modelo PCA-QDA das folhas de todas as cultivares combinadas, separadas de acordo com o tipo sexual do mamoeiro, femininas e hermafroditas.

		Teste		Validação externa	
Out Class	Hermafrodita	TP = 41,2% (n=21)	FP = 5,9% (n=3)	TP = 41,2% (n=60)	FP = 5,9% (n=9)
	Feminina	FN = 7,8% (n=4)	TN = 45,1% (n=23)	FN = 7,8% (n=19)	TN = 45,1% (n=44)
		Hermafrodita	Feminina	Hermafrodita	Feminina
		Target Class		Target Class	

TP = True Positive; FP=False Positive; FN=False Negative; TN=True Negative

Figuras

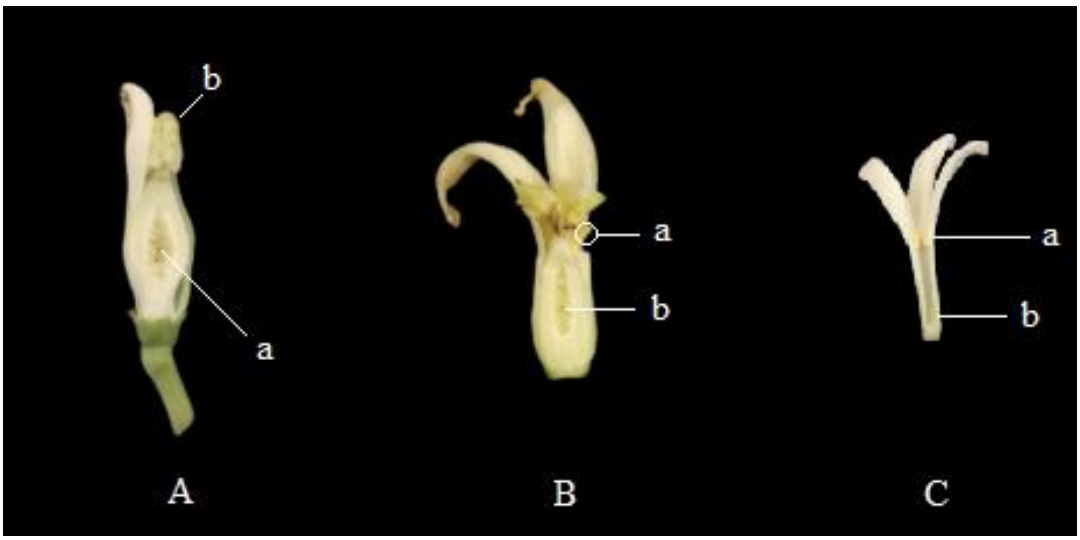

Figura 1. Método do método de referência para a identificação do tipo sexual dos mamoeiros. (A) flores femininas (a – ovário grande e em formado arredondado; b – estigma em forma de leque), (B) flores hermafroditas (a – androceu; b – ovário) e (C) flores masculinas (a – androceu; b – ovário rudimentar).

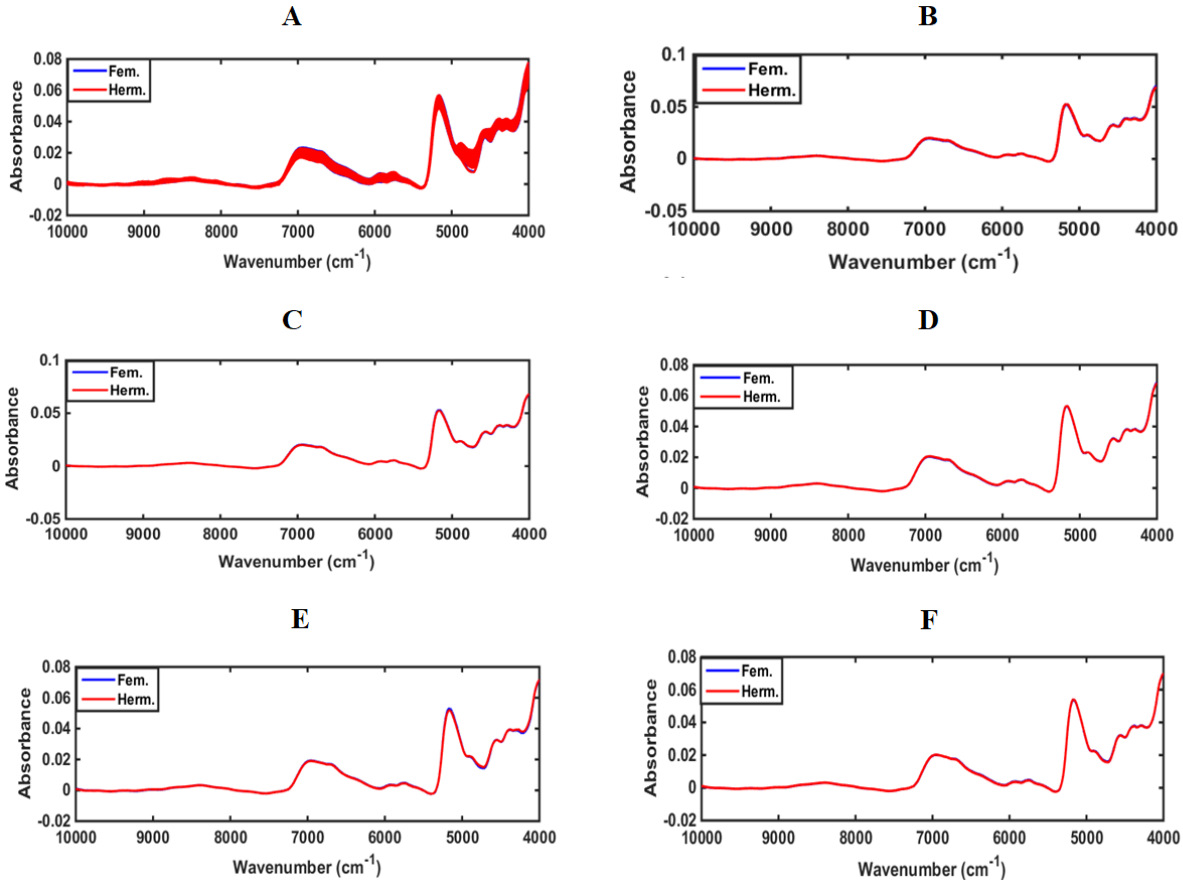

Figura 2. Espectros NIR brutos médios das sementes de mamoeiros. (A) espectros de todas as cultivares combinadas e espectros das cultivares (B) ‘Calimosa’, (C) ‘Formosa’, (D) ‘T2’, (E) ‘Ouro’ e (F) ‘THB’ separadas de acordo com o tipo sexual do mamoeiro (femininas e hermafroditas).

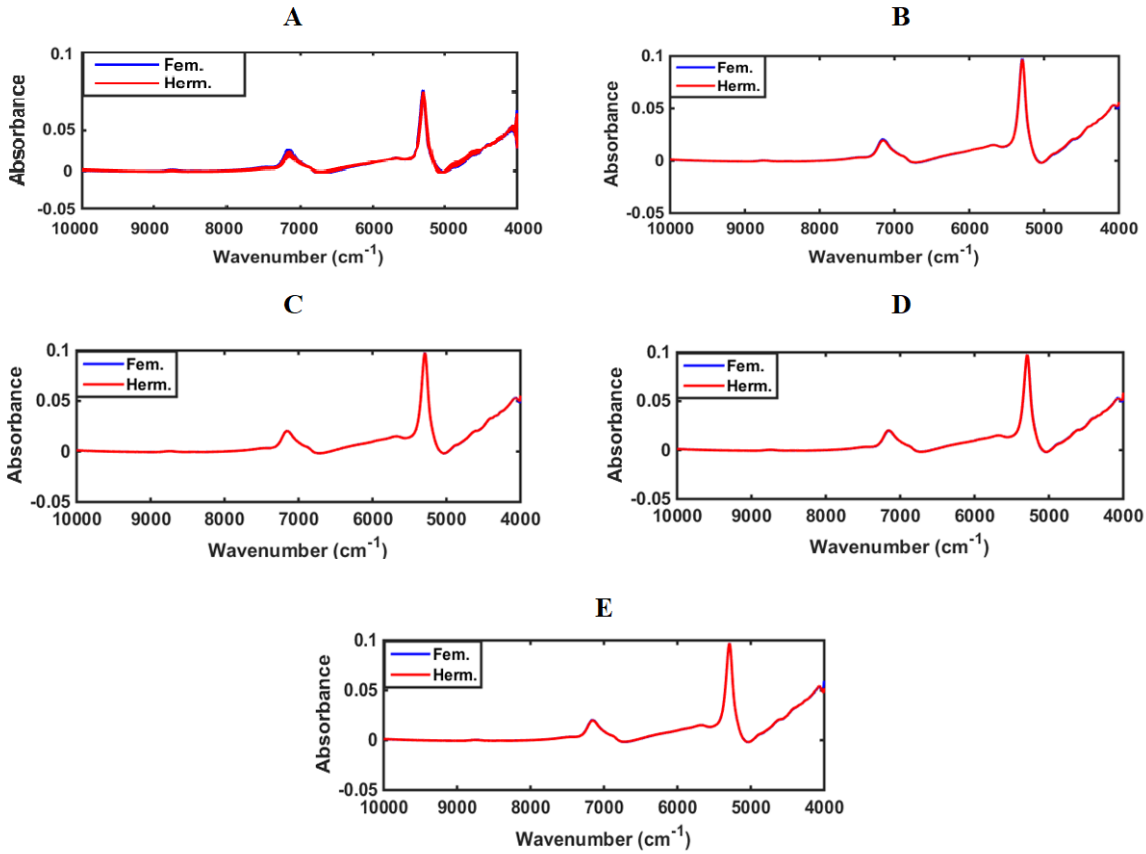

Figura 3. Espectros NIR brutos médios das folhas de mamoeiros. (A) espectros de todas as cultivares combinadas e espectros médios das cultivares (B) ‘Calimosa’, (C) ‘Formosa’, (D) ‘T2’ e (E) ‘THB’ separadas de acordo com o tipo sexual do mamoeiro (femininas e hermafroditas).

CAPÍTULO 3 – Short communication: Classificação de cultivares de mamoeiros por espectroscopia no infravermelho próximo combinada à PLS-DA e machine learning

Resumo - O mamoeiro (*Carica papaya* L.) é uma importante espécie frutífera tropical cujas cultivares pertencem a dois grupos, o Solo e o Formosa. As cultivares apresentam características distintas que são geralmente diferenciadas por meio de descritores morfológicos. Todavia, este método é considerado limitado para a cultura do mamoeiro, pois a baixa diversidade genética desta espécie dificulta a identificação fenotípica das plantas. Desta forma, a espectroscopia no infravermelho próximo (NIR) poderia ser alternativa à classificação destes materiais, pois os espectros NIR são fontes de informações para a identificação qualitativa de amostras. Assim, este trabalho teve por objetivo verificar a possibilidade de se utilizar a espectroscopia NIR em amostras de sementes e folhas de *seedlings* de diferentes cultivares de mamoeiros visando a sua classificação. Foram utilizadas sementes e folhas dos *seedlings* das cultivares 'T2', 'Formosa' e 'Calimosa', do grupo Formosa, e 'THB' e 'Ouro', do grupo Solo. Ao se utilizar as sementes, foi possível obter bons valores de F-score apenas para as cultivares 'THB' (0,85) e 'Ouro' (0,77) no grupo de validação externa utilizando regressão por mínimos quadrados parciais e análise discriminante (PLS-DA). Para as demais cultivares não foi possível obter bons modelos de classificação utilizando as sementes e as folhas. Conclui-se que é possível a classificação das cultivares de mamoeiro 'THB' e 'Ouro' utilizando sementes por meio da espectroscopia NIR combinada à técnica multivariada de calibração PLS-DA. Os modelos gerados podem ser utilizados para a certificação da origem do material vegetal, bem como no controle de fraudes. Entretanto, são necessários mais estudos visando melhorar o desempenho dos modelos desenvolvidos, por exemplo com uso de seleção de variáveis.

Palavras-chaves: *Carica papaya* L, NIR, sementes, folhas, fraudes, autenticação.

1. Introdução

O mamoeiro (*Carica papaya* L.) pertence à família Caricaceae e é o único membro do gênero *Carica* (Paull and Duarte, 2011). Esta é uma das espécie frutífera tropicais de maior interesse econômico e nutricional (Orrillo et al., 2019) e, em 2019, foram produzidas 13,73 milhões de toneladas (t) deste fruto em todo o mundo (Faostat, 2021b). Apesar de ter se originado na América tropical, atualmente a Índia é o maior produtor mundial de mamões (6,05 milhões de t), seguido da República Dominicana (1,17 milhões de t) e do Brasil (1,16 milhões de t), (Faostat, 2021b).

As cultivares existentes de mamoeiros são classificadas em dois grupos, ou seja, o grupo Solo e o Formosa (Dantas et al., 2013). Os mamoeiros do grupo Solo são os mais cultivados no mundo em função das altas produtividades das plantas, boas características organolépticas de seus frutos, além deste serem de tamanho médio (350 - 600 g) preferido pelos consumidores. Os mamoeiros do grupo Formosa também são bastante produtivos e seus frutos têm ótima qualidade organoléptica, porém estes são bem maiores com massa variando de 800 a 1.100 g (Dantas et al., 2013).

As cultivares destes grupos apresentam características distintas e podem ser oriundas de linhagens ou linhas puras com elevado nível de homozigose, bem como de híbridos resultante do cruzamento entre parentais diferentes. Dessa forma, para testar e garantir que as cultivares são inédita e até mesmo para proteger os direitos dos desenvolvedores, são utilizados descritores morfológicos como caracterização descritiva agronômica destes materiais (Sandoval et al., 2017).

Os descritores morfológicos são tradicionalmente utilizados por constituírem um método simples e confiável de identificação de cultivares (Xu et al., 2009). Todavia, este método é considerado limitado para a cultura do mamoeiro, pois a baixa diversidade genética desta espécie dificulta a identificação fenotípica das plantas (Santana et al., 2004). Desta forma, métodos mais acurados de diferenciação das cultivares de mamoeiros vem sendo empregados. Por exemplo, o uso de marcadores moleculares, tais como Random Amplification of Polymorphic DNA (RAPD) e microssatélites (Inter Simple Sequence Repeats – ISSR e Simple Sequence Repeats - SSR) foram relatados por Degel Barbosa et al. (2011) e Ming et

al. (2007a), respectivamente. Contudo, estes métodos são destrutivos, geram resíduos químicos e demandam tempo e necessitam de laboratórios especializados para sua execução.

Neste sentindo, o desenvolvimento de métodos rápidos, não destrutivos, robustos e precisos para identificar as diferentes cultivares de mamoeiro se faz necessário. A espectroscopia no infravermelho próximo (NIR) poderia ser uma técnica analítica a ser utilizada, pois os espectros NIR são fontes de informações para a identificação qualitativa de amostras (Pasquini, 2003). Dessa formar, estudo recentes tem demonstrado a possibilidade de diferenciação de cultivares de soja, arroz e feijão (Singh et al., 2018), de cultivares de milho doce (Qiu et al., 2019) e nozes de macadâmia (Rahman et al., 2021).

Entretanto, na literatura há ausência de trabalhos utilizando a espectroscopia NIR para a identificação de cultivares de mamoeiros. Da mesma forma, para uma melhor discriminação das cultivares é necessário o desenvolvimento de modelos multivariados de classificação (Pasquini, 2003). Assim, às informações espectrais devem ser combinada a análises quimiométricas, tais como a regressão por mínimos quadrados parciais e análise discriminante (PLS-DA), pois esta é um método de classificação linear (Ballabio and Consonni, 2013) supervisionado aplicado à espectroscopia NIR que apresentar forte correlação entre a característica a ser predita (cultivares) e o espectro NIR (Kusumaningrum et al., 2017). Da mesma forma, por ser um método de classificação não linearizado (Balabin and Lomakina, 2011), a máquina de suporte de vetores (SVM) também pode ser empregado para gerar modelos de classificação.

Desta forma, este trabalho teve por objetivo verificar a possibilidade de se utilizar a espectroscopia NIR em amostras de sementes e folhas de *seedlings* de diferentes cultivares de mamoeiros visando a sua classificação. Os modelos gerados poderiam ser utilizados para a certificação da origem do material vegetal, bem como no controle de fraudes.

2. Material e Métodos

2.1 Material Vegetal

Sementes de mamoeiro (*Carica papaya* L.) das cultivares ‘T2’, ‘Formosa’ e ‘Calimosa’, do grupo Formosa, e ‘THB’ e ‘Ouro’, do grupo Solo, foram fornecidas pela empresa Feltrin® Sementes (Farroupilha, Brasil). Estas foram divididas em dois lotes visando constituir o conjunto de calibração (treinamento) e de validação externa (Tabela 1). Além das sementes, foram utilizados também as folhas recém maduras que apresentavam três folíolos dos *seedlings* de cada semente.

2.2 Aquisição dos espectros NIR

A aquisição dos espectros NIR foi realizada nas sementes e folhas de seus respectivos *seedlings* no conjunto de calibração (treinamento). Em seguida, o segundo lote de sementes e das folhas dos respectivos *seedlings* foram utilizados para a obtenção dos espectros NIR (conjunto de validação externa). Os espectros foram coletados da seguinte forma:

2.2.1 Sementes

Para a coleta dos espectros NIR foram utilizadas 150 sementes de cada uma das cultivares anteriormente descritas, o que totalizou 750 sementes para o conjunto de treinamento. Para o conjunto de validação externa foram coletados 350 espectros das sementes, 70 de cada cultivar. Estas foram agrupadas por cultivares e armazenadas em becker de vidro, sendo transferidos para um dessecador de vidro por 24 horas à temperatura ambiente (~25 °C) visando a uniformização do teor de umidade das sementes.

Em seguida, cada semente, previamente identificada com numeração única, foi transferida para um suporte metálico acoplado ao acessório de reflectância difusa *Near Infrared Reflectance Accessory* (NIRA) do espectrofotômetro FT-IR Spectrum 100N (PerkinElmer, Shelton, Estados Unidos) previamente calibrado, e os espectros NIR foram obtidos duas vezes, sendo as sementes reviradas aleatoriamente após cada leitura. O suporte foi colocado diretamente na saída do feixe de luz do NIRA no intuito de se evitar a entrada de luz externa.

2.2.2 Folhas

Após a obtenção dos espectros NIR das sementes, estas foram individualmente semeadas a dois centímetros de profundidade em sacos plásticos pretos medindo 15 cm de diâmetro x 20 cm de altura, contendo o substrato comercial composto de cascas de pinus e eucaliptos (Multiplant® Citrus), e estes foram identificados visando associar cada semente ao seu respectivo *seedlings*.

À medida que os *seedlings* apresentavam as folhas totalmente expandidas contendo três folíolos, esses foram limpos e secas com papel macio e depois encaminhado para laboratório e mantidas à temperatura ambientes (~25 °C) antes de todas as coletas espectrais. Posteriormente, procedeu as coletadas de dois espectros NIR em folhas alternadas, nas posições da segunda lâmina foliar e próximo da nervura central, utilizando o acessório de fibra óptica do espectrofotômetro FT-IR Spectrum 100N (PerkinElmer, Shelton, Estados Unidos), colocando o mais próximo possível da superfície das folhas.

As amostras de sementes e folhas foram submetidas a análise NIR e os dados espectrais gerados ocorreram por meio do detector FT-IR (*Fourier-Transform Near-Infrared*). Previamente à coleta de espectros, o equipamento foi calibrado utilizando o padrão Spectralon® e após 150 leituras espectrais, realizava-se uma nova calibração do equipamento. Os dados foram obtidos na faixa espectral entre 4.000 a 10.000 cm^{-1}, sendo realizados 64 *scans*, com resolução espectral de 16 cm^{-1} e um intervalo de 2 cm^{-1}. Os espectros foram coletados como log(1/R), onde R é a refletância relativa (Miralbés, 2008).

2.3 Quimiometria

Os dados espectrais foram todos foram todos importados e processados em ambiente MATLAB® versão 8.4 (R2014b) (The MathWorks Inc., Natick, MA, EUA) juntamente com a plataforma PLS Toolbox versão 7.9.3 (Eigenvector Research, Inc, Manson, WA, EUA) e algoritmos feitos no laboratório de pesquisa.

Previamente a construção dos modelos quimiométricas multivariados, os dados espectrais das amostras referentes às sementes e às folhas foram submetidos a pré-processamentos, sendo estes avaliados pelos resultados para o conjunto de validação cruzada pelo método venetian blinds (CV VB) com 5 splits.

Para as sementes, o melhor conjunto de validação interna foi obtido através da suavização espectral de Savitzky-Golay (Savitzky and Golay, 1964) com 21 pontos de janela e ordem polinomial de segunda ordem, seguido de uma correção de linha de base do tipo AWLS com normalização vetorial, sendo este o pré-processamento selecionado. Já para os dados obtidos a partir das folhas, foi utilizada a primeira derivada SG (31 pontos de janela, segunda ordem polinomial), seguido de AWLS correção de linha de base com normalização vetorial.

Para os dados de sementes e folhas, todo o espectro na região de 10.000 – 4.000 cm^{-1} foi utilizado. Para todos os casos, a média espectral foi utilizado como espectro representativo para cada amostra. Similarmente, foram utilizados um conjunto de treinamento, composto de 70% dos dados iniciais e o conjunto de validação interna, composto de 30% dos dados (Pasquini, 2018). O primeiro subconjunto incluiu a maioria das amostras e foi o conjunto utilizado para a construção e otimização dos modelos supervisionados. O segundo subconjunto, foi composto pelas demais amostras e serviu para a avaliação parcial da performance dos modelos.

Os dados foram organizados de forma a se obter uma matriz final com a presença das cinco cultivares, tanto para o conjunto espectral das sementes e folhas. A seleção das amostras foi realizada a partir do clássico algoritmo de seleção de Kennard and Stone (1969). Este foi aplicado, com o objetivo de reduzir o viés na divisão entre os conjuntos de treinamento e validação interna dos modelos. A divisão foi realizada de forma a se obter porcentagens semelhantes, para as duas classes, em cada uma das cultivares, construindo assim um modelo que continha informação proporcional relacionada tanto às diferentes cultivares (Tabela 1).

Para testar a real eficácia dos modelos de sementes e folhas, foi utilizado integralmente o segundo lote amostral, no processo de validação externa, sendo os resultados estatísticos para este conjunto os mais importantes e representativos da real eficácia dos modelos.

2.4 Modelos supervisionados

Para o desenvolvimento dos modelos supervisionados foi utilizado a regressão por mínimos quadrados parciais com análise discriminante (PLS-DA) e

máquinas de vetores de suporte (SVM). A análise PLS-DA foi escolhida por ser uma técnica multivariada de classificação linear (Wong et al., 2013) que separa as amostras em grupos conhecidos e prevê classes de novas amostras (Sjöström et al., 1986). Os modelos de classificação baseado no PLS-DA utilizam a relação entre uma matriz **X** multivariada independente e um vetor de resposta que permite um valor 0 ou 1 que indica a qual classe a amostra pertence, ou seja, se a classe pertencer a amostra é classificada com 1, caso contrário é classificada como 0 (Ballabio and Consonni, 2013).

Adotou-se também a classificação por meio SVM pois nem sempre a PLS-DA produz resultados positivos quando se trata de matrizes complexas (Cortes and Vapnik, 1995). A SVM assume uma relação não-linearidade dos dados, efetivando uma mudança de espaço através de uma função de núcleo, assim modificando a estrutura dos dados, classificando as amostras através de hiperplanos que têm o objetivo de maximizar as diferenças entre os grupos. Neste estudo, apenas a função de base radial (RBF) foi utilizada por ser costumeiramente utilizada em problemas de classificação binária, apresentando resultados satisfatórios em diversos casos. Para esta função, temos a seguinte equação (Mazumder et al., 2011):

$$\mathrm{k}(\mathbf{x}_\mathrm{i}, \mathbf{z}_\mathrm{j}) = \exp(-\gamma \left\| \mathbf{x}_\mathrm{i} - \mathbf{z}_\mathrm{j}^2 \right\|) \quad (1)$$

Na qual, $\mathbf{x}_\mathrm{i}$ e $\mathbf{z}_\mathrm{j}$ são vetores de medidas para as amostras e γ é um parâmetro para ajuste para espaçamento da função RBF.

Já o classificador SVM é obtido através da seguinte função:

$$\mathrm{f(x)} = \mathrm{sign}\left(\sum_{\mathrm{i}=1}^{\mathrm{N_{SV}}} \alpha_\mathrm{i} \mathrm{y_i k}(\mathbf{x}_\mathrm{i}, \mathbf{z}_\mathrm{j}) + \mathrm{b}\right) \quad (2)$$

Em que, N_{SV} é o número de vetores de suporte; α_i representa o multiplicador de Lagrange, y_i é a classe, podendo assumir o valor de ± 1, $\mathrm{k}(\mathbf{x}_\mathrm{i}, \mathbf{z}_\mathrm{j})$ sendo uma função de núcleo, definido pela equação (x) e b é a tendência (bias) do parâmetro. Funcionando muito bem para determinação de classes a partir de muitas variáveis, é rápido, também é eficiente para separações não lineares. Contudo, a desvantagem desse método é a demora no processamento quando se tem um grande conjunto de repetições, não trabalhando bem com variáveis com muitos ruídos.

2.4.1 Avaliação dos modelos

Os modelos foram avaliados quanto a sua eficácia através de parâmetros estatísticos. Tais parâmetros medem a capacidade de predição de um dado modelo através de equações, onde seus parâmetros são obtidos por meio dos resultados encontrados para um dado conjunto de teste, que, neste caso, foram considerados o conjunto de validação cruzada (teste) e de validação externa. Dentre estes parâmetros, foram escolhidos a sensibilidade e especificidade, que se relacionam à proporção de amostras positivas e negativas. Além destes, foi avaliado a média harmônica entre sensibilidade e especificidade, conhecida como F-score, representando a acurácia real do modelo, quando é levada em conta os diferentes tamanhos das classes, calculada a partir dos valores de sensibilidade e especificidade. Estes parâmetros foram determinados a partir das seguintes equações:

$$Sensibilidade\ (\%) = \left(\frac{TP}{TP+FN}\right)\ x\ 100 \quad (3)$$

$$Especificidade\ (\%) = \left(\frac{TN}{TN+FP}\right)\ x\ 100 \quad (4)$$

$$F-score\ (\%) = \frac{2\ x\ Sensitivity\ x\ Specificity}{Sensitivity+Specificity} \quad (5)$$

Na qual os valores de TP e TN representam o número de verdadeiros positivos (TP) e negativos (TN), enquanto FP e FN equivalem ao número de falsos positivos (FP) e negativos (FN), respectivamente. Todos estes valores foram obtidos a partir do conjunto de calibração (treinamento) e de validação externa.

3. Resultados

3.1 Sementes

3.1.1 Espectros NIR

Os espectros NIR brutos das sementes das diferentes cultivares de mamoeiros apresentaram semelhanças entre si sem mostrarem diferenças em função das cultivares (Figura 1A). Por outro lado, foi possível constatar que os espectros NIR médios das três cultivares do grupo Formosa (‘T2’, ‘Formosa’ e ‘Calimosa’) apresentaram-se diferentes em relação às cultivares ‘Ouro’ e ‘THB’ que são do grupo Solo (Figura 1B).

Em relação as posições e intensidades das bandas de absorção, foram observados três picos nas regiões situadas à 6.922 cm^{-1}, 5.178 cm^{-1} e 4.000 cm^{-1}. Dessa forma, as maiores diferenças entre as amostras foram observadas em 6.922 cm^{-1} e 5.178 cm^{-1}, ou seja, a amostras da cultivar ‘THB’ apresentaram maiores absorbâncias em 6.922 cm^{-1} e as da cultivar ‘Ouro’ as menos em 5.178 cm^{-1}, o que as diferenciou das demais (Figura 1B).

3.2.2 Desenvolvimento dos modelos de classificação

Não foi possível obter bons modelos de classificação para todas as cultivares, porém para as cultivares do grupo Solo (‘THB’ e ‘Ouro’) foram observados excelentes resultados de sensibilidade, especificidade e F-score ao se utilizar o PLS-DA para o conjunto de validação cruzada (Tabela 2). Todavia, ao se utilizar o PLS-DA no conjunto de validação externa foi observado uma piora na performance na discriminação das cultivares, sendo observados valores de F-score de 0,74, 0,00, 0,60, 0,77 e 0,85 para as cultivares ‘Calimosa’, ‘Formosa’, ‘T2’, ‘Ouro’ e ‘THB’, respectivamente (Tabela 2). Vale destacar que os valores de sensibilidade, ou seja, o número de verdadeiros positivos (TP) foi de 0,97 para as cultivares ‘Calimosa’ e ‘THB’ (Tabela 2).

Ao se aplicar o SVM os resultados de F-score foram ainda mais baixos, sendo observados valores de 0,29, 0,06, 0,36, 0,68 e 0,62 para as cultivares ‘Calimosa’, ‘Formosa’, ‘T2’, ‘Ouro’ e ‘THB’, respectivamente (Tabela 2). Todavia, os valores de sensibilidade para as cultivares ‘Calimosa’ (0,96) e ‘THB’ (0,95) foram muito bons (Tabela 2).

3.2 Folhas

3.2.1 Espectros NIR

Os espectros NIR brutos das folhas dos *seedlings* das diferentes cultivares de mamoeiros se mostraram bastante semelhantes apresentando dois picos nos números de onda de 6.922 cm^{-1} e 5.178 cm^{-1} (Figura 2A). Os espectros NIR médios também não apresentaram diferenças entre as cultivares dos dois grupos de mamoeiros e os picos em 6.922 cm^{-1} e 5.178 cm^{-1} estão associados às bandas de ligações e de vibrações assimétricas do OH, de estiramento e ao primeiro sobretom

da ligação OH, relacionados à presença de água nas folhas (Purcell et al., 2009). Vale ressaltar que não foi possível obter espectros NIR das folhas da cultivar ‘Ouro’ devido à alta infestação por ácaros que provocou alta taxa de mortalidade nos *seedlings* desta cultivar.

3.2.2 Desenvolvimento dos modelos de classificação

Os modelos de classificação para as cultivares utilizado as folhas dos *seedlings* apresentou performance inferior em relação aos desenvolvidos a partir das sementes, não sendo possível a discriminação das amostras em função das cultivares (Tabela 3). Os melhores resultados foram obtidos ao se utilizar o PLS-DA, mesmo assim os valores de F-score foram muito baixos e variaram entre 0,00 (‘Calimosa’) a 0,46 (‘T2’), Tabela 3. O uso do SVM não obteve melhores resultados, pelo contrário, com esta técnica os valores de F-score foram inferiores (0,00 a 0,11), apesar de excelente resultados de sensibilidade terem sido observados para as cultivares ‘Calimosa’ (0,99), ‘Formosa’ (0,98) e ‘THB (1,00) no conjunto de validação externa (Tabela 3).

4. Discussão

4.1 Sementes

Apesar das diferenças nos espectros NIR médios entre as cultivares do grupo Solo (‘Ouro’ e ‘THB’) em relação às do grupo Formosa (‘T2’, ‘Formosa’ e ‘Calimosa’), não foi possível obter bons modelos de classificação para todas as cultivares ao se utilizar o conjunto de validação externa (Tabela 2). Como as maiores diferenças foram observadas nos números de onda de 5.178 cm^{-1} para a cultivar ‘Ouro’ e de 6.922 cm^{-1} para a ‘THB’ (Figura 1B), e os modelos foram desenvolvidos sem a seleção de variáveis, isto pode ter influenciado a performance dos modelos em classificar corretamente as cultivares de mamoeiro.

Mesmo usando o PLS-DA que é um método de classificação supervisionado aplicado à espectroscopia NIR por apresentar forte correlação entre a característica a ser predita (cultivares) e o espectro NIR (Kusumaningrum et al., 2018), esta deveria ter sido associada à seleção de variáveis. Fernandes et al. (2011) relataram

que o uso de seleção de variáveis, ou seja, entre dois e oito comprimentos de onda, associados ao método de algoritmos das projeções sucessivas e regressão linear múltipla (SPA-MLR) resultou em melhores modelos de classificação de biodiesel e erros quadrados médios de predição (RMSEP) de apenas 0,95%. Da mesma forma, Qiu et al. (2019) relataram que a seleção de variáveis não reduziu a performance dos modelos de classificação de sementes de duas cultivares de milho doce ao usarem PLS-DA, sendo observados valores de acurácia de 99,19%.

No tocante a sementes de mamoeiros, estas foram discriminadas corretamente com uma acurácia de 100% em relação a sementes de pimenta-do-reino (*Piper nigrum* L.) com uso da modelagem independente e flexível por analogia de classes (SIMCA) e imagens hiperespectrais no infravermelho próximo (Orrillo et al., 2019). Isto demonstra que estas sementes apresentam padrões espectrais únicos, porém os espectros NIR das sementes das diferentes cultivares foram bastantes semelhantes, o que pode ter influenciado nos resultados.

4.2 Folhas

As folhas dos *seedlings* das diferentes cultivares apresentaram espectros NIR brutos e médios muito semelhantes entre si, sendo estes dominados pelas bandas de vibrações assimétricas do OH (5.158 cm^{-1}) e de estiramento e ao primeiro sobretom da ligação OH (6.984 cm^{-1}) que se relacionam à presença de água nas folhas. Lopes Cruz et al. (2004) relataram que as folhas de mamoeiro apresentam em média 96,9 - 98,6% de umidade. Desta forma, os modelos de classificação das cultivares utilizado as folhas dos *seedlings* apresentaram resultados inferiores (Tabela 3) se comparados aos desenvolvidos com as sementes (Tabela 2).

O mesmo comportamento foi relatado por Haq et al. (2018) ao utilizarem espectros NIR de folhas de mamoeiros sadios e infectados por begomovírus. Todavia, estes autores observaram que as folhas infectadas apresentaram um pico no número de onda 2.391,54 cm^{-1} que não era detectado nas folhas sadias. Assim, os modelos PLS-DA com a seleção desta variável foram eficientes em classificar as folhas sadias em relação às infectadas pelo vírus. Contudo, as folhas dos *seedlings* das diferentes cultivares não apresentaram diferenças espectrais o que pode ter levado aos baixos valores de sensibilidade, especificidade e F-score (Tabela 2).

5. Conclusão

É possível a classificação das cultivares de mamoeiro ‘THB’ e ‘Ouro’ utilizando sementes por meio da espectroscopia no infravermelho próximo (NIR) combinada à técnica multivariada de calibração PLS-DA. No entanto, não foi possível classificar corretamente as demais cultivares utilizando tanto as sementes quanto as folhas dos *seedlings.*

Os modelos gerados podem ser utilizados para a certificação da origem do material vegetal, bem como no controle de fraudes. Entretanto, são necessários mais estudos visando melhorar o desempenho dos modelos desenvolvidos, por exemplo com uso de seleção de variáveis.

6. Referências

Balabin, R.M., Lomakina, E.I., 2011. Support vector machine regression (SVR/LS-SVM) - An alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst 136, 1703–1712. https://doi.org/10.1039/c0an00387e

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: Linear models. PLS-DA. Anal. Methods. https://doi.org/10.1039/c3ay40582f

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/bf00994018

Dantas, J.L.L., Junghans, D.T., Lima, J.F. de, 2013. Mamão: o produtor pergunta, a Embrapa responde., 2ª rev. e atual. ed. Embrapa, Brasília, DF .

Degel Barbosa, C., Pio Viana, A., Silva, S., Quintal, R., Pereira, M.G., 2011. CD Barbosa et al. Brazilian Society of Plant Breeding. Printed in Brazil, Crop Breeding and Applied Biotechnology.

FAOSTAT, F. and A.O. of the U.N.S., 2021. Major tropical fruits - Statistical compendium 2021., in: Statistics Division. Rome, pp. 20–31.

Fernandes, D.D.S., Gomes, A.A., Costa, G.B. Da, Silva, G.W.B.D., Véras, G., 2011. Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection. Talanta 87, 30–34. https://doi.org/10.1016/j.talanta.2011.09.025

Ferreira, M.M.C., 2015. Quimiometria: conceitos, métodos e aplicações, in:

UNICAMP (Ed.), .

Haq, Q.M.I., Mabood, F., Naureen, Z., Al-Harrasi, A., Gilani, S.A., Hussain, J., Jabeen, F., Khan, A., Al-Sabari, R.S.M., Al-khanbashi, F.H.S., Al-Fahdi, A.A.M., Al-Zaabi, A.K.A., Al-Shuraiqi, F.A.M., Al-Bahaisi, I.M., 2018. Application of reflectance spectroscopies (FTIR-ATR & FT-NIR) coupled with multivariate methods for robust in vivo detection of begomovirus infection in papaya leaves. Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 198, 27–32. https://doi.org/10.1016/j.saa.2018.02.065

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

Kusumaningrum, D., Lee, H., Lohumi, S., Mo, C., Kim, M.S., Cho, B.K., 2018. Non-destructive technique for determining the viability of soybean (Glycine max) seeds using FT-NIR spectroscopy. J. Sci. Food Agric. 98, 1734–1742. https://doi.org/10.1002/jsfa.8646

Lopes Cruz, J., Ferreira Coelho, E., Pelacani, C.R., Coelho Filho, M.A., Tosta Dias, A., Taluana Dos Santos, M., 2004. Growth and dry matter and carbon partition in papaya plants in response to nitrogen nutrition. Bragantia 63, 351–361. https://doi.org/10.1590/s0006-87052004000300005

Mazumder, R., Friedman, J.H., Hastie, T., 2011. SparseNet: Coordinate descent with nonconvex penalties. J. Am. Stat. Assoc. 106, 1125–1138. https://doi.org/10.1198/jasa.2011.tm09738

Ming, R., Yu, Q., Moore, P.H., 2007. Sex determination in papaya. Semin. Cell Dev. Biol. https://doi.org/10.1016/j.semcdb.2006.11.013

Miralbés, C., 2008. Discrimination of European wheat varieties using near infrared reflectance spectroscopy. Food Chem. 106, 386–389. https://doi.org/10.1016/j.foodchem.2007.05.090

Orrillo, I., Cruz-Tirado, J.P., Cardenas, A., Oruna, M., Carnero, A., Barbin, D.F., Siche, R., 2019. Hyperspectral imaging as a powerful tool for identification of papaya seeds in black pepper. Food Control 101, 45–52. https://doi.org/10.1016/j.foodcont.2019.02.036

Pasquini, C., 2018. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. Anal. Chim. Acta. https://doi.org/10.1016/j.aca.2018.04.004

Pasquini, C., 2003. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc. https://doi.org/10.1590/S0103-50532003000200006

Paull, R.E., Duarte, O., 2011. Tropical fruits. CABI.

Purcell, D.E., O'shea, M.G., Johnson, R.A., Kokot, S., 2009. Near-infrared spectroscopy for the prediction of disease ratings for fiji leaf gall in sugarcane clones. Appl. Spectrosc. 63, 450–457. https://doi.org/10.1366/000370209787944370

Qiu, G., Lü, E., Wang, N., Lu, H., Wang, F., Zeng, F., 2019. Cultivar classification of single sweet corn seed using fourier transform near-infrared spectroscopy combined with discriminant analysis. Appl. Sci. 9, 1530. https://doi.org/10.3390/app9081530

Rahman, A., Wang, S., Yan, J., Xu, H., 2021. Intact macadamia nut quality assessment using near-infrared spectroscopy and multivariate analysis. J. Food Compos. Anal. 102, 104033. https://doi.org/10.1016/j.jfca.2021.104033

Sandoval, K.V., Vila, D.D., Gracia, T.J.H., 2017. Estudio del mercado de papaya mexicana: un análisis de su competitividad (2001-2015) (Study of the Mexican Papaya Market: An Analysis of Its Competitiveness (2001–2015)). undefined.

Santana, L.R.R., Matsuura, F.C.A.U., Cardoso, R.L., 2004. Genótipos melhorados de mamão (Carica papaya L.): avaliação sensorial e físico-química dos frutos. Ciência e Tecnol. Aliment. 24, 217–222. https://doi.org/10.1590/s0101-20612004000200010

Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047

Singh, S., Patel, S., Litoria, N., Gandhi, K., Faldu, P., Patel, K.G., 2018. Comparative Efficiency of Conventional and NIR Based Technique for Proximate Composition of Pigeon Pea, Soybean and Rice Cultivars. Int. J. Curr. Microbiol. Appl. Sci. 7, 773–782. https://doi.org/10.20546/ijcmas.2018.701.094

Sjöström, M., Wold, S., Söderström, B., 1986. PLS DISCRIMINANT PLOTS, in: Pattern Recognition in Practice. Elsevier, pp. 461–470. https://doi.org/10.1016/b978-0-444-87877-9.50042-x

Wong, K.H., Razmovski-Naumovski, V., Li, K.M., Li, G.Q., Chan, K., 2013. Differentiation of Pueraria lobata and Pueraria thomsonii using partial least square discriminant analysis (PLS-DA). J. Pharm. Biomed. Anal. 84, 5–13. https://doi.org/10.1016/j.jpba.2013.05.040

Xu, H.R., Yu, P., Fu, X.P., Ying, Y. Bin, 2009. On-site variety discrimination of tomato plant using visible-near infrared reflectance spectroscopy. J. Zhejiang Univ. Sci. B 10, 126–132. https://doi.org/10.1631/jzus.B0820200

Tabelas

Tabela. 1. Número de sementes e folhas utilizadas no conjunto de calibração (treinamento), validação cruzada (teste) das cultivares de mamoeiro.

Cultivares	Sementes		Folhas	
	Treinamento	Teste	Treinamento	Teste
Calimosa	95	40	33	14
Formosa	105	45	30	13
T2	116	49	46	20
Ouro	9	3	0	0
THB	23	10	12	4
Total	348	147	121	51

Tabela. 2. Performance dos modelos de classificação desenvolvidos com os espectros NIR das sementes de cinco cultivares de mamoeiro para os conjuntos calibração (treinamento), validação cruzada (teste) e validação externa.

Parâmetros	Cultivares	[1]PLS-DA (10 LVs)			[2]SVM		
		Sensibilidade	**Especificidade**	**F-score**	**Sensibilidade**	**Especificidade**	**F-score**
Treinamento	Calimosa	0.71	0.83	0.77	0.40	0.87	0.55
	Formosa	0.58	0.90	0.71	0.70	0.79	0.74
	T2	0.79	0.92	0.85	0.79	0.81	0.78
	Ouro	1.00	0.98	0.99	0.89	1.00	0.94
	THB	0.96	0.99	0.97	0.39	0.99	0.56
Validação Cruzada	Calimosa	0.63	0.79	0.70	0.37	0.81	0.51
	Formosa	0.51	0.87	0.64	0.54	0.74	0.62
	T2	0.75	0.89	0.81	0.72	0.76	0.74
	Ouro	0.67	0.98	0.80	0.11	1.00	0.20
	THB	0.87	0.98	0.92	0.09	1.00	0.17
Predição	Calimosa	0.68	0.84	0.75	0.38	0.94	0.54
	Formosa	0.62	0.83	0.71	0.73	0.68	0.70
	T2	0.78	0.94	0.85	0.74	0.82	0.78
	Ouro	0.67	1.00	0.80	0.67	1.00	0.80
	THB	0.90	0.98	0.94	0.30	1.00	0.46
Validação externa	**Calimosa**	**0.60**	**0.97**	**0.74**	**0.17**	**0.96**	**0.29**
	Formosa	**0.00**	**0.83**	**0.00**	**0.03**	**0.65**	**0.06**
	T2	**0.48**	**0.81**	**0.60**	**0.23**	**0.83**	**0.36**
	Ouro	**0.77**	**0.77**	**0.77**	**0.65**	**0.71**	**0.68**
	THB	**0.75**	**0.97**	**0.85**	**0.46**	**0.95**	**0.62**

[1]PLS-DA = regressão por mínimos quadrados parciais e análise discriminante; [2]SVM = máquinas de vetor de suporte.

Tabela. 3. Performance dos modelos de classificação desenvolvidos com os espectros NIR das folhas dos seedlings de cinco cultivares de mamoeiro para os conjuntos calibração (treinamento), validação cruzada (teste) e validação externa.

Parâmetros	Cultivares	[1]PLS-DA (10 LVs)			[2]SVM		
		Sensibilidade	Especificidade	F-score	Sensibilidade	Especificidade	F-score
Treinamento	Calimosa	0.39	0.84	0.53	0.09	0.84	0.16
	Formosa	0.43	0.88	0.58	0.10	0.93	0.18
	T2	0.61	0.79	0.69	0.80	0.23	0.36
	Ouro	0.75	0.84	0.79	0.00	1.00	0.00
Validação Cruzada	Calimosa	0.36	0.73	0.48	0.00	1.00	0.00
	Formosa	0.27	0.82	0.41	0.00	1.00	0.00
	T2	0.37	0.65	0.47	1.00	0.00	0.00
	Ouro	0.33	0.87	0.48	0.00	1.00	0.00
Predição	Calimosa	0.43	0.68	0.53	0.14	0.92	0.24
	Formosa	0.23	0.95	0.37	0.00	1.00	0.00
	T2	0.60	0.71	0.65	0.90	0.10	0.18
	Ouro	0.25	0.87	0.39	0.00	1.00	0.00
Validação externa	**Calimosa**	**0.10**	**0.94**	**0.18**	**0.00**	**0.99**	**0.00**
	Formosa	**0.00**	**0.99**	**0.00**	**0.06**	**0.98**	**0.11**
	T2	**0.52**	**0.42**	**0.46**	**0.98**	**0.04**	**0.08**
	Ouro	**0.29**	**0.62**	**0.40**	**0.00**	**1.00**	**0.00**

[1]PLS-DA = regressão por mínimos quadrados parciais e análise discriminante; [2]SVM = máquinas de vetor de suporte.

Figuras

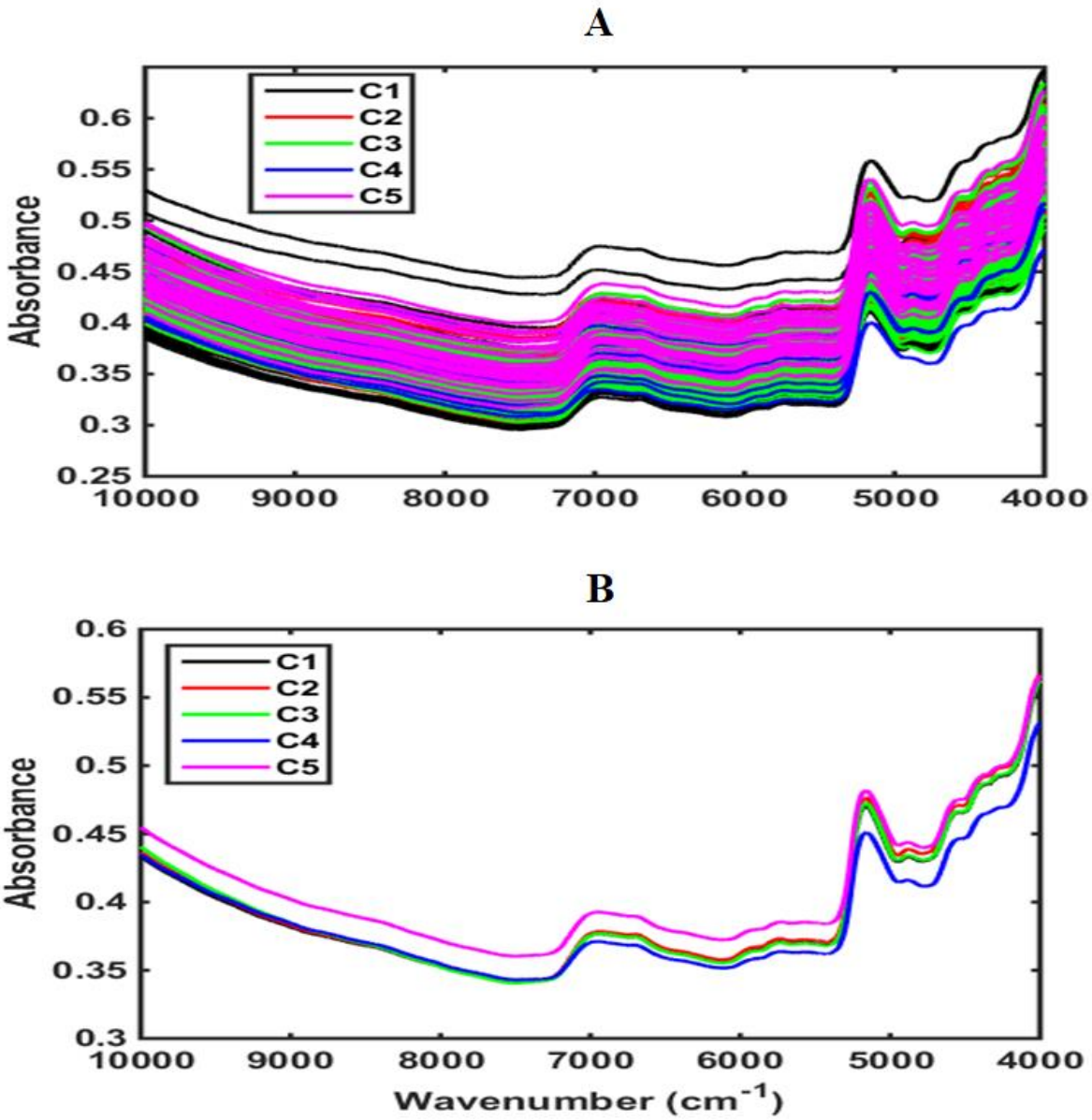

Figura. 1. Espectros NIR brutos totais (A) e médios (B) das sementes de mamoeiros. Espectros de todas as cultivares combinadas e médios das cultivares ‘Calimosa’ (C1), ‘Formosa’ (C2), ‘T2’ (C3), ‘Ouro’ (C4) e ‘THB’ (C5).

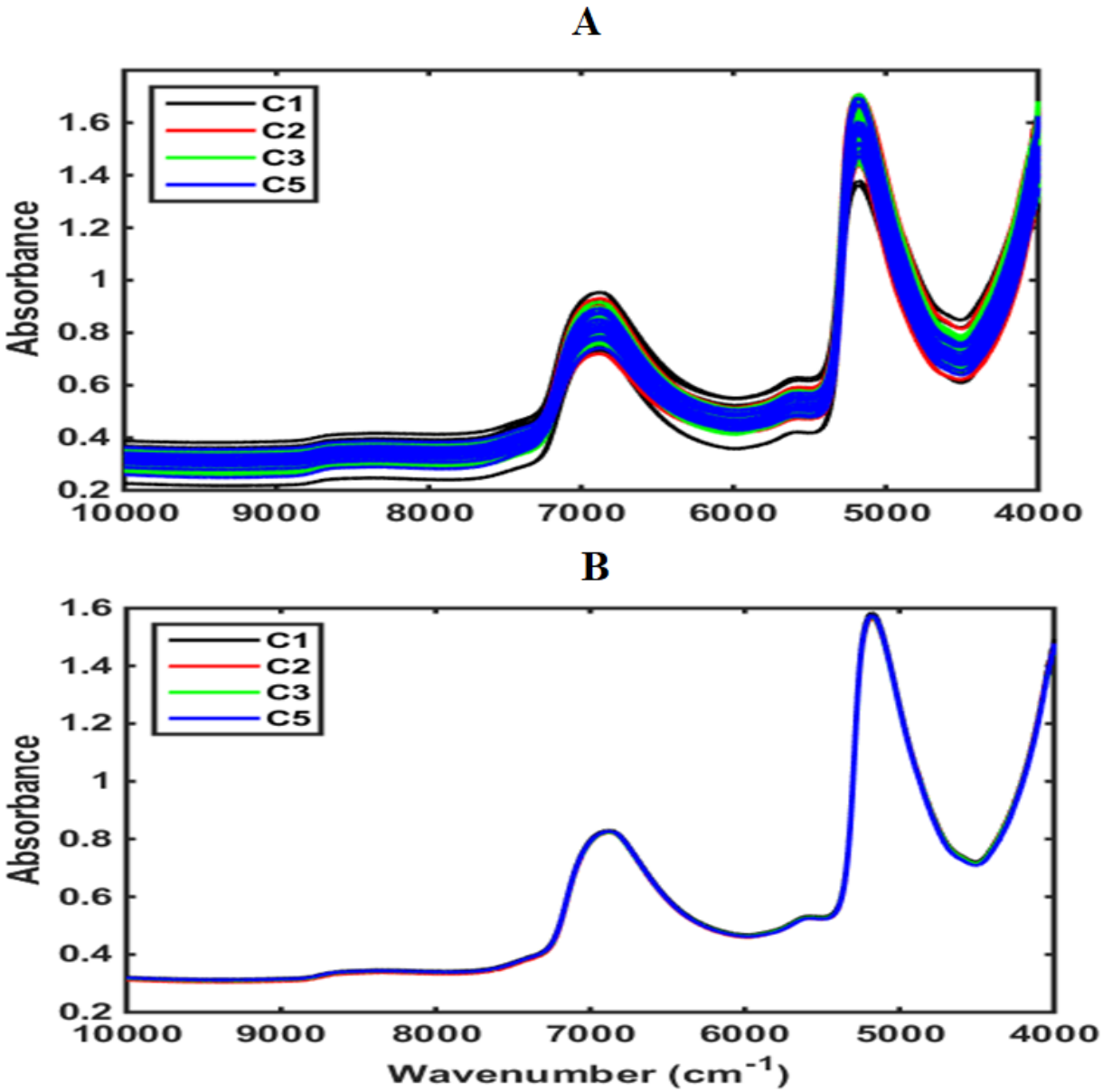

Figura. 2. Espectros NIR brutos totais (A) e médios (B) das folhas de mamoeiros. Espectros de todas as cultivares combinadas e médios das cultivares ‘Calimosa’ (C1), ‘Formosa’ (C2), ‘T2’ (C3), e ‘THB’ (C5).

CAPÍTULO 4 – Considerações Finais

A realização deste projeto de pesquisa permitiu a identificação do tipo sexual e classificação de cultivares de mamoeiros por meio da espectroscopia no infravermelho próximo (NIR) combinadas à diferentes ferramentas matemáticas.

Quando comparadas aos métodos analíticos baseados em análises laboratoriais (marcadores moleculares e/ou de compostos orgânicos) a classificação do tipo sexual e de cultivares com a espectroscopia NIR apresentou como vantagens a sua característica não destrutiva, a possibilidade de se analisar uma grande quantidade de amostras em um tempo reduzido, bem como a redução de custos laboratoriais e operacionais.

Os modelos desenvolvidos poderiam ser utilizados por empresas que produzem e comercializam sementes de mamoeiros, pois com isto seria possível a venda de sementes dos diferentes tipos sexuais separadamente, com consequente agregação de valor. Da mesma forma, os modelos poderiam ser utilizados para a averiguação de fraudes, especialmente no tocante ao material genético (cultivares).

Similarmente, os modelos poderiam ser transferidos para espectrômetros NIR portáteis e assim, seria possível a classificação das mudas de mamoeiros em condições de viveiro tanto em função do tipo sexual quanto em relação às cultivares.

Contudo, é importante salientar que pesquisas futuras precisam ser desenvolvidas visando testar o desempenho dos modelos desenvolvidos com novas fontes de variação, tais como: diferentes cultivares, variações das safras e condições de cultivo. Além disso, os fatores ambientais, incluindo o clima, solo e atmosfera, também podem afetar as características fenotípicos das cultivares. Desta forma, estudos futuros também devem considerar esses fatores.

Por fim, a realização deste trabalho pode identificar lacunas a serem respondidas em trabalhos futuros, como: incorporação de amostras de plantas com inflorescência masculinas; determinação de compostos químicos presentes nas sementes, desenvolvimentos de modelos multivariados não lineares e seleção de variáveis.