



**UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA**

TESE DE DOUTORADO

**FENOTIPAGEM NÃO DESTRUTIVA USANDO ESPECTROSCOPIA NO
INFRAVERMELHO PRÓXIMO E QUIMIOMETRIA EM SEMENTES DE MAMONA**

Maria Betania Hermenegildo dos Santos

**João Pessoa – PB - Brasil
Fevereiro/2013**



**UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA**

TESE DE DOUTORADO

**FENOTIPAGEM NÃO DESTRUTIVA USANDO ESPECTROSCOPIA NO
INFRAVERMELHO PRÓXIMO E QUIMIOMETRIA EM SEMENTES DE MAMONA**

Maria Betania Hermenegildo dos Santos^{*}

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

Orientador: Prof. Dr. Mário César Ugulino de Araújo

2º Orientador: Prof. Dr. Everaldo Paulo de Medeiros

^{*} Bolsista (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior)

**João Pessoa – PB - Brasil
Fevereiro/2013**

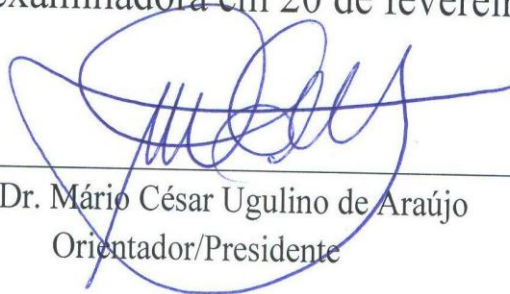
S237f Santos, Maria Betania Hermenegildo dos.
Fenotipagem não destrutiva usando espectroscopia no infravermelho próximo e quimiometria em sementes de mamona / Maria Betania Hermenegildo dos Santos.-- João Pessoa, 2013.
95f. : il.
Orientadores: Mário César Ugulino de Araújo, Everaldo Paulo de Medeiros
Tese (Doutorado) – UFPB/CCEN
1. Química. 2. Espectroscopia NIR. 3. Semente de mamona - classificação. 4. Ricina. 5. Calibração multivariada.

UFPB/BC

CDU: 54(043)

Fenotipagem não Destrutiva usando Espectroscopia no Infravermelho Próximo e Quimiometria em Sementes de Mamona.

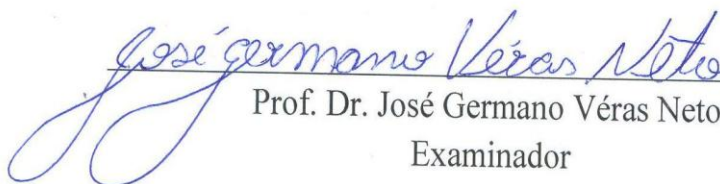
Tese de Doutorado de Maria Betania Hermenegildo dos Santos aprovada pela banca examinadora em 20 de fevereiro de 2013:



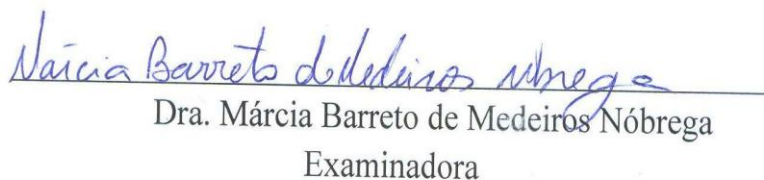
Prof. Dr. Mário César Ugulino de Araújo
Orientador/Presidente



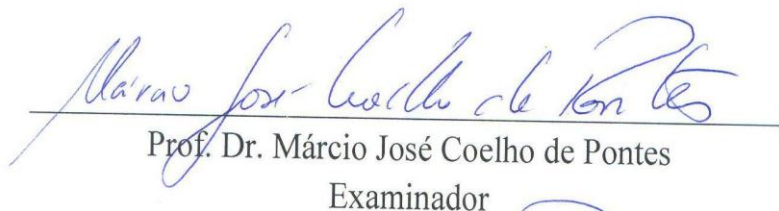
Prof. Dr. Everaldo Paulo de Medeiros
2º Orientador



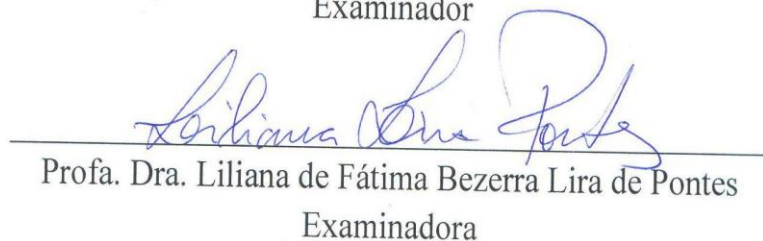
Prof. Dr. José Germano Vêras Neto
Examinador



Dra. Márcia Barreto de Medeiros Nóbrega
Examinadora



Prof. Dr. Márcio José Coelho de Pontes
Examinador



Profa. Dra. Lilitiana de Fátima Bezerra Lira de Pontes
Examinadora

De forma bem especial dedico este trabalho a meu esposo, **Marconi Coelho dos Santos**, por acreditar na minha capacidade, por sempre me incentivar e apoiar e pela compreensão nos diversos momentos em que precisou de mim e eu estava ausente.

Nós conseguimos!

AGRADECIMENTOS

A **Deus**, por me fazer forte, ajudando-me a vencer mais uma etapa.

Ao meu esposo, **Marconi Coelho**, pela paciência e sabedoria transmitidas em momentos difíceis e pelo incentivo em momentos de desânimo e tristeza.

A meus pais, **Maria do Carmo e José Mauricio**, e meus irmãos, **Gutemberg, Danilo, Karla, Kalberta e Karina**, pelo apoio e incentivo durante minha vida.

Ao Professor **Dr. Everaldo Paulo de Medeiros**, pela confiança, orientação, paciência, atenção, conselhos, ensinamentos e, acima de tudo, pela oportunidade na execução deste trabalho.

Ao Professor **Dr. Mário Ugulino**, pela oportunidade de trabalho, orientação, apoio e confiança.

À **Embrapa Algodão**, pela oportunidade de desenvolver este trabalho e por aprimorar meus conhecimentos.

À equipe LATECQ, **João Paulo, Edjane Valéria, Adenilton Silva, Katcilanya Almeida, Lígia Sampaio, Talita Farias, Wesley Pereira, Gustavo Paula, Lydiane Nascimento, Germana Rosy e Clebia França**, em especial a **Pollyne Almeida, Welma Vilar, Ademir Medeiros e Iranilma Maciel**, pela enorme ajuda durante as análises no NIR, irrigação das plantas e análises cromatográficas; sem vocês não teria conseguido.

Aos pesquisadores da Embrapa Algodão, **Máira Milani, Márcia B. M. Nóbrega e Francisco P. de Andrade**, pela colaboração e disponibilidade durante o experimento no campo.

A todos que fazem o LAQA, em especial a **Renato Andrade, Paulo Diniz, Fátima Sanches, Williane Ribeiro, Sófacles Soares, Inakã Barreto e Karla Melo**, pelas sugestões acadêmicas.

A **Adriano Araújo**, pela enorme ajuda nas análises quimiométricas.

Aos **professores** do Departamento de Química da UFPB, em especial aos professores **Mário Ugulino, Sherlan, Edvan Cirino, Regiane Ugulino, Ilda Toscano, Juliana Alves, Ércules Teotônio, Márcio Coelho e Wallace Fragoso**.

Aos **funcionários** do Departamento de Química da UFPB, em especial a **Marcos Pequeno** e a **Danila**.

Aos professores do Departamento de Química da UEPB, em especial aos professores **Germano Véras, Antônio Augusto, Verônica Evangelista, Edilâne Laranjeira e Mary Cristina**.

À **Capex**, pela bolsa concedida durante um ano;

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

MUITO OBRIGADA!

SUMÁRIO

Lista de Figuras.....	xi
Lista de Tabelas.....	xiii
Lista de Abreviaturas e Siglas.....	xiv
Resumo.....	xv
Abstract.....	xvi
1. INTRODUÇÃO	1
1.1. Caracterização do Problema	1
1.2. Objetivos Gerais.....	2
2. FUNDAMENTAÇÃO TEÓRICA	4
2.1. A Cultura da Mamona	4
2.1.1. Origem e Denominação	4
2.1.2. Produtos da Mamona.....	4
2.1.3. Ricina.....	5
2.2. Espectroscopia na Região do Infravermelho	7
2.2.1. Espectroscopia na Região do Infravermelho Próximo	9
2.3. Quimiometria	12
2.4. Pré- Processamento	12
2.5. Métodos de Reconhecimento de Padrões	13
2.5.1. PCA	14
2.5.2. SIMCA	16
2.5.3. LDA.....	18
2.6. Seleção de Variáveis e Amostras	18
2.6.1. Algoritmo para Seleção de Variáveis	18
2.6.1.1. Algoritmo das Projeções Sucessivas.....	19
2.6.2. Seleção de Amostras.....	22
2.7. Calibração Multivariada	24
2.7.1. Regressão Linear Múltipla	25
2.7.2. Regressão em Componentes Principais.....	26

2.7.3. Regressão em Mínimos Quadrados Parciais.....	27
3. CLASSIFICAÇÃO DE SEMENTES DE MAMONA.....	30
3.1. Introdução.....	30
3.2. Objetivos Específicos.....	31
3.3. Experimental.....	32
3.3.1. Aquisição das Amostras	32
3.3.2. Instrumentação	32
3.3.3. Aquisição dos espectros NIR.....	33
3.3.4. Programas Computacionais.....	34
3.3.5. Tratamento Quimiométrico dos Dados	34
3.3.5.1. Pré-processamento	34
3.3.5.2. Reconhecimento de Padrões	35
3.3.6. Método de Referência – Plantio no Campo Experimental	35
3.4. Resultados e Discussão.....	36
3.4.1. Espectros NIR.....	36
3.4.2. Análise Exploratória dos Dados.....	38
3.4.3. Reconhecimento de Padrões Supervisionados	40
3.4.3.1. Construção e Validação dos Modelos SIMCA	40
3.4.3.2. Construção e Validação do Modelo SPA-LDA.....	42
3.4.3.3. Aplicação dos Modelos ao Conjunto de Teste.....	45
3.4.4. Aplicação do Modelo SIMCA as Sementes Plantadas no Campo Experimental.....	46
3.5. Considerações Finais.....	47
4. MODELO DE CALIBRAÇÃO DE RICINA EM SEMENTES DE MAMONA.....	49
4.1. Introdução.....	49
4.2. Objetivo Específico.....	50
4.3. Experimental.....	50
4.3.1. Aquisição de Amostras	50
4.3.2. Instrumentação	50
4.3.3. Preparo de Amostra e Aquisição dos Espectros NIR.....	51
4.3.4. Programas Computacionais.....	52

4.3.5. Tratamento Quimiométricos dos Dados.....	52
4.3.6. Extração, Purificação e Determinação do Teor de Ricina.....	53
4.3.6.1. Obtenção do Extrato Proteico.....	53
4.3.6.2. Purificação da Ricina	54
4.3.6.3. Preparação da Curva de Calibração.....	55
4.4. Resultados e Discussão	55
4.4.1. Espectros NIR.....	55
4.4.2. Pré-processamento dos espectros	56
4.4.3. Construção dos Modelos de Calibração Multivariada	57
4.4.3.1. Modelo de Calibração por PLS.....	57
4.4.3.2. Modelo de calibração por SPA-MLR.....	58
4.4.3.2. Avaliação dos Modelos no Conjunto de Predição	59
4.5. Considerações Finais	61
5. CONCLUSÕES.....	63
5.1. Propostas Futuras	63
REFERÊNCIAS.....	64

Lista de Figuras

Figura 1 -	Estrutura molecular do ácido ricinoleico.....	5
Figura 2 -	2 (a) - Estrutura tridimensional da ricina (2,5 Å). Em verde, a cadeia B; em vermelho, as α -hélices da cadeia A; em laranja, as folhas- β da cadeia A; em cinza, as alças da cadeia A. 2 (b) - Estrutura tridimensional da ricina. Acima, a cadeia A; abaixo a cadeia B; em vermelho, as galactoses; e em verde, as pontes dissulfeto.....	6
Figura 3 -	3 (a) - Diagrama de energia potencial para os osciladores harmônico e 3 (b) – anarmônico.....	8
Figura 4 -	Modos de medição utilizados em espectroscopia NIR. 4 (a) - transmitância; 4 (b) - transflectância; 4 (c) - reflectância difusa, através do meio de dispersão.....	10
Figura 5 -	Sementes das cultivares de mamona, BRS Nordestina e BRS Paraguaçu.....	32
Figura 6 -	Espectrofotômetro VIS-NIR.....	33
Figura 7 -	7 (a) - Célula de quartzo; 7 (b) - Tampas reflexivas para a célula de quartzo.....	33
Figura 8 -	Padrão de reflectância.....	33
Figura 9 -	Plantio das cultivares BRS Paraguaçu e BRS Nordestina no campo experimental.....	36
Figura 10 -	Espectros Originais NIR de reflectância difusa das sementes de mamona, BRS Nordestina e BRS Paraguaçu.....	37
Figura 11 -	Espectros NIR de reflectância difusa pré-processados das 600 sementes de mamona.....	37
Figura 12 -	Gráfico dos escores (PC1 vs PC2) para o conjunto das 600 amostras de sementes de mamona (●) BRS Nordestina e (■) BRS Paraguaçu.....	38
Figura 13 -	Gráfico de pesos de PC1 e PC2.....	39
Figura 14 -	Gráfico de escores (PC1 vs PC2) para o conjunto das 600 amostras de sementes de mamona (●) BRS Nordestina e (■) BRS Paraguaçu; entre parêntese estão indicadas a variância explicada, (a) faixa 1: 1340 – 1460 nm, (b) faixa 2: 1850 - 1930	

	nm, (c) faixa 3: 2110 – 2155 nm e (d) faixa 4: 2200 - 2277 nm.....	40
Figura 15 -	Gráfico dos escores para classe (a) BRS Nordestina e para (b) classe BRS Paraguaçu.....	41
Figura 16 -	Gráfico da porcentagem de variância explicada versus número de PCs incluída no modelo para as classes de (a) BRS Nordestina e (b) BRS Paraguaçu.....	41
Figura 17 -	Gráfico da função do custo associado à seleção de variáveis com o SPA-LDA.....	43
Figura 18 -	Espectro médio das amostras de treinamento. A faixa cinza corresponde ao intervalo usado nos modelos SIMCA e (o) a variável selecionada pelo SPA-LDA.....	43
Figura 19 -	Espectros derivados, com destaque para variável selecionada pelo SPA-LDA.....	44
Figura 20 -	Sinal analítico em 2152,5 nm versus índice das amostras para o conjunto das amostras de treinamento (o) BRS Nordestina e (□) BRS Paraguaçu e validação (o) BRS Nordestina e (□)BRS Paraguaçu. A linha tracejada representa a fronteira de decisão.....	44
Figura 21 -	Sinal analítico em 2152,5 nm versus índice das amostras para o conjunto de teste (o) BRS Nordestina e (□) BRS Paraguaçu, e a linha azul representa a fronteira de decisão estimada para o conjunto de teste.....	46
Figura 22 -	22 (a) - Cultivar BRS Paraguaçu; 22 (b) - Cultivar BRS Nordestina...	46
Figura 23 -	Teste de Germinação das sementes de mamona escarificadas com ácido sulfúrico.....	52
Figura 24 -	Cromatográfico de exclusão molecular da BIO-RAD.....	54
Figura 25 -	Perfil cromatográfico para uma amostra de extrato proteico de um endosperma da mamoneira.....	54
Figura 26 -	Espectro do endosperma da semente da mamona.....	55
Figura 27 -	Conjunto dos 69 espectros das amostras do endosperma da mamona.....	56

Figura 28 -	Espectros derivativos das amostras do endosperma da mamona..	57
Figura 29 -	Gráfico da função de custo SPA-MLR (a) validação externa e (b) validação cruzada.....	58
Figura 30 -	Variáveis selecionadas pelo SPA-MLR (a) validação externa e (b) validação cruzada.....	59
Figura 31 -	Elipse de confiança para os modelos (a) PLS, (b) SPA-MLR, utilizando validação externa e (c) PLS, (d) SPA-MLR, utilizando validação cruzada.....	60

Lista de Tabelas

Tabela 1 -	Número de amostras de treinamento, validação e teste selecionadas pelo algoritmo KS para classes Nordestina e Paraguaçu.....	35
Tabela 2 -	Número de erros de classificação obtido pelos modelos SIMCA utilizando-se o conjunto de amostras de validação das sementes de mamona nos níveis de significância do Teste – F(1%, 5%, 10% e 25%). O número de PCs é indicado entre parênteses.....	42
Tabela 3 -	Resumo da aplicação dos modelos SIMCA e SPA-LDA no conjunto de teste.....	45
Tabela 4 -	Resumo da aplicação dos modelos SIMCA (5% de nível de significância) SPA-LDA no conjunto de sementes plantadas no campo experimental.....	47
Tabela 5 -	Parâmetros da calibração do modelo PLS.....	58
Tabela 6 -	Parâmetros da calibração do modelo SPA-MLR.....	59
Tabela 7 -	Parâmetros estatísticos da predição.....	60

Lista de Abreviaturas e Siglas

- EVD - Decomposição em Autovalores
- FAR - Infravermelho Distante
- F_{cal} - Valor Calculado para o Teste F
- F_{crit} - Valor Crítico Adotado para o Teste F
- HCA - Análise Hierárquica de Agrupamentos
- iPLS: Regressão pelos Mínimos Quadrados Parciais por Intervalo
- KS - Algoritmo Kennard-Stone
- LDA - Análise Discriminante Linear
- MIR - Infravermelho Médio
- MLR - Regressão Linear Múltipla
- NIPALS - Mínimos Quadrados Parciais Iterativos não-lineares
- NIR - Infravermelho Próximo
- PCA - Análise de Componentes Principais
- PCR - Regressão por Componentes Principais
- PCs - Componentes Principais
- PLS - Regressão por Mínimos Quadrados Parciais
- R - Coeficiente de Correlação
- RMSECV - Raiz quadrada do Erro Médio Quadrático de Validação Cruzada
- RMSEP - Raiz quadrada do Erro Médio Quadrático de Predição
- RMSEV - Raiz quadrada do Erro Médio Quadrático de Validação
- SIMCA - Modelagem Independente Flexível por Analogia de Classe
- siPLS- Mínimos Quadrados Parciais em Intervalos Sinérgicos
- SPA - Algoritmo das Projeções Sucessivas
- SPA-LDA - Algoritmo das Projeções Sucessivas em Análise Discriminante Linear
- SPA-MLR - Algoritmo das Projeções Sucessivas em Regressão Linear Múltipla
- SPXY - Partição de Amostra Baseado na Distância de X-y
- SVD - Decomposição em Valores Singulares
- UV-VIS - Ultravioleta – Visível
- VIS-NIR - Visível - Infravermelho Próximo
- OLS – Mínimos Quadrados Ordiniais
- R – coeficiente de correlação

RESUMO

Neste trabalho utilizaram-se a espectroscopia do infravermelho próximo (Near Infrared-NIR) e técnicas quimiométricas para desenvolver modelos de classificação de duas diferentes cultivares comerciais de mamoneira BRS Nordestina e BRS Paraguaçu. Estudou-se também a viabilidade de modelos de calibração para predição do teor de ricina em sementes de três cultivares comerciais de mamoneira (BRS Nordestina, BRS Paraguaçu e BRS Energia). Os espectros de reflectância difusa foram registrados na região de 400 a 2500 nm. Para os modelos de classificação foram utilizadas 350 sementes intactas para cada cultivar. Na calibração o conjunto de amostras foi formado por 69 sementes escarificadas, sendo 25 da BRS Energia, 25 da BRS Nordestina e 19 da BRS Paraguaçu. As leituras foram feitas em quatro posições, para cada semente. Os espectros foram pré-processados com algoritmo Savitzky-Golay com janela de 15 pontos, primeira derivada para correção de linha de base. Com base na PCA (Principal Component Analysis) a região espectral correspondente à faixa de 2110 a 2155 nm, foi selecionada por apresentar distinção entre as cultivares. O modelo SIMCA (Soft Independent Modelling of Class Analogy) forneceu resultados promissores na classificação das sementes para os níveis de significância 1, 5 e 10%. O SPA-LDA (Sucessive Projections Algorithm-Linear Discriminant Analysis) foi eficiente selecionando apenas uma variável na faixa espectral NIR das medidas e classificando corretamente todas as amostras do conjunto de teste. Ao avaliar a precisão dos modelos de calibração SPA-MLR (Sucessive Projections Algorithm-Multiple Linear Regresssion) e PLS (Partial Least Square), usando-se a região elíptica de confiança percebe-se que os mesmos contêm o ponto ideal, quando a técnica utilizada foi a validação externa, isso permite inferir, nesses modelos a ausência de erros sistemáticos significativos. Ao analisar estes modelos usando a técnica de validação cruzada, nota-se que os mesmos não contêm o ponto ideal de acordo com a região elíptica de confiança. Os métodos propostos são promissores para determinar características fenotípicas de forma não destrutiva em genótipos de mamoneira.

Palavras-chave: semente, mamoneira, ricina, espectroscopia NIR, calibração multivariada.

Abstract

In this work we used the near infrared spectroscopy (NIR) and chemometric tools to develop e classification models of two different cultivars of castor bean BRS Nordestina (N) and BRS Paraguaçu (P). It was also studied the feasibility of calibration models for ricin content in seeds prediction of three cultivars of castor bean (BRS Nordestina, BRS Paraguaçu and BRS Energia). Diffuse reflectance spectra were recorded in the region of 400-2500 nm. For classification models were used 350 intact seeds for each cultivar. In the calibration sample set was formed by 69 scarified seeds, 25 of BRS Energia, 25 of BRS Nordestina and 19 of BRS Paraguaçu. Measurements were made at four positions for each seed. The spectra are pre-processed with Savitzky-Golay algorithm with a 15 points window, first derived for baseline correction. Based on PCA (Principal Component Analysis) models, the region corresponding to the spectral range from 2110 to 2155 nm, was selected because it has good distinction between cultivars. SIMCA (Soft Independent Modeling of Class Analogy) model provided promising results in the classification of seed for the significance levels 1, 5 and 10%. The SPA-LDA (Sucessive Projections Algorithm-Linear Discriminant Analysis) was efficient, selecting only one variable in the NIR spectral range of measures, correctly classifying all samples of the test set. When evaluating the accuracy of the calibration models SPA-MLR (Sucessive Projections Algorithm- Multiple Linear Regression) and PLS (Partial Least Square) using the elliptical confidence region it is perceived that they contain the ideal point, when the technique used was the external validation, it allows us to infer, these models lack of significant systematic errors. By analyzing these models using the cross-validation technique, we note that they do not contain the ideal point according to the elliptical region of confidence. The proposed methods are promising for determining phenotypic characteristics in a nondestructively way in castor bean genotypes.

Keywords: seed, castor bean, ricin, NIR spectroscopy, multivariate calibration

CAPÍTULO 1

Introdução e Objetivos

1. INTRODUÇÃO

1.1. Caracterização do Problema

As sementes apresentam duas importantes funções: implantação da cultura e matéria-prima para a indústria. Dentre outros fatores o uso de sementes de boa qualidade e cultivares melhoradas podem definir a produção e a produtividade de uma cultura. Portanto, sementes representam uma tecnologia que envolve, no caso de cultivares, um direito de propriedade intelectual que pode ser de alto valor de mercado (PESKE; BARROS, 2012).

A preservação da variabilidade ou a conservação dos recursos genéticos é considerada uma das questões primordiais para a sobrevivência da humanidade. Esta preservação necessita de classificações para posterior utilização dos genótipos armazenados (MILANI; MIGUEL JÚNIOR; SOUSA, 2009). Segundo os mesmos autores, a correta classificação e identificação das plantas a partir de sementes é uma ferramenta relevante para o melhoramento e desenvolvimento de cultivares que atendam aos diversos agroecossistemas.

A mamona destaca-se dentre as oleaginosas utilizadas para a produção de biodiesel e na indústria química, principalmente como cultura promissora para o semiárido do Brasil em decorrência do alto teor de óleo (45% a 50%), precocidade na produção e relativa resistência ao estresse hídrico (AZEVEDO et al., 2007; CÉSAR; BATALHA, 2010; SEVERINO et al., 2012).

O óleo extraído da semente da mamoneira é matéria-prima para a fabricação de diversos produtos elaborados tais como: cosméticos, sabões, lubrificantes, tintas, plásticos biodegradáveis, fibras sintéticas, além de produtos farmacêuticos. Na biomedicina, este óleo entra na composição de próteses e de implantes e substitui o silicone, como ocorre em cirurgias ósseas, de mama e de próstata. Apesar do mercado ricinoquímico garantir a demanda por este óleo, sua expansão em larga escala se deve ao campo energético dos biocombustíveis (BELTRÃO et al., 2011; SEVERINO et al., 2012).

O principal coproduto gerado a partir da extração do óleo é a torta ou farelo de mamona e, por ser uma excelente fonte de nitrogênio é utilizada como adubo de qualidade. Atualmente, um grande desafio tem sido produzir torta ou farelo para ração animal. Porém, em sua forma natural ela é imprópria, pois apresenta

compostos tóxicos e alergogênicos, que são: a proteína tóxica ricina, o alcaloide ricinina e um complexo alergogênico CB - 1A (LIMA et al., 2011; SEVERINO et al., 2012).

A cultura da mamona possui baixa expansão e um dos entraves é a baixa qualidade do material utilizado para implantação da cultura, pois o cultivo ainda é realizado com sementes dos próprios agricultores, as quais possuem alto grau de heterogeneidade, diversidade e alto polimorfismo. Com isto, ocorrem problemas na produtividade, surgimento de doenças e pragas, maior demanda nos tratamentos culturais e maior tempo gasto na colheita e no beneficiamento (FREIRE et al., 2007).

Diante do exposto surge a necessidade de metodologias analíticas rápidas, não destrutivas, não invasivas, de alta frequência analítica e de baixo custo para fenotipagem de componentes das sementes. Dentre estas metodologias se destaca a espectroscopia no infravermelho próximo (NIR), por ser uma técnica que atende essas características (SIMÕES, 2008; OZAKI, 2012). A espectroscopia NIR é considerada uma poderosa ferramenta para análises quantitativas e qualitativas de variáveis químicas e físicas, podendo ser aplicada às amostras de vários tipos, tais como da indústria de fármacos, de polímeros, produtos petroquímicos, alimentos e agrícolas (OZAKI, 2012).

1.2. Objetivos Gerais

- ✓ Desenvolver uma metodologia para classificação de duas diferentes cultivares comerciais de mamoneira.
- ✓ Estudar a viabilidade de modelos de calibração com medidas no NIR para predição de ricina em sementes de mamoneira.

CAPÍTULO 2

Fundamentação Teórica

2. FUNDAMENTAÇÃO TEÓRICA

2.1. A Cultura da Mamona

2.1.1. Origem e Denominação

A origem da mamoneira (*Ricinus Communis* L.) é incerta em razão de sua ampla adaptação as mais distintas condições climáticas. Apesar de ser uma cultura de regiões áridas e semiáridas, é também encontrada em outros locais (WEISS, 1983; FORNAZIERI JÚNIOR, 1986; SEVERINO et al., 2012).

Alguns estudiosos propõem o continente asiático como provável centro de origem ao passo que outros consideram a África intertropical. A hipótese mais aceita é de que esta espécie seja originária do Nordeste Africano, provavelmente da antiga Abissínia, hoje Etiópia, em virtude da presença de uma elevada diversidade desta planta neste local (BELTRÃO; AZEVEDO, 2007; CHIERICE; CLARO NETO, 2007; ANJANI, 2012).

No Brasil, a introdução da mamoneira se deu durante a colonização portuguesa por ocasião da vinda dos escravos africanos (TÁVORA, 1982; MOREIRA et al., 1996; COSTA et al., 2006; ANJANI, 2012). No País, a planta se adaptou de forma espontânea e asselvajada em várias regiões chegando a ser confundida com uma planta nativa (BELTRÃO et al., 2011).

A mamoneira é uma planta pertencente família Euphorbiaceae, gênero *Ricinus* e espécie *Ricinus communis* L, conhecida no Brasil pelas denominações de mamoneira, carrapateira, palma-de-cristo, enxerida; em inglês, *castor bean* e *castor seed*, em alemão, *wunder-baun* (FORNAZIERI JÚNIOR, 1986; SEVERINO et al., 2012).

2.1.2. Produtos da Mamona

A mamoneira é explorada devido ao óleo que é extraído de suas sementes, além de ser o único solúvel na natureza em álcool metílico e etílico, contêm em sua composição o ácido graxo ricinoleico variando de 80% a 90% (GAJERA et al., 2010; FERNÁNDEZ-CUESTA et al., 2011; YADAVA et al., 2012).

A estrutura química do ácido ricinoleico possui a particularidade de três grupos funcionais altamente reativos: o grupo carbonila, no primeiro carbono; a dupla ligação no nono carbono e o grupo hidroxila, no décimo segundo carbono (**Figura 1**). Esses grupos funcionais fazem com que o óleo de mamona possa ser submetido a diversos processos químicos, nos quais podem ser obtidos muitos produtos ([CANGEMI, SANTOS; CLARO NETO, 2010](#)).

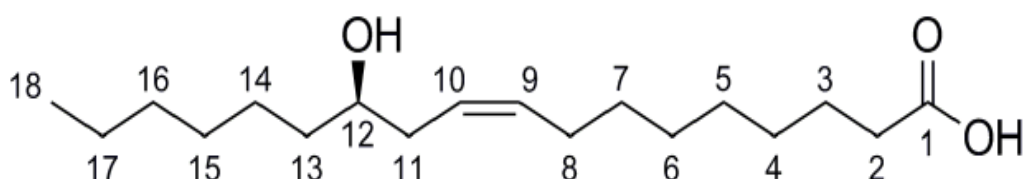


Figura 1 - Estrutura molecular do ácido ricinoleico ([ALBUQUERQUE, 2010](#)).

Apesar da alta toxicidade das sementes o óleo de rícino não é tóxico visto que a ricina não é solúvel em lipídios. Desta forma, todo componente tóxico fica restrito à torta ou farelo ([SEVERINO et al., 2012](#)). Em virtude desta toxicidade a torta da mamona apesar de possuir alto teor de proteínas, não pode ser utilizada diretamente como alimento para animais ([HOFFMAN et al., 2007](#); [SEVERINO et al., 2012](#); [FERNANDES et al., 2012](#)).

A torta vem sendo utilizada como fertilizante de cobertura sem nenhum tipo de tratamento. Ela possui alto teor de fibras, rápida mineralização e é excelente fonte de nitrogênio, fósforo, potássio e cálcio, além de promover o controle de algumas espécies de nematoides ([LIMA et al., 2011](#); [FERNANDES et al., 2012](#)).

2.1.3. Ricina

A ricina é classificada como uma lectina glicoproteína composta de duas cadeias, A e B, unidas por uma ligação de dissulfeto. Ela possui cerca de 60 KDa e representa de 1 a 5% da massa da torta de mamona ([GREENFIELD et al., 2002](#); [SEVERINO et al., 2012](#)).

A estrutura tridimensional da ricina pode ser visualizada na **Figura 2 (a)** e **2 (b)**. Na **Figura 2 (a)**, em verde encontra-se a cadeia B; em vermelho, as α -hélices da cadeia A; em laranja, as folhas- β da cadeia A; em cinza, as alças da cadeia A.

A cadeia A, também chamada RTA, possui predominância do padrão α -hélice (36%) e é dividida em três domínios, visualizados na **Figura 2 (b)**: do resíduo 01 ao resíduo 117 (acinzentado); do resíduo 118 ao 210 (branca) e do resíduo 211 ao 267 (pontilhado). A estrutura secundária folha- β é a de maior quantidade (37%) na cadeia B (RTB). Esta pode ser dividida em dois domínios iguais tridimensionalmente, cada um possui dois pares de pontes dissulfeto e uma galactose (**Figura 2 (b)**) (HALLING et al., 1985; MONTFORT et al., 1987; HARTLEY; LORD, 2004; AUDI et al., 2005).

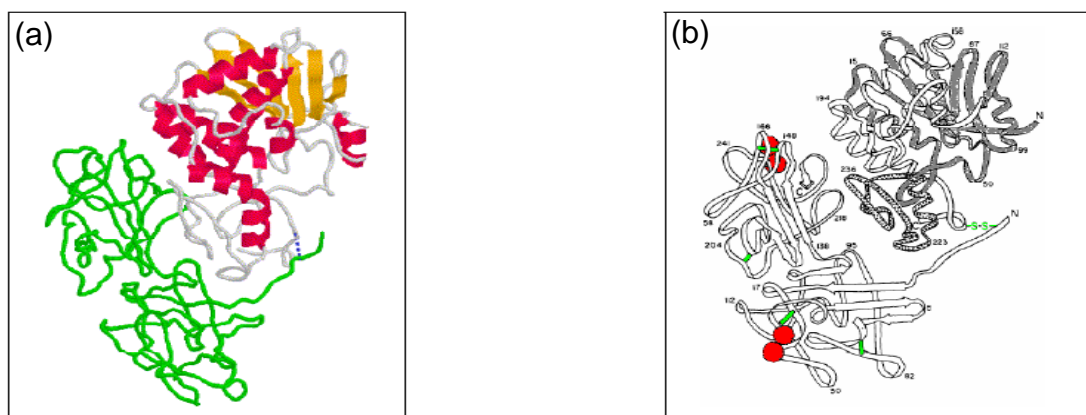


Figura 2 – 2 (a) - Estrutura tridimensional da ricina (2,5 Å). Em verde, a cadeia B; em vermelho, as α -hélices da cadeia A; em laranja, as folhas- β da cadeia A; em cinza, as alças da cadeia A. **2 (b)** - Estrutura tridimensional da ricina. Acima, a cadeia A; abaixo a cadeia B; em vermelho as galactoses e em verde, as pontes dissulfeto (HARTLEY; LORD, 2004).

Enquanto a maioria dos genes envolvidos na síntese e no volume do óleo de rícino são cópias simples o número de genes da família ricina é muito maior do que o que se pensava antes do sequenciamento do genoma da mamona (CHAN et al., 2010). Devido a isto, os programas de melhoramento genéticos se deparam com a dificuldade de desenvolvimento de variedades com baixo teor de ricina. Já que é difícil mutagenizar vários desses genes, simultaneamente, sem causar alterações fenotípicas indesejáveis (HALLING et al., 1985; BALDONI, 2010).

A ricina é uma proteína que tem a função de armazenamento nas sementes, fornecendo nutrientes durante a germinação e atuando como proteína de defesa. Ela é sintetizada como preproricina no desenvolvimento das sementes e se

encontra no lúmen do retículo endoplasmático (RE), quando o peptídico é removido, formando a proricina. No RE é formada uma ligação de dissulfeto intramolecular entre as subunidades A e B, juntando o heterodímero maduro para posterior remoção do propeptideo, gerando o dímero maduro (MALTMAN et al., 2007).

Segundo Baldoni et al. (2011) foi observado entre 20 acessos do banco germoplasma da Embrapa Algodão, teores de ricina entre 3,5 e 32,2 g Kg⁻¹.

A ricina é considerada uma arma potencial de bioterrorismo em razão da sua alta toxicidade e facilidade de produção em laboratório simples (Doan, 2004; Audi et al., 2005). Neste sentido, o Centro Britânico de Controle e Preservação de Doenças classifica a ricina como uma substância de ameaça moderada (tipo B) (CANGEMI, SANTOS, CLARO NETO, 2010).

2.2. Espectroscopia na Região do Infravermelho

A região do infravermelho compreende a radiação eletromagnética com comprimento de onda de 780 a 1.000.000 nm, sendo subdivida em três sub-regiões: infravermelho próximo - NIR (780 – 2.500 nm), infravermelho médio - MIR (2.500 – 50.000 nm) e infravermelho distante - FAR (50.000 – 1.000.000 nm) (SKOOG; HOLLER; NIEMAN, 2009).

A radiação infravermelha causa alteração nos modos rotacionais e vibracionais das moléculas (BARBOSA, 2008). Portanto, é uma técnica que se limita para espécies moleculares com pequenas diferenças de energia entre diversos estados vibracionais e rotacionais (SKOOG; HOLLER; NIEMAN, 2009; EWING, 2011).

Para que uma molécula absorva a radiação infravermelha ela deve possuir uma variação no momento de dipolo, durante seu movimento rotacional ou vibracional (SKOOG; HOLLER; NIEMAN, 2009). Nessas circunstâncias, o campo elétrico alternado da radiação pode interagir com a molécula e ocasionar variações na amplitude de um de seus movimentos.

Moléculas diatômicas heteronucleares como, por exemplo, o cloreto de hidrogênio, possuem um momento de dipolo significativo, isto é, modos vibracionais de absorção ativos no infravermelho. O contrário ocorre em espécies

homonucleares, como O₂, N₂ ou Cl₂, as quais não possuem variação no momento de dipolo tendo, como consequência, a não absorção da radiação infravermelha (BARBOSA, 2008; SKOOG; HOLLER; NIEMAN, 2009; EWING, 2011).

Devido às vibrações e rotações de diferentes tipos que ocorrem nas ligações da molécula, as posições relativas aos átomos, não são fixas, mas oscilam continuamente; assim, essas vibrações podem ser classificadas em estiramento e deformação (SKOOG; HOLLER; NIEMAN, 2009).

A característica da vibração de estiramento é a variação contínua na distância interatômica ao longo do eixo da ligação entre dois átomos; já as vibrações de deformação envolvem uma variação no ângulo entre duas ligações e são de quatro tipos: deformação simétrica no plano, deformação assimétrica no plano, deformação simétrica fora do plano e deformação assimétrica fora do plano (SKOOG; HOLLER; NIEMAN, 2009).

A vibração molecular pode ser descrita por um modelo simples, similar ao de um oscilador harmônico, conforme a **Figura 3 (a)** (PASQUINI, 2003).

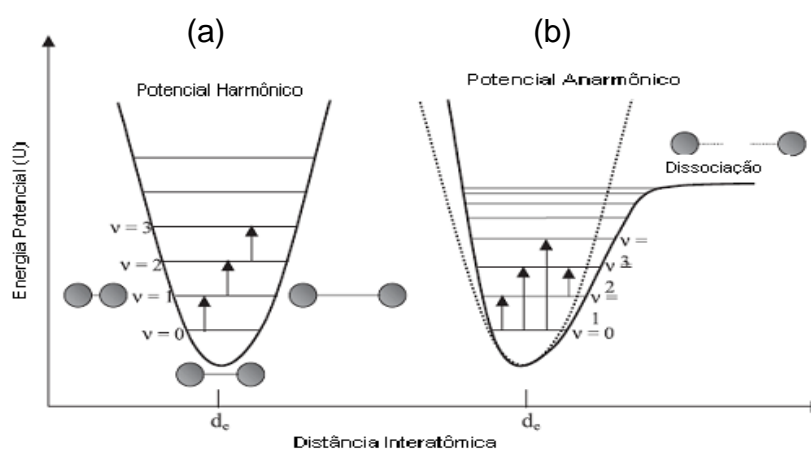


Figura 3 – 3 (a) - Diagrama de energia potencial para os osciladores harmônico e **3 (b)** – anarmônico.

De acordo com o tratamento da mecânica quântica referente ao modelo simples do oscilador harmônico, o nível de energia vibracional entre dois átomos de uma molécula é quantizado segundo a **Equação 1**.

$$E_v = hv \left(1 + \frac{1}{2} \right) \quad (1)$$

Em que:

E_v - energia vibracional,

h - constante de Plank;

ν - frequência vibracional clássica.

O modelo harmônico impõe uma restrição adicional na qual o número quântico vibracional só poderá variar de uma unidade, $\Delta = \pm 1$, ficando proibidas transições entre mais de um nível de energia (PASQUINI, 2003).

Em temperatura ambiente a maioria das moléculas se encontra no nível vibracional fundamental $\nu = 0$ e as transições permitidas $\nu = 0 \rightarrow \nu = 1$, são denominadas transição fundamental ou 1^o harmônico em que este domina o espectro de absorção do infravermelho (SCAFI, 2000; 2005).

Embora o modelo harmônico possa ser útil para entender a espectroscopia vibracional, este modelo não consegue explicar o comportamento de moléculas reais. A principal limitação é não considerar as forças coulômbicas de atração e repulsão nem a dissociação da ligação (NUNES, 2008)

A partir de evidências experimentais as moléculas se comportam como osciladores anarmônicos (**Figura 3 (b)**); neste modelo são permitidas a ocorrência de sobretons (transições com $\Delta\nu \geq \pm 2, \pm 3 \dots$) e a existência de bandas de combinação (CHAGAS, 2006).

2.2.1. Espectroscopia na Região do Infravermelho Próximo

A região espectral NIR compreende um tipo de espectroscopia vibracional que utiliza energia do fóton na faixa de energia de $2,65 \times 10^{-19}$ a $7,96 \times 10^{-20}$ J. Neste intervalo as bandas de absorção são de sobretons ou combinações de vibrações fundamentais de estiramento, que envolvem os grupos funcionais cujas ligações são polarizadas, como C-H, N-H, O-H e S-H (PASQUINI, 2003; SKOOG; HOLLER; NIEMAN, 2009).

A intensidade das bandas de absorção no NIR é cerca de 10 a 1000 vezes mais fracas que sua banda fundamental na região do infravermelho médio (MIR). Isto poderia ocasionar uma desvantagem devido à diminuição da sensibilidade

analítica (LIMA et al., 2009). Entretanto, tal dificuldade pode ser superada com o uso de fontes de radiação interna e detectores de alta eficiência que contribuem para o aumento da relação sinal/ ruído (HONORATO, 2006).

Uma vantagem do NIR é sua baixa absortividade, a qual permite melhor penetração da radiação em amostras sólidas e análises diretas de fortes absorventes como, por exemplo, de líquidos turvos ou sólidos nos modos de reflectância, transmitância ou transfectância, conforme a **Figura 4**, sem necessidade de pré-tratamento da amostra (SIMÕES, 2008; LIRA, 2010).

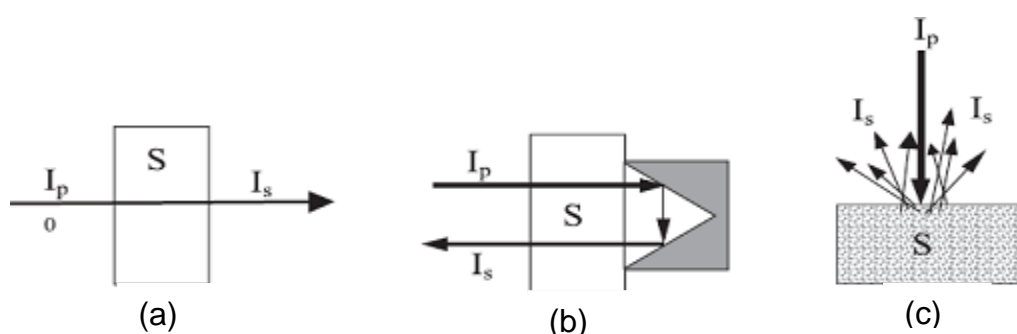


Figura 4 - Modos de medição utilizados em espectroscopia NIR. **4 (a)** transmitância; **4 (b)** transfectância e **4 (c)** reflectância difusa, através do meio de dispersão (PASQUINI, 2003).

Na **Figura 4 (a)** observa-se o modo de transmitância, muito usada na espectrometria UV – VIS convencional. As amostras são medidas em cubetas de vidro ou quartzo com percurso óptico variando de 1 a 50 mm.

O modelo de transfectância é representado na **Figura 4 (b)**. Durante esse tipo de medida usam-se feixes de fibra óptica ou dispositivos para este fim, diferenciando-se das medidas de transmitância pelo caminho óptico duplo.

As medidas de reflectância difusa de amostras sólidas (**Figura 4 (c)**) formam a base das medidas NIR, com predominância dos fenômenos de espalhamento e absorção de partículas sólidas. Para descrever esse comportamento, Kubelka-Munk (KULBELKA; MUNK, 1931) propuseram um modelo empírico que descreve esse tipo de medida, conforme a **Equação 2** mas ela não se aplica no caso de materiais opacos de espessura infinita e não são descritos na lei de Beer.

$$f(C) = \frac{(1-R)^2}{2R} \quad (2)$$

Em que:

C - concentração;

R - reflectância difusa, obtida por:

$$R = \frac{I_R}{I_{R_0}} \quad (3)$$

Sendo:

I_R - intensidade da radiação refletida pela amostra;

I_{R_0} - intensidade refletida por um material de referência padrão.

Este padrão deve ser um material não absorvente, estável, com reflectância absoluta elevada e relativamente constante na região espectral do NIR. Em geral, são empregados, para esta finalidade, o brometo de potássio, o teflon, o sulfato de bário e o óxido de magnésio.

Na prática, a equação de Kubelka-Munk, tal como a lei de Beer, é limitada, sendo aplicada apenas em bandas de absorção de baixa intensidade; no caso do NIR ocorre desvio de linearidade já que não é possível separar a absorção do analito da absorção da matriz. Assim, deve-se substituir a **Equação 2** pela **Equação 4**, em que é utilizada a aplicação de uma relação entre a concentração e a reflectância:

$$f(C) = \text{Log} \frac{1}{R} \quad (4)$$

A **Equação 4** é muito utilizada para o desenvolvimento de métodos analíticos baseados em medidas de reflectância e não se afasta muito da previsão de Kubelka-Munk. Para pequenas alterações na reflectância (R) convencionou-se um comportamento linear com a concentração do analito.

2.3. Quimiometria

A Quimiometria se propõe a solucionar problemas de interesse e origem na química, ainda que as ferramentas de trabalho provenham principalmente da matemática, estatística e computação (BEEBE; PELL; SEASHOLTZ, 1998; FERREIRA et al., 1999).

As abordagens da quimiometria envolvem: planejamento e otimização de experimentos; pré-processamento de dados espectrais; reconhecimento de padrões; seleção de variáveis e amostras; calibração multivariada e transferência de calibração (BEEBE; PELL; SEASHOLTZ, 1998).

2.4. Pré- Processamento

Os dados originais provenientes de técnicas instrumentais podem apresentar alterações não desejadas, como ruídos instrumentais, intensidade com magnitudes diferentes e variação sistemática da linha de base. Essas alterações espectrais, não possuem, normalmente, relação com a composição da amostra e, portanto, não contribuem para os modelos multivariados, sendo necessário sua remoção por meio de técnicas de pré-processamento (MASSART et al., 1997; FERREIRA et al., 1999; BUENO, 2011).

A maior contribuição desta variação pode ser atribuída à falta de estabilidade do instrumento, ao espalhamento da radiação durante a realização das medidas ou à variabilidade das propriedades físicas da amostra (BEEBE; PELL; SEASHOLTZ, 1998).

As técnicas mais usadas no pré-processamento de dados aplicadas no domínio das amostras são: normalização, ponderação, suavização e correção da linha de base (BEEBE; PELL; SEASHOLTZ, 1998).

A normalização é efetuada dividindo-se cada variável por uma constante, a partir de uma análise preliminar dos dados. Na ponderação se atribuem as amostras mais importantes, pesos proporcionais, multiplicando-se cada elemento do vetor amostra pelo seu peso. A técnica de suavização de ruído é usada para aumentar a relação sinal/ruído. Com esta finalidade podem ser utilizados os seguintes filtros digitais: Savitzky-Golay (SAVITZKY; GOLAY, 1964; BEEBE; PELL; SEASHOLTZ,

1998), transformada de Fourier (CERQUEIRA; POPPI; KUBOTA, 2000) e transformada Wavelet (GALVÃO et al., 2001).

As variações sistemáticas não relacionadas com a propriedade de interesse analítico são descritas como feições da linha de base. Elas podem dominar a análise, se não removidas. Para sua correção pode-se usar: derivação e correção multiplicativa de sinais (MSC) (BEEBE; PELL; SEASHOLTZ, 1998).

Nas variáveis podem ser aplicadas três técnicas de pré-processamento: centralização dos dados na média, o escalonamento e o auto-escalonamento. A centralização dos dados na média pode ser definida como a subtração dos elementos de cada linha pela média da sua respectiva coluna. No escalonamento cada elemento de uma linha é dividido pelo desvio padrão da sua respectiva variável, fazendo com que todos os eixos da coordenada sejam conduzidos à mesma escala. O auto-escalonamento consiste em centralizar os dados na média e efetuar o escalonamento. Utilizam-se o escalonamento e o auto-escalonamento quando se pretende atribuir a mesma importância às variáveis do sistema de investigação (MASSART et al., 1997, BEEBE; PELL; SEASHOLTZ, 1998).

2.5. Métodos de Reconhecimento de Padrões

As técnicas de reconhecimento de padrões têm, por finalidade, identificar as semelhanças e diferenças presentes nos diversos tipos de amostras. Essas técnicas se fundamentam nas seguintes suposições: amostras do mesmo tipo são semelhantes, existem diferenças entre tipos variados de amostras. As semelhanças e diferenças são expressas nas medidas utilizadas para caracterizar as amostras (GONZÁLEZ, 2007).

Dentre as técnicas de reconhecimento de padrões, Beebe; Pell; Seasholtz (1998) e González (2007) as classificam em:

✓ Não – supervisionadas: são aquelas usadas para avaliar a existência de similaridade ou diferenças entre as amostras, sem utilizar o conhecimento prévio dos membros das classes. Os principais métodos deste tipo são: análise de agrupamento hierárquico (*Hierarchical Cluster Analysis* – HCA) e análise de componentes principais (*Principal Component Analysis* – PCA) (BARROS NETO; SCARMINIO; BRUNS, 2006).

✓ Supervisionadas: são aquelas usadas para prever se uma amostra desconhecida pertence a uma classe conhecida, a várias classes ou a nenhuma. Para isto, é conveniente uma informação adicional sobre os membros das classes, ou seja, é necessário um conjunto de treinamento com objetos de categorias conhecidas para a elaboração de modelos que sejam capazes de identificar amostras desconhecidas. Dentre as técnicas de reconhecimento de padrões supervisionadas podem ser citadas: a modelagem independente e flexível por analogia de classes (*Soft Independent Modeling of Class Analogy* - SIMCA) e a análise discriminante linear (*Linear Discriminant Analysis* - LDA) (BRUNS; FAIGLE, 1985; DERDE; MASSART, 1988).

2.5.1. PCA

A PCA é um dos métodos multivariados mais comuns empregada na análise dos dados (BROWN, 1995; FERREIRA, 2002). Ela permite a interpretação multivariada de conjuntos de dados complexos e com grande número de variáveis como, por exemplo, espectros no infravermelho próximo, por meio de gráficos bi ou tridimensionais. Esses gráficos apresentam informações que expressam a existência de correlação entre diversas variáveis facilitando a interpretação multivariada do comportamento da amostra (BEEBE; PELL; SEASHOLTZ, 1998; SABIN; FERRÃO; FURTADO, 2004).

A utilização da PCA visa reduzir a dimensionalidade do conjunto de dados original ou pré-processados, minimizando a covariância entre as variáveis, sem perda de informações, permitindo a observação de semelhança e diferença entre as amostras. Esta redução é obtida por meio do estabelecimento de novas variáveis ortogonais entre si, denominadas componentes principais (PCs) que são combinações lineares das variáveis originais (MARTENS; NAES, 1989; GELADI; KOWALSKI, 1986, FERREIRA, 2002).

De acordo Beebe; Pell; Seasholtz (1998); Sena et al. (2000); Sabin; Ferrão; Furtado (2004); Souza; Poppi (2012), na análise de componentes principais a matriz X é decomposta em um produto de duas matrizes, como ilustrado na **Equação 5**:

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E} \quad (5)$$

Em que:

\mathbf{T} - escores (scores);

\mathbf{P} - pesos (loadings, P);

\mathbf{E} – resíduos.

Os escores representam as coordenadas das amostras no sistema de eixos formados pelas componentes principais. Cada PC é constituída pela combinação linear das variáveis originais e os coeficientes da combinação são denominados pesos. Os pesos são os cossenos dos ângulos entre as variáveis originais e as componentes principais (PCs). A PC1 é definida na direção da máxima variação no conjunto de dados, a PC2 é traçada perpendicular à primeira, com o intuito de descrever a maior porcentagem da variação não explicada pela PC1 e assim por diante.

Enquanto os escores representam as relações de similaridade entre as amostras, os pesos representam a contribuição de cada variável para a formação das PCs. Por meio da análise conjunta do gráfico de escores e pesos é possível verificar quais variáveis são responsáveis pelas diferenças observadas entre as amostras. Uma das ferramentas utilizadas para determinar o número de PC a ser utilizado no modelo PCA é a porcentagem de variância explicada acumulada (WOLD; ESBENSEN; GELADI, 1987; KAMAL-ELDIN; ANDERSSON, 1997; SABIN; FERRÃO; FURTADO, 2004).

Segundo Souza; Poppi (2012) existem diversos algoritmos disponíveis para a realização da PCA e quatro deles aparecem com frequência na literatura: o algoritmo dos mínimos quadrados parciais iterativo não linear (Non linear Iterative Partial Least Squares - NIPALS) (GERADI; KOWALSKI, 1986), decomposição em valores singulares (Singular Value Decomposition - SVD) (MARTENS; NAES, 1989; BRERETON, 2007), os quais utilizam a matriz de dados X, decomposição em autovalores (Eigenvalue Decomposition - EVD) e POWER que trabalham produto cruzado $X'.X$. (LATHAUWER; MOOR; VANDEWALLET, 2000; BRERETON, 2007).

2.5.2. SIMCA

É um método de reconhecimento de padrões supervisionado que considera informações sobre a distribuição de um conjunto de amostras. O SIMCA estima um grau de confiança da classificação podendo prever novas amostras como pertencentes a uma ou mais classes ou a nenhuma classe. Para isto este método se baseia no uso da PCA para modelar a forma e a posição do objeto definido pelas amostras no espaço linha visando à definição de uma classe (WOLD, 1976; BRUNS; FAIGLE, 1985; DASZYKOWSKI et al., 2007; STUMPE et al., 2012).

Um modelo PCA é construído e delimitado a uma região espacial multidimensional para cada classe ou grupo de amostras conhecida. Esses modelos são totalmente independentes. O número de componentes principais necessários para descrever os dados pode variar de uma classe para outra dependendo do grau da complexidade da estrutura dos dados em cada classe (BARROS NETO; SCARMINIO; BRUNS, 2006; FLATEN; GRUNG; KVALHEIM, 2004; FLUMIGNAN, 2010).

Após seu estabelecimento, os modelos são utilizados para classificar amostras futuras como pertencendo a uma das classes. Isto ocorrerá quando a amostra apresentar características semelhantes que a permitam ser inserida neste espaço multidimensional de uma das classes (BEEBE; PELL; SEASHOLTZ 1998; SCAFI, 2000).

O SIMCA baseia-se no cálculo da distância da amostra ao modelo, utilizando-se, para isto, a variância residual para cada amostra da classe X (S_i) (Equação 6) e a variância residual total, S_o (Equação 7) (SCAFI, 2000; FLATEN; GRUNG; KVALHEIM, 2004; BRANDEN; HUBERT, 2005; POVIA, 2007).

$$S_i^b = \sqrt{\frac{\sum_{j=1}^M (res_j^b)^2}{M - A_b}} \quad (6)$$

$$S_0^b = \sqrt{\frac{\sum_{i=1}^{N_b} \sum_{j=1}^M (res_{ij}^b)^2}{(N_b - A_b - 1) \cdot (M - A_b)}} \quad (7)$$

Em que:

N_b - número de amostras pertencentes ao conjunto de treinamento da classe b ;

A_b - número de componentes principais utilizados pela classe b ;

M - número de variáveis,

i e j - índices das amostras e variáveis, respectivamente.

A localização da amostra em relação ao modelo é verificada por meio de um teste F , o qual compara o valor obtido pela **Equação 8** (F_{cal}) com um valor crítico (F_{crit}) que pode ser obtido empiricamente ou tabelado para determinado nível de confiança. No caso da amostra investigada apresentar um valor de F_{cal} menor que o obtido pelo F_{crit} , esta amostra pertencerá, então, à classe em consideração (WOLD; SJOSTROM, 1977; BLANCO et al., 1998; SCAFI, 2000; FLATEN; GRUNG; KVALHEIM, 2004; POVIA, 2007).

$$F_{cal} = \frac{(S_i^b)^2}{(S_0^b)^2} \cdot \frac{N_b}{N_b - A_b - 1} \quad (8)$$

De acordo com Beebe; Pell; Seasholtz (1998) a classificação SIMCA pode ser expressa por dois tipos de erro:

- ✓ **Tipo I:** a amostra não é classificada em sua classe verdadeira;
- ✓ **Tipo II:** a amostra é classificada em uma classe distinta da sua.

Com base nesses tipos de erro, uma mesma amostra poderá não ser classificada na sua classe verdadeira e ser ou não classificada em outra(s) classe(s).

2.5.3. LDA

A análise discriminante linear é uma técnica de classificação probabilística que consiste em estimar uma combinação linear de duas ou mais variáveis independentes. Ela obtêm funções discriminantes lineares as quais maximizam a variância entre as classes e minimizam a variância dentro de cada classe. No caso da existência desta função pode-se dizer que os pontos pertencentes às duas classes são linearmente separáveis (BRUNS; FAIGLE, 1985; MASSART et al., 1997; BALABIN; SAFIEVA, 2008; CASALE et al., 2010; DINIZ et al., 2012).

A LDA se assemelha à PCA, pois ambas buscam reduzir a dimensionalidade dos dados na matriz de variáveis, enquanto a PCA busca encontrar uma direção que tenha a máxima variância dos dados e um mínimo de dimensões relacionada; a LDA tem a finalidade de selecionar uma direção por meio da qual se alcance a separação máxima entre as classes avaliadas (YU; YANG, 2001; PONTES, 2009).

Apesar das diversas aplicações, a LDA possui, quando comparada com os outros métodos de reconhecimento de padrões supervisionados, duas desvantagens: a primeira é com relação à limitação de uso em conjunto de dados de pequena dimensão e a segunda é a colinearidade dos dados (YU; YANG, 2001; SOARES et al., 2013).

Diante das desvantagens expostas a aplicação da LDA, em dados espectrométricos é limitada pela geração de diversas variáveis por amostra. Assim, o uso de procedimentos de redução de dimensionalidade e/ou seleção de variáveis é uma maneira de superar esta dificuldade (CASALE et al., 2010; DINIZ et al., 2012).

2.6. Seleção de Variáveis e Amostras

2.6.1. Algoritmo para Seleção de Variáveis

Segundo Vasconcelos (2011) técnicas de reconhecimento de padrões e calibração multivariada possuem limitações quando aplicadas a conjuntos de dados com grande número de variáveis, visto que muitas dessas variáveis são irrelevantes e apresentam alguma correlação.

As técnicas de seleção de variáveis envolvem a utilização de métodos computacionais cuja finalidade é encontrar um subconjunto de variáveis capazes de melhorar os resultados. Em último caso, mantê-los constante em termos de erro a partir dos dados originais ou transformados. Os métodos de seleção de variáveis buscam, ainda, produzir modelos mais simples ou parcimoniosos, por meio da remoção de variáveis não informativas e da minimização da multicolinearidade entre as variáveis (SOARES, 2010; GOMES, 2012).

De acordo com Gomes (2012) a busca por este subconjunto de variáveis consiste de um problema de otimização combinatorial guiado por uma função objetivo. Em geral, usa-se o erro de validação cruzada ou o erro para um conjunto externo de amostras. As restrições impostas às combinações e às funções de custo, definem a estratégia do algoritmo de seleção.

Existe vários algoritmos de seleção de variáveis dentre os quais se destacam: Busca Exaustiva (FERREIRA; MONTANARI; GAUDIO, 2002), Algoritmo Genético (COSTA FILHO; POPPI, 1999; LUCASIUS; KATEMAN, 1993), Método de eliminação de variáveis não informativas (CENTER et al., 1996), Jack-Knife (EFRON, 1982; MARTENS; MARTENS, 2000), Colônia de formigas (SHAMSIPUR, 2006), PLS em intervalos – iPLS (NORGAARD et al., 2000), Backward PLS (PIERNA et al., 2009), PLS em intervalos sinérgicos – siPLS (NORGAARD, 2005), OPS-PLS (TEOFILO, 2009), Busca de Tabu (GLOVER, 1989), Ponderação Iterativa dos Preditores (FORINA; CASOLINO; MILLAN, 1999) e Algoritmo das Projeções Sucessivas (ARAÚJO et al., 2001).

O algoritmo das projeções sucessivas tem sido muito utilizado em diversos trabalhos de pesquisa e foi adotado neste trabalho.

2.6.1.1. Algoritmo das Projeções Sucessivas

O algoritmo SPA (Sucessive Projections Algorithm) foi proposto por Araújo et al. (2001) como método de seleção de variáveis no âmbito de regressão linear múltipla e aplicado a dados espectroscópicos.

Segundo Gomes (2012) o SPA é uma técnica do tipo *forward* com a restrição de que a variável incorporada em cada iteração deve ser a menos multicolinear possível com as variáveis previamente selecionadas.

O SPA é composto por três fases (GALVÃO et al., 2008). Gomes (2012) as descrevem da seguinte forma:

Na primeira fase são geradas as cadeias de variáveis minimamente redundantes empregando-se somente a matriz X_{cal} , geralmente centrada na média das colunas. A etapa seguinte (Fase 2 do SPA), consiste em avaliar a correlação das cadeias com o variável de interesse.

A terceira e última fase consiste em eliminar as variáveis que não apresentam melhoria em termos de valor PRESS (*Predicted Residual Error Sum of Squares*), com base em um teste F . Para isto, a cada variável é associado um “fator de relevância” dado pelo produto dos desvios padrões amostral e do módulo do coeficiente de regressão desta variável. Posteriormente, os modelos MLR são construídos incluindo-se as variáveis em ordem decrescente de importância e a cada nova variável calcula-se o valor de PRESS. O menor número de variáveis para qual o valor de PRESS não difere do mínimo global empregando um teste F a 75% de confiança é empregado no modelo MLR final.

O algoritmo SPA foi adaptado por Pontes et al. (2005) para atuar como ferramenta de seleção de variáveis em problemas de classificação. Com a finalidade de melhorar o desempenho da análise discriminante linear (LDA), que também é afetada por problema de colinearidade (NAES; MEVIK, 2001).

Em geral, o procedimento de seleção de variáveis SPA - LDA utiliza três conjuntos de dados: treinamento, validação e teste e compreende duas fases, conforme descrito por Soares et al. (2013):

Na Fase 1 os dados de treinamento são centrados na média de cada classe. Na Fase 2 os dados centrados na média são empregados para calcular uma matriz de covariância. Nesta, os subconjuntos de variáveis são avaliados de acordo com uma função de custo (**Equação 9**) relacionada com o risco médio de classificação incorreta sobre a validação definida.

$$J_{cost} = \frac{1}{N_{val}} \sum_{n=1}^{N_{val}} g_n \quad (9)$$

Sendo:

$$g_n = \frac{MD^2[x_{val,n}, \bar{x}(I_n)]}{\min_{I_j \neq I_n} MD^2[x_{val,n}, \bar{x}(I_j)]} \quad (10)$$

Na **Equação 10** o numerador $MD^2[x_{val,n}, \bar{x}(I_n)]$ é o quadrado da distância de Mahalanobis (MAESSCHALCK; JOUAN-RIMBAUD; MASSART, 2000) entre a amostra $x_{val,n}$ (com índice de classe I_n) e a média $\bar{x}(I_n)$ de sua verdadeira classe (ambos os vetores linha) calculado sobre o conjunto de validação, distância dada pela **Equação 11**:

$$MD^2[x_{val,n}, \bar{x}(I_n)] = [x_{val,n}, \bar{x}(I_n)] S^{-1} [x_{val,n}, \bar{x}(I_n)]^T \quad (11)$$

Em que:

S - matriz de covariância calculada para o conjunto de validação (PEREIRA et al., 2008).

O denominador na **Equação 10** corresponde ao quadrado da distância Mahalanobis entre a amostra $x_{val,n}$ e o centro da classe incorreta mais próxima. Um valor pequeno de g_n indica que a amostra $x_{val,n}$ está próxima do centro de sua verdadeira classe e distante dos centros das demais classes. A função de custo J_{cost} é definida como o valor médio de g_n sobre todas as amostras de validação ($n = 1, 2, \dots, N_{val}$), de modo que o menor valor dos resultados de J_{cost} resulta em uma separação melhor das amostras, de acordo com sua verdadeira classe.

Após as variáveis serem selecionadas a classificação de uma nova amostra, x_{new} , pode ser realizada por meio do cálculo da distância de Mahalanobis em relação ao vetor médio de cada classe. A amostra é, então, atribuída à classe com menor distância de Mahalanobis. Observa-se que os vetores de médias e de matriz de covariância agrupada são calculados sobre o conjunto de treinamento usando-se as variáveis selecionadas.

Pontes et al. (2012) relatam que a divisão das amostras em três conjuntos restringe o uso do SPA – LDA, no caso em que o número de amostra disponível é

limitante. Para superar esta dificuldade, os autores sugerem utilizar o conjunto de treinamento para realizar a validação e assim orientar a seleção de variáveis no SPA – LDA.

Aplicações bem sucedidas do SPA envolvendo calibração multivariada, foram empregadas na quantificação de biodiesel em diesel (FERNANDES et al., 2011), determinação de parâmetros de qualidade de óleos isolantes (PONTES et al., 2011a), determinação simultânea de compostos aromáticos em água (LIMA; RAIMUNDO; PIMENTEL, 2011) e na determinação de dipirona em ampolas fechadas (SANCHES et al., 2012). Outros artigos reportam o uso do SPA associado à classificação de diferentes tipos de amostras com em cigarros (MOREIRA et al., 2009); óleos vegetais (GAMBARRA NETO et al., 2009), diesel/biodiesel (PONTES et al., 2011b), álcool combustível (SILVA et al., 2012) e cerveja (GHASEMI-VARNAMKHAHI et al., 2012).

2.6.2. Seleção de Amostras

O algoritmo para seleção de amostras desenvolvido pelos pesquisadores KENNARD e STONE em 1969, denominado KS, é o mais conhecido entre os químicos analíticos (KENNARD; STONE, 1969; GALVÃO et al., 2005; DANTAS FILHO, 2007).

Segundo Galvão et al. (2005); Dantas Filho (2007) e Marreto (2010) este algoritmo visa selecionar um subconjunto representativo de um conjunto de N amostras, com a finalidade de assegurar uma distribuição uniforme do subconjunto de amostras representadas pelo espaço de dados baseado na resposta instrumental X. O KS segue um procedimento orientado, no qual novas seleções são realizadas em regiões do espaço, distantes das amostras selecionadas. Para isto, o algoritmo emprega a distância Euclidiana $d_x(p,q)$ entre os vetores \mathbf{x} de cada par (p,q) de amostras calculadas conforme descrito na **Equação 12**:

$$d_x(p, q) = \sqrt{\sum_{j=1}^J [x_p(j) - x_q(j)]^2}; \quad p, q \in [1, N] \quad (12)$$

Em que:

$x_p(j)$ e $x_q(j)$ - respostas instrumentais no j -ésimo comprimento de onda para as amostras " p " e " q ", respectivamente;

" j " - número de comprimento de onda no espectro.

O procedimento inicia-se pela escolha do par (p_1, p_2) de amostras para as quais a distância $d_x(p_1, p_2)$ seja a maior. Em cada iteração subsequente o algoritmo seleciona a amostra que apresenta a maior distância em relação à amostra selecionada, procedimento este repetido até o número de amostras especificado ser alcançado.

Galvão et al. (2005) propuseram uma extensão do KS denominada SPXY, cuja função é aumentar a distância Euclidiana (d_x) com a distância da variável no espaço \mathbf{y} . A distância $d_y(p, q)$ pode ser calculada para cada par de amostras p e q , conforme a **Equação 13**:

$$d_y(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{y}_p - \mathbf{y}_q)^2} = |\mathbf{y}_p - \mathbf{y}_q|; \quad \mathbf{p}, \mathbf{q} \in [1, N] \quad (13)$$

Com o objetivo de atribuir a mesma importância na distribuição de amostras em \mathbf{x} e no espaço \mathbf{y} , as distâncias $d_x(\mathbf{p}, \mathbf{q})$ e $d_y(\mathbf{p}, \mathbf{q})$ são divididas pelos seus valores máximos no conjunto de dados. E desta maneira, a distância \mathbf{xy} será normalizada segundo a **Equação 14**:

$$d_{xy}(\mathbf{p}, \mathbf{q}) = \frac{d_x(\mathbf{p}, \mathbf{q})}{\max_{\mathbf{p}, \mathbf{q} \in [1, N]} d_x(\mathbf{p}, \mathbf{q})} + \frac{d_y(\mathbf{p}, \mathbf{q})}{\max_{\mathbf{p}, \mathbf{q} \in [1, N]} d_y(\mathbf{p}, \mathbf{q})} \quad \mathbf{p}, \mathbf{q} \in [1, N] \quad (14)$$

Um procedimento de seleção similar ao algoritmo KS pode ser aplicado com $d_{xy}(\mathbf{p}, \mathbf{q})$ ao invés de $d_x(\mathbf{p}, \mathbf{q})$ sozinho.

2.7. Calibração Multivariada

Pimentel; Galvão; Araújo (2008) definiram calibração como um procedimento matemático e estatístico usado para relacionar valores medidos com grandezas analíticas caracterizando os tipos de analito e suas quantidades ou concentrações.

Segundo Braga; Poppi (2004) entre os métodos de calibração existentes os mais difundidos são os métodos univariados em que se tem apenas uma medida instrumental para cada uma das amostras de calibração. Esses métodos são relativamente fáceis de serem aplicados e validados. Porém em muitas situações a medida de uma única variável não é capaz de descrever o sistema, a exemplo da calibração baseada em dados espectroscópicos e cromatográficos.

Na calibração multivariada duas ou mais respostas instrumentais são relacionadas à propriedade de interesse. Esses métodos possibilitam análises, mesmo na presença de interferentes, desde que estejam presentes nas amostras de calibração; determinações simultâneas, análises com baixa resolução, entre outros. Isto permite que modelos de calibração multivariada sejam uma alternativa quando métodos univariados não podem ser aplicados (BRAGA ; POPPI, 2004).

Soares et al. (2013) relatam que o processo de calibração multivariada consiste, basicamente em duas etapas: calibração e validação. De acordo com Soares (2010) busca-se na etapa de calibração, estabelecer uma relação matemática entre a matriz de resposta instrumental (Matriz X_{cal} – contém as variáveis independentes) com um vetor contendo a variável dependente, ou seja, aquele que possui as propriedades de interesse determinadas pelos métodos de referência (y_{cal}).

Na segunda etapa, conhecida como validação do modelo, é oportuno verificar se a relação entre a matriz X_{cal} e o vetor y_{cal} é satisfatória para determinação da propriedade de interesse. Segundo Brereton (2000) esta etapa de validação pode ser realizada de duas formas diferentes: validação cruzada (Cross-validation) ou validação externa por série de teste.

Soares (2010) relata que existem algumas métricas capazes de avaliar se os valores preditos a partir das medidas X são condizentes com os de y , entre elas pode-se citar: PRESS (Predicted Residual Error Sum of Squares) (MARTENS; NAES (1989); BEEBE; PELL; SEASHOLTZ (1998); BRERETON (2000)), RMSE (Root

Mean Squares Error) (NAES et al. (2002)) e o RSEP (Relative Standard Error of Prediction) (NAES et al. (2002)).

Vários métodos de regressão vêm sendo utilizados visando à construção de modelos de calibração multivariada, tais como: Regressão Linear Múltipla (MLR), Regressão por Componentes Principais (PCR) e Regressão por Mínimos Quadrados Parciais (PLS) (NAES; MARTENS, 1984; FERREIRA et al., 1999).

2.7.1. Regressão Linear Múltipla

A MLR (Multiple Linear Regressssion) é a mais simples dos métodos de calibração, no qual se assume que cada variável dependente do vetor y relaciona-se linearmente com as variáveis independentes da matriz X (NAES; MARTENS, 1984; BEEBE; PELL; SEASHOLTZ, 1998; FERREIRA, et al., 1999) como ilustrado na **Equação 15**.

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{MLR} + \mathbf{E} \quad (15)$$

Sendo:

X - matriz dos sinais de m amostras, medidos em j variáveis;

y - matriz dos q parâmetros de m amostras;

\mathbf{b}_{MLR} - matriz dos coeficientes lineares de regressão;

E - resíduo não modelado em y .

O vetor de regressão b é estimado na etapa de calibração empregando-se o método mínimos quadrados, conforme a **Equação 16**.

$$\mathbf{b}_{mlr} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \quad (16)$$

Entretanto, de acordo com Gomes (2012) a resolução da **Equação 16** requer, para obter o vetor dos coeficientes de regressão (b) a inversa da matriz $(\mathbf{X}^t\mathbf{X})$ e esta operação algébrica envolve algumas suposições acerca dos dados:

✓ O número de amostras de calibração deve ser maior ou igual ao número de variáveis ($m > n$), caso contrário, o sistema de equações será indeterminado.

✓ As variáveis devem ser vetores linearmente independentes. A violação desta suposição pode levar a uma matriz singular.

Tais suposições impossibilitam o uso da calibração MLR, em medidas que possuam muitas variáveis sem a realização de uma seleção prévia das mesmas (Gomes, 2012).

2.7.2. Regressão em Componentes Principais

A PCR (*Principal Components Regression*) é um método de calibração que faz uso de uma transformação ortogonal da matriz X , de maneira a se obter um novo conjunto de variáveis linearmente independentes. Para tanto, não necessita de seleção de variáveis para contornar o problema de multicolinearidade dos dados (VALDERRAMA, 2009).

A decomposição da matriz X é baseada no conceito de análise de componentes principais em que uma matriz de alta dimensão é decomposta em duas matrizes menores, chamadas escores (T) e pesos (P) (NAES et al., 2002; BRERETON, 2003) de acordo com a **Equação 17**.

$$X = TP^t + E \quad (17)$$

Em que:

E - parte do resíduo deixado pela modelagem;

A regressão PCR faz uso da matriz T , que é ortogonal, para obter o vetor dos coeficientes de regressão b_{PCR} empregando-se o método dos mínimos quadrados (OLS) similar ao MLR, de acordo com a **Equação 18** (Gomes, 2012).

$$y = T_{(mxk)} b_{PCR} + F \quad (18)$$

Em que:

k - número de componentes principais empregados na obtenção dos coeficientes de regressão;

F - resíduos não modelados.

2.7.3. Regressão em Mínimos Quadrados Parciais

O método de calibração multivariada PLS(Partial Least Square) foi desenvolvido por Herman Wold e colaboradores, no período de 1975 a 1982. Na modelagem PLS a matriz X também é decomposta, assim como ocorre na PCR, porém este método utiliza tanto as informações da matriz de dados independentes (Matriz X), como as informações da matriz de referências (Y) (WOLD, 2001; SIMÕES, 2008).

Ao considerar a determinação de mais de uma espécie de interesse, as matrizes X_{cal} e Y_{cal} são decompostas em suas matrizes de pesos e escores, respectivamente, como indicado nas **Equações 19 e 20**.

$$X = TP^t + E \quad (19)$$

$$Y = UQ^t + F \quad (20)$$

Em que:

T e U - matrizes dos escores;

P e Q - matrizes dos pesos das matrizes X e Y ;

E - matriz de resíduos espectrais;

F - matriz dos resíduos de concentração.

Por fim, o modelo resultante da PLS consiste em relacionar linearmente os escores da matriz X com os escores da matriz Y (SIMÕES, 2008) de acordo com as **Equações 21 e 22:**

$$U = BT^t + G \quad (21)$$

$$Y = BTQ^t + H \quad (22)$$

Em que:

B - matriz dos coeficientes de regressão;

G - matriz de resíduos dos escores;

H - matriz de resíduos de concentração.

A obtenção dos parâmetros de um modelo PLS pode ser realizada empregando-se diferentes tipos de algoritmo (ANDERSSON, 2009), com destaque para o algoritmo de escores não ortogonalizados (MARTENS; NAES, 1989) e o NIPALS (BRERETON, 2000).

CAPÍTULO 3

Classificação de sementes de mamona

3. CLASSIFICAÇÃO DE SEMENTES DE MAMONA

3.1. Introdução

As plantas da mamoneira possuem grande variabilidade em diversas características, como hábito de crescimento, cor das folhas e do caule, tamanho, cor e teor de óleo das sementes. Pode-se, portanto, encontrar tipos botânicos com porte baixo ou arbóreo, ciclo anual ou semiperene, como folhas e caule verde, vermelho ou rosa, com a presença ou ausência de cera no caule, com frutos inermes ou com espinhos, deiscentes ou indeiscentes, com sementes de diversos tamanhos, colorações, teores de óleo e de ricina (SAVY FILHO, 2005; BELTRÃO; AZEVEDO, 2007).

Apesar disto, nem sempre é possível identificar qual o genótipo por inspeção visual das sementes. Em geral, o procedimento de identificação de algumas cultivares é feito por meio do plantio da semente e espera-se, no mínimo, um mês para que, através do seu crescimento e desenvolvimento, ocorra sua identificação morfológica. Técnicas de marcadores moleculares também são empregadas para esta classificação e identificação (VECCHIA; SILVA; SOBRINHO TERCENIANO, 1998; FERREIRA; 2003; VIDAL et al., 2005). Porém são difíceis de serem implantadas em escala de rotina, destroem a semente, inviabilizando-as para futuros testes; são lentas e necessitam de pessoal com alta qualificação técnica.

Esses desafios podem ser superados por meio do desenvolvimento de métodos analíticos baseados no uso da espectrometria de reflectância no infravermelho próximo (NIR) e das técnicas quimiométricas.

A aplicação da espectroscopia NIR e das técnicas de classificação, tem sido utilizadas em diversos tipos de matrizes, como: biodiesel (VERAS et al., 2010; BALABIN; SAFIRA, 2011; INSAUSTI et al., 2012), gasolina (BALABIN, R.; SAFIEVA; LOMAKINAC, 2010); cigarros (MOREIRA et al., 2009), cerveja (EGIDIO et al., 2011; GHASEMI-VARNAMKHAZI et al., 2012); madeira (CARNEIRO, 2008); vagens de soja (SIRISOMBOON; HASHIMOTO; TANAKA, 2009); azeitonas (CASALE et al., 2010), azeite (SINELLI et al., 2010; GALTIER et al., 2011); vinhos (RIOVANTO et al., 2011); mel (CHEN et al., 2012) e gás liquefeito de petróleo (DANTAS et al. 2013).

Apesar dos diversos artigos com aplicações bem sucedidas da espectroscopia NIR e técnicas de classificação, a análise de sementes utilizando

essa associação ainda é pouco explorada na literatura e apenas dois trabalhos foram encontrados. Estes serão detalhados a seguir.

LEE; CHOUNG (2011) desenvolveram um estudo para avaliar o potencial da espectroscopia NIR na classificação de sementes de soja geneticamente modificada (GM) e não-GM. Espectros NIR foram coletados a partir das sementes individuais em que cada semente foi colocada em um suporte que permitiu que a radiação fosse refletida de um lado da semente. Todos os dados espectrais foram registrados como o logaritmo do inverso da reflectância ($\log 1 / R$) na região espectral de 400 a 2500 nm, com resolução de 2 nm e média de 32 varreduras. As técnicas quimiométricas utilizadas foram análise de componentes principal (PCA) e análise discriminante por mínimos quadrados parciais (PLS-DA). O modelo PLS-DA usando os dados pré-processados, obteve a melhor calibração e um certo na classificação de 97%. De acordo com os autores, os resultados com a espectroscopia NIRA em conjunto com técnicas quimiométricas, podem ser usado para identificar soja GM evitando, assim, análises demoradas, destrutivas e trabalhosas.

VITALE et al. (2013) estudaram o potencial da espectroscopia NIR acoplada a técnicas quimiométricas (SIMCA, PLS-DA) para verificar a origem de sementes de pistache (*Pistacia vera* L.). Foram analisadas 483 amostras de seis diferentes origens. Os espectros foram registrados entre 10.000 e 4000 cm^{-1} , média de 82 varreduras em uma resolução nominal de 4 cm^{-1} , em sementes cortadas ao meio de forma longitudinalmente, no modo de reflectância. Os resultados demonstraram que mais de 95% das amostras de validação foram corretamente classificadas utilizando o PLS-DA. Resultados similares foram obtidos utilizando-se a técnica SIMCA. Os autores concluíram que a associação da espectroscopia NIR e técnicas de classificação pode ser uma valiosa ferramenta para rastrear a origem de pistache, proporcionando uma autenticação confiável de forma rápida e barata.

3.2. Objetivos Específicos

- ✓ Aplicar a espectroscopia NIR e a técnica PCA na discriminação de cultivares duas de mamona;
- ✓ Utilizar medidas NIR com modelos SIMCA e SPA-LDA para a classificação de duas cultivares de mamona.

3.3. Experimental

3.3.1. Aquisição das Amostras

Duas cultivares de mamona (BRS Nordestina e BRS Paraguaçu) foram utilizadas neste trabalho (**Figura 5**). Para cada cultivar foram empregadas trezentas e cinquenta amostras de sementes de alta qualidade genética, cedidas pela Embrapa Algodão, na cidade de Campina Grande, Paraíba, Brasil. As amostras foram sempre acondicionadas a uma temperatura de 21°C e umidade relativa de 70%.

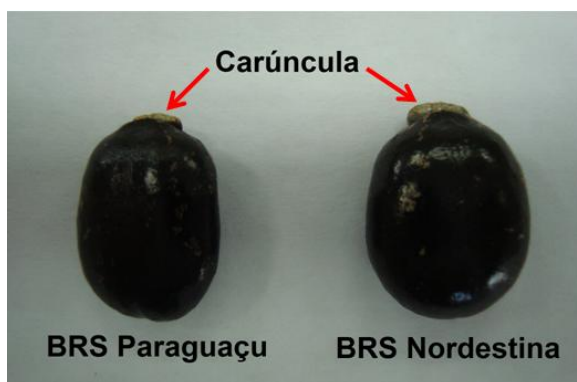


Figura 5 – Sementes das cultivares de mamona, BRS Nordestina e BRS Paraguaçu.

3.3.2. Instrumentação

As medidas de reflectância difusa foram obtidas em um espectrofotômetro VIS-NIR modelo XDS Rapid Content™ Analyser (Foss Analytical, Hogans, Sweden) conforme ilustrado na **Figura 6**.

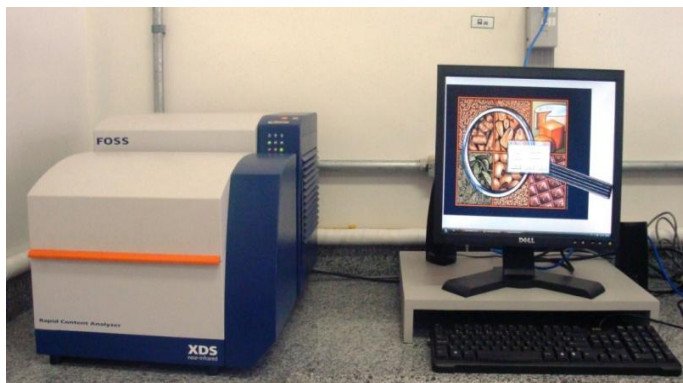


Figura 6 - Espectrofotômetro VIS-NIR.

Na **Figura 7 (a)** observa-se a célula de quartzo circular de 3 cm de diâmetro usada para posicionar a amostra a ser analisada. Para bloquear a radiação espúria do ambiente, foram usadas tampas reflexivas na célula de amostragem, **Figura 7 (b)**.

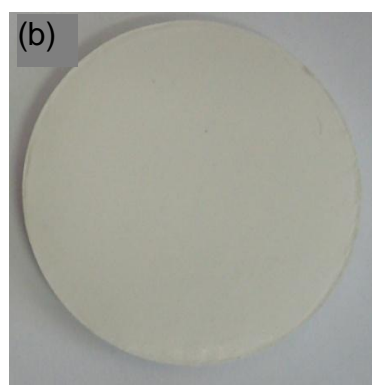
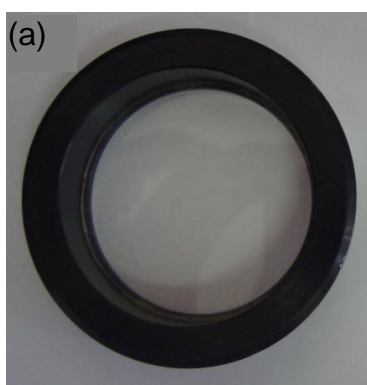


Figura 7 – 7 (a) Célula de quartzo; **7 (b)** Tampas reflexivas para a célula de quartzo.

3.3.3. Aquisição dos espectros NIR

Para o registro do sinal de base utilizou-se um padrão de reflectância conforme observado na **Figura 8**.



Figura 8 - Padrão de reflectância

Os espectros de reflectância foram obtidos diretamente sem nenhum tratamento químico das sementes. As medidas foram realizadas em quatro posições em relação a carúncula da semente de mamona (0, 90, 180 e 360⁰) (**Figura 5**).

As amostras sempre foram dispostas na célula, da mesma maneira para assegurar a uniformidade das medidas. Cada espectro foi obtido a partir de 32 varreduras na faixa de 400 a 2500 nm em intervalos de 0,5 nm. No total, obtiveram-se 1200 espectros para cada cultivar de mamona. Um espectro médio para cada semente foi calculado posteriormente, a partir das quatro posições de amostragem.

3.3.4. Programas Computacionais

O pré-processamento dos espectros originais e a aplicação das técnicas de reconhecimento de padrões não supervisionado (Análise de Componentes Principais-PCA) e supervisionado (SIMCA) foram realizados utilizando-se o programa Unscrambler® 9.8. A aplicação do algoritmo Kennard-Stone utilizado para seleção de amostras e a modelagem SPA-LDA foram realizadas em ambiente Matlab R2008a.

3.3.5. Tratamento Quimiométrico dos Dados

3.3.5.1. Pré-processamento

A região espectral de 400 a 1099 nm foi descartada pois não continha informação relevante para a construção dos modelos de classificação. Portanto, a faixa compreendida entre 1100 a 2500 nm foi selecionada a priori como a região de trabalho para classificação de sementes de mamoneira.

As técnicas de suavização Savitzky-Golay, correção multiplicativa de sinais (MSC), correção de linha de base e derivação, foram avaliadas no pré-processamento dos espectros.

3.3.5.2. Reconhecimento de Padrões

Realizou-se uma análise exploratória utilizando-se a PCA com o objetivo de observar a formação de agrupamentos.

O algoritmo Kennard-Stone foi aplicado separadamente aos espectros de cada cultivar de mamona com a finalidade de dividir as amostras em conjuntos de treinamento (50%), validação (25%) e teste (25%), conforme a **Tabela 1**. Esses conjuntos foram utilizados na análise de classificação SIMCA e na modelagem SPA-LDA.

Tabela 1 – Número de amostras dos conjuntos de treinamento, validação e teste, selecionadas pelo algoritmo KS, para as classes Nordestina e Paraguaçu.

Classe	Conjuntos			Total
	Treinamento	Validação	Teste	
Nordestina	150	75	75	300
Paraguaçu	150	75	75	300

Na etapa de seleção de variáveis pelo algoritmo SPA foram utilizadas as amostras dos conjuntos de treinamento e validação. Na seleção do número ótimo de variáveis, com base na minimização da função do custo G, foi utilizado o conjunto de validação e, para avaliar a eficiência dos modelos de classificação, o conjunto de teste.

3.3.6. Método de Referência – Plantio no Campo Experimental

Com intuito de testar a habilidade de classificação do modelo SIMCA, 50 sementes de cada cultivar foram rotuladas com numeração de 1 até 100 e analisadas no NIR conforme descrito na seção 3.3.3.

Após a realização deste ensaio as sementes foram plantadas no campo experimental, onde a semeadura foi realizada a uma profundidade de 5 cm, utilizando-se apenas uma semente por cova. A emergência ocorreu em média nove dias após o plantio (**Figura 9**). Porém algumas sementes não germinaram, sendo necessário a realização do plantio de novas sementes.



Figura 9 – Plantio das cultivares BRS Paraguaçu e BRS Nordestina no campo experimental.

Devido ao clima quente, irrigações diárias foram efetuadas até o vigésimo dia após o plantio. Realizou-se também a limpeza manual da área para o controle de plantas daninhas.

3.4. Resultados e Discussão

3.4.1. Espectros NIR

Na **Figura 10** são observados os espectros originais de 600 amostras das duas diferentes cultivares de mamoneira: BRS Nordestina (N) e BRS Paraguaçu (P) obtidos entre 1100 a 2500 nm. Nesses espectros não há ruído instrumental evidente. Contudo, uma alteração do perfil de linha de base pode ser observado. Esta foi corrigida empregando-se a primeira derivada, com o filtro de Savitzky-Golay, polinômio de segunda ordem e uma janela de 15 pontos.

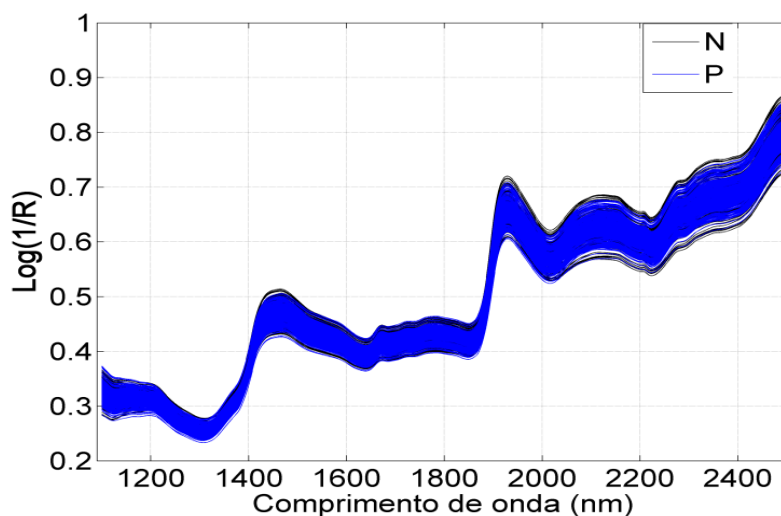


Figura 10 - Espectros Originais NIR de reflectância difusa das sementes de mamona, BRS Nordestina e BRS Paraguaçu.

Os espectros derivativos das 600 amostras de sementes de mamoneira são visualizados na **Figura 11**, observando-se a correção do incremento de linha de base com o procedimento empregado. Nota-se que todas as amostras analisadas possuem perfis espectrais semelhantes e sobrepostos, sendo observadas transições correspondentes às bandas de combinação de grupos funcionais, típicos de ROH, CONH₂ e RNH₂ presentes nas sementes de mamoneira.

A complexidade do sinal obtido e a semelhança existente na composição química das sementes impossibilitam a distinção visual das sementes das cultivares P e N. Neste contexto, torna-se necessário o uso de ferramentas quimiométricas.

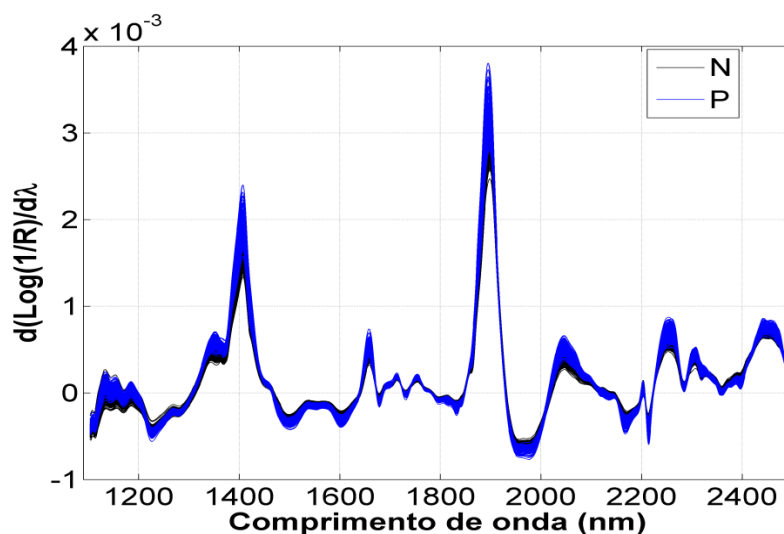


Figura 11 - Espectros NIR de reflectância difusa pré-processados das 600 sementes de mamona.

3.4.2. Análise Exploratória dos Dados

Na **Figura 12** observa-se o gráfico dos escores como resultado da aplicação da PCA (PC1 versus PC2) aos espectros pré-processados.

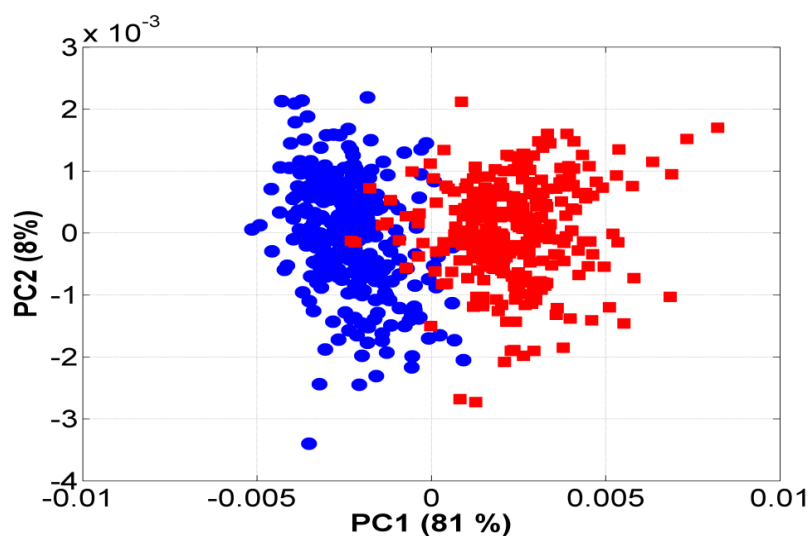


Figura 12 - Gráfico dos escores (PC1 vs PC2) para o conjunto das 600 amostras de sementes de mamona (●) BRS Nordestina e (■) BRS Paraguaçu.

Com base no gráfico de escores há uma tendência de separação entre as amostras das cultivares N e P em PC1. Contudo, também ocorre uma sobreposição entre as classes o que, possivelmente, pode vir a comprometer o desempenho dos modelos de classificação.

Observa-se, ainda, no gráfico dos escores, que não existem amostras isoladas e, portanto, as 600 amostras foram utilizadas.

A discriminação entre as amostras de sementes de mamona ocorre praticamente na PC1. Com a finalidade de verificar os principais comprimentos de onda, responsáveis por tal efeito, foi examinado, o gráfico de pesos de PC1 e PC2 (**Figura 13**).

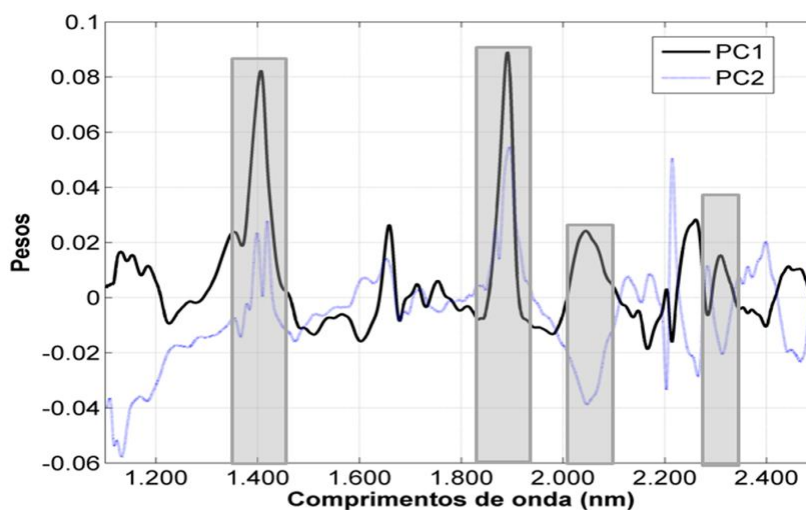


Figura 13 - Gráfico de pesos de PC1 e PC2.

Com base no gráfico de pesos (PC1 vs PC2) observam-se quatro regiões do espectro com influência em ambas as PCs. A primeira na região, em torno de 1400 nm, referente ao segundo sobretom de OH; a segunda, na região de 1890 nm, refere-se ao primeiro sobretom de OH, SH, CH, CH₂ e CH₃; a terceira, por volta de 2100 nm, caracteriza-se pela provável presença de bandas de combinação de ROH, RNH₂, CONH₂, CHO e CC e a quarta região de 2300 nm, evidencia bandas de combinação de CH, CH₂ e CH₃ (XIAOBO et al., 2010). Essas regiões espectrais foram analisadas separadamente por meio de uma PC. Os resultados, em termos de gráfico dos escores, estão ilustrados na **Figura 14**.

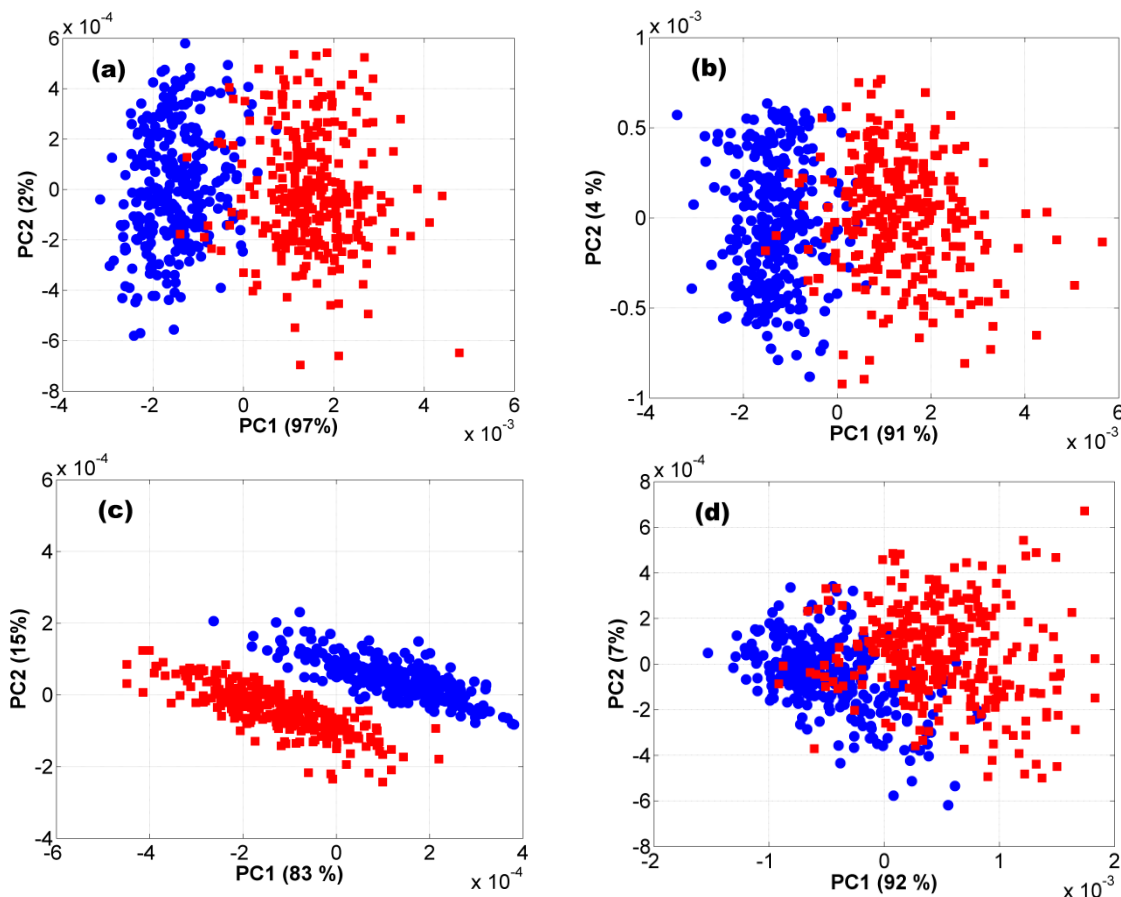


Figura 14 - Gráfico de escores (PC1 vs PC2) para o conjunto das 600 amostras de sementes de mamona (●) BRS Nordestina e (■) BRS Paraguaçu; entre parêntese estão indicadas a variância explicada, (a) faixa 1: 1340 – 1460 nm, (b) faixa 2: 1850-1930 nm, (c) faixa 3: 2110 – 2155 nm e (d) faixa 4: 2200-2277 nm.

Observa-se, com base nos gráficos dos escores (**Figura 14**), uma tendência geral de aumento de variância explicada em PC1 na PCA por faixa, quando comparado ao modelo PCA global (**Figura 12**). A construção de modelos PCA para as faixas espectrais indicadas possibilitou encontrar uma região de boa separação entre as cultivares de mamona. Esta região corresponde à faixa de 2110 a 2155 nm, cujo gráfico de escores é ilustrado na **Figura 14 (c)**. Essa região espectral será utilizada em todos os modelos subsequentes empregando-se a PCA.

3.4.3. Reconhecimento de Padrões Supervisionados

3.4.3.1. Construção e Validação dos Modelos SIMCA

Modelos PCA foram construídos para duas classes, separadamente, e validados empregando-se um conjunto externo de amostras. Na **Figura 15** é ilustrado o gráfico dos escores das amostras de treinamento e validação dos modelos PCA de cada classe.

Observa-se que as classes surgem como grupos homogêneos e as amostras de validação aparecem internas ao conjunto de treinamento. Esse resultado evidencia a seleção de amostras realizadas com o algoritmo KS.

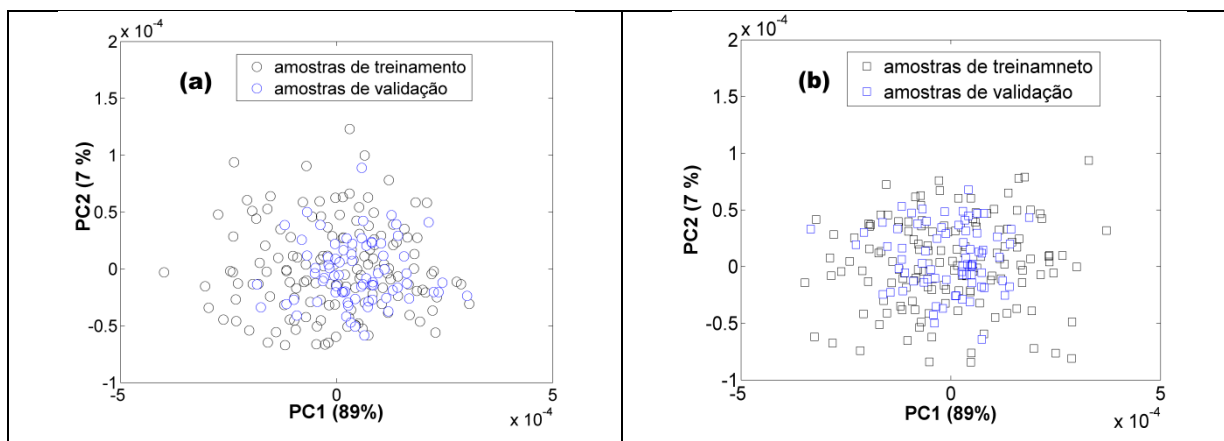


Figura 15 - Gráfico dos escores para classe (a) BRS Nordestina e para (b) classe BRS Paraguaçu.

O gráfico de variância explicada para o conjunto de validação versus o número de PCs foi usado como uma das ferramentas de diagnóstico para escolha do número de PCs (**Figura 16**). Além desta observou-se a rotina do programa Unscrambler.

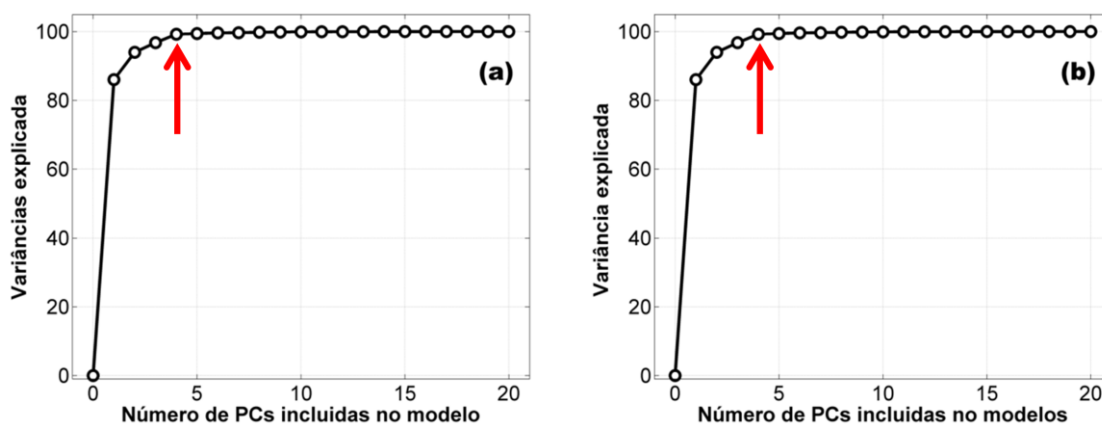


Figura 16 - Gráfico da porcentagem de variância explicada versus número de PCs incluída no modelo para as classes de (a) BRS Nordestina e (b) BRS Paraguaçu.

No total, quatro PCs foram selecionadas para ambas as classes. Com base nos gráficos da **Figura 16**, estas explicam 99,4% do modelo para a classe Nordeste e 99,2% para a classe Paraguaçu. Concordante com o número da rotina do programa. O número ótimo de PCs na PCA será usado na classificação SIMCA.

Na **Tabela 2** são ilustrados os erros de classificação do conjunto de validação com o objetivo de avaliar o desempenho dos modelos construídos. Os valores localizados nas células com tonalidade cinza correspondem ao erro do Tipo I.

Tabela 2 - Número de erros de classificação obtido pelos modelos SIMCA utilizando-se o conjunto de amostras de validação das sementes de mamona nos níveis de significância do Teste – F(1%, 5%, 10% e 25%). O número de PCs é indicado entre parênteses.

	Modelos							
	Nordestina				Paraguaçu			
	(4 PCs)				(4 PCs)			
Nível (%)	1	5	10	25	1	5	10	25
Nordestina	-	-	-	22	1	1	1	-
Paraguaçu	3	3	3	2	-	-	-	18

Os resultados de erros de classificação são promissores, exceto para o modelo com 25% como nível de significância. Neste observa-se 22 erros do Tipo I para cultivar BRS Nordeste e 18 para cultivar BRS Paraguaçu. Os erros do Tipo II são menos frequentes e maiores para cultivar a BRS Paraguaçu nos demais níveis de significância. Neste particular, vale considerar a complexidade da matriz e o número de amostras analisadas. Portanto, os modelos foram considerados validados e o nível de significância escolhido foi o de 5%, por ser o mais utilizado na literatura.

3.4.3.2. Construção e Validação do Modelo SPA-LDA

O número ideal de variáveis para o SPA-LDA foi determinado a partir do mínimo da função de custo G, exibido na **Figura 17**. Como observa-se, um mínimo bem localizado é obtido para um único comprimento de onda. Essa variável corresponde a 2152,5 nm.

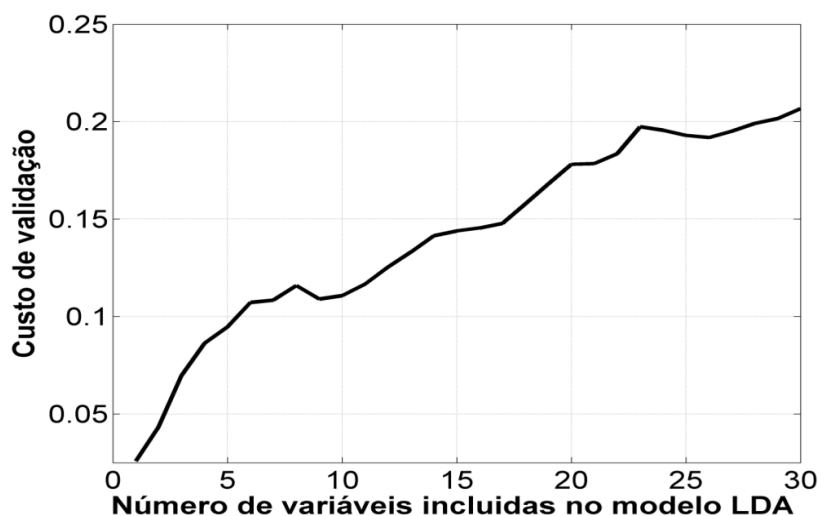


Figura 17 - Gráfico da função do custo associado à seleção de variáveis com o SPA-LDA

Portanto, o modelo resultante é parcimonioso. A partir da variável selecionada, basta estabelecer a fronteira de decisão entre as classes. Em classificação binária, o limiar entre as classes é dado pela média dos centroides das amostras de treinamento das duas classes.

Na **Figura 18** observa-se o espectro médio das amostras de treinamento ao qual foi indexada a variável selecionada pelo SPA-LDA e o intervalo usado na construção dos modelos SIMCA.

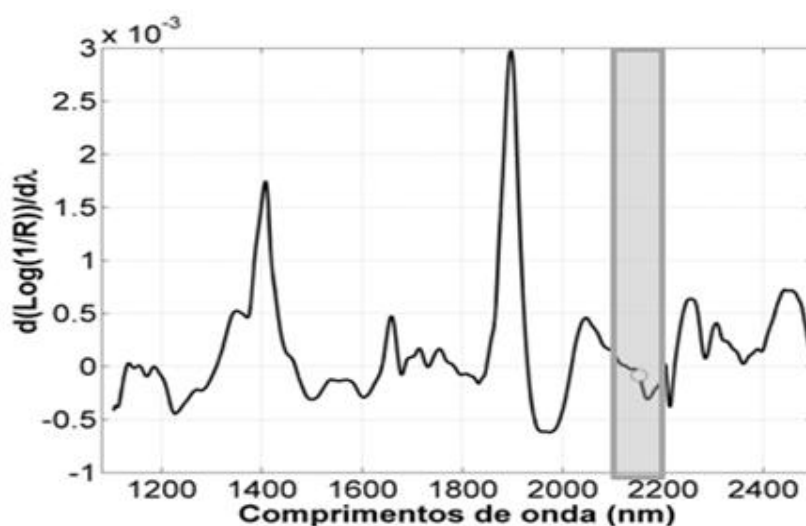


Figura 18 - Espectro médio das amostras de treinamento. A faixa cinza corresponde ao intervalo nos modelos SIMCA e (o) à variável selecionada pelo SPA-LDA.

A variável 2152,5 nm foi selecionada na faixa espectral de 2100 - 2155 nm (**Figura 14 (c)**) por ser portadora da informação capaz de discriminar os dois tipos

de semente de mamona. Nessa faixa certamente as transições vibracionais são distintas para cada tipo de cultivar.

Na **Figura 19** é ilustrado o gráfico dos espectros derivativos das amostras de treinamento e validação, com destaque para a variável selecionada pelo SPA-LDA.

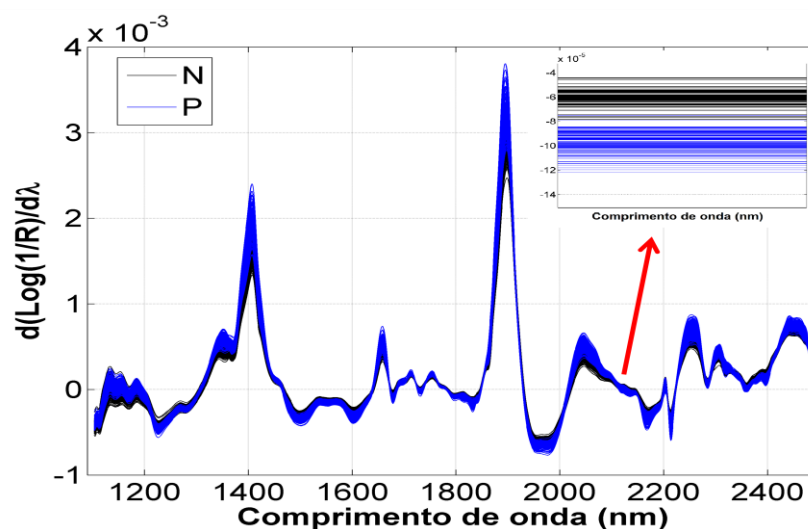


Figura 19 - Espectros derivados, com destaque para variável selecionada pelo SPA-LDA.

A distinção ocorre entre os dois tipos de sementes de mamona (N e P) em que apenas uma amostra, P, está sobreposta aos espectros das classes N. A fronteira de decisão foi calculada após definição da capacidade discriminante da variável selecionada.

Um gráfico contendo o sinal analítico medido no comprimento de onda selecionado pelo SPA-LDA versus o índice das amostras, é fornecido na **Figura 20**.

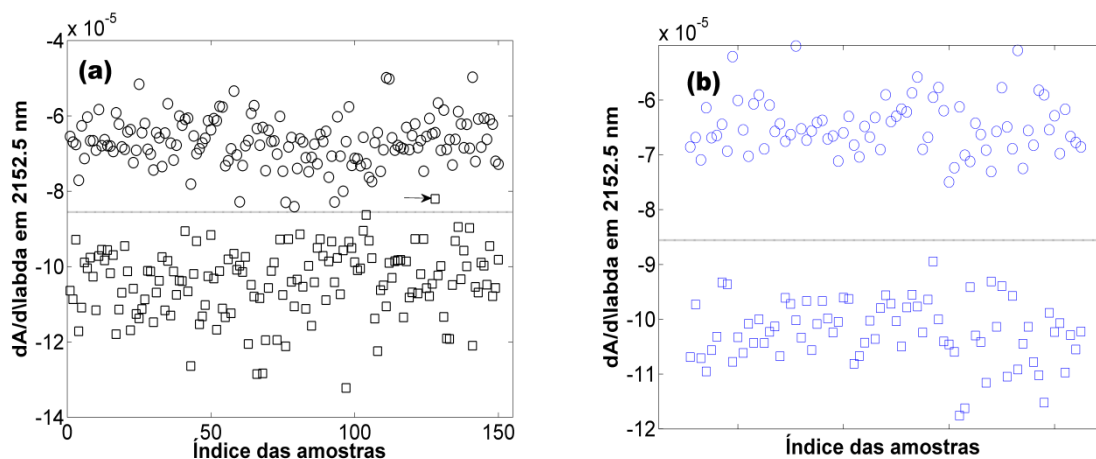


Figura 20 - Sinal analítico em 2152,5 nm versus índice das amostras para o conjunto das amostras de treinamento (○) BRS Nordestina e (□) BRS Paraguaçu e validação (○) BRS Nordestina e (□) BRS Paraguaçu. A linha tracejada representa a fronteira de decisão.

Ao analisar a **Figura 20**, fica evidenciada a separação das duas cultivares das sementes de mamona para o conjunto das amostras de treinamento **Figura 20 (a)** e validação **Figura 20 (b)**. Nota-se ainda nesta figura que uma amostra da cultivar Paraguaçu foi classificada incorretamente. Esta amostra está indicada com uma seta na **Figura 20 (a)**, correspondendo ao espectro sobreposto na **Figura 19**.

3.4.3.3. Aplicação dos Modelos ao Conjunto de Teste

Na **Tabela 3** são descritos os erros de classificação obtidos pelos modelos SIMCA e SPA-LDA em um conjunto externo como forma de verificar o desempenho de ambos. Os parâmetros utilizados nos modelos SIMCA foram: quatro PCs, para cada classe e 5% de nível de significância estatística.

Tabela 3 - Resumo da aplicação dos modelos SIMCA e SPA-LDA no conjunto de teste

Classes	MODELOS			
	SIMCA		SPA-LDA	
	Nordestina	Paraguaçu	Nordestina	Paraguaçu
Nordestina	-	3	-	-
Paraguaçu	1	-	-	-

A capacidade de discriminação das etapas de treinamento e validação de ambos os modelos foi comprovada no conjunto externo de amostras. Contudo, o modelo SPA-LDA para esse tipo de matriz empregando-se apenas uma variável espectral, mostrou-se eficaz classificando corretamente todas as amostras do conjunto de teste.

O resultado do SPA-LDA também é demonstrado por meio do gráfico do sinal na variável selecionada versus os índices das amostras (**Figura 21**). A discriminação do conjunto de amostras teste é observada entre as duas classes de semente de mamona.

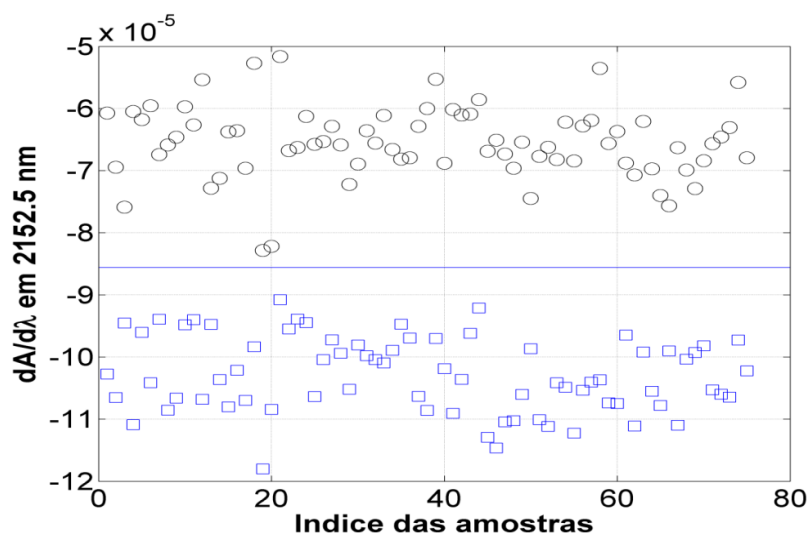


Figura 21 - Sinal analítico em 2152,5 nm versus índice das amostras para o conjunto de teste (○) BRS Nordestina e (□) BRS Paraguaçu, e a linha azul representa a fronteira de decisão estimada para o conjunto de teste.

3.4.4. Aplicação do Modelo SIMCA as Sementes Plantadas no Campo Experimental

Das cem sementes plantadas no campo experimental, dezesseis não germinaram, mesmo após o plantio de novas sementes.

A identificação das cultivares foi realizada dois meses após o plantio observando a formação do pigmento roxo na cultivar BRS Paraguaçu (**Figura 22 (a)**) e ausência dessa cor na BRS Nordestina (**Figura 22 (b)**).

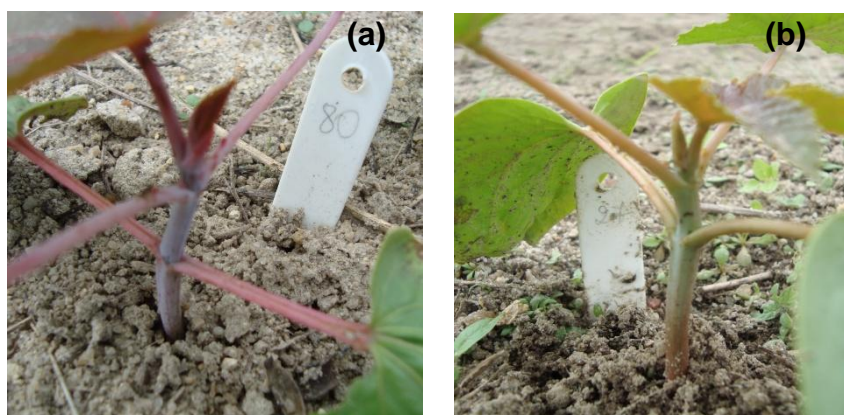


Figura 22 – **22 (a)** - Cultivar BRS Paraguaçu; **22 (b)** - Cultivar BRS Nordestina.

Na **Tabela 4** visualizam-se os erros de classificação obtidos pelos modelos SIMCA e SPA-LDA em um conjunto de sementes plantadas no campo experimental como forma de avaliar o desempenho de ambos.

Tabela 4 – Resumo da aplicação dos modelos SIMCA (5% de nível de significância) SPA-LDA no conjunto de sementes plantadas no campo experimental.

Classes	MODELOS			
	SIMCA		SPA-LDA	
	Nordestina	Paraguaçu	Nordestina	Paraguaçu
Nordestina	5	4	4	4
Paraguaçu	-	3	-	-

Com base nos resultados apresentados pelos modelos SIMCA, pode-se perceber que o número de erros do Tipo I é mais frequente (8 erros). E que os erros Tipo II ocorre apenas; para cultivar BRS Nordeste. A modelagem SIMCA classificou corretamente 86% das sementes plantadas no campo experimental. O SPA-LDA apresentou 4 erros do Tipo I (sementes da BRS Nordeste não classificada como pertencente ao modelo Nordeste), que, conseqüentemente, também se qualifica como erro do Tipo II (sementes da BRS Nordeste classificada como BRS Paraguaçu).

3.5. Considerações Finais

A PCA permitiu discriminar as cultivares de mamona BRS Nordeste e BRS Paraguaçu na região espectral correspondente à faixa de 2110 a 2155 nm.

O modelo SIMCA forneceu erros de 4% e 1,3% para as classes BRS Nordeste e BRS Paraguaçu nos níveis de significância 1, 5 e 10%, para os conjuntos de validação e teste, classificando 86% das sementes utilizadas no ensaio de campo experimental.

O SPA – LDA mostrou-se eficiente, selecionando uma variável espectral classificando corretamente todas as amostras do conjunto teste e 90% das sementes utilizadas no ensaio de campo experimental.

CAPÍTULO 4

Modelo de calibração de ricina em sementes de mamona

4. MODELO DE CALIBRAÇÃO DE RICINA EM SEMENTES DE MAMONA

4.1. Introdução

A ricina é uma proteína exclusiva do endosperma das sementes da mamoneira não sendo detectada em nenhuma outra parte da planta. Esta proteína é a principal responsável pela toxidez das sementes e da torta de mamona estando entre as proteínas mais letais, conhecidas pelo homem ([JACKSON; TOLLESON; CHIRTEL, 2006](#); [BELTRÃO; OLIVEIRA, 2009](#)).

Segundo [Ler; Lee; Gopalakrishnakone \(2006\)](#) a ricina é tóxica a humanos, animais e insetos. Uma vez dentro da célula, uma única cadeia A é capaz de inativar mais de 1500 ribossomos por minuto, o que resulta em morte celular ([FRANZ; JAAX, 1997](#); [DEMANT, 2008](#)).

Devido a grande disponibilidade de matéria-prima e alta toxicidade da ricina, esta proteína é considerada arma química de fácil preparo ([AUDI et al., 2005](#); [CHAKRAVARTULA; GUTTARLA, 2008](#)). De acordo com [Xie; Kirby; Keasling \(2012\)](#) a preocupação com segurança em relação à exposição a ricina, tem levado a uma proibição de plantio generalizado nos Estados Unidos.

Segundo [McGrath et al. \(2011\)](#) os métodos atuais para detecção de ricina podem ser classificados em três categorias: 1) métodos que detectam a presença de ricina por meio de interações imunogênicas; 2) métodos que exploram a atividade enzimática da ricina e 3) métodos que detectam a presença do DNA da mamona. [Lubelli et al. \(2006\)](#) e [Severino et al. \(2012\)](#) descreveram diversas técnicas de detecção de ricina. Porém estas técnicas são limitadas por serem caras, pouco seguras, demoradas e destrutivas.

A necessidade de genótipos com baixo teor de ricina a fim de reduzir sua toxicidade visando aumentar as diversas aplicações econômicas, principalmente para as indústrias de óleo e seus derivados, tem sido um dos principais desafios da pesquisa agrícola da mamona. Para que isto ocorra é indispensável o uso de métodos que não destruam suas sementes para uso posterior mas que também combinem viabilidade, eficiência, precisão e segurança para detecção desta toxina ([SEVERINO et al., 2012](#)).

Tais características podem ser encontradas em métodos analíticos baseados no uso da espectroscopia de refletância no infravermelho próximo (NIR), pois estes métodos podem ser capazes de associar tanto às propriedades químicas, como as

propriedades físicas das amostras, de forma não invasiva, pouco laboriosa, rápida e precisa, sem produzir resíduos químicos.

Aplicações bem sucedidas associando o uso da espectroscopia NIR e modelos quimiométricos têm sido desenvolvidos para determinação de ácidos graxos, teor de aminoácidos, umidade, proteínas, açúcares solúveis em sementes de diferentes oleaginosas: soja (PATIL et al., 2010); colza (KIM et al., 2007; CHEN et al., 2011); milho (BAYE; PEARSON; SETTLES, 2006; TALLADA; PALACIOS-ROJAS; ARMSTRONG, 2009); girassol (PÉREZ-VICHA; VELASCO; FERNÁNDEZ-MARTÍNEZ, 1998, FASSIO; COZZOLINO, 2004; CANTARELLI et al., 2009; GRUNVALD, 2012.); algodão (QUAMPAH, et al., 2012; HUANG, et al., 2013); canola (PETISCO et al., 2010); amendoim (TILLMAN; GORBET; PERSON, 2006; RAO et al., 2009.)

4.2. Objetivo Específico

✓ Desenvolver modelos PLS e SPA-MLR utilizando a espectroscopia NIR para predição do teor de ricina, de forma não destrutiva em sementes escarificadas de mamona.

4.3. Experimental

4.3.1. Aquisição de Amostras

O conjunto de amostras utilizado para determinação da ricina em sementes de mamoneira foi formado por três cultivares (BRS Energia, BRS Nordestina e BRS Paraguaçu), as quais foram cedidas pela Embrapa Algodão.

4.3.2. Instrumentação

Esta descrição foi relatada na seção 3.3.2.

4.3.3. Preparo de Amostra e Aquisição dos Espectros NIR

Realizou-se um estudo com objetivo de determinar: o melhor agente esscarificante (ácido sulfúrico ou peróxido de hidrogênio); e o melhor tempo de contato das sementes com este agente (5, 10 ou 20 minutos).

Inicialmente foram realizadas análises no NIR de 350 sementes intactas de cada cultivar de mamona (BRS Paraguaçu, BRS Nordestina e BRS Energia), em seguida iniciou-se o processo de esscarificação com ácido sulfúrico de 50 sementes para cada cultivar e para cada tempo já relatado, totalizando 450 sementes.

O procedimento da esscarificação envolveu as seguintes etapas: 1) Adição 5 mL de ácido sulfúrico P.A. concentrado a um balão Randall devidamente identificado; 2) Imersão de semente individual da mamona no balão; 3) Agitação automática por 5 minutos em mesa agitadora (modelo Tecnal TE – 424); 4) Retirada da semente com auxílio de uma pinça e lavagem em água corrente; 5) Secagem em estufa de circulação e renovação de ar (modelo SL 102) durante 4 h, a temperatura em torno de 25 °C; 6) Retirada manual da casca e obtenção do endosperma. Este procedimento foi repetido com alteração do tempo de agitação (Etapa 3) para 10 e 20 minutos. Utilizou-se a mesma metodologia para o peróxido de hidrogênio.

As sementes esscarificadas com o peróxido de hidrogênio desenvolveram apenas alteração na coloração das sementes, não sendo possível a retirada da casca. Adotou-se assim o ácido sulfúrico como agente esscarificante.

Após as etapas descritas realizou-se a aquisição dos espectros NIR de maneira individual para os 450 endospermas obtidos pela esscarificação com o ácido sulfúrico, conforme descrito na seção 3.3.3.

Para avaliar qual o melhor tempo de contato do ácido com as sementes realizou-se um teste de germinação com os 450 endospermas. Para este utilizou-se papel para germinação de semente pH neutro e água destilada autoclavada, conforme visualiza-se na **Figura 23**.

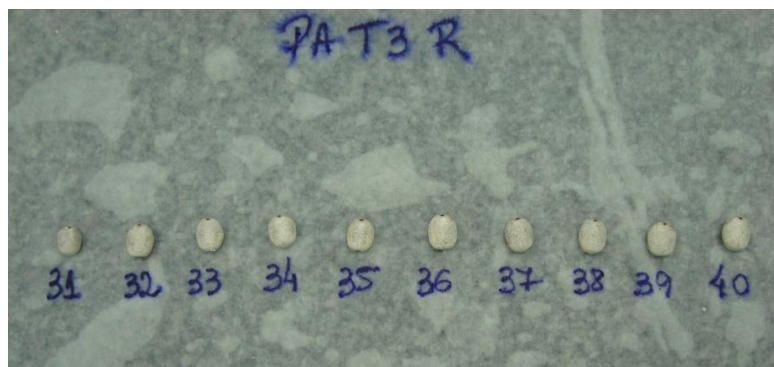


Figura 23 – Teste de germinação das sementes de mamona escarificadas com ácido sulfúrico.

Os 450 endospermas foram mantidos em germinador regulado a 25° C, durante 7 dias. Neste último foi realizada a contagem e o melhor resultado baseado no número de germinação foi obtido para o tempo de 5 minutos.

4.3.4. Programas Computacionais

O pré-processamento dos espectros originais e aplicação das técnicas de calibração (PLS e SPA-MLR) foi realizado utilizando-se o programa Unscrambler® 9.8. A aplicação dos algoritmos SPXY e SPA foi realizada em ambiente Matlab R2008a.

4.3.5. Tratamento Quimiométricos dos Dados

Como descrito na seção 3.5.1., a região espectral de 400 a 1099 nm foi descartada e a faixa compreendida entre 1100 a 2500 nm foi selecionada como a região de trabalho para calibração de ricina nas sementes de mamoneira.

As técnicas utilizadas para correção do efeito de linha de base foram: 1) primeira derivada Savitzky-Golay com janela de 15 pontos e polinômio de segunda ordem; 2) segunda derivada Savitzky-Golay com janela de 15 pontos e polinômio de segunda e 3) correção por offset. Após o pré-processamento, os espectros foram particionados empregando-se o algoritmo SPXY, em calibração (41 amostras), validação (15 amostras) e predição (13 amostras). Para cada tipo de correção de

linha de base empregada foi construído um modelo PLS com validação externa e o menor valor de RMSEP foi usado para otimização do pré-processamento.

4.3.6. Extração, Purificação e Determinação do Teor de Ricina

O procedimento para análise de ricina em sementes de mamona foi adaptado de [ANIMASHAUN; TOGUN; HUGHES,1994](#). As etapas necessárias para execução do ensaio são descritas a seguir.

4.3.6.1. Obtenção do Extrato Proteico

Para obtenção do extrato proteico utilizou-se 50 endospermas de cada cultivar e o procedimento envolveu as seguintes etapas: 1) maceração individual do endosperma, com auxílio de almofariz e pistilo de porcelana; 2) Pesagem do conteúdo marcerado em tubos de centrifugação e identificação; 3) Remoção da fração lipídica por meio da adição de hexano P.A. na proporção de 1:3 (m/ v) e agitação automática em temperatura 25 °C durante 12 h, em uma incubadora refrigerada com agitação (modelo TE - 424); 4) Centrifugação por 15 min a 4.000 rpm para obtenção do extrato bruto delipidado (centrífuga modelo 3 – 16 PK); 5) Com o auxílio de uma pipeta realizou-se a remoção do hexano e do óleo deixando-se no tubo apenas o farelo; 6) Colocação do farelo em uma placa Petri, com uma pinça metálica e secagem em estufa com circulação e renovação de ar por 4 h a 25 °C; 7) Pesagem de 250 mg de farelo (obtido a partir de endosperma individual) em tubo de Eppendorf[®] devidamente identificado; 8) Adição, à amostra de 1 mL de água destilada em cada tubo de Eppendorf; 9) Agitação dos tubos por 10 min em equipamento do tipo Vórtex (modelo AP 56); 10) Centrifugação dos tubos durante 15 min a 14.000 rpm (centrífuga modelo MCD 2.000) e 11) Transferência do sobrenadante para um tubo de Eppendorf, limpo e identificado.

4.3.6.2. Purificação da Ricina

A purificação da ricina foi realizada por meio da identificação de frações proteicas em sistema de cromatografia de exclusão molecular (**Figura 24**) cuja fase móvel foi o ácido trifluoroacético 0,1% (TFA), fase estacionária de Sephadex G-50 e detecção em 254 nm. Obtendo-se, assim, o perfil cromatográfico (**Figura 25**).



Figura 24 - Cromatográfico de exclusão molecular da BIO-RAD.

As frações correspondentes às proteínas com mais de 60 KDa, entre elas a ricina, foram coletadas sempre no primeiro pico, conforme destacado na **Figura 25** e armazenadas em vidros âmbar, devidamente identificados para posterior quantificação.

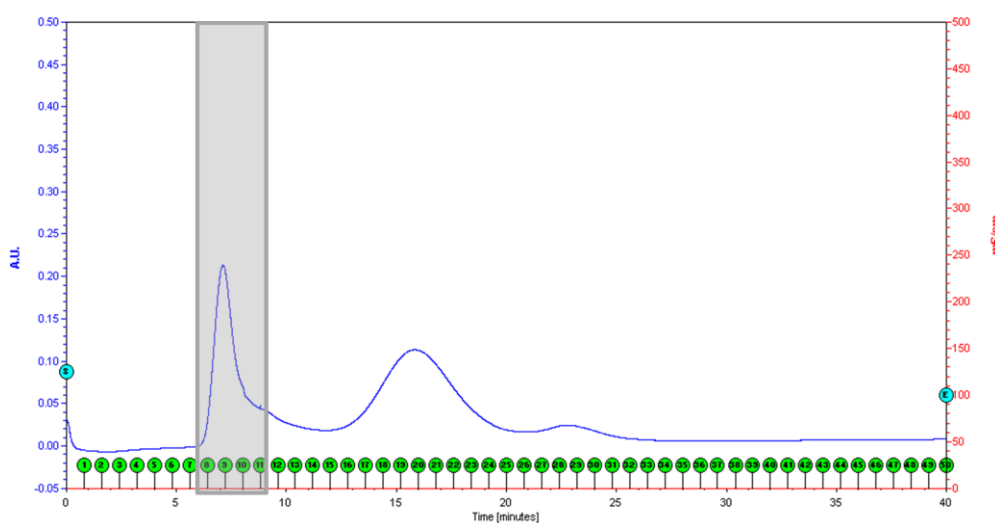


Figura 25 - Perfil cromatográfico para uma amostra de extrato proteico do endosperma da mamoneira.

4.3.6.3. Preparação da Curva de Calibração

O método de Bradford foi utilizado para quantificação de ricina (BRADFORD, 1976). Na construção da curva analítica para dosagem da ricina usou-se, como proteína padrão, a albumina bovina $1\mu\text{g}/\mu\text{L}$, a qual possui massa de 67 KDa.

As soluções de albumina bovina foram preparadas nas concentrações de $0,1\mu\text{g}/\mu\text{L}$, $0,08\mu\text{g}/\mu\text{L}$, $0,06\mu\text{g}/\mu\text{L}$, $0,05\mu\text{g}/\mu\text{L}$, $0,04\mu\text{g}/\mu\text{L}$, $0,02\mu\text{g}/\mu\text{L}$ e $0,01\mu\text{g}/\mu\text{L}$.

A cada $0,500\text{ mL}$ de amostra foram adicionados $2,0\text{ mL}$ do reagente comercial de Bradford (Sigma-Aldrich). A mistura com o reagente permaneceu em contato por 10 min com ausência de luz.

As leituras foram feitas em triplicata no comprimento de onda de 595 nm para o qual foi construída uma curva analítica de calibração.

4.4. Resultados e Discussão

4.4.1. Espectros NIR

Os espectros na região NIR foram obtidos no modo reflectância em que o perfil característico para os endospermas das sementes de mamona nessa região está representado na **Figura 26**.

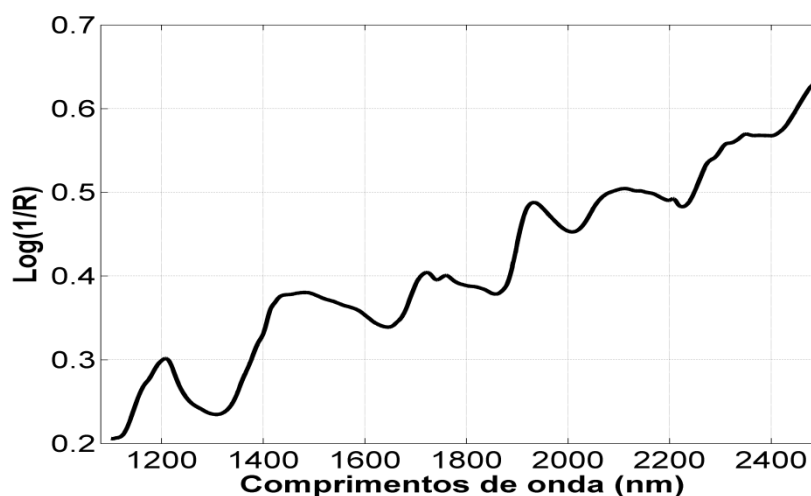


Figura 26 - Espectro do endosperma da semente de mamona.

4.4.2. Pré-processamento dos espectros

Ao analisar os perfis cromatográficos dos 150 endospermas, notou-se que a formação do primeiro pico (correspondente a ricina), ocorreu apenas em 69 endospermas. Deste total, 25 são oriundos da BRS Energia, 25 BRS Nordestina e 19 BRS Paraguaçu.

Na **Figura 27** é ilustrado o conjunto dos 69 espectros dos endospermas das sementes de mamona. É possível observar, nos espectros, um desvio sistemático de linha de base.

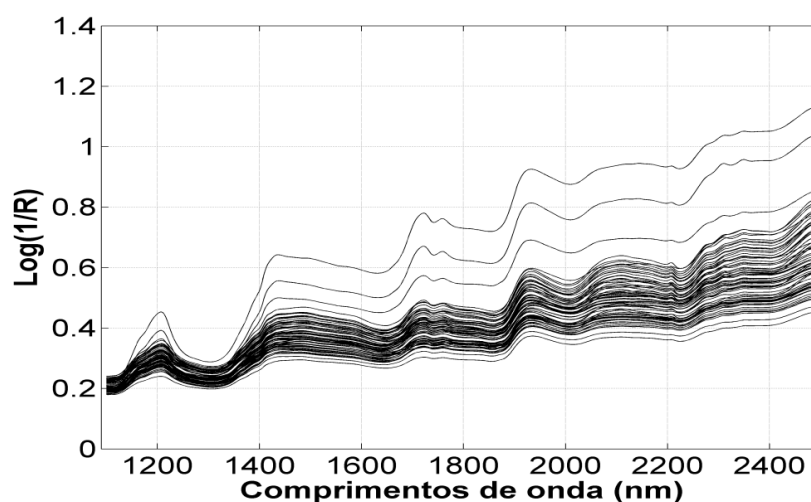


Figura 27 - Conjunto dos 69 espectros das amostras do endosperma da mamona.

O melhor pré-processamento dos espectros foi obtido com a aplicação da primeira derivada Savitzky-Golay com janela de 15 pontos e polinômio de segunda ordem cujos espectros derivativos são ilustrados na **Figura 28**.

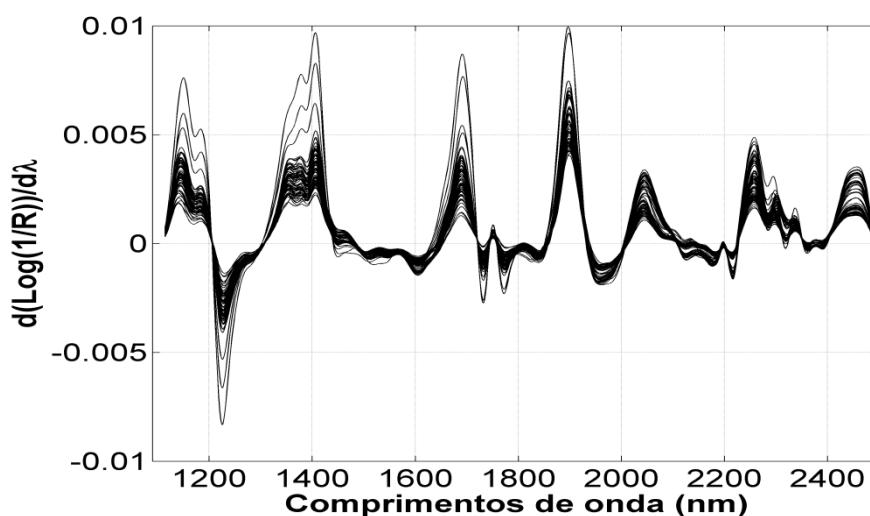


Figura 28 - Espectros derivativos das amostras do endosperma da mamona.

Observando os espectros derivativos, o perfil de linha de base devido ao efeito do espalhamento no modo de reflectância foi corrigido. Este conjunto de dados passou a ser usado em todos os cálculos subsequentes.

4.4.3. Construção dos Modelos de Calibração Multivariada

Os espectros NIR foram relacionados à concentração de ricina obtida pelo método de referência das amostras para construção do modelo de calibração. Para construção dos modelos PLS e SPA-MLR foram empregadas e avaliadas duas técnicas: validação externa e validação cruzada. A faixa de calibração variou entre 0,8 a 3,0 % (m/ m).

4.4.3.1. Modelo de Calibração por PLS

Os modelos PLS construídos com as duas técnicas citadas forneceram os parâmetros descritos na **Tabela 5**.

Tabela 5 - Parâmetros da calibração do modelo PLS.

Modelo	RMSEC (%m/m)	RMSECV/RMSEV (%m/m)	bias(val)	Nº Variáveis Latentes
PLS	0.2	0.4	0.17	10
PLS(CV)	0.2	0.6	0.01	10

O modelo PLS desenvolvido com a técnica de validação externa obteve o mesmo valor de RMSEC e número de variáveis do modelo que utilizou validação cruzada, este último modelo apresentou um maior erro para as amostras do conjunto de validação.

4.4.3.2. Modelo de calibração por SPA-MLR

Na **Figura 29** são apresentados os gráficos da função de custo associado à seleção de variáveis, usando-se o SPA-MLR. Na **Figura 29 (a)** visualiza-se que 17 variáveis foram selecionadas ao utilizar a técnica de validação externa, quando foi aplicado a técnica de validação cruzada apenas 9 variáveis foram selecionadas (**Figura 29 (b)**).

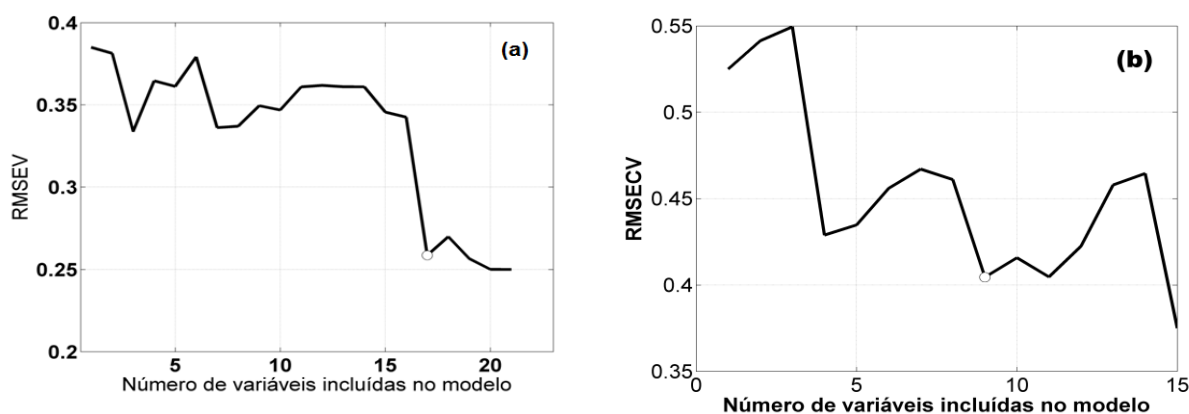


Figura 29 - Gráfico da função de custo SPA-MLR (a) validação externa e (b) validação cruzada.

O ponto em destaque na **Figura 29** sinaliza o mínimo local que não possui diferença estatística do mínimo global sendo, portanto, a quantidade de variáveis selecionadas pelo SPA-MLR.

Na **Figura 30** os comprimentos de onda selecionados são indexados no espectro médio das amostras de calibração.

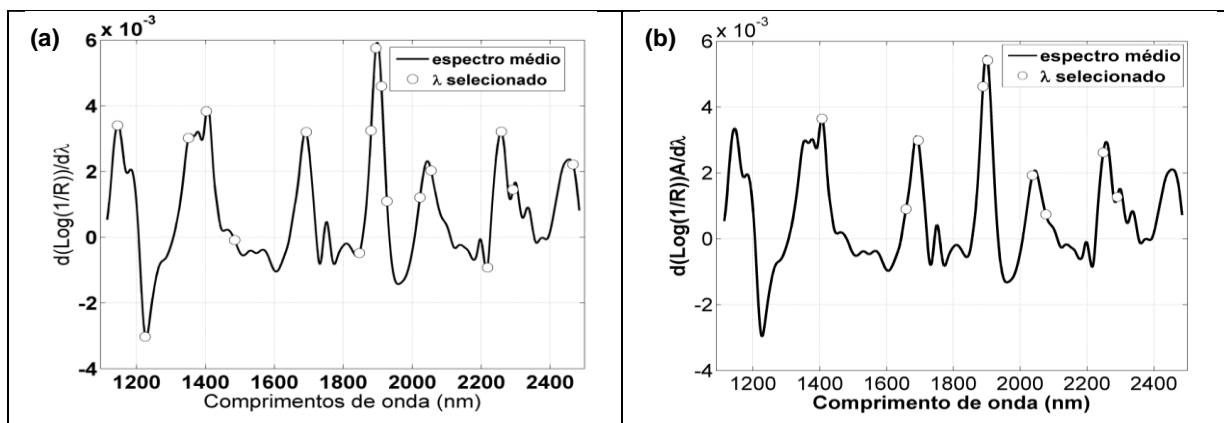


Figura 30 - Variáveis selecionadas pelo SPA-MLR (a) validação externa e (b) validação cruzada.

O gráfico da **Figura 30 (a)** mostra que as variáveis selecionadas se distribuem por toda a faixa espectral. Ao utilizar a validação cruzada **Figura 30 (b)** percebe-se que essa faixa torna-se menor.

Na **Tabela 6** são apresentados os parâmetros estatísticos do modelo SPA-MLR. Foram avaliadas as técnicas de validação externa e validação cruzada.

Tabela 6 - Parâmetros da calibração do modelo SPA-MLR.

Modelo	RMSEC (%m/m)	RMSECV/RMSEV (%m/m)	bias(val)	Nº Variáveis
SPA-MLR	0.2	0.3	0.16	17
SPA-MLR (CV)	0.3	0.4	0.05	9

Ao analisar a **Tabela 6** observa-se que apesar do número de variáveis ser menor no modelo SPA-MLR (CV) os valores dos erros para o conjunto de calibração e validação são similares.

4.4.3.2. Avaliação dos Modelos no Conjunto de Predição

Na **Tabela 7** é ilustrado o resumo da predição para os modelos PLS e SPA-MLR, utilizando as duas técnicas de validação. Observa-se que o modelo SPA-MLR forneceu resultado de RMSEP similar ao modelo PLS e com maior coeficiente de correlação, quando foi utilizada a validação externa. Ao analisar o RMSEP dos dois modelos obtidos a partir da técnica de validação cruzada observa-se que os valores foram iguais, porém maiores quando comparados com os obtidos com a validação

externa. Pode-se notar também que o modelo SPA-MLR obtido com a técnica de validação cruzada apresentou uma menor correlação.

Tabela 7 - Parâmetros estatísticos da predição

MODELO	RMSEP (%m/m)	r	bias
PLS	0.24	0.6	0.07
PLS(CV)	0.35	0.6	0.09
SPA-MLR	0.22	0.8	0.09
SPA-MLR (CV)	0.35	0.5	0.22

A precisão dos modelos foi avaliada por meio da região elíptica de confiança (FRANCO et al., 2002). O resultado desse teste é apresentado na **Figura 31**. Nesta é possível observar que, a partir da elipse de confiança, ambos os modelos obtidos utilizando-se a validação externa contêm o ponto ideal. Isso permite inferir, nesses modelos a ausência de erros sistemáticos significativos. Porém percebe-se que os modelos obtidos a partir da validação cruzada não contêm o ponto ideal.

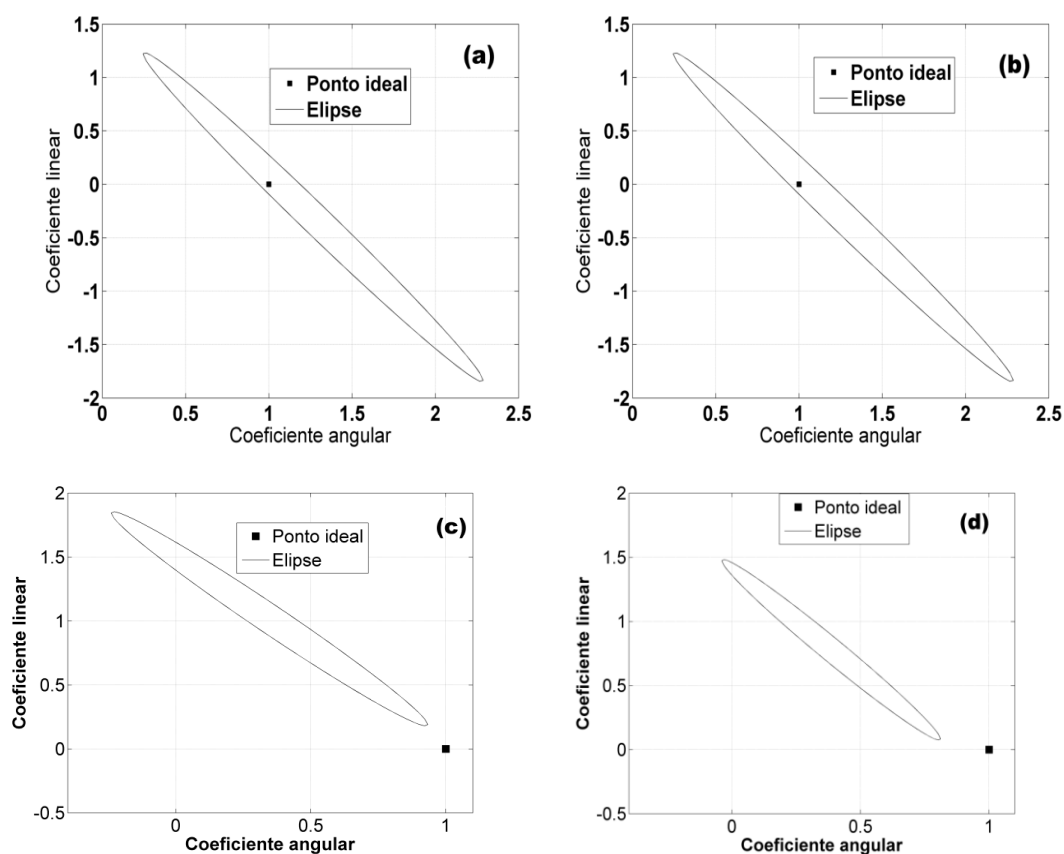


Figura 31 - Elipse de confiança para os modelos (a) PLS, (b) SPA-MLR, utilizando validação externa e (c) PLS, (d) SPA-MLR, utilizando validação cruzada.

4.5. Considerações Finais

O modelo SPA-MLR forneceu resultado de RMSEP similar ao PLS e melhor correlação ao utiliza-se a validação externa.

Os resultados RMSEP obtidos com a validação cruzada foram maiores independente dos modelos utilizados e modelo SPA-MLR obteve ainda uma menor correlação.

Ao avaliar os modelos usando-se a região elíptica de confiança, os mesmos não evidenciam erros sistemáticos significativos quando obtidos com a validação externa.

CAPÍTULO 5

Conclusões

5. CONCLUSÕES

A espectroscopia NIR aliada aos modelos SIMCA e SPA-LDA forneceu desempenho eficiente para classificação de sementes individuais, intactas e com alta frequência analítica de duas cultivares comerciais de mamona.

Em modelos de calibração para predição de ricina em sementes escarificadas de mamona, a espectroscopia NIR e as técnicas de PLS e SPA-MLR, são precisas, menos laboriosas que o método de referência, não destrutivas, rápidas e com menor custo para alta demanda de ensaios.

Os métodos de classificação e de calibração desenvolvidos são estratégias promissoras para seleção assistida e expedita de características fenotípicas em genótipos de mamona sob melhoramento genético.

5.1. Propostas Futuras

- ✓ Explorar outras técnicas quimiométricas, tais como SPA, com algoritmo genético e busca exaustiva, dentre outros modelos, para a classificação de cultivares de mamona.
- ✓ Explorar as técnicas de imagens para prospecção de genótipos com características de baixo teor de ricina e distribuição do perfil de composição.
- ✓ Estudar a viabilidade de empregar ricina purificada por exclusão molecular e liofilizada na etapa de calibração do método de dosagem do teor de ricina.

REFERÊNCIAS

- ALBUQUERQUE, A. R. **Autoxidação de Ésteres Metílicos de Ácidos Graxos: Estudo Teórico-Experimental**. 2010. 120 f. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa, 2010.
- ANDERSSON, M. A comparison of nine PLS1 algorithms. **Journal Chemometrics**, 23: 518, 2009.
- ANIMASHAUN, T ; TOGUN, RA.; HUGHES, RC. Characterization of isolectins in tetracarpidium-conophorum seeds (nigerian walnut). **Glycoconjugate Journal**, 11: 299, 1994.
- ANJANI, K. Castor genetic resources: A primary gene pool for exploitation. **Industrial Crops and Products**, 35: 1, 2012.
- ARAÚJO, M. C. U. et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. **Chemometrics and Intelligent Laboratory Systems**, 57:65, 2001.
- AUDI, J. et al. Ricin poisoning. A comprehensive review. **JAMA**, 294:2342, 2005.
- AZEVEDO, D. M. P. et al. Manejo Cultural. In: AZEVEDO, D. M. P. de; LIMA, E. F. (Eds). **O agronegócio da Mamona no Brasil**. Brasília: Embrapa Comunicação para Transferência de Tecnologia, 2007. cap. 10, p.223-253.
- BALABIN, R. M., SAFIEVA, R. Z. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. **Fuel**, 87: 1096, 2008.
- BALABIN, R. M.; SAFIEVA, R. Z. Biodiesel classification by base stock type (vegetable oil) using near infrared spectroscopy data. **Analytica Chimica Acta**, 689: 190, 2011.
- BALABIN, R. M.; SAFIEVA, R. Z.; LOMAKINAC, E. I. Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. **Analytica Chimica Acta**, 671: 27, 2010.
- BALDONI, A. B. **Acúmulo de ricina em sementes de mamona e silenciamento do gene em planta geneticamente modificada**. 2010. 82 f. Tese (Doutorado em Biologia Molecular) – Universidade de Brasília, Brasília, 2010.
- BALDONI, A. B. et al. Variability of ricin content in mature seeds of castor bean. **Pesquisa Agropecuária Brasileira**, 46: 776, 2011.
- BARBOSA, L.C. de A. **Espectroscopia no infravermelho na caracterização de compostos orgânicos**. Viçosa: UFV, 2008.
- BARROS NETO, B.; SCARMINIO, I. S.; BRUNS, R. E. 25 Anos de quimiometria no Brasil. **Química Nova**, 29: 1401, 2006.

BAYE, T. M.; PEARSON, T. C.; SETTLES, A. M. Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. **Journal of Cereal Science**, 43: 236, 2006.

BEEBE, K.R.; PELL, R.J; SEASHOLTZ, M.B. **Chemometrics A Practical Guide**. New York: John Wiley & Sons, 1998.

BELTRÃO, N. E. et al. Ecofisiologia da mamoneira (*Ricinus communis* L.). In: BELTRÃO, N. E. de M.; OLIVEIRA, M. I. P. **Ecofisiologia das culturas de algodão, amendoim, gergelim, mamona, pinhão-manso e sisal**. Brasília: Embrapa Comunicação para Transferência de Tecnologia, 2011. cap. 5, p.195-256.

BELTRÃO, N. E. M. OLIVEIRA, M. I. P. **Detoxicação e Aplicações da Torta de Mamona**. Campina Grande: Embrapa Algodão, 2009. 35p. Documento, 217.

BELTRÃO, N. E. M.; AZEVEDO, D. M. P. Fitologia. In: AZEVEDO, D. M. P. de; BELTRÃO, N. E. de M. (Eds). **O agronegócio da mamona no Brasil**. 2. ed. Brasília: Embrapa Comunicação para Transferência de Tecnologia, 2007. cap. 5, p.119-137.

BLANCO, M. et al. Near-infrared spectroscopy in the pharmaceutical industry. **Analyst**, 123: 135, 1998.

BRADFORD, M. M. A Rapid and sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein-dye binding. **Analytical Biochemistry**, 72: 248, 1976.

BRAGA, J. W. B.; POPPI, R. J. Validação de modelos de calibração multivariada: uma aplicação na determinação de pureza polimórfica de carbamazepina por espectroscopia no infravermelho próximo. **Química Nova**, 27: 1004, 2004.

BRANDEN, K. V.; HUBERT, M. Robust classification in high dimensions based on the SIMCA Method. **Chemometrics and Intelligent Laboratory Systems**, 79: 10, 2005.

BRERETON, R. **Chemometrics for Pattern Recognition**. John Wiley & Sons: Chichester, 2007.

BRERETON, R. G. **Chemometrics: data Analysis for the laboratory and chemical plant**. New York: John Wiley & Sons, 2003.

BRERETON, R. G. Introduction to multivariate calibration in analytical chemistry. **Analyst**, 125: 2125, 2000.

BROWN, S.D. Chemical systems under indirect observation: Latent properties and chemometrics. **Applied Spectroscopy**, 49: 14, 1995.

BRUNS, R. R.; FAIGLE, J. F. G. Quimiometria. **Química Nova**, 8:84, 1985.

BUENO, A. F. **Desenvolvimento de um analisador de processo por espectroscopia no infravermelho próximo (NIR) para precisão de propriedades de derivados de petróleo.** 2011. 264 f. Tese (Doutorado Química) - Universidade Estadual de Campinas, Campinas, 2011.

CAETANO, V. F. et al. Prediction of mechanical properties of poly(ethylene terephthalate) using infrared spectroscopy and multivariate calibration. **Journal of Applied Polymer Science**, 127: 3441, 2013.

CANGEMI, J. M.; SANTOS, A. M.; CLARO NETO, S. A revolução verde da mamona. **Química Nova Na Escola: Química e Sociedade**, 32:3, 2010.

CANTARELLI, M. A. et al. Determination of oleic acid in sunflower seeds by infrared spectroscopy and multivariate calibration method. **Talanta**, 80: 489, 2009.

CARNEIRO, M. E. **Classificação de lâminas de madeira de pinus spp por espectroscopia óptica.** 2008, 97 f. Dissertação (Mestrado em Engenharia Florestal) – Universidade Federal do Paraná, Curitiba, 2008.

CASALE, M. et al. Characterisation of table olive cultivar by NIR spectroscopy. **Food Chemistry**, 122: 1261, 2010.

CENTER, V. et al. Elimination of Uninformative Variables for Multivariate Calibration. *Analytical Chemistry*, 68: 3851, 1996.

CERQUEIRA, E. O.; POPPI, R. J.; KUBOTA, L. T., Utilização de filtro de transformada de fourier para a minimização de ruídos em sinais analíticos. **Química Nova**, 23:690, 2000.

CÉSAR, A.S., BATALHA M.O. Biodiesel production from castor oil in Brazil: A difficult reality. **Energy Policy**, 38:4031, 2010.

CHAGAS, I. P. **Desenvolvimento de um Fotômetro Portátil NIR Para Determinação do Teor de Água no Álcool Combustível e do Teor de Etanol na Gasolina.** 2006. 151 f. Tese (Doutorado em Química) - Universidade Estadual de Campinas, Campinas, 2006.

CHAKRAVARTULA, S. V. S.; GUTTARLA, N. Amino acids of ricin and its ptypeptides. **Natural Product Research**, 22:258, 2008.

CHAN, A. P. et al. Draft genome sequence of the oilseed species *Ricinus communis*. **Nature biotechnology**, 28:9, 2010.

CHEN, G. L. et al. Nondestructive assessment of amino acid composition in rapeseed meal based on intact seeds by near-infrared reflectance spectroscopy. **Animal Feed Science and Technology**, 165:111, 2011.

CHEN, L. et al. Classification of Chinese honeys according to their floral origin by Near Infrared Spectroscopy. **Food Chemistry**, 135:338, 2012.

CHIERICE, G. O.; CLARO NETO, S. Aplicação Industrial do óleo, In: AZEVEDO, D. M. P. de; LIMA, E. F. (Eds). **O agronegócio da Mamona no Brasil**. Brasília: Embrapa Comunicação para Transferência de Tecnologia, 2001. cap. 18, p.419-447.

COSTA FILHO, C. A., POPPI, R. J. Algoritmo Genético em química. **Química Nova**, 22: 405, 1999.

COSTA, M.N. et al. Genetic divergence on castor bean accesses and cultivars through multivariate analysis. **Pesquisa Agropecuária Brasileira**, 41: 1617, 2006.

DANTAS FILHO, H. A. **Desenvolvimento de técnicas quimiométricas de compressão de dados e de redução de ruído instrumental aplicadas a óleo diesel e madeira de eucalipto usando espectroscopia NIR**. 2007. 158 f. Tese (Doutorado em Química) - Universidade Estadual de Campinas, Campinas, 2007.

DANTAS, H. V. et al. An automatic flow system for NIR screening analysis of liquefied petroleum gas with respect to propane content. **Talanta**, 106: 158, 2013.

DASZYKOWSKI, M. et al. Robust statistics in data analysis — A review Basic concepts. **Chemometrics and Intelligent Laboratory Systems**, 85: 203, 2007.

DEMANT, C. A. R. **Metodologia para quantificar ricina em sementes de mamona com o uso de *Caenorhabditis elegans***. 2008. 54 f. Tese (Doutorado em Agronomia) - Universidade Estadual Paulista “Júlio de Mesquita Filho, Botucatu, 2008.

DERDE, M.P.; MASSART, D.L. Comparison of the Performance of the Class Modelling Techniques UNEQ, SIMCA, and PRIMA. **Chemometrics and Intelligent Laboratory Systems**, 4: 65, 1988.

DINIZ, P. H. G. D. et al. Using a simple digital camera and SPA-LDA modeling to screen teas. **Analytica Methods**, 4: 2648, 2012.

DOAN, L. G. Ricin: Mechanism of Toxicity, Clinical Manifestations, and Vaccine Development. A Review. **Journal of Toxicology Clinical Toxicology**, 42: 201, 2004.

EFRON, B. **The Jack-knife, the bootstrap and other resampling plans**. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 1982.

EGIDIO, V. et al. Confirmation of brand identity in foods by near infrared transreflectance spectroscopy using classification and class-modelling chemometric techniques — The example of a Belgian beer. **Food Research International**, 44: 544, 2011.

EWING, G.W. **Métodos instrumentais de análise química**. Tradução: ALBANESE, A. G.; CAMPOS, J. T. de S. São Paulo: Blucher, 2011.

FASSIO, A.; COZZOLINO, D. Non-destructive prediction of chemical composition in sunflower seeds by near infrared spectroscopy. **Industrial Crops and Products**, 20: 321, 2004.

FERNANDES, D. D. S., et al. Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection. **Talanta**, 87: 30, 2011.

FERNANDES, K.V. et al. Simultaneous allergen inactivation and detoxification of castor bean cake by treatment with calcium compounds. **Brazilian Journal of Medical and Biological Research**, 45: 1002, 2012.

FERNÁNDEZ-CUESTA, A. I.; FERNÁNDEZ-MARTÍNEZ J. M. VELASCO, L. Identification of High Oleic Castor Seeds by Near Infrared Reflectance Spectroscopy. **Journal of the American Oil Chemists Society**, 89:431, 2012.

FERNÁNDEZ-CUESTA, A.; FERNÁNDEZ-MARTÍNEZ, J. M.; VELASCO, L. Identification of High Oleic Castor Seeds by Near Infrared Reflectance Spectroscopy. **Journal of the American Oil Chemists' Society**, 89: 431, 2011.

FERREIRA, M. A. J. da F. **Utilização das técnicas de marcadores moleculares na genética de populações, na genética quantitativa e no melhoramento de plantas**. Boa Vista: Embrapa Roraima, 2003. 63p. Documento, 1.

FERREIRA, M. M. C. Multivariate QSAR. **Journal of the Brazilian Chemical Society**, 13: 742, 2002.

FERREIRA, M. M. C., MONTANARI, C. A., GAUDIO, A. C. Seleção de Variáveis em QSAR. **Química Nova**, 3: 439, 2002.

FERREIRA, M.M.C.; et. al. Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, 22:724, 1999.

FLATEN, G. R.; GRUNG, B.; KVALHEIM, O. M. A method for validation of reference sets in SIMCA modeling. **Chemometrics and Intelligent Laboratory Systems**, 72: 101, 2004.

FLUMIGNAN, D. L. **Caracterização da qualidade e precisão dos parâmetros físico-químicos de gasolinas comerciais brasileiras através da aplicação de métodos quimiométricos em perfis (fingerprintings) espectroscópicos de ressonância magnética nuclear**. 2010. 225f. Tese (Doutorado em Química). Universidade Estadual Paulista, Araraquara, 2010.

FORINA, M., CASOLINO, C., MILLAN, C. P. Iterative predictor weighting (IPW) pls: a technique for the elimination of useless predictors in regression problems. **Journal Chemometrics**, 13: 165, 1999.

FORNAZIERI JÚNIOR, A. F. **Mamona: uma rica fonte de óleo e de divisas**. São Paulo: Cone, 1986. 72p.

FRANCO, V. G. et al. **Teaching Chemometrics with a Bioprocess: Analytical Methods Comparison Using Bivariate Linear Regression**. *Chemical Educator*, 7: 265, 2002.

FRANZ, D.R.; JAAX, N.K. U.S. Army Medical Research Institute of Infectious Disease, Fort Detrick, Frederick, Maryland, Chapter 32, RicinToxin. 1997.

FREIRE, E. C.; et al. Melhoria Genética. In: AZEVEDO, D. M. P. de; BELTRÃO, N. E. de M. (Eds). **O agronegócio da mamona no Brasil**. 2. ed. Brasília: Embrapa informação tecnológica, 2007. cap. 8, p.171-194.

GAJERA, B. B. et al. Assessment of genetic diversity in castor (*Ricinus communis* L.) using RAPD and ISSR markers. **Industrial Crops and Products**, 32: 491, 2010.

GALTIER, O. et al. Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions. **Vibrational Spectroscopy**, 55: 132, 2011.

GALVÃO, R. K. H, et al. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. **Chemometrics and Intelligent Laboratory Systems**, 92: 83, 2008.

GALVÃO, R. K. H. et al. A method for calibration and validation subset partitioning. **Talanta**, 67: 736, 2005.

GALVÃO, R. K. H. et al. Estudo comparativo sobre filtragem de sinais instrumentais usando transformadas de Fourier e Wavelet. **Química Nova**, 24: 874, 2001.

GAMBARRA-NETO, F. F. et al. Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis. **Talanta**, 77: 1660, 2009.

GELADI, P., KOWALSKI, B. R. Partial Least-square: A tutorial. **Analytica Chimica Acta**, 185: 17, 1986.

GHASEMI-VARNAMKHAJASTI, M. et al. Screening analysis of beer ageing using near infrared spectroscopy and the Successive Projections Algorithm for variable selection. **Talanta**, 89: 286, 2012.

GLOVER, F. Tabu Search — Part I. **Journal on Computing**, 1: 190, 1989.

GOMES, A. de A. **Algoritmo das Projeções Sucessivas aplicado à seleção de variáveis em regressão PLS**. 2012. 121 f. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa, 2012.

GONZÁLEZ, A. G. Use and misuse of supervised pattern recognition methods for interpreting compositional data. **Journal of Chromatography A**, 1158:215, 2007.

GREENFIELD, R.A. et al. Microbiological, biological, and chemical weapons of warfare and terrorism. **The American Journal of the Medical Sciences**, 323: 326, 2002.

GRUNVALD, A. K. Discriminant Analysis of Sunflower Seeds for Fatty Acid Composition Using NIR Spectroscopy. **Journal of the American Oil Chemists Society**, 89:995, 2012.

HALLING, K. C. et al. Genomic cloning and characterization of ricin gene from ricinus. **Communis Nucleic Acids Research**, 13:8019, 1985.

HARTLEY, M. R.; LORD, J. M. Cytotoxic ribosome-inactivating lectins from plants. *Biochimica et Biophysica Acta (BBA)*. **Proteins & Proteomics**, 1701:1, 2004.

HOFFMAN, L.V.et al. **Ricina: Um Impasse para Utilização da Torta de Mamona e suas Aplicações**. Campina Grande: Embrapa Algodão, 2007. 26p. Documento, 174.

HONORATO, F. A. et al. Robust modeling for multivariate calibration transfer by the successive projections algorithm. **Chemometrics and Intelligent Laboratory Systems**, 76:65, 2005.

HONORATO, F. A. **Previsão das propriedades das gasolinas do Nordeste empregando espectroscopia NIR/MID e transferência de calibração**. 2006. 106 f. Tese (Doutorado em Química) – Universidade Federal de Pernambuco, Recife, 2006.

HUANG, Z. et al. Feasibility study of near infrared spectroscopy with variable selection for non-destructive determination of quality parameters in shell-intact cottonseed. **Industrial Crops and Products**, 43: 654, 2013.

INSAUSTI, M. et al. Screening analysis of biodiesel feedstock using UV–vis, NIR and synchronous fluorescence spectrometries and the successive projections algorithm. **Talanta**, 97: 579, 2012.

JACKSON, L. S.; TOLLESON, W. H.; CHIRTEL, S. J. Thermal Inactivation of Ricin Using Infant Formula as a Food Matrix. **Journal of Agricultural and Food Chemistry**, 54: 7300, 2006.

KAMAL-ELDIN, A.; ANDERSSON, R. A Multivariate Study of the Correlation Between Tocopherol Content and Fatty Acid Composition in Vegetable Oils. **Journal of the American Oil Chemists' Society**, 74: 375, 1997.

KENNARD, R. W.; STONE, L. A. Computer-aided design of experiments, **Technometrics**, 11:137, 1969.

KIM, K. S. et al. Use of Near-Infrared Spectroscopy for Estimating Fatty Acid Composition in Intact Seeds of Rapeseed. **Journal of Crop Science and Biotechnology**, 10: 15, 2007.

KUBELKA, P.; MUNK, F. Ein beitrage zur optik der farbanstriche. **Z. Technische Physik**, 12:593, 1931.

LATHAUWER, L., MOOR, B., VANDEWALLET, L. A multilinear singular value decomposition. **Journal on Matrix Analysis and Applications**, 21: 1253, 2000.

LEE, J. H.; CHOUNG, M. G. Nondestructive determination of herbicide-resistant genetically modified soybean seeds using near-infrared reflectance spectroscopy. **Food Chemistry**, 126: 368, 2011.

LER, S. G.; LEE, F. K.; GOPALAKRISHNAKONE P. Trends in detection of warfare agents Detection methods for ricin, staphylococcal enterotoxin B and T-2 toxin. **Journal of Chromatography A**, 1133:1, 2006.

LIMA, K. M. G. et al. Sensores ópticos com detecção no infravermelho próximo e médio. **Química Nova**, 32: 1635, 2009.

LIMA, K. M.G.; RAIMUNDO JÚNIOR, I. M., PIMENTEL, M. F. Simultaneous determination of BTX and total hydrocarbons in water employing near infrared spectroscopy and multivariate calibration. **Sensors and Actuators B**, 160: 691, 2011.

LIMA, R. L. S. et al. Blends of castor meal and castor husks for optimized use as organic fertilizer. **Industrial Crops and Products**, 33: 364, 2011.

LIRA, L. F. B. **Desenvolvimento de métodos analíticos para monitoramento da qualidade do biodiesel e suas misturas**. 2010. 145 f. Tese (Doutorado em Química) – Universidade Federal de Pernambuco, Recife, 2010.

LUBELLI, C. et al. Detection of ricin and other ribosome-inactivating proteins by an immuno-polymerase chain reaction assay. **Analytical Biochemistry**, 355: 102, 2006.

LUCASIU, C.B., KATEMAN, G. Understanding and using genetic algorithms Part 1. Concepts, properties and context. **Chemometrics and Intelligent Laboratory Systems**, 19: 1, 1993.

MAESSCHALCK, R.; JOUAN-RIMBAUD, D.; MASSART, D.L. Tutorial - The Mahalanobis distance. **Chemometrics and Intelligent Laboratory Systems**. 50: 1, 2000.

MILANI, M.; MIGUEL JÚNIOR, S. R.; SOUSA, R. de L. **Subespécies de mamona**. Campina Grande: Embrapa Algodão, 2009. 23p. Documento, 230.

MALTMAN, D. J. et al. Differential proteomic analysis of the endoplasmic reticulum from developing and germinating seeds of castor (*Ricinus communis*) identifies seed protein precursors as significant components of the endoplasmic reticulum. **Proteomics**, 7: 1513, 2007.

MARRETO, P. D. **Determinação simultânea de íons metálicos utilizando voltametria de redissolução anódica e métodos de calibração multivariada.** 2010. 176 f. Tese (Doutorado em Química) - Universidade Federal de São Carlos, São Carlos, 2010.

MARTENS, H., MARTENS, M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). **Food Quality and Preference**, 11: 5, 2000.

MARTENS, H., NAES T. **Multivariate Calibration.** John Wiley: New York, 1989.
MASSART, D. L. et al. **Journal Handbook of Chemometrics and Qualimetrics: Parte B,** Amsterdam: Elsevier, 1997.

MCGRATH, S. et al. Detection and Quantification of Ricin in Beverages Using Isotope Dilution Tandem Mass Spectrometry. **Analytical Chemistry**, 83: 2897, 2011.

MONTFORT, W. et al. The Three-dimensional Structure of Ricin at 2.8Å. **Journal of Biological Chemistry**, 262: 5398, 1987.

MOREIRA, E. D. T. et al. Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection. **Talanta**, 79: 1260, 2009.

MOREIRA, J.A.N. et al. **Melhoramento da mamoneira (Ricinus communis L.).** Campina Grande: Embrapa Algodão, 1996. 29p. Documento, 44.

NAES, et al. **A User-Friendly Guide to Multivariate Calibration and Classification.** Chichester, UK: NIR Publications, 2002.

NAES, T.; MARTENS, H. Multivariate calibration. II. Chemometric methods. **Trends in Analytical Chemistry**, 3: 266, 1984.

NAES, T.; MEVIK, B. H. Understanding the collinearity problem in regression and classification. **Journal of Chemometrics**, 15:413, 2001.

NORGAARD, L. **iToolbox Manual**, 2005.

NORGAARD, L., et al. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. **Applied Spectroscopy**, 54: 413, 2000.

NUNES, P. G. A. **Uma nova técnica para seleção de variáveis em calibração multivariada aplicada às Espectrometrias UV-VIS E NIR.** 2008. 121 f. Tese (Doutorado em Química) – Universidade Federal da Paraíba, João Pessoa, 2008.

OZAKI, Y. Near-Infrared Spectroscopy—Its Versatility in Analytical Chemistry. **Analytical Sciences**, 28: 545, 2012.

PAIVA, H. M. et al. A graphical user interface for variable selection employing the Successive Projections Algorithm. **Chemometrics and Intelligent Laboratory Systems**, 118: 260, 2012.

PASQUINI, C. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, 14:198, 2003.

PATIL, A.G. et al. Nondestructive estimation of fatty acid composition in soybean [Glycine max (L.) Merrill] seeds using Near-Infrared Transmittance Spectroscopy. **Food Chemistry**, 120: 1210, 2010.

PEREIRA, A. F. C., et al. NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection. **Food Research International**, 41: 341, 2008.

PÉREZ-VICHA, B.; Velasco, L.; FERNÁNDEZ-MARTÍNEZ, J. M. Determination of Seed Oil Content and Fatty Acid Composition in Sunflower Through the Analysis of Intact Seeds, Husked Seeds, Meal and Oil by Near-Infrared Reflectance Spectroscopy. **Journal of the American Oil Chemists Society**, 75: 547, 1998.

PESKE, S. T.; BARROS, A. C. S. A. Produção de Sementes. In: PESKE, S. T.; LUCCA, O. F.; BARROS, A. C. S. A. **Sementes: Fundamentos científicos e tecnológicos**. 3. ed. Pelotas: UFPel, 2012. cap. 1, p. 12-91.

PETISCO, C. et al. Measurement of quality parameters in intact seeds of Brassica species using visible and near-infrared spectroscopy. **Industrial Crops and Products**, 32:139, 2010.

PIERNA, J. A. F. et al. A Backward Variable Selection method for PLS regression (BVSPLS). **Analytica Chimica**, 642: 89, 2009.

PIMENTEL, M. F.; GALVÃO, R. K. H.; ARAÚJO, M. C. U. Recomendações para calibração em química analítica parte 2. Calibração multianálito. **Química Nova**, 31: 462, 2008.

PONTES, et al. Internal and external validation in SPA-LDA: A comparative study involving diesel/biodiesel blends. **NIR News**, 23:6, 2012.

PONTES, et al. Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification. **Talanta**, 85: 2159, 2011b.

PONTES, M. J. C. de. **Algoritmo das projeções sucessivas para a seleção de variáveis espectrais em problemas de classificação**. 2009. 144 f. Tese (Doutorado em Química) – Universidade Federal da Paraíba, João Pessoa, 2009.

PONTES, M. J. C., et al. Determining the quality of insulating oils using near infrared spectroscopy and wavelength selection. **Microchemical Journal**, 98: 254, 2011a.

PONTES, M. J. C., et al. The successive projections algorithm for spectral variable selection in classification problems. **Chemometrics and Intelligent Laboratory Systems**, 78: 11, 2005.

POVIA, G. S. **Determinação dos Parâmetros de Qualidade de Detergentes em Pó Utilizando Espectroscopia no Infravermelho Próximo**. 2007. 84 f. Dissertação (Mestrado em Química) - Universidade Estadual de Campinas, Campinas, 2007.

QUAMPAH, A. et al. Estimation of Oil Content and Fatty Acid Composition in Cottonseed Kernel Powder Using Near Infrared Reflectance Spectroscopy. **Journal of the American Oil Chemists Society**, 89: 567, 2012.

RAO, Y. et al. Quantitative and qualitative determination of acid value of peanut oil using near-infrared spectrometry. **Journal of Food Engineering**, 93: 249, 2009.

RIOVANTO, R. et al. Discrimination between Shiraz Wines from Different Australian Regions: The Role of Spectroscopy and Chemometrics. **Journal of Agricultural and Food Chemistry**, 2011.

SABIN, J. G.; FERRÃO, M. F.; FURTADO, J. C. Análise multivariada aplicada na identificação de fármacos antidepressivos. Parte II: Análise por componentes principais (PCA) e o método de classificação SIMCA. **Revista Brasileira de Ciências Farmacêuticas**, 40: 387, 2004.

SANCHES, F. A. C. et al. Near-infrared spectrometric determination of dipyrone in closed ampoules. **Talanta**, 92: 84, 2012.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least-squares procedures. **Analytical Chemistry**, 36: 1627, 1964.

SAVY FILHO, A. **Mamona: tecnologia agrícola**. Campinas: EMOPI, 2005. 105 p.

SCAFI, S. H. F. **Espectroscopia no Infravermelho Próximo para identificação de medicamentos falsificados**. 2000. 139 f. Dissertação (Mestrado em Química) - Universidade Estadual de Campinas, Campinas, 2000.

SCAFI, S. H. F. **Sistema de Monitoramento em Tempo Real de Destilações de Petróleo e Derivados Empregando a Espectroscopia no Infravermelho Próximo**. 2005. 214 f. Tese (Doutorado em Química) - Universidade Estadual de Campinas, Campinas, 2005.

SENA, M. M. et al. Avaliação do uso de métodos quimiométricos em análise de solos. **Química Nova**, 23:4, 2000.

SEVERINO, L. S. et al. A Review on the Challenges for Increased Production of Castor. **Agronomy Journal**, 104: 853, 2012

SHAMSIPUR, M., et al. Ant colony optimisation: a powerful tool for wavelength selection. **Journal Chemometrics**, 20: 146, 2006.

SILVA, A. C. et al. Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods. **Talanta**, 93: 129, 2012.

SIMÕES, S. S. **Desenvolvimento de métodos validados para a determinação de captopril usando espectrometria NIR e calibração multivariada**. 2008. 98 f. Tese (Doutorado em Química) – Universidade Federal da Paraíba, João Pessoa, 2008.

SINELLI, N. et al. Varietal discrimination of extra virgin olive oils by near and mid infrared spectroscopy. **Food Research International**, 43: 2126, 2010.

SIRISOMBOON, P.; HASHIMOTO, Y.; TANAKA, M. Study on non-destructive evaluation methods for defect pods for green soybean processing by near-infrared spectroscopy. **Journal of Food Engineering**, 93: 502, 2009.

SKOOG, D.A.; HOLLER, F.J.; NIEMAN, T.A. **Princípios de análise Instrumental**. 6. ed. Porto Alegre: Bookman, 2009.

SOARES, S. F. C. et al. The successive projections algorithm, **Trends in Analytical Chemistry**, 42: 84, 2013.

SOARES, S. F. C. **Um novo critério para seleção de variáveis usando o Algoritmo das Projeções**. 2010. 107f. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa, 2010.

SOUZA, A. M. e POPPI, R. J. 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. **Química Nova**, 35: 223, 2012

STUMPE, B. et al. Application of PCA and SIMCA Statistical Analysis of FT-IR Spectra for the Classification and Identification of Different Slag Types with Environmental Origin. **Environmental Science Technology**, 46: 3964, 2012.

SUNDARAM, J. et al. Sensing of Moisture Content of In-Shell Peanuts by NIR Reflectance Spectroscopy. **Journal of Sensor Technology**, 2: 1, 2012.

TALLADA, J. G.; PALACIOS-ROJAS, N.; ARMSTRONG, P. R. Prediction of maize seed attributes using a rapid single kernel near infrared instrument. **Journal of Cereal Science**, 50:381, 2009.

TÁVORA, F. J. A. F. **A cultura da mamona**. Fortaleza: EPACE, 1982. 111 p.

TEOFILO, R. F., Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. **Journal Chemometrics**, 23: 32, 2009.

TILLMAN, B. L.; GORBET, D. W.; PERSON, G. Predicting oleic and linoleic acid content of single peanut seeds using near-infrared reflectance spectroscopy. **Crop Science**, 46: 2121, 2006.

VALDERRAMA, P. **Calibração multivariada de primeira e segunda ordem e figuras de mérito na quantificação de enantiômeros por espectroscopia**. 2009. 230 f. Tese (Doutorado em Química) - Universidade Estadual de Campinas, Campinas, 2009.

VASCONCELOS, F. V. C. de. **Uso da região espectral de sobretons para determinação do teor de biodiesel e classificação de misturas diesel/biodiesel adulteradas com óleo vegetal**. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa, 2011.

VECCHIA, P. T. D.; SILVA, C.A.R.; SOBRINHO TERCENIANO P. Use of molecular marker techniques in seed testing by brazilian seed companies. **Scientia Agricola**, 55: 79, 1998.

VERAS, G. et al. Classification of biodiesel using NIR spectrometry and multivariate techniques. **Talanta**, 83:565, 2010.

VIDAL, M. S. et al. **Seleção de Marcadores do Tipo Rapd para Caracterização Genética Ricinus communis L.** Campina Grande: Embrapa Algodão, 2005. 5p. Documento, 90.

VITALE, R. et al. A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics. **Chemometrics and Intelligent Laboratory Systems**, 121: 90, 2013.

WEISS, E. A. **Oilseed crops**. London: Longman, 1983. 660p.

WOLD, S. Pattern recognition by means of disjoint principal components models. **Pattern Recognition**, 8:127, 1976.

WOLD, S. Personal memories of the early PLS development. **Chemometrics and Intelligent Laboratory Systems**, 58:83, 2001.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal Component Analysis. **Chemometrics and Intelligent Laboratory Systems**, 2:37, 1987.

WOLD, S.; SJOSTROM, M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In **Chemometrics: Theory and Applications**, Washington: American Chemical Society, 1977. cap. 12, p 243-282.

XIAOBO, Z. et al. Variables selection methods in near-infrared spectroscop. **Analytica Chimica Acta**, 667: 14, 2010.

XIE, X.; KIRBY, J.; KEASLING, J. D. Functional characterization of four sesquiterpene synthases from Ricinus communis (Castor bean). **Phytochemistry**, 78: 20, 2012.

YADAVA, D. K. et al. **Technological Innovations in Major World Oil Crops**. New York, NY: Springer New York, 2012.

YU, H.; YANG, J. A direct LDA algorithm for high-dimensional data with application to face recognition. **Pattern Recognition**, 34: 2067, 2001.