



**UEPB**

**UNIVERSIDADE ESTADUAL DA PARAÍBA  
CAMPUS CAMPINA GRANDE  
CENTRO DE CIÊNCIAS E TECNOLOGIAS  
DEPARTAMENTO DE QUÍMICA  
CURSO DE GRADUAÇÃO EM QUÍMICA INDUSTRIAL**

**JULIANA LIMA BRASIL DA SILVA**

**ESPECTROSCOPIA NIR ALIADA A MÉTODOS DE RECONHECIMENTO DE  
PADRÕES PARA DISTINÇÃO ENTRE SEMENTES DE ALGODÃO DAS  
VARIEDADES BRS AROEIRA E BRS 368 RF**

**CAMPINA GRANDE  
2020**

JULIANA LIMA BRASIL DA SILVA

**ESPECTROSCOPIA NIR ALIADA A MÉTODOS DE RECONHECIMENTO DE  
PADRÕES PARA DISTINÇÃO ENTRE SEMENTES DE ALGODÃO DAS  
VARIEDADES BRS AROEIRA E BRS 368 RF**

Trabalho de Conclusão de Curso em Química Industrial da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Química Industrial.

**Área de concentração:** Química Analítica.

**Orientador:** Profa. Dra. Simone da Silva Simões.

**CAMPINA GRANDE  
2020**

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

S586e Silva, Juliana Lima Brasil da.  
Espectroscopia NIR aliada a métodos de reconhecimento de padrões para distinção entre sementes de algodão das variedades BRS Aroeira e BRS 368 RF [manuscrito] / Juliana Lima Brasil da Silva. - 2020.  
37 p. : il. colorido.  
Digitado.  
Trabalho de Conclusão de Curso (Graduação em Química Industrial) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2020.  
"Orientação : Profa. Dra. Simone da Silva Simões , Departamento de Química - CCT."  
1. Algodão. 2. Espectroscopia NIR. 3. Química analítica. I.  
Título  
21. ed. CDD 543

JULIANA LIMA BRASIL DA SILVA

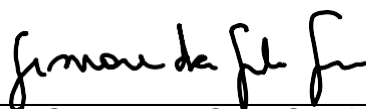
ESPECTROSCOPIA NIR ALIADA A MÉTODOS DE RECONHECIMENTO DE  
PADRÕES PARA DISTINÇÃO ENTRE SEMENTES DE ALGODÃO DAS  
VARIEDADES BRS AROEIRA E BRS 368 RF

Trabalho de Conclusão de Curso em  
Química Industrial da Universidade  
Estadual da Paraíba, como requisito  
parcial à obtenção do título de  
Bacharel em Química Industrial.

Área de concentração: Química  
Analítica.

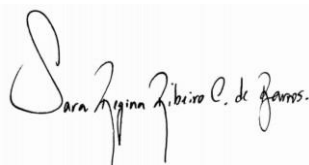
Aprovada em: 07/outubro/2020.

**BANCA EXAMINADORA**



---

Profa. Dra. Simone da Silva Simões (Orientador)  
Universidade Estadual da Paraíba (UEPB)



---

Profa. Dra. Sara Regina Ribeiro Carneiro de Barros  
Universidade Estadual da Paraíba (UEPB)



---

Dr. Everaldo Paulo de Medeiros  
Pesquisador da Embrapa Algodão

A meu bom Deus; a minha família que em todos os momentos me incentivaram e que são meu maior exemplo de vida, em especial meus pais Joseane e José Severino, e minha avó D. Miriam (*in memoriam*).

Com carinho e amor **DEDICO**.

## AGRADECIMENTOS

A Deus, que me dá sempre a força necessária para nunca desistir.

Aos meus pais Joseane e José Severino, por me apoiarem e se esforçarem ao máximo para que esse sonho pudesse ser concretizado.

Aos meus familiares por todo incentivo e compreensão, em especial minha avó D. Miriam (*in memoriam*) que sempre acreditou em mim, mas partiu sem ver essa grande conquista.

A minha orientadora Profa. Dra. Simone da Silva Simões pela orientação não apenas de TCC, mas de vida e pela dedicação para garantir sempre os melhores resultados.

A Ingrid e Priscila pela ajuda na realização das análises e pelos momentos de amizade e apoio.

Aos professores e funcionários da UEPB, especialmente os que compõem o CCT.

A todos os amigos que tive o prazer de conhecer na UEPB e que trilharam esse longo caminho ao meu lado.

A minha amiga Thainara e sua família por todo carinho com que acolheram em sua casa.

A Embrapa Algodão por ceder seus laboratórios para realização das análises.

A Dr. Everaldo, Edjane e Joabson parte integrante da Embrapa por toda orientação e ajuda na realização desse trabalho.

A CNPq pelo auxílio financeiro para realização da pesquisa.

## RESUMO

O algodão é uma planta de porte ereto que possui flores hermafroditas e seus frutos verdes são chamados de “maçãs”. A produção de algodão é uma das atividades agrícolas que possui maior relevância, devido aos principais usos dessa cultivar. Para evitar a contaminação por pragas e aumento de produtividade, métodos de melhoramento vem sendo aplicados. Os métodos utilizados para verificação de transgenia são caros e poucos acessíveis. Neste sentido, a espectroscopia no infravermelho próximo (NIR), que tem sido muito usada na área agrícola, pode se apresentar como uma ferramenta para a distinção entre espécies convencionais e transgênicas. Este trabalho tem por objetivo utilizar a espectroscopia NIR como uma metodologia eficiente para distinção de sementes de algodão transgênicas e convencionais, a fim de evitar a contaminação em cultivos convencionais. As medidas foram realizadas em sementes fornecidas pela sede da Embrapa Algodão, localizada no município de Campina Grande. Foram utilizadas sementes de dois cultivares de algodão, BRS Aroeira e BRS 368 RF, convencional e transgênica, respectivamente. As medidas espectrais foram adquiridas por meio do espectrômetro VIS/NIR em que os espectros foram registrados na região de 400-2500 nm, com resolução de 0,5 nm e 32 varreduras, realizou-se 10 leituras para cada classe. Com os espectros registrados, os modelos foram construídos usando as técnicas de reconhecimento de padrões por análise de componentes principais (PCA), modelagem independente e flexível por analogia de classes (SIMCA) e a análise discriminante por mínimos quadrados parciais (PLS-DA). Estas ferramentas quimiométricas serviram para os testes de prova de conceito para a discriminação dos cultivares de algodão das classes convencional e transgênica. Os modelos construídos utilizando os dados obtidos de reflectância de espectros NIR permitiram verificar tendências de agrupamento e distinguir entre as duas classes de forma não destrutiva, rápida e precisa. Deste modo, o método desenvolvido pode ser considerado viável, eficiente e envolver mais cultivares para classificação das sementes de algodão.

**Palavras-chave:** Algodão. Classificação. Espectroscopia NIR.

## ABSTRACT

Cotton is an erect sized plant that has hermaphrodite flowers and its green fruits are called "apples". Cotton production is one of the most relevant agricultural activities, due to the main uses of this cultivar. In order to avoid contamination by pests and increase in productivity, improvement methods have been applied. The methods used to verify transgenics are expensive and few accessible. In this sense, spectroscopy in the near infrared (NIR), which has been widely used in the agricultural area, can present itself as a tool for the distinction between conventional and transgenic species. This work aims to use NIR spectroscopy as an efficient methodology to distinguish transgenic and conventional cotton seeds, in order to avoid contamination in conventional crops. The measurements were carried out on seeds supplied by the Embrapa Cotton headquarters, located in the municipality of Campina Grande. Seeds from two cotton cultivars, BRS Aroeira and BRS 368 RF, conventional and transgenic, were used respectively. The spectral measurements were acquired using the VIS / NIR spectrometer in which the spectra were recorded in the region of 400-2500 nm, with 0.5 nm resolution and 32 scans, 10 readings were made for each class. With the registered spectra, the models were built using the pattern recognition techniques by principal component analysis (PCA), independent and flexible modeling by class analogy (SIMCA) and the discriminant analysis by partial least squares (PLS-DA). These chemometric tools were used for proof of concept tests for the discrimination of cotton cultivars of the conventional and transgenic classes. The models constructed using the data obtained from reflectance of NIR spectacles allowed to verify grouping trends and to distinguish between the two classes in a non-destructive, fast and precise way. Thus, the method developed can be considered viable, efficient and involve more cultivars for the classification of cotton seeds.

**Keywords:** . Classification. Cotton. NIR Spectroscopy.



## LISTA DE ILUSTRAÇÕES

<b>Figura 1</b> - Mostra o espectro eletromagnético. A região de infravermelho é dividida em três partes; o infravermelho próximo (Near), infravermelho médio (Mid) e a região do infravermelho distante (Far) .....	13
<b>Figura 2</b> - Representação esquemática de aquisição dos espectros .....	14
<b>Figura 3</b> - Métodos utilizados para classificar os tomates transgênicos (triângulos rosa) dos tomates não transgênicos (quadrados pretos). (A) Análise de componentes principais (PCA). (B) Trama pontuação tridimensional no modelo de PC. (C) Análise discriminante para tomates transgênicos (triângulos rosa) e tomates não transgênicos (quadrados pretos). (D) Análise de distância parcial dos mínimos quadrados .....	16
<b>Figura 4</b> - Interdisciplinaridade da quimiometria .....	17
<b>Figura 5</b> - Representação do PLS1-DA .....	22
<b>Figura 6</b> - Representação do PLS2-DA .....	23
<b>Figura 7</b> - Sementes de BRS 368 RF e BRS Aroeira .....	24
<b>Figura 8</b> - Espectrofotômetro XDS near-infrared Rapid Content TM Analyser da FOSS Analytical .....	25

## LISTA DE GRÁFICOS E TABELAS

<b>Gráfico 1</b> - Espectros brutos (A) e pré-processados com (B, C e D).....	26
<b>Gráfico 2</b> - Análise de componentes principais (PCA) referente aos espectros brutos (A) e pré-processados com (B) Baseline offset, (C) SNV + baseline offset e (D) SNV + baseline offset + suavização com janela de 9 pontos e polinômio Savitzky Golay de 2ª ordem.....	27
<b>Gráfico 3</b> - Poder de modelagem por variável .....	29
<b>Gráfico 4</b> - Poder de Discriminação por variável, destaque relacionado a região relacionada a estrutura alfa-hélice do peptídeo.....	30
<b>Gráfico 5</b> - Espectros característicos das amostras .....	30
<b>Gráfico 6</b> - Curvas ROC e respostas estimadas para a calibração .....	31
<b>Gráfico 7</b> - Curvas ROC e respostas estimadas para a predição.....	32
<b>Gráfico 8</b> - Validação do modelo PLS-DA .....	32
<b>Tabela 1</b> - Regiões espectrais do infravermelho.....	12
<b>Tabela 2</b> - Classificação SIMCA. Fonte: Software UnscramblerX.....	29

## SUMÁRIO

<b>1 INTRODUÇÃO</b> .....	10
<b>2 FUNDAMENTAÇÃO TEÓRICA</b> .....	12
2.1 ESPECTROSCOPIA NO INFRAVERMELHO .....	12
2.1.1 Espectroscopia NIR .....	13
2.1.2 Uso da espectroscopia NIR para detecção de transgenia....	14
2.2 QUIMIOMETRIA .....	16
2.2.1 Análise de Componentes Principais (PCA).....	18
2.2.2 Reconhecimento de padrões .....	19
2.2.2.1 <i>Soft Independent Modeling by Class Analogy (SIMCA)</i> .....	20
2.2.2.2 <i>A análise discriminante por mínimos quadrados parciais (PLS-DA)</i> .	21
<b>3 METODOLOGIA</b> .....	24
3.1 LOCAL DO EXPERIMENTO.....	24
3.2 AMOSTRAS.....	24
3.3 INSTRUMENTAÇÃO .....	25
3.3.1 Espectroscopia NIR .....	25
<b>4 RESULTADOS E DISCUSSÃO</b> .....	26
4.1 ANÁLISE EXPLORATÓRIA.....	26
4.2 MODELOS DE RECONHECIMENTO DE PADRÕES SUPERVISIONADOS.....	28
4.3 MODELO SIMCA .....	28
4.4 MODELO PLS-DA .....	31
<b>5 CONCLUSÃO</b> .....	34
<b>REFERÊNCIAS</b> .....	35

## 1 INTRODUÇÃO

O algodão é uma planta de porte ereto, sua produção pode ser anual ou perene, possui flores hermafroditas, que se abrem a cada 3 ou 6 dias. Seus frutos quando verdes são chamados de “maçãs”, e quando se abrem de “capulhos” (ZABOT, 2007). O algodoeiro não é somente uma planta fibrosa e oleaginosa, é sim um organismo oleoproteíginoso e fibroso. Após o processo do beneficiamento que é a separação nos descaroadores das fibras das sementes, a fibra é enfardada para ser comercializada e ficam as sementes, comercialmente denominadas de caroço de algodão (BELTRÃO; PEREIRA, 2000).

A produção de algodão é uma das atividades agrícolas que possui maior relevância, devido a sua importância no âmbito social e econômico, movimentando dessa forma o setor agrícola em geral. Essa importância se dá devido aos seus principais usos, como para a produção de fibras empregadas na indústria têxtil e na utilização de sua semente desde o uso direto para alimentação de bovinos até a industrialização, com a produção de óleo, manteiga vegetal, concentrado protéico e outros produtos.

Atualmente, a transferência de genes permite o desenvolvimento de organismos de diferentes espécies por vias não sexuais, previamente caracterizados e geneticamente manipulados em laboratório. Essa ferramenta da genética molecular permite a introdução de características desejáveis em espécies de interesse, levando ao desenvolvimento de organismos geneticamente modificados ou transgênicos, sejam eles plantas, animais ou microrganismos (BARROS; MOREIRA, 2000). A transgenia é uma ferramenta a mais à disposição da agricultura, que permite a incorporação de genes de interesse ao genoma das plantas produtivas, ampliando consideravelmente as possibilidades de melhoramento genético das espécies (BÉLOT; GILBANDI; SILVIE, 2019).

Tendo em vista a grande evolução que o desenvolvimento da biotecnologia trouxe para a agricultura, principalmente no que diz respeito aos transgênicos, especificadamente o algodão, buscou-se neste trabalho uma metodologia alternativa utilizando-se de medidas na região NIR para diferenciar os genótipos de algodoeiro transgênico do convencional, com o propósito de avançar mais rapidamente em etapas de melhoramento genético e na melhor detecção de contaminação em cultivos que apresentem riscos biológicos, como em sistemas orgânico e

agroecológico, uma vez que as metodologias mais comuns para detecção dos transgênicos são caracterizados pela presença de um ou mais segmentos de DNA exógenos, podendo ou não acomodar a expressão de novas proteínas, esses são os chamados métodos convencionais, que são caros, onerosos e destrutivos (PETIT et al., 2003).

Dessa forma, os métodos baseados na Espectroscopia no Infravermelho Próximo (NIR) possuem características para que esses propósitos sejam alcançados, pois, apresenta as vantagens de ser mais rápido, não invasivo, não destrutivo e de baixo custo. Assim, a base dessa tecnologia na classificação e identificação do cultivar transgênico ou convencional é poder identificar alterações fenotípicas causadas por alterações genotípicas de seu genótipo.

## 2 FUNDAMENTAÇÃO TEÓRICA

### 2.1 ESPECTROSCOPIA NO INFRAVERMELHO

A Espectroscopia no Infravermelho ocorre devido à interação da radiação, no caso Infravermelha (IV), com a matéria, e com o acoplamento do campo elétrico oscilante da vibração molecular e o da radiação incidente (FORATO et al., 2010).

A radiação infravermelha no infravermelho próximo (NIR) foi descoberta por Herschel, em 1800, e por volta de 1900 Coblentz obteve espectros de absorção de grande número de compostos orgânicos em estado sólido, líquido e vapor (SALA, 2008).

A região espectral do infravermelho compreende radiações com números de onda no intervalo de aproximadamente  $12.800$  a  $10 \text{ cm}^{-1}$  ou comprimentos de onda de  $0,78$  a  $1.000 \text{ }\mu\text{m}$ . Do ponto de vista tanto da aplicação como da instrumentação, o espectro infravermelho é convenientemente dividido em radiação no infravermelho próximo, médio e distante (SKOOG; HOLLER; NIEMAN, 2016). No entanto, a região que será abordada nesse trabalho será a NIR.

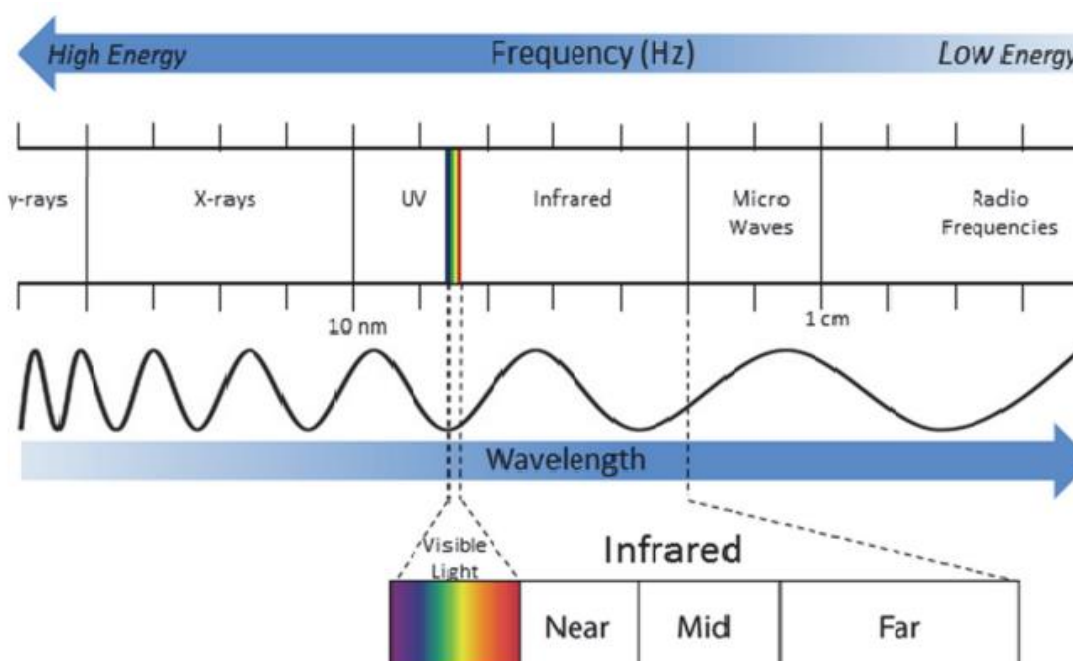
A primeira aplicação da espectroscopia no infravermelho foi para monitorar a qualidade e controlar a produção na indústria petroquímica, por ocasião da segunda guerra mundial. A espectroscopia no infravermelho próximo não foi considerada inicialmente como uma técnica analítica com algum valor prático, sendo originalmente, uma extensão da região do visível que não foi explorada até por volta de 1970. Atualmente, o valor, a funcionalidade e os benefícios oferecidos pela espectroscopia no infravermelho, principalmente a espectroscopia no infravermelho próximo, são incontestáveis (VALDERRAMA, 2005). A tabela 1 e a figura 1 mostram a divisão com os limites aproximados para cada região.

**Tabela 1** - Regiões espectrais do infravermelho.

Região	Intervalo de número de onda ( $\hat{\nu}$ ) – ( $\text{cm}^{-1}$ )	Região em comprimento de onda ( $\lambda$ ) – (nm)	Região de frequência ( $\nu$ ) – (Hz)
Próximo (NIR)	12800 a 4000	780 a 2500	$3,8 \times 10^{14}$ a $1,2 \times 10^{14}$
Médio (MID)	4000 a 200	2500 a 5000	$1,2 \times 10^{14}$ a $6,0 \times 10^{12}$
Distante (FAR)	200 a 10	5000 a 100000	$6,0 \times 10^{12}$ a $3,0 \times 10^{11}$

Fonte: SKOOG; HOLLER; NIEMAN, 2016.

**Figura 1** - Espectro eletromagnético. A região de infravermelho dividida em três partes; o infravermelho próximo (Near), infravermelho médio (Mid) e a região do infravermelho distante (Far).



Fonte: ESKILDSEN, 2016.

### 2.1.1 Espectroscopia NIR

A espectroscopia no infravermelho próximo (NIR) refere-se à faixa do espectro eletromagnético cujas ondas (ou fótons) apresentam um comprimento de onda que se situa entre 750 nm e 2.500 nm, equivalente à faixa de número de onda entre  $4.000\text{ cm}^{-1}$  e  $13.333\text{ cm}^{-1}$ , frequência entre  $4,0 \times 10^{14}\text{ Hz}$  e  $1,2 \times 10^{14}\text{ Hz}$  e energia entre 1,65 eV e 0,50 eV ou  $160\text{ kJ mol}^{-1}$  e  $48\text{ kJ mol}^{-1}$ , respectivamente (PASQUINI, 2018).

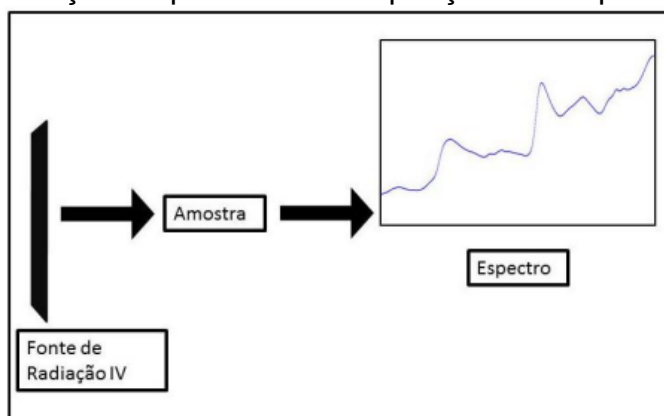
A espectroscopia NIR atua através do princípio de emissão de radiação eletromagnética no qual necessita de uma única padronização para efetuar as análises (MARCHESE et al., 2018). Na região do NIR ocorre a absorção de *overtones* ou combinações de vibrações fundamentais de estiramento das moléculas, e os espectros dependem da composição química e características físicas destas. As ligações mais comuns são de C-H, N-H, O-H e S-H (CALEGARI, 2018).

A técnica NIR se fundamenta no desenvolvimento de modelos de regressão multivariada e é vantajosa por ser rápida e não requerer trabalho intensivo no

processamento das amostras, permitindo a grande escala de amostragem, além de não demandar reagentes e não destruir as amostras (STUTH; JAMA; TOLLESON, 2003).

Os instrumentos para aquisição de espectros no infravermelho são compostos basicamente de uma fonte de luz, monocromador ou interferômetro e detector. Todos os espectrofotômetros de infravermelho possuem um registrador, onde o espectro é mostrado, geralmente, em escalas de comprimento de onda ou número de onda no eixo horizontal e absorbância (A) ou transmitância (T) percentual no eixo vertical. Os instrumentos mais modernos permitem a comparação do espectro obtido com os de bibliotecas de espectros (AMORIM; KLIER; ANGELIS, 2013). A figura 2 mostra de forma esquemática como é adquirido o espectro a partir da fonte de radiação, ou seja, o espectrofotômetro.

**Figura 2** – Representação esquemática de aquisição dos espectros.



Fonte: MENDES, 2014.

### 2.1.2 Uso da espectroscopia NIR para detecção de transgenia

Existem aplicações do NIR nas áreas agrícola, alimentícia, ambiental, farmacêutica e biomédica, nas indústrias têxtil, de polímeros, de óleo e gás, entre outras. Análises que podem ser realizadas no NIR compreendem a composição de soluções químicas complexas, a configuração molecular, cristalinidade de sólidos, tamanho e forma de meios heterogêneos, entre outras (MONTEIRO, 2011).

O Brasil, é um dos principais produtores de alimentos do mundo e com crescente desenvolvimento no setor do agronegócio e exportação de alimentos, logo, necessita se ajustar à realidade internacional no que diz respeito a legislação e



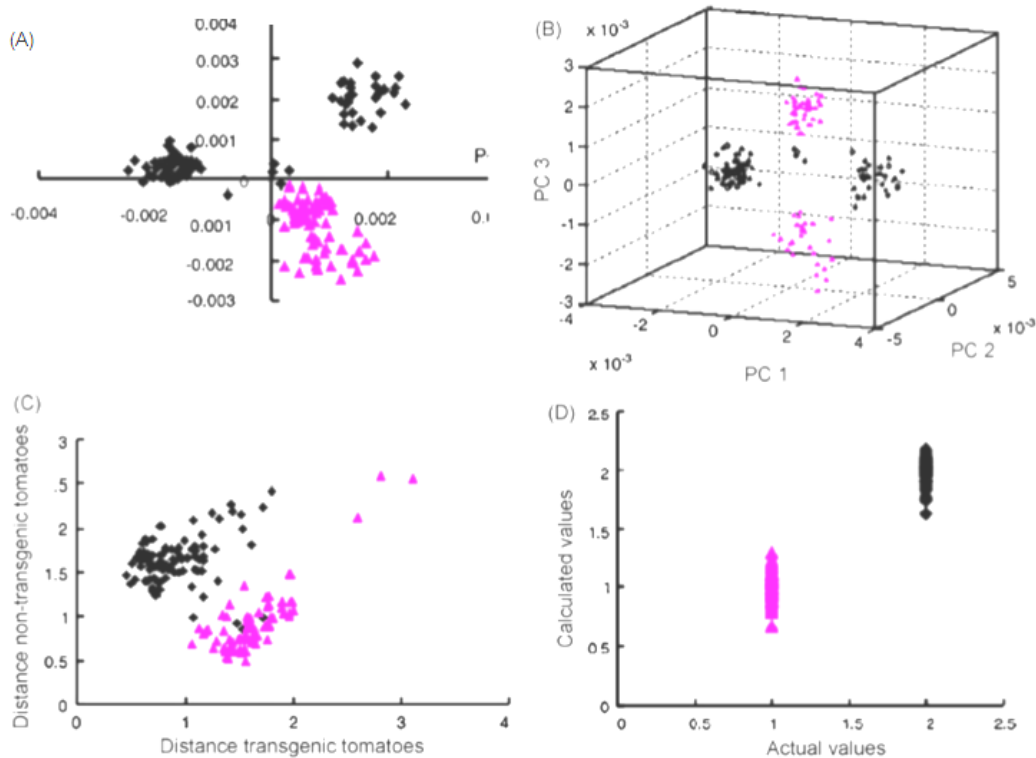
as determinações dos países importadores desses produtos. Assim, um grande desafio para a pesquisa e para o setor produtivo é manter-se atualizado, em termos metodológicos, visando atender às exigências dos parceiros comerciais. A detecção e quantificação de Organismos Geneticamente Modificados (OGMs) é exigida em quase todos os países para os quais o Brasil exporta alimentos, sendo cada vez mais uma exigência global (CONCEIÇÃO; MOREIRA; BINSFELD, 2006).

As técnicas analíticas para a detecção e quantificação de OGMs baseadas na HPLC e nos microarranjos de DNA possuem as desvantagens de serem lentas, de custo elevado e requerem o pré-tratamento da amostra. A espectroscopia NIR, baseada na mensuração do comprimento de onda e na intensidade de absorção de luz infravermelha próxima, apresenta potencialidade para ser empregada na detecção e quantificação desses organismos, por apresentar as vantagens de ser uma técnica rápida e que não requer o pré-tratamento da amostra. A detecção de OGMs em alimentos através de NIR é altamente eficaz, no entanto, a quantificação ainda precisa ser validada (CONCEIÇÃO; MOREIRA; BINSFELD, 2006).

Lee; Choung, (2011) classificaram sementes de soja geneticamente modificadas (GM) e não GM utilizando a espectroscopia NIR em conjunto com técnicas de Análise de Componentes Principais (PCA) e Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA).

Xie et al., (2007) utilizaram dessa ferramenta combinada com a análise multivariada para classificar tomates transgênicos e não-transgênicos. Neste estudo, setenta tomates transgênicos com *LeETR* anti-sentido2 e 94 de seus pais, os não-transgênicos foram medidos. Análise de componentes principais (PCA), análise discriminante (DA) e análise discriminante de mínimos quadrados parciais (PLSDA) foram aplicadas para classificar os tomates em dois grupos. Os resultados mostraram diferenças entre os tomates transgênicos e não transgênicos e uma ótima classificação foi obtida após a otimização do pré-tratamento espectral. As classificações corretas para os tomates foram de 100% usando PLSDA após pré-tratamento espectral derivado. Os resultados mostram que essa técnica oferece o benefício de evitar análises químicas e sensoriais demoradas, caras e trabalhosas. Alguns dos métodos utilizados na pesquisa podem ser visto na figura 3.

**Figura 3** – Métodos utilizados para classificar os tomates transgênicos (triângulos rosa) dos tomates não transgênicos (quadrados pretos). (A) Análise de componentes principais (PCA). (B) Trama pontuação tridimensional no modelo de PC. (C) Análise discriminante para tomates transgênicos (triângulos rosa) e tomates não transgênicos (quadrados pretos). (D) Análise de distância parcial dos mínimos quadrados.



Fonte: XIE et al., 2007.

CUI et al., (2012) relataram a aplicação de métodos infravermelho próximo (NIR) e reconhecimento de padrões de discriminação rápida e automática dos genótipos (pai, transgênicos, e híbrido pai-transgênica) de plantas de algodão. Os resultados demonstram que os melhores modelos parciais para análise de mínimos quadrados discriminante (PLSDA) obtido de precisão e classificação total de 100% e 97,6% para as sementes e folhas, respectivamente.

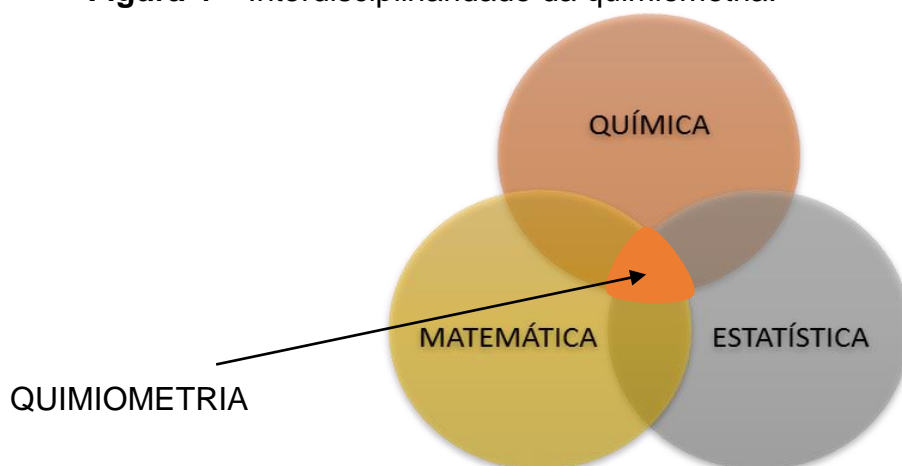
## 2.2 QUIMIOMETRIA

A denominação “quimiometria” foi usada pela primeira vez em 1971 pelo químico orgânico Svante Wold, mas só se firmou definitivamente quando o computador – por si só ou como componente de instrumentos de medição – invadiu o laboratório químico. No sentido amplo, há quem argumente que a combinação de

química com estatística começou muito antes, com os trabalhos do químico cervejeiro que se denominava Student, o do famoso teste t. Apesar da conservadora política de informática adotada pelo governo brasileiro na época, a invasão dos computadores não demorou a alcançar os laboratórios nacionais (FERREIRA, 2015; NETO; SCARMINIO; BRUNS, 2006).

Na literatura há diversas definições do que vem a ser quimiometria e isso é devido a sua interdisciplinaridade, contudo, é consenso geral hoje que a quimiometria é uma intersecção de três grandes áreas (FERREIRA, 2015), tal como representado na figura 4.

**Figura 4 – Interdisciplinaridade da quimiometria.**



**Fonte:** Adaptada de FERREIRA, 2015.

A quimiometria é considerada uma das áreas mais recentes da química analítica. Desde seu surgimento, foram desenvolvidos métodos que vêm tornando possível o processamento e interpretação de dados que antes seriam impossíveis de serem analisados. Um exemplo da utilização da quimiometria é a análise realizada na região do infravermelho próximo, onde sem a utilização de modelos de calibração multivariada não seria possível realizar determinações quantitativas (OLIVEIRA, 2008).

Na quimiometria as técnicas utilizadas podem ser divididas de forma geral em 3 grupos: planejamento e otimização de experimentos, reconhecimento de padrões (análise exploratória e classificação) e calibração. Este trabalho irá destacar as técnicas de reconhecimento de padrões.

### 2.2.1 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais foi introduzida por Karl Pearson em 1901. O tratamento formal do método, todavia, é devido ao trabalho de Hotelling, divulgado na década de 30 e que causou uma revolução no uso de métodos multivariados na área da psicologia (FERREIRA, 2015).

A PCA é um método não-supervisionado de reconhecimento de padrões utilizado para reduzir a dimensionalidade do conjunto de dados original, preservando a maior quantidade de informação (variância) possível. Essa redução é obtida através da introdução de novas variáveis ortogonais entre si, denominadas componentes principais (PCs). Organizadas em ordem decrescente de importância, as PCs são combinações lineares das variáveis originais. A apresentação dos resultados experimentais são na forma de gráficos, o que facilita a interpretação dos dados, os gráficos obtidos representam as amostras em um sistema cartesiano onde os eixos são as PCs. A avaliação das PCs pode auxiliar no estabelecimento de uma assinatura química particular para cada grupo de amostras segregado após a PCA. Esse é o objetivo principal dos estudos de reconhecimento de padrões, que busca encontrar uma maneira de relacionar a identidade de uma amostra com suas características químicas (CORREIA; FERREIRA, 2007).

A ideia da PCA, representada na Equação 1, é aproximar a matriz de dados originais ( $X$ ) pelo produto de duas matrizes, as matrizes de scores ( $T$ ), que são as projeções das amostras na direção das componentes principais, e de loadings ou “pesos” ( $L$ ), que são os cossenos dos ângulos formados entre as PCs e cada variável original, e indicam a colaboração de cada variável nos eixos das PCs.  $T$  significa a transposta da matriz, e  $E$  é a matriz do erro que determina o produto entre scores e loadings:

$$X = T L^T + E \quad (1)$$

A aplicação desta técnica na modelagem de resultados analíticos, como cromatogramas e dados espectroscópicos, permite observar as similaridades e as diferenças entre as amostras analisadas de uma forma que os resultados univariados obtidos em laboratório não conseguem proporcionar (ARAÚJO, 2016).

A matriz dos escores de PCA está diretamente associada com as concentrações dos compostos das amostras, enquanto os pesos estão correlacionado com as variáveis responsáveis pela maior variedade dos dados. O

principal intuito desta técnica é obter informações sobre o conjunto de dados através de: (i) número de PCs independentes que melhor descrevem as informações no conjunto de dados, ou seja, encontrar as fontes de variação dos dados; (ii) o comportamento dos escores, ou seja, as similaridades e diferenças entre amostras; e (iii) o perfil dos pesos, ou seja, as variáveis originais que mais contribuem para a variabilidade e como estão relacionadas com as amostras (SILVA, 2017).

### **2.2.2 Reconhecimento de padrões**

A técnica de reconhecimento de padrões é um método através do qual pode-se encontrar tendências e agrupamentos em um conjunto de amostras e com base nos padrões obtidos, é possível interpretar resultados e tomar decisões (FERREIRA, 2015). Existe 2 casos particulares de técnicas de reconhecimento de padrões usadas para agrupar e classificar amostras: supervisionada e não supervisionada.

As técnicas não supervisionadas são ferramentas que tentam representar semelhanças e diferenças entre as amostras sem conhecimento prévio das classes às quais pertencem. Já as técnicas supervisionadas por sua vez, compreendem basicamente técnicas classificatórias, onde um modelo é construído a partir de um conjunto de amostras de classes conhecidas e, em seguida, uma amostra de classe desconhecida pode ser classificada de acordo com a semelhança/diferença entre ela e as amostras do conjunto de treinamento (SILVA, 2017).

Os métodos de reconhecimento de padrões não supervisionados consistem basicamente na Análise de Agrupamentos Hierárquicos (HCA: Hierarchical Cluster Analysis) e na Análise de Componentes Principais (PCA: Principal Component Analysis) que podem ser abordados também como métodos de análise exploratória dos dados, onde o objetivo é extrair informações, estabelecendo relações entre objetos e variáveis, sem estabelecer modelos, além de não fazer uso de nenhum conjunto de treino para classificar as amostras em pesquisa.

No caso dos métodos de reconhecimento de padrões supervisionados pode-se citar a Modelagem Independente e Flexível por Analogia de Classes (SIMCA) e a Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) que são utilizados na construção dos modelos de classificação e faz uso de um conjunto de treino para classificar as amostras.

Como ferramentas para o desenvolvimento desta pesquisa foi utilizada a espectroscopia no infravermelho próximo associado ao método de PCA e ao método de classificação SIMCA e PLS-DA.

### ***2.2.2.1 Soft Independent Modeling by Class Analogy (SIMCA)***

O método SIMCA é construído a partir de modelos baseados em componentes principais, onde cada classe corresponde a um conjunto de treinamento. Um número ótimo de componentes principais é determinado independentemente para cada classe e o modelo final é obtido através da definição da fronteira entre cada PCA modelado, como efeito uma hipercaixa é obtida. Ao definir as hipercaixas é importante verificar se existem classes com sobreposição. O poder discriminatório do modelo SIMCA se baseia na não existência de sobreposição entre duas classes diferentes, bem como na maior distância interclasses possível. A classificação de uma amostra externa empregada na modelagem SIMCA é baseada nas diferentes projeções possíveis no espaço dos escores avaliando-se se a mesma encontra-se dentro das fronteiras definidas para a respectiva classe em todas as projeções obtidas (SABIN; FERRÃO; FURTADO, 2004).

A modelagem independente e flexível por analogia de classes se beneficia dos dois conceitos que lhe dão nome. Soft modelling, ou modelagem flexível, permite que as classes se sobreponham, sem fazer distinção rígida entre elas. Assim, uma amostra pode pertencer a uma classe, a mais de uma classe ou a nenhuma das classes. Já a modelagem independente (Independent modelling) é uma característica ainda mais útil aos modelos quimiométricos. Por causa da modelagem independente, cada classe é modelada separadamente e não é influenciada por amostras de outras classes. Isso também é útil para atualização de modelos, pois pode-se incluir novas classes ao modelo SIMCA sem alterar as classes já modeladas (PEREIRA, 2018).

No SIMCA, a classificação da amostra é feita se esta estiver dentro da caixa construída pelo modelo. Logo, cada classe é modelada separadamente e a classificação pode ser feita pela comparação da variância residual (Equação 2) da

amostra a ser classificada com a variância residual total da classe (Equação 3) (OLIVEIRA, 2018).

$$S_i = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^p (\text{res}_{ij})^2}{(N-A-1)(p-A)}} \quad (2)$$

$$S_i = \sqrt{\frac{\sum_{j=1}^p (\text{res}_j)^2}{p-A}} \quad (3)$$

Onde N é o número de amostras do conjunto de treinamento da classe; A é o número de componentes principais utilizadas pela classe; p é o número de variáveis; i e j representam os índices das amostras e variáveis, respectivamente. A localização da amostra em relação ao modelo pode ser feita utilizando um teste F, onde o valor calculado na Equação 4 ( $F_{\text{cal}}$ ) é comparado com um valor crítico ( $F_{\text{crit}}$ ), determinado de modo empírico ou tabelado dentro de certos níveis de confiança e graus de liberdade. Se o valor de  $F_{\text{cal}}$  da amostra investigada for menor que o  $F_{\text{crit}}$ , ela pertencerá à classe em consideração (OLIVEIRA, 2018).

$$F_{\text{cal}} = \frac{(S_i)^2}{(S_t)^2} \cdot \frac{N}{N-A-1} \quad (4)$$

Assim, nesse método de classificação supervisionada, o modelo obtido faz a atribuição de amostra, definindo se esta pertence ou não a classe modelada.

#### **2.2.2.2 A análise discriminante por mínimos quadrados parciais (PLS-DA)**

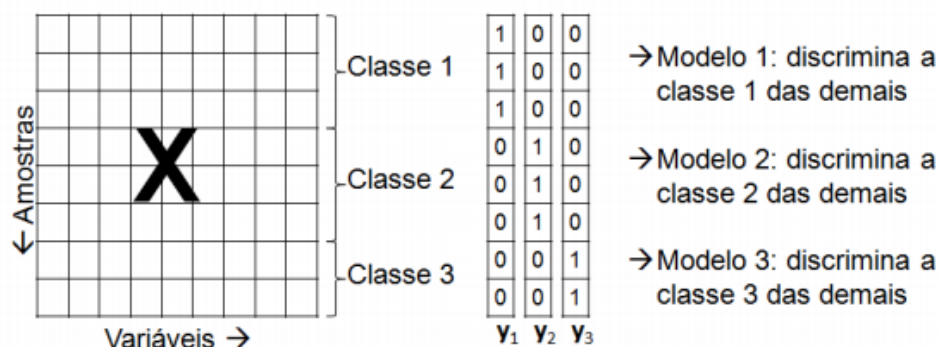
A análise discriminante por mínimos quadrados parciais (PLS-DA) também chamada de classificação supervisionada tem como objetivo criar um método para classificar amostras em grupos pré-definidos (BERGO, 2014). Esse método se baseia no modelo de calibração multivariada PLS, no qual se faz a decomposição simultânea dos espaços X e Y. A partir da decomposição simultânea, ocorre uma pequena perda de ortogonalidade em relação à decomposição por PCA. Assim, o novo sistema de coordenadas é formado pelas denominadas variáveis latentes (VL) (PEREIRA, 2018).

O número de variáveis latentes tem grande importância, uma vez que altera a quantidade de informação que será modelada. Se o número de VL for muito pequeno, ocorre o subajuste e o modelo não é capaz de explicar toda informação

sistemática dos dados, não apresentando uma boa capacidade preditiva. No caso contrário, modelando um número muito grande de VL, ocorre o sobreajuste e parte da informação aleatória e ruído instrumental, é incorporada ao modelo. Sendo assim, os resultados da calibração são supermodelados apresentando erros de previsão não realistas, e, portanto, incompatíveis com a capacidade de previsão para novas amostras independentes. Dessa forma, o número de VL é escolhido através do procedimento de validação cruzada, com base em processos de reamostragem, buscando o menor erro para amostras externas ao modelo previstas durante esta etapa (PEREIRA, 2018).

O PLS-DA possui duas abordagens possíveis. Na primeira, chamada de PLS1-DA realiza-se uma calibração qualitativa, onde o interesse é determinar se as amostras pertencerem ou não a uma determinada classe. São atribuídos os valores das classes ao vetor  $y$ , sendo o valor 1 para as amostras que serão discriminadas e o valor 0 para as demais classes. Dessa forma, o modelo é construído para cada classe. Conforme representado na Figura 5, devem ser construídos três modelos: o modelo 1 que diferencia a Classe 1 das Classes 2 e 3, o modelo 2 que diferencia a Classe 2 das Classes 1 e 3 e por fim, o modelo 3 que diferencia a Classe 3 das Classes 1 e 2 (COLARES, 2016).

**Figura 5** – Representação do PLS1-DA.



Fonte: PEREIRA, 2018.

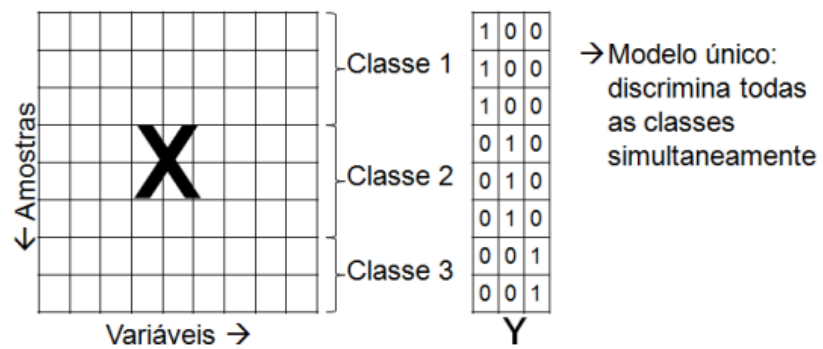
Na segunda abordagem, chamada de PLS2-DA utiliza-se uma matriz  $Y$  que possui as informações de classes, resultando em um único modelo, que discrimina todas as classes simultaneamente, como mostra a Figura 6.

Ambas as abordagens são eficientes, porém a PLS1-DA fornece melhores resultados justamente devido ao número de variáveis latentes, pois esta altera a



quantidade de informação que será modelada para cada classe, podendo dessa forma ser otimizado para o modelo de discriminação de cada classe.

**Figura 6** – Representação do PLS2-DA.



Fonte: PEREIRA, 2018.

### 3 METODOLOGIA

#### 3.1 LOCAL DO EXPERIMENTO

A parte experimental deste trabalho foi realizada no Laboratório Avançado de Tecnologia Química- (LATECQ) localizada na Embrapa Algodão, em Campina Grande, PB. E o desenvolvimento dos modelos foi realizado nas dependências do CCT/UEPB/campus I.

#### 3.2 AMOSTRAS

As amostras de sementes das cultivares de algodões convencionais e OGM's foram cedidas pela Embrapa Algodão, localizada na cidade de Campina Grande-PB. A espécie convencional é denominada BRS Aroeira e a OGM é denominada BRS 368 RF. Foram coletadas 10 porções (aproximadamente 20 sementes) de amostras de cada uma delas totalizando uma quantidade de 20 porções de amostras. Após a coleta das amostras na forma de plumas de algodão contendo a semente, foi realizada a limpeza desta deixando-a com o mínimo de resíduo de fibra.

Os aspectos das duas cultivares, convencional e OGM, são semelhantes, sendo impossível distingui-las visualmente, como mostra a Figura 7.

**Figura 7** - Sementes de BRS 368 RF e BRS Aroeira.



Fonte: PRÓPRIA, 2019.

### 3.3 INSTRUMENTAÇÃO

#### 3.3.1 Espectroscopia NIR

As medidas espectrais de reflectância difusa na região do visível e infravermelho próximo foram adquiridas por meio do espectrômetro VIS/NIR modelo XDS Analyser (Foss Analytical, Hogans, Sweden) Figura 8. Os espectros foram registrados na região de 400-2500 nm, com resolução de 0,5 nm e 32 varreduras. Para a análise das sementes por espectroscopia NIR foram utilizadas uma quantidade de sementes que preenchesse o recipiente de amostra do equipamento (em torno de 20 sementes), como mostra a figura 7.

**Figura 8** - Espectrofotômetro XDS near-infrared Rapid Content TM Analyser da FOSS Analytical.



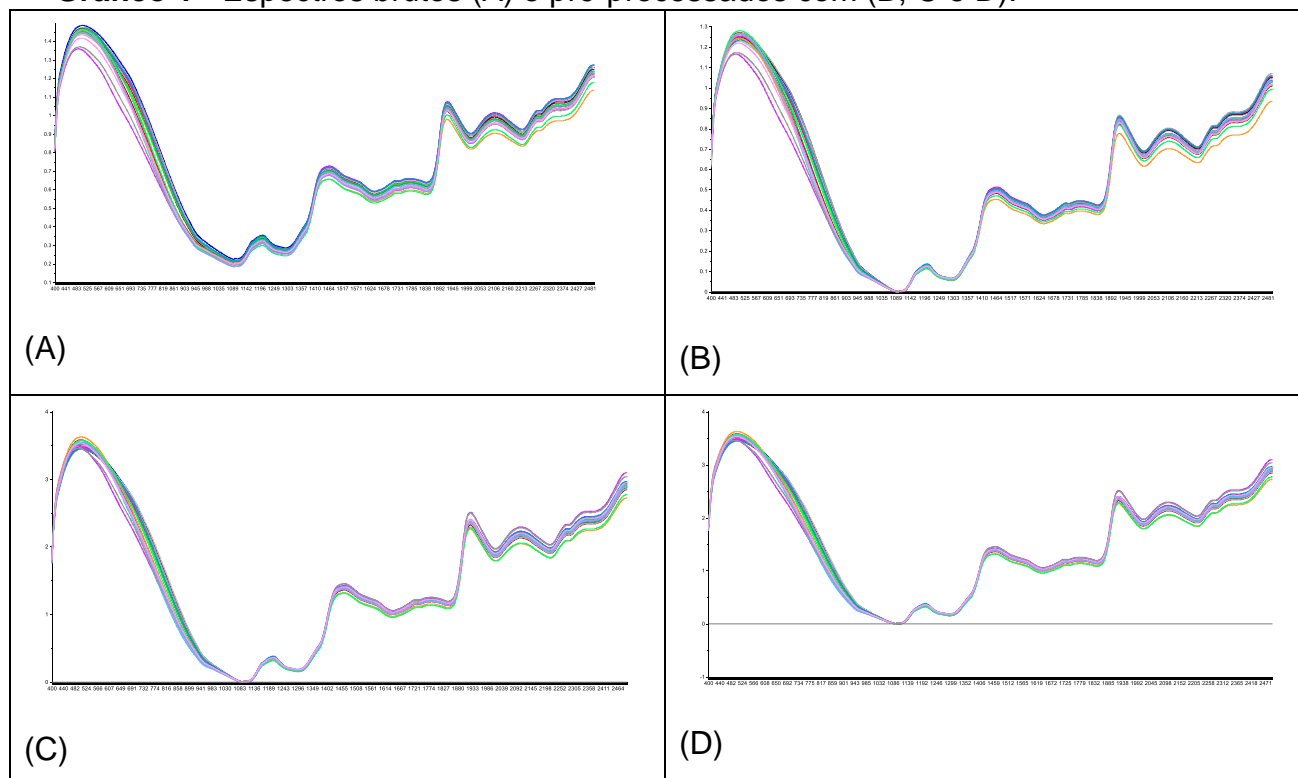
Fonte: PRÓPRIA, 2019.

## 4 RESULTADOS E DISCUSSÕES

### 4.1 ANÁLISE EXPLORATÓRIA

O gráfico 1 apresenta os espectros brutos das 20 amostras. Pode-se observar que não é possível encontrar visualmente características espectrais que possam diferenciar as amostras transgênicas e convencionais. Também pode-se observar um deslocamento de linha de base e um espalhamento no sinal espectral devido à natureza da amostra. Deste modo, alguns pré-processamentos espectrais foram aplicados a fim de remover estas características relacionadas aos aspectos físicos das amostras. Os pré-processamentos testados foram baseline offset, SNV (standard normal deviation), suavização com janela de 9 pontos e polinômio Savitzky Golay de segunda ordem, e a combinação destes. O gráfico 1 apresenta os espectros brutos (A) e pré-processados com (B) Baseline offset, (C) SNV + baseline offset e (D) SNV + baseline offset + suavização com janela de 9 pontos e polinômio Savitzky Golay de 2ª ordem.

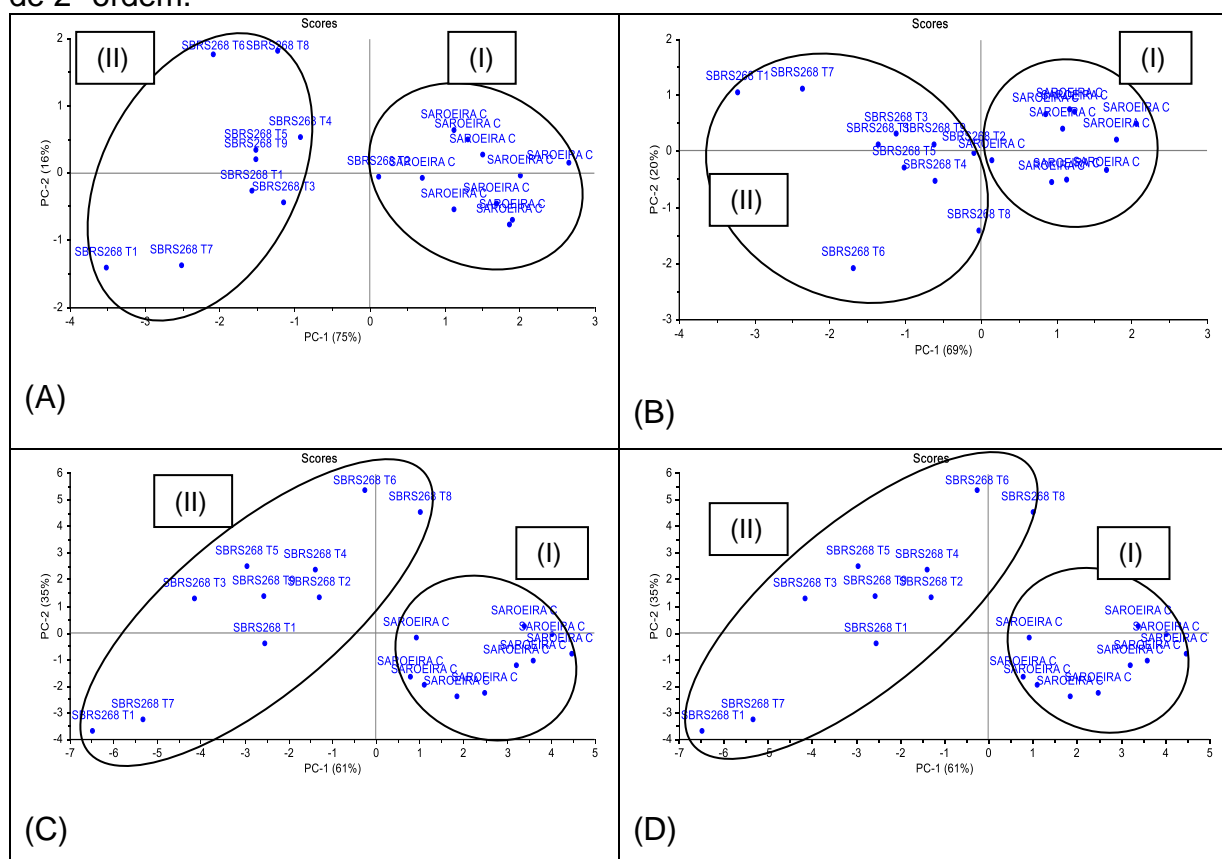
**Gráfico 1** - Espectros brutos (A) e pré-processados com (B, C e D).



Fonte: PRÓPRIA, 2019.

A análise de componentes principais (PCA) foi aplicada a fim de realizar uma análise exploratória dos dados brutos e pré-processados. Tanto a PCA dos dados brutos quanto dos pré-processados apresentou uma tendência de separação entre a classe das sementes convencionais (I) e OGM (II), como mostra gráfico 2.

**Gráfico 2** - Análise de componentes principais (PCA) referente aos espectros brutos (A) e pré-processados com (B) Baseline offset, (C) SNV + baseline offset e (D) SNV + baseline offset + suavização com janela de 9 pontos e polinômio Savitzky Golay de 2ª ordem.



Fonte: PRÓPRIA, 2019.

De modo geral, as amostras OGM se localizaram nos escores negativos de PC1, enquanto as amostras convencionais se localizaram nos escores positivos de PC1. Também pode-se observar que as amostras convencionais apresentam uma menor variabilidade espectral, apresentando uma menor distância intraclasse. Já as amostras transgênicas apresentam uma maior variabilidade e consequentemente uma maior distância intraclasse.

Apesar dos dados brutos terem apresentado uma boa distinção entre as duas classes, alguns pré-processamentos espectrais foram aplicados a estes dados a fim de remover informações físicas (como espalhamento e ruído espectral) que não estão relacionados com o problema em questão e que podem prejudicar a capacidade preditiva dos modelos que serão construídos. A aplicação da correção

de linha de base diminuiu a distância entre as duas classes e conseqüentemente a distinção entre elas, o que pode levar a classificações equivocadas na etapa de previsão dos modelos supervisionados. Os dados tratados com a combinação dos pré-processamentos SNV, baselineoffset, suavização foram os que apresentaram a melhor distinção entre as duas classes. Estes resultados indicam que os dados em questão fornecerão modelos de reconhecimento de padrões supervisionados com uma boa capacidade preditiva.

#### 4.2 MODELOS DE RECONHECIMENTO DE PADRÕES SUPERVISIONADOS

Foram construídos modelos de reconhecimento de padrões supervisionados utilizando os métodos SIMCA e PLS-DA para estudos de prova de conceito para identificação das classes da BRS Aroeira e BRS 368 RF. Levando em conta os resultados da análise exploratória, os modelos SIMCA e PLS-DA foram construídos utilizando os dados tratados com SNV+Baselineoffset+suavização com janela de 9 pontos e polinômio Savitzky Golay com polinômio de segunda ordem. Para a construção dos modelos 6 amostras de cada classe foram utilizadas para a etapa de treinamento e 4 amostras de cada classe na etapa de teste. Deste modo, o conjunto de treinamento foi composto de 12 amostras e o conjunto de teste foi composto por 8 amostras. Visando a representatividade e reprodutibilidade da amostragem, as amostras dos conjuntos de treinamento e teste foram selecionadas com o auxílio do algoritmo Kennard Stone.

#### 4.3 MODELO SIMCA

A construção do modelo SIMCA, foi calculada uma PCA para a classe convencional e uma para a classe OGM, separadamente, utilizando-se as 6 amostras de treinamento para cada classe. Os modelos foram construídos utilizando 4 componentes principais. Em seguida o modelo SIMCA foi testado utilizando as amostras de 8 amostras do conjunto de teste, 4 convencionais (amostras de 1 a 4) e 4 OGM (amostras de 5 a 8).

A Tabela 2 mostra o resultado da etapa de teste para o modelo SIMCA. Nela é apresentada a classificação de cada amostra nas respectivas classes.

**Tabela 2 - Classificação SIMCA.**

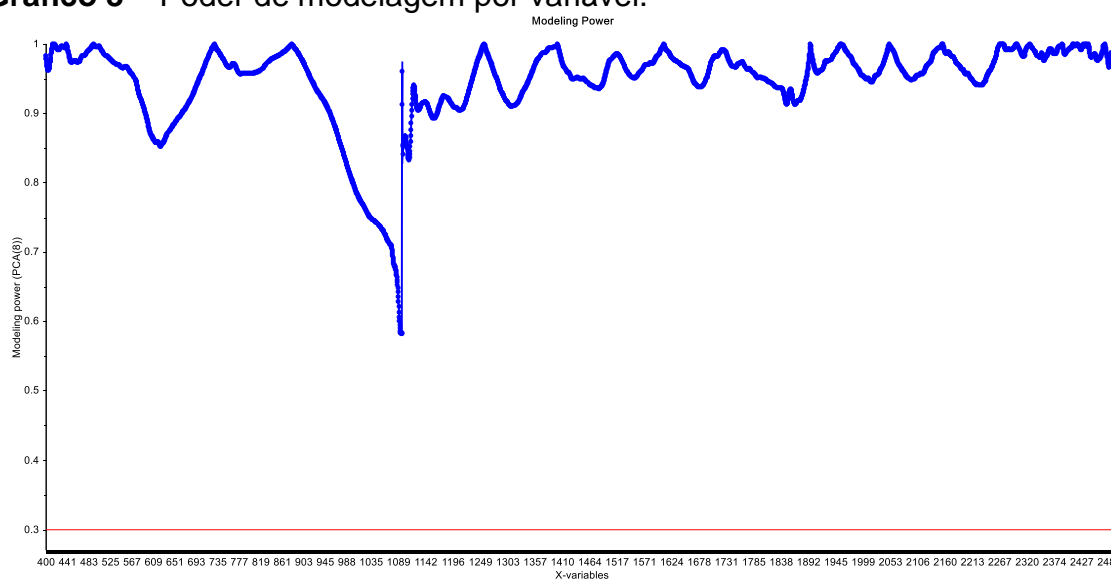
Amostra	Classe convencional	Classe OGM
Amostra_1 (C)	*	
Amostra_2 (C)	*	
Amostra_3 (C)	*	
Amostra_4 (C)	*	
Amostra_5 (T)		*
Amostra_6 (T)		*
Amostra_7 (T)		*
Amostra_8 (T)		*

Fonte: PRÓPRIA, 2019.

Como pode ser observado, foi alcançado 100% de classificação correta, a um nível de confiança de 95%, não sendo apresentados resultados falsos positivos ou falsos negativos.

O gráfico 3, mostra o poder de modelagem (PM) e o poder discriminatório (PD) para cada variável espectral, respectivamente.

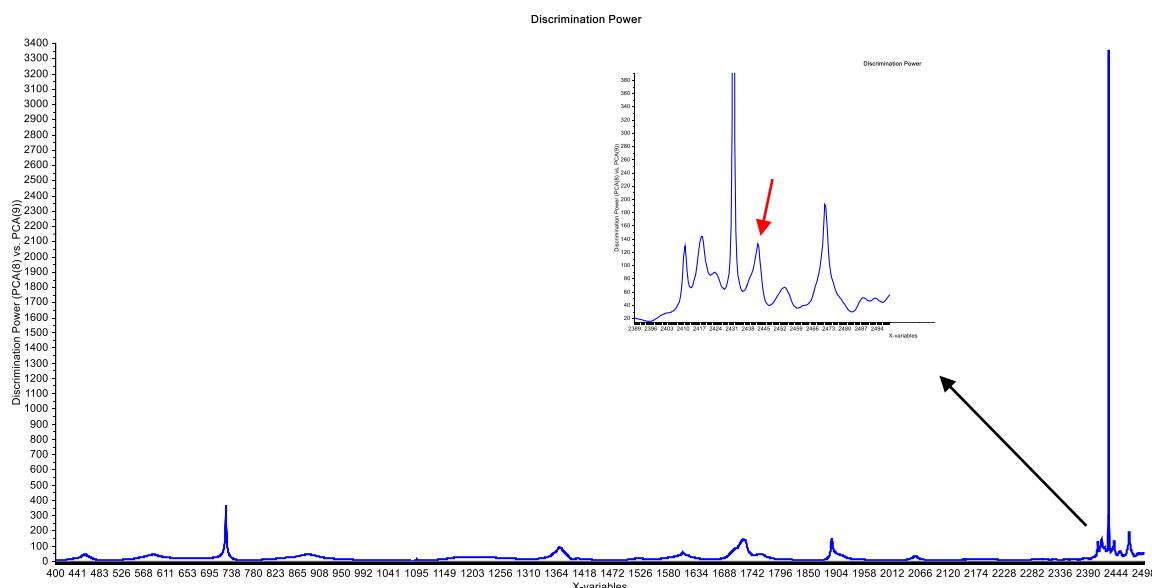
**Gráfico 3 – Poder de modelagem por variável.**



Fonte: PRÓPRIA, 2019.

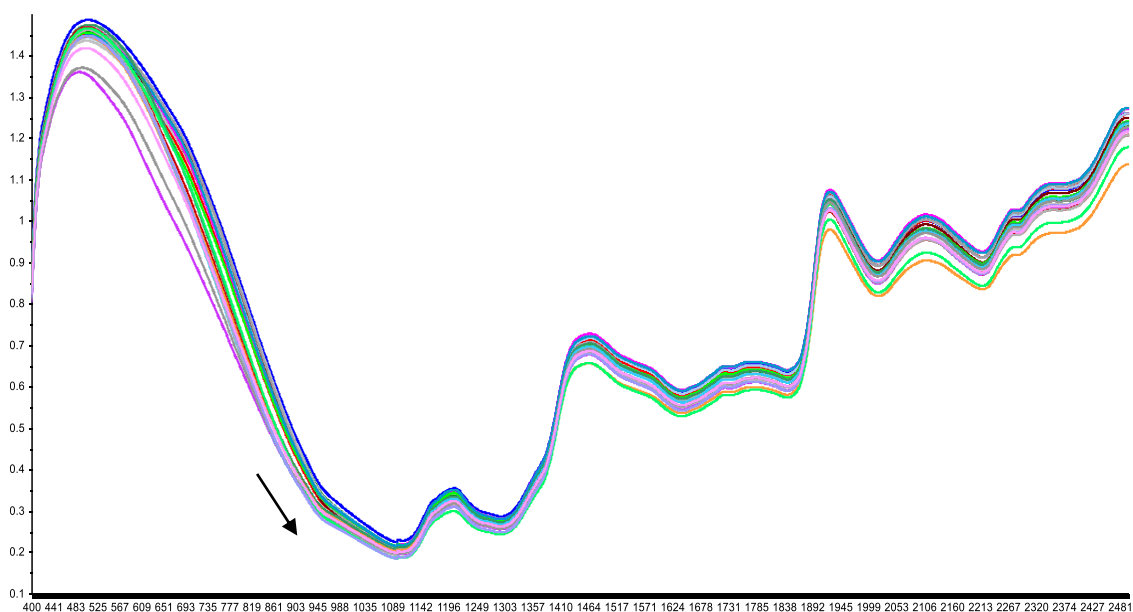
Variáveis que tem PM próximo a unidade são importantes para o modelo, já aquelas com PM inferior a 0.3 tem pouca importância para o modelo.

**Gráfico 4 - Poder de Discriminação por variável, destaque relacionado a região relacionada a estrutura alfa-hélice do peptídeo.**



Fonte: PRÓPRIA, 2019.

**Gráfico 5 - Espectros característicos das amostras.**



Fonte: PRÓPRIA, 2019.

Comparando os gráficos, 3 e 4, com os espectros característicos das amostras, gráfico 5, pode-se observar que a maioria das variáveis espectrais tem

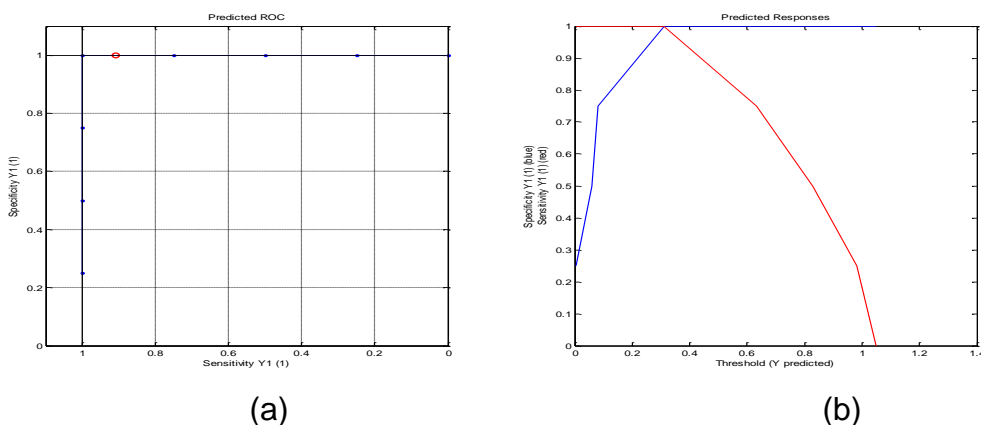


PM > 0.8. Apenas a variável ~1100 nm (seta preta em destaque no gráfico 5) apresentou PM próximo a 0.5. Em relação ao poder discriminatório, gráfico 4, que mostra o quanto cada variável contribui para a separação dos dois modelos, pode-se observar que a variável 2432 nm possui um alto valor de PD. O comprimento de onda 2432 nm pode ser atribuído ao grupo funcional aril (C-H aromático C-H). A região de 2400 a 2493 nm apresenta variáveis com alto valor de PD, dentre as variáveis desta região se encontra o comprimento de onda 2445 nm (seta vermelha) que pode ser atribuído ao grupo funcional CONH<sub>2</sub>, especialmente devido a ligação hidrogênio C=O com N-H do peptídeo denominada estrutura alfa-hélice (WORKMAN, 2007). A estrutura  $\alpha$ -hélice está presente na estrutura secundária dos níveis de organização das proteínas e se assemelha a uma escada em espiral. Nesta estrutura o esqueleto de polipeptídeo está estreitamente enrolado ao longo do maior eixo da molécula e os grupos R dos resíduos de aminoácido projetam-se para fora do esqueleto helicoidal. A estabilização se dá pela presença das ligações de hidrogênio entre os grupamentos NH e CO da cadeia principal. O grupamento CO de cada aminoácido forma ponte de hidrogênio com o grupamento NH do aminoácido que está situado a quatro unidades adiante na sequência linear, sendo que todos os grupamentos NH e CO formam pontes de hidrogênio (WORKMAN, 2007).

#### 4.4 MODELO PLS-DA

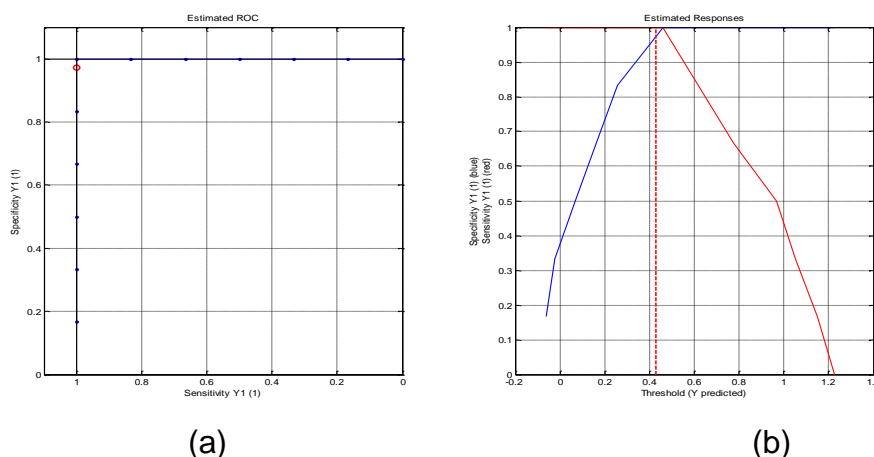
O modelo PLS-DA foi construído utilizando as amostras selecionadas pelo algoritmo Kennard Stone. Os gráficos 6 (a e b) e 7 (a e b) apresentam as estimativas da curva ROC e das respostas estimadas, nas etapas de treinamento e predição.

**Gráfico 6** - Curvas ROC e respostas estimadas para a calibração.



Fonte: PRÓPRIA, 2019.

**Gráfico 7 - Curvas ROC e respostas estimadas para a predição.**

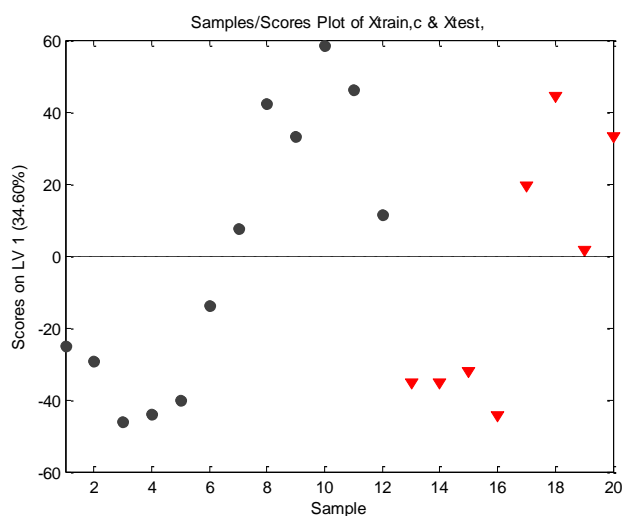


Fonte: PRÓPRIA, 2019.

Pode-se observar no gráfico 6a, que na calibração (em azul) foi obtida uma especificidade e sensibilidade igual a 1, já na etapa de validação interna (em vermelho) foi obtida uma especificidade de 0,97. No gráfico 6b podemos observar o limiar de separação entre as classes de 0,43. Este limiar foi determinado pelo método do teorema Bayesiano, que é uma maneira mais flexível e que se adapta melhor as características dos conjuntos de dados. Na etapa de predição, gráfico 7, foi obtida uma especificidade de 1 e uma sensibilidade de 0,92 (gráfico 7a), a separação das classes foi alcançada a um limiar de 0,37 (gráfico 7b).

A validação do modelo PLS-DA, gráfico 8, foi realizada utilizando as amostras do conjunto de teste selecionadas pelo algoritmo Kennard Stone.

**Gráfico 8 - Validação do modelo PLS-DA.**



Fonte: PRÓPRIA, 2019.

Nele pode-se observar a separação entre as duas classes, tanto na etapa de treinamento quanto na etapa de teste, utilizando 1 VL, que explica 34,60% da variância dos espectros (X) e 82,45% da variância da variável de classe (Y). As amostras convencionais estão localizadas nos escores positivos de LV1 e as OGM nos escores negativos de LV1.

## 5 CONCLUSÃO

A aplicação preliminar da tecnologia NIR para classificação das cultivares de algodão BRS Aroeira e BRS 368 RF como ferramenta de diagnóstico rápido e não destrutivo é viável de ser desenvolvida como método alternativo. Deste modo, o desenvolvimento de metodologias alternativas aos métodos convencionais para identificação de OGM's que sejam rápidos, confiáveis e menos onerosos economicamente é uma realidade. A espectroscopia NIR mostrou-se uma boa alternativa aos métodos convencionais por ser uma técnica não destrutiva, rápida e que não necessita do consumo de reagentes químicos. O método de reconhecimento de padrões não supervisionados, a partir da técnica de Análise de Componentes Principais, realizou uma análise exploratória dos dados onde foi observada uma tendência de separação entre as duas classes de amostras, bem como auxiliou na escolha no melhor pré-processamento a ser utilizado no desenvolvimento dos modelos de reconhecimento de padrões supervisionados. Os modelos de reconhecimento de padrão supervisionado, SIMCA e PLS-DA, foram desenvolvidos com os dados pré processados utilizando a correção de linha de base, associada ao SNV e a suavização utilizando janela de 9 pontos e polinômio Savitzky Golay de segunda ordem. As técnicas não supervisionadas (PCA) e as técnicas supervisionadas (SIMCA e PLS-DA) foram adequadas para distinguir entre as duas classes estudadas com uma taxa de classificação correta de 100% a um nível de confiança de 95%.

## REFERÊNCIAS

- AMORIM, S. R.; KLIER, A. H.; ANGELIS, L. H. **Controle de qualidade na indústria farmacêutica: identificação de substâncias por espectroscopia no infravermelho**. Rev. Bras. Farm, 2013.
- ARAÚJO, H. O. M. **Aditivação de querosene em diesel s10 b7 e estudo da estabilidade do combustível através de métodos físico-químicos e quimiométricos**. Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio Grande do Norte, Centro de Ciências Exatas e da Terra, Instituto de Química, Natal, 2016.
- BARROS, E. G.; MOREIRA, M. A. **Alimentos transgênicos**. Química e Derivados. n.385, p. 12-20, Agosto/2000.
- BÉLOT, J.L.; GILBANDI, M.; & SILVIE, P. **Transgenia é ferramenta para maximizar retornos**. Disponível em: <https://www.esalq.usp.br/visaoagricola/sites/default/files/va06-melhoramento-genetico03.pdf>. Acesso em 04 de julho de 2019.
- BELTRÃO, N. E. de M.; PEREIRA, J. R. **A semente do algodoeiro I. Uso in natura na alimentação animal**. Comunicado técnico 120: Embrapa Algodão, Campina Grande, 2000.
- BERGO, M. C. J. **Transferência de calibração na discriminação de mogno e espécies semelhantes utilizando NIRS e PLS-DA**. Dissertação (mestrado) - Universidade de Brasília, Instituto de Química, Programa de Pós-Graduação em Química, Brasília, 2014.
- CALEGARI, M. A.; **Espectroscopia na região do infravermelho próximo (NIR) e calibração multivariada: desenvolvimento de modelos PLS para a determinação da atividade antioxidante em amostras de própolis**. Dissertação (Mestrado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Tecnologia de Processos Químicos e Bioquímicos. Pato Branco, 2018.
- COLARES, C. J. G. **Estudo da homogeneidade e discriminação de madeiras através de espectroscopia de imagens, imagens digitais e ferramentas quimiométricas**. Tese (Doutorado) - Instituto de Química da Universidade de Brasília, Brasília, 2016.
- CONCEIÇÃO, F.R.; MOREIRA, A. N.; BINSFELD, P. C. **Detecção e quantificação de organismos geneticamente modificados em alimentos e ingredientes alimentares**. Ciência Rural, Santa Maria, v.36, n.1, p.315-324, jan-fev, 2006.
- CORREIA, P. R. M; FERREIRA, M. M. C. **Reconhecimento de padrões por métodos não supervisionados: explorando procedimentos quimiométricos para tratamento de dados analíticos**. Quim. Nova, Vol. 30, No. 2, 481-487, 2007.

ESKILDSEN, C. E. A. **Prediction of milk quality parameters using vibrational spectroscopy and chemometrics – opportunities and challenges in milk phenotyping.** University of Copenhagen, 2016.

FERREIRA, M. M. C. **Quimiometria: conceitos, métodos e aplicações** [online]. Campinas, SP: Editora da Unicamp, 2015.

FORATO, L. A. et al. **A espectroscopia na região do infravermelho e algumas aplicações.** 1ª ed. São Carlos: Embrapa Instrumentação Agropecuária, 2010.

CUI et al. **Automatic and rapid discrimination of cotton genotypes by near infrared spectroscopy and chemometrics.** Hindawi Publishing Corporation Journal of Methods. Analíticos em Química Volume 2012, Artigo 793468, 7 Páginas Analíticos em Química Volume 2012, Artigo 793468, 7 Páginas doi: 10,1155 / 2012/793468.

LEE, J. H.; CHOUNG, M. G. **Nondestructive determination of herbicide-resistant genetically modified soybean seeds using near infrared reflectance spectroscopy.** Food Chemistry, v. 126, p. 368-373, 2011.

MARCHESE, N. R. et al. **Espectroscopia NIR para análise da qualidade de soja inativada termicamente.** Tecnol. & Ciên. Agropec. João Pessoa, v.12, n.4, p.57-63, 2018.

MENDES, R. A. G. **Utilização da espectroscopia em reflectância no infravermelho próximo para discriminação de espécies da família Myrtaceae.** Dissertação (Mestrado) - Universidade de Brasília. Brasília-DF, 2014.

MONTEIRO, A. R. D. **Estudo de aspectos estatísticos da espectroscopia na região do infravermelho próximo.** Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Química. Rio de Janeiro, 2011.

NETO. B. B; SCARMINIO. I. S; BRUNS. R. E. **25 anos de quimiometria no Brasil.** Quím. Nova. vol.29 no.6. São Paulo, 2006.

OLIVEIRA, I. K. **Aplicação de quimiometria e espectroscopia no infravermelho no controle de qualidade de biodiesel e mistura biodiesel/diesel.** Dissertação (Mestrado) - Universidade Estadual de Campinas, Instituto de Química. Campinas-SP, 2008.

OLIVEIRA. V. S. **Análise da autenticidade de cédulas de real utilizando espectroscopia NIR portátil e quimiometria.** Dissertação (Mestrado) – Universidade Federal de Pernambuco, CCEN, Química Fundamental. Recife, 2018.

PASQUINI, C. **Espectroscopia no infravermelho próximo para avaliar indicadores de qualidade tecnológica e contaminantes de grãos.** 1ª ed. Brasília-DF: Embrapa, 2018.

PEREIRA, L. S. A. **Análise multivariada e espectroscopia no infravermelho aplicadas em análises forenses : drogas e medicamentos.** Tese (doutorado) – Universidade Federal de Minas Gerais – Departamento de Química. Belo Horizonte, 2018.

PETIT, L. et al. **Screening of genetically modified organisms and specific detection of Bt176 maize in flours and starches by PCR-enzyme linked immunosorbent assay.** European Food Research and Technology, v.217, p.83-89, 2003.

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.** Rev. Bras. Cienc. Farm. vol.40 n.3 São Paulo, 2004.

SALA, O. **Fundamentos da espectroscopia Raman e no infravermelho.** 2ª ed. São Paulo: Editora UNESP, 2008.

SILVA, C. S. **Espectroscopia no infravermelho para aplicações forenses: documentoscopia e identificação de sêmen em tecidos.** Tese (doutorado) – Universidade Federal de Pernambuco. CCEN. Química fundamental. Recife, 2017.

SKOOG, D. A.; HOLLER, F. J.; NIEMAN, T. A. **Princípios de Análise Instrumental.** 5ª. ed. Porto Alegre: Bookman, 2002.

SOARES, D. X. **Planejamento e otimização de processos multirespostas: tratamento quimiométrico.** Tese (doutorado) – Universidade Estadual de Campinas. Campinas - SP, 2014.

STUTH, J.; JAMA, A.; TOLLESON, D. **Direct and indirect means of predicting forage quality through near infrared reflectance spectroscopy.** Field Crops Research, v.84, p.45-56, 2003.

VALDERRAMA, P. **Avaliação de figuras de mérito em calibração multivariada na determinação de parâmetros de controle de qualidade em indústria alcooleira por espectroscopia no infravermelho próximo.** Dissertação (Mestrado) – Universidade Estadual de Campinas, Instituto de Química. Campinas, 2005.

XIE, L. et al. **Discrimination of transgenic tomatoes based on visible/near-infrared spectra.** Analytica Chimica Acta, volume 584, edição 2, 2007, p: 379-384.

ZABOT, L. **A cultura do algodão: (Gossypium hirsutum L.).** Santa Maria, RS: Centro de Ciências Rurais, 2007.