UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA
DE ALIMENTOS
NÍVEL MESTRADO ACADÊMICO

SABRINA FORTINI SPOLADORE

CLASSIFICAÇÃO DE GENÓTIPOS DE TRIGO USANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO E QUIMIOMETRIA

DISSERTAÇÃO DE MESTRADO

CAMPO MOURÃO
2019

SABRINA FORTINI SPOLADORE

CLASSIFICAÇÃO DE GENÓTIPOS DE TRIGO USANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO E QUIMIOMETRIA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Tecnologia de Alimentos da Universidade Tecnológica Federal do Paraná como requisito para a obtenção do título de Mestre em Tecnologia de Alimentos.

Orientador: Profº. Dr. Evandro Bona.

Coorientadora: Profª Drª. Maria Brígida dos Santos Scholz.

CAMPO MOURÃO

2019

S762c Spoladore, Sabrina Fortini.
Classificação de genótipos de trigo usando espectroscopia de infravermelho próximo e quimiometria. / Sabrina Fortini Spoladore – Campo Mourão, 2019.
53 f.: il. color., 30 cm.

Orientador: Profº. Dr. Evandro Bona.
Coorientador: Profª. Drª. Maria Brígida dos Santos Scholz .
Dissertação (Mestrado) - Universidade Tecnológica Federal do Paraná, Programa de Pós-Graduação em Tecnologia de Alimentos. Campo Mourão, 2019.
Inclui bibliografia.

1. Trigo. 2. Espectroscopia de infravermelho. 3. Quimiometria 4. Alimentos - Dissertações I. Bona, Evandro, orient. II. Scholz, Maria Brígida dos Santos, coorient. III. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Tecnologia de Alimentos. IV. Título.

CDD (22 ed.) 664

Ficha catalográfica elaborada por Caroline Felema dos Santos Rocha CRB: 9/1880
Biblioteca da UTFPR-Dois Vizinhos

TERMO DE APROVAÇÃO

CLASSIFICAÇÃO DE GENÓTIPOS DE TRIGO USANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO E QUIMIOMETRIA

Por

SABRINA FORTINI SPOLADORE

Essa dissertação foi apresentada às 14 horas, do dia 03 de Abril de 2019, como requisito parcial para a obtenção do título de Mestre em Tecnologia de Alimentos, Linha de Pesquisa Ciência e Tecnologia de Produtos Alimentícios, no Programa de Pós-Graduação em Tecnologia de Alimentos - PPGTA, da Universidade Tecnológica Federal do Paraná. A candidata foi arguida pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho APROVADO.

_____

Prof. Dr. Evandro Bona (Orientador – PPGTA)

_____

Prof. Dr. Paulo de Tarso Carvalho (Membro Externo – UTFPR-LD)

_____

Prof. Dr. Fábio Luiz Melquiades (Membro Externo – UEL)

* A via original com as assinaturas encontra-se na secretaria do programa.

AGRADECIMENTOS

A Deus, pela concessão divina da graça da vida.

Agradeço imensamente ao meu orientador, Prof. Dr. Evandro Bona por ter confiado em mim para a execução desse projeto. Agradeço por todo suporte dado, pela paciência e disponibilidade sem tamanhos, pela amizade adquirida que tornou tudo mais leve e fácil, e principalmente por todo ensinamento compartilhado. Foi uma honra muito grande ser orientada por você, a você toda admiração e gratidão.

À minha coorientadora Profª Drª. Maria Brígida dos Santos Scholz, por toda disponibilidade para o enriquecimento do trabalho.

À minha família e amigos que me deram suporte e apoio nos momentos difíceis e não duvidaram do meu êxito.

Aos meus colegas de mestrado, em especial a Tamires, que se tornou uma grande amiga e esteve ao meu lado em todas as dificuldades.

À banca examinadora pelas sugestões e atenção dedicadas a este estudo.

Aos professores do Programa de Pós-Graduação em Tecnologia de Alimentos da Universidade Tecnológica Federal do Paraná (UTFPR), por todo ensinamento, conselhos e experiências compartilhadas.

Ao IAPAR-Londrina, que cedeu as amostras e deu todo suporte necessário para a realização desse trabalho.

À CAPES pelo financiamento dos meus estudos.

Agradeço a todos que diretamente ou indiretamente contribuíram para realização deste projeto.

RESUMO

SPOLADORE, Sabrina Fortini. Classificação de genótipos de trigo usando espectroscopia de infravermelho próximo e quimiometria. 2019. f.54. Dissertação de Mestrado – Programa de Pós-Graduação em Tecnologia de Alimentos, Universidade Tecnológica Federal do Paraná, Campo Mourão, 2019.

O trigo, *Triticum aestivum L.*, é uma das mais importantes culturas de cereais, representando cerca de 30% da produção mundial. As interações entre os genótipos de trigo e as condições ambientais definem a qualidade do grão e, consequentemente sua utilização industrial. A espectroscopia por reflexão no infravermelho próximo (NIRS) possui entre suas principais aplicações inúmeros usos para a análise de qualidade do trigo. Entre as vantagens do NIRS estão o baixo tempo de análise, simples preparo da amostra e a obtenção de espectros com informações químicas relacionadas com a composição da farinha de trigo. Métodos quimiométricos auxiliam na interpretação de dados analíticos instrumentais, tais como dados espectrais. Assim, o objetivo geral do estudo é empregar a espectroscopia de infravermelho próximo combinada com a quimiometria para discriminar genótipos de trigo. Foram analisadas 180 amostras de farinha de trigo (8 genótipos, 17 cidades de cultivo e 2 safras) cedidas pelo programa de Cereais de Inverno do Instituto Agronômico do Paraná (IAPAR-Londrina). Foram utilizados métodos quimiométricos (PCA, HCA, *k-means* e PLS-DA) para analisar os espectros NIR dessas amostras. O emprego da NIRS combinado com a análise exploratória PCA e de agrupamento, HCA e *k-means*, indicou que os grupos são formados em função dos genótipos. O local de cultivo e a safra não apresentam um efeito importante na formação de agrupamentos. Através da análise PCA foi observada uma separação dos oito genótipos de trigo em três grupos. Esses genótipos se separaram principalmente nas bandas características da umidade, proteína e cinzas. A dureza dos genótipos de trigo foi um fator importante para a formação desses grupos, onde os genótipos de textura macia “soft” se separaram notavelmente dos de textura dura “hard”, semidura e muito dura. O modelo PLS-DA usando os espectros NIR também classificou assertivamente os genótipos com valores médios de 95,83% para a sensibilidade e 99,53% para a seletividade. Foi possível notar uma clara separação das classes LD121102, LD132210 e LD131102.

Portanto, foi possível classificar os genótipos de trigo através dos espectros NIR combinado com os métodos quimiométricos.

Palavras-chave: *Triticum aestivum L*, NIR, PCA, HCA e PLS-DA.

ABSTRACT

SPOLADORE, Sabrina Fortini. Classification of wheat genotypes using near infrared spectroscopy and chemometrics. 2019. f.54. Defesa de Dissertação de Mestrado – Programa de Pós-Graduação em Tecnologia de Alimentos, Universidade Tecnológica Federal do Paraná, Campo Mourão, 2019.

Wheat, *Triticum aestivum* L., is one of the most important cereal crops, accounting for about 30% of world production. The interactions between wheat genotype and environmental conditions define the quality of the grain and consequently its industrial use. Each type of industrial application requires wheat flour with specific physicochemical and rheological characteristics. Near infrared spectroscopy (NIRS) has among its main applications numerous uses for the analysis of wheat quality. One of the advantages of NIRS is the low time of analysis and simple sample preparation, the spectra obtained contains chemical information that may be related to the properties of wheat flour. Chemometrics methods aid in the interpretation of instrumental analytical data, such as spectral data. Thus, the overall objective of the project is to employ near infrared spectroscopy combined with chemometrics to discriminate wheat genotypes. A total of 180 samples (8 genotypes, 17 crop cities and 2 harvests) were analyzed by the IAPAR-Londrina Institute of Agronomic Institute's Winter Cereals program. NIR spectra were collected from wheat flours extracted in an experimental mill at the Laboratory of Plant Physiology of IAPAR. The spectra were pretreated and chemometrics methods (PCA and PLS-DA) were used to classify the samples. The use of the NIRS combined with the PCA exploratory analysis, and by the HCA and k-means cluster analyzes indicate that the groups are formed by the genotypes; the place of cultivation and the crop do not present an important effect in the formation of these groups. Through the PCA analysis, a separation of the eight wheat genotypes in three groups was observed. These genotypes separated mainly in the characteristic bands of moisture, protein and ashes, presenting higher humidity and lower content of ashes and proteins than the other genotypes. The hardness of the wheat genotypes was an important factor for the formation of these groups, where the soft genotypes severely separated from those of hard texture, semi hard and very hard. The PLS-DA model using the NIR spectra also assertively classified the samples

with mean values of 95.83% for sensitivity and 99.53% for selectivity. Through this model it was also possible to notice a clear separation of classes LD121102, LD132210 and LD131102. Thus, it was possible to classify the wheat samples through the NIR spectra in tandem with the chemometrics methods PCA, HCA and PLS-DA.

Keywords: *Triticum aestivum L*, NIR, PCA, HCA e PLS-DA.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS DE SIGLAS E SÍMBOLOS

A - Número de Variáveis Latentes mais um

AUC - Área abaixo da curva ROC – do inglês *Area Under the Curve ROC*

c - Concentrações

E - Matriz de erro

FN - Falso Negativo – do inglês *False Negative*

FP - Falso Positivo – do inglês *False Positive*

HCA - Análise Hierárquica de Cluster – do inglês *Hierarchical Cluster Analysis*

I - Número de amostras utilizadas na validação cruzada

LV - Variável Latente – do inglês *latent variable*

MSC - Correção Multiplicativa de Espalhamento – do inglês *Multipicative Scatter Correction*

NIRS - Espectroscopia de Infravermelho Próximo – do inglês *Near-infrared Spectroscopy*

P - Matriz de *Loadings*

PC - Componente Principal – do inglês *Principal Component*

PCA - Análise de Componentes Principais – do inglês *Principal Component Analysis*

PCC - Porcentagem de Classificação Correta- do inglês *Percentage of Correct Classification*

PCR - Regressão pelo Método das Componentes Principais – do inglês *Principal Component Regression*

PLS-DA - Análise Discriminante pelo Método de Quadrados Mínimos Parciais – do inglês *Partial Least Squares Discriminant Analysis*

RMSE - Raiz Quadrada do Erro Quadrático Médio– do inglês *Root Mean Square Error*

RMSEC- Raiz Quadrada do Erro Quadrático Médio de Calibração – do inglês *Root Mean Square Error of Calibration*

RMSEP - Raiz Quadrada do Erro Quadrático Médio de Previsão - do inglês *Root Mean Square Error of Prediction*

ROC – Características de Operação do Receptor – do inglês *Receiver Operation Characteristics*

T - Matriz de *Scores*

TN - Verdadeiro Negativo – do inglês *True Negative*

TP - Verdadeiro Positivo – do inglês *True Positive*

VIS-NIR - Espectroscopia Visível e de Infravermelho Próximo

X - Matriz original

y_i- valor de referência

$\hat{y}$ - valor previsto para cada amostra

SUMÁRIO

1 INTRODUÇÃO

O trigo, *Triticum aestivum* L., é um dos cereais mais cultivado em todo o mundo e, devido principalmente ao seu potencial calórico, é considerado um alimento básico em muitos países por ser importante fonte de energia. A farinha de trigo é um ingrediente muito popular incorporado aos hábitos alimentares da maioria da população mundial. É o cereal mais utilizado em panificação e outros produtos à base de cereais, como biscoitos, massas, bolos, etc. (Correia *et al.*, 2017; Guo *et al.*, 2018)

O grão de trigo é classificado de acordo com suas características de dureza. Os grãos podem apresentar textura dura (*hard*) e textura macia (*soft*). A dureza é resultado da agregação entre amido e proteína no endosperma do grão e é avaliada pela produção de farinha de quebra durante a moagem. Nos trigos de textura macia a separação entre o pericarpo e o endosperma é mais fácil e há maior formação de farinha de quebra e poucas etapas de redução para atingir o rendimento de farinha rentável. Assim, farinha final tem baixo teor de farelo (Choy, Walker e Panozzo, 2015).

A capacidade panificadora da farinha de trigo está diretamente relacionada à sua complexa composição de proteínas. As proteínas do glúten, gliadina e glutenina, conferem qualidade ao pão, devido à sua capacidade de absorção de água e coesão, viscosidade e características de extensibilidade. A qualidade e a quantidade dessas proteínas dependem de vários fatores, como variedade de trigo e condições ambientais (Bardini *et al.*, 2018).

A aplicação da espectroscopia no infravermelho próximo (NIRS) na análise de alimentos tem sido utilizada amplamente em grãos, farinhas e produtos industrializados. A determinação da proteína tem sido realizada pela NIRS em substituição ao método Kjeldahl, que por mais que tenha precisão, é um método lento e ainda gera resíduos químicos (Ferrão *et al.*, 2004).

A NIRS é muito utilizada na análise da farinha de trigo, por possuir capacidade de gerar resultados extremamente rápidos, podendo mesmo ser utilizada antes de descarregar a farinha nos silos. Com isso, as farinhas podem ser analisadas quanto à conformidade com as especificações, caso o produto seja inadequado, essas cargas são rejeitadas antes de entrarem no sistema de produção e comprometerem a qualidade da linha de produtos. Os parâmetros de qualidade da farinha que podem ser determinados pela NIRS são proteína, umidade, tamanho de partícula, cinza, cor, dano de amido e absorção de água (Burns e Ciurczak, 2007).

Estudos avaliaram a difusão da água em seções únicas de grãos de trigo ao longo do tempo, utilizando imagens hiperespectrais de infravermelho próximo (Lancelot *et al.*, 2017), relataram a determinação da proteína total e do glúten úmido da farinha de trigo através do NIR (Chen, Zhu e Zhao, 2017). Esta técnica foi utilizada para caracterizar o processo de tratamento térmico de farinha de trigo para bolos (Verdú, Ivorra, *et al.*, 2016) e classificar variedades de trigo (Ziegler *et al.*, 2016).

Os métodos espectrais proporcionam uma elevada quantidade de dados que devem ser processados para fornecer informações práticas. A quimiometria se refere à extração de informações relevantes de dados químicos com ferramentas matemáticas e estatísticas. Dentre os métodos quimiométricos mais utilizados destacam-se a análise de componentes principais (PCA), a regressão pelo método das componentes principais (PCR) e a regressão pelo método dos mínimos quadrados parciais (PLS). Todas essas técnicas fornecem uma descrição resumida de conjuntos de dados multivariados, podendo assim ser considerada a análise do espectro completo (Ranzan *et al.*, 2014).

Embora a NIRS seja mais utilizada na determinação de parâmetros químicos como umidade, proteína e teor de cinzas, o objetivo do presente estudo foi a aplicação da NIRS para a classificação de diferentes genótipos de trigo. O desenvolvimento de modelos de discriminação baseou-se na avaliação quimiométrica dos espectros NIR obtidos de amostras de farinhas de trigo. Por apresentar resultados rápidos, o investimento ser baixo, pequena quantidade de amostra, esta técnica se mostra viável para a classificação de genótipos, sendo benéfica até mesmo na indústria de farinha de trigo, no controle de qualidade e processamento do trigo.

OBJETIVOS

2 OBJETIVO GERAL

Aplicar a espectroscopia de infravermelho próximo em conjunto com os métodos quimiométricos para classificar amostras de trigo de diferentes genótipos, safras e locais de cultivo.

2.1 Objetivos específicos

- Coletar os espectros NIR das amostras e fazer os pré-tratamentos necessários;
- Fazer uma análise exploratória dos espectros NIR empregando PCA;
- Realizar análises de agrupamento usando HCA e *k-means* nos espectros NIR;
- Construir modelos PLS-DA para discriminar os diferentes genótipos de trigo com base nos espectros NIR.

3 REVISÃO BIBLIOGRÁFICA

3.1 Farinha de trigo

O trigo, *Triticum aestivum* L., é utilizado para diversos alimentos e é considerada uma das mais importantes culturas de cereais. Cerca de 43 países e 35 % da população mundial consomem regularmente produtos alimentícios à base de farinha de trigo (Misra *et al.*, 2015; Zhang *et al.*, 2014). Em 2018/2019 a previsão de comercialização mundial de trigo é de aproximadamente 172 milhões de toneladas. Os mais importantes exportadores de trigo são Argentina, Austrália, Canadá, União Europeia, Cazaquistão, Federação Russa, Ucrânia e Estados Unidos (FAO, 2019).

A farinha de trigo é fonte de carboidratos e proteínas importantes para o metabolismo humano. A farinha de trigo é composta principalmente pelo amido que representa de 70 a 75% e por proteínas que representam cerca de 8 a 14% da composição total. Existem alguns componentes secundários como os 2% de lipídios, 2 a 3% de polissacárideos não-amiláceos cerca de 2 a 3%, os minerais, as vitaminas e os antioxidantes. Outros nutrientes também estão presentes na farinha de trigo integral (Guo *et al.*, 2018).

A composição do trigo e seus derivados é reflexo dos componentes biodisponíveis presentes nos solos onde são cultivados (González-Martín *et al.*, 2014).

A farinha de trigo é um ingrediente base para diversos produtos alimentícios, incluindo pães, bolos, biscoitos, *doughnuts*, macarrão e massas. A qualidade da farinha de trigo é comumente descrita principalmente pela quantidade de proteína total e pela composição e quantidade de proteínas formadoras do glúten. A partir desses parâmetros, é possível selecionar a farinha mais apropriada para determinado produto (Chen, Zhu e Zhao, 2017).

O comportamento tecnológico da farinha não depende somente do teor de proteína e glúten, mas também é resultado de complexas interações entre macromoléculas responsáveis pelo desempenho da massa. Assim, a farinha de trigo é classificada por alguns parâmetros, mais comumente determinados por análises reológicas que fornecem informação quantitativa das propriedades mecânicas (Dobraszczyk e Morgenstern, 2003). Uma das mais importantes estruturas formadoras de produtos de panificação são as proteínas do glúten, que contribuem para as características de elasticidade, coesividade, extensibilidade e viscosidade da massa,

que determinam a qualidade dos produtos à base de farinha de trigo. A rede de glúten é composta por duas principais frações proteicas, gliadinas (monoméricas) que conferem propriedades viscosas à massa e gluteninas (poliméricas) que conferem força e elasticidade (Hu, Wang e Li, 2017).

Os trigos são classificados de acordo com as características de formação de glúten e pelo número de queda *Falling Number*. Cada classe de trigo é adequada para um tipo de produto, possuindo um determinado atributo funcional. Para a panificação é preferível as farinhas de variedades de trigo fortes, pois possuem uma forte rede de glúten. Já para a produção de biscoitos e bolos, o trigo mole/brando é mais apropriado por apresentar uma rede de glúten fraca (Dobraszczyk e Morgenstern, 2003). Porém, os lotes de farinha possuem uma grande variação nos parâmetros reológicos, devido as variedades de trigo, condições climáticas e técnicas agronômicas (como controle de doenças, etc), por isso é difícil manter a qualidade constante da farinha (Li Vigni *et al.*, 2013).

A dureza do trigo é um parâmetro que determina o comportamento de moagem e é geralmente utilizada para diferenciar classes de trigo para fins específicos como, por exemplo, para a produção de pães, biscoitos, macarrão, massas e outros alimentos feitos a partir de grãos de trigo (Choy, Walker e Panozzo, 2015; Morris, 2002). A facilidade com que ocorre a fratura dos grãos, influencia no grau de dano do amido, na distribuição do tamanho das partículas da farinha e no rendimento da quebra (Choy, Walker e Panozzo, 2015). A classificação do grão trigo pode ser dividida de acordo com a dureza em textura macia, semimacia, dura, semidura e muito dura (Morris, 2002). A dureza do trigo é geralmente medida usando o índice de tamanho de partícula (PSI) ou sistema de caracterização de um único grão (SKCS) (Choy, Walker e Panozzo, 2015).

Os grãos de trigo mole são mais facilmente quebrados, obtendo elevada quantidade de grânulos de amido intactos e uma farinha mais fina, com pouco dano ao amido. A farinha com maior granulometria é produzida por trigo duro, devido a fratura de grânulos de amido e consequentemente maior dano ao amido, consumindo mais energia no moinho de farinha. A diferença física mais relevante é entre o endosperma de trigo duro e mole, onde o trigo mole possui grânulos de amido ligados com matriz protéica ao redor desses grânulos (Pasha, Anjum e Morris, 2001).

Em geral, o trigo duro geralmente produz uma farinha com cor mais branca e produz uma maior absorção de água, por ser amido fraturado e danificado. Para

panificação, esta característica geralmente resulta em maior rendimento de pão (Van Der Borght et al. 2005). Já biscoitos, bolos e doces é indicado usar farinha de trigo mole, por apresentar menor teor de proteína e glúten fraco (Pasha, Anjum e Morris, 2001).

3.2 Espectroscopia no Infravermelho Próximo

A NIRS tornou-se uma alternativa às técnicas químicas padrões que são utilizadas para analisar alimentos. Uma das primeiras revisões sobre a espectroscopia por reflexão no infravermelho próximo (NIRS) ocorreu nos anos 80, onde foram relatadas as principais aplicações na avaliação de trigo, bem como da farinha e seus produtos industrializados. Foram também apresentadas vantagens na utilização do NIRS, como o baixo tempo de análise das técnicas de reflexão no controle de qualidade da farinha de trigo (Osborne, 1981).

A região do espectro eletromagnético que corresponde ao infravermelho próximo está compreendida na faixa de comprimento de onda entre 780 e 2500 nm. Nessa região é possível encontrar informações referentes às ligações C–H, N–H e O–H, as quais estão presentes nos principais componentes estruturais das moléculas orgânicas. É possível identificar origens geográficas de vários produtos com base nas respostas vibracionais de ligações químicas à radiação na região do infravermelho próximo (Zhao *et al.*, 2013).

O espectro NIR é composto por bandas harmônicas e combinações de vibrações fundamentais. Esta técnica é considerada sensível a uma variedade de grupos químicos e interações moleculares, possuindo grande número de aplicações, na agricultura, indústrias farmacêuticas e petróleo. Até então, a espectroscopia NIRS é considerada uma boa técnica para a análise quantitativa (Shi e Yu, 2017).

As análises usando a NIRS oferecem informações para o estudo de mudanças que ocorrem na farinha de trigo. A espectroscopia de infravermelho na região do visível e infravermelho próximo VIS-NIR tem sido utilizado em estudos dentro da área de processamento de cereais, analisando a variabilidade e as propriedades das farinhas se baseando tanto na proteína quanto no amido. Esta técnica foi utilizada para detectar farinha de trigo adulterada com outros cereais (Verdú *et al.*, 2015; Verdú, Vásquez, *et al.*, 2016; Xing *et al.*, 2011), caracterização do processo de tratamento térmico de farinha de trigo para bolos (Verdú, Ivorra, *et al.*, 2016), determinação da

proteína total e do glúten úmido na triagem da farinha comercial para adequar ao processamento desejado (Chen, Zhu e Zhao, 2017). Também foi avaliada a difusão da água em grãos de trigo ao longo do tempo (Lancelot *et al.*, 2017) e a farinha de trigo destinada ao processo de panificação (Verdú *et al.*, 2015).

Nesse contexto, a espectroscopia no infravermelho próximo apresenta-se como um método instrumental com grande potencial para a análise do trigo. Entretanto, é necessário salientar que o emprego da quimiometria é imprescindível para o tratamento da informação contida nos espectros NIR (Ahmad *et al.*, 2016; Ferrão *et al.*, 2004). A espectroscopia de infravermelho associada à quimiometria demonstrou ser uma técnica não destrutiva e rápida em pesquisas ambientais e alimentícias (Shi e Yu, 2017). Para o trigo, a técnica NIRS já foi aplicada para identificar a origem geográfica (Zhao *et al.*, 2013).

3.3 Quimiometria

O surgimento da quimiometria foi na década de 1970, onde seu nome foi apresentado pela primeira vez em um artigo de Svante Wold (Brereton, 2018). O termo “quimiometria” (*chemometrics*) se tornou reconhecido na década de 1980, quando a revista *Analytical Chemistry* substituiu o título “Statiscal and Mathematical Methods in Analytical Chemistry” por “Chemometrics”, em sua revisão bianual (Ferreira, 2015). A quimiometria teve origem baseda em três áreas fundamentais: estatística aplicada (análise exploratória e planejamento experimental), estatística em química analítica e físico-químca e computação científica (Brereton, 2018).

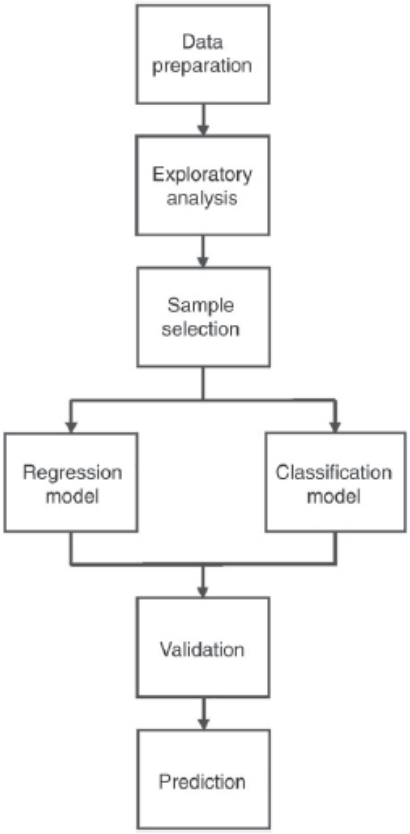
Na análise exploratória são avaliados conjuntos de dados químicos complexos, com medições laboratoriais englobando o registro de diversas variáveis por amostra. Um espectro pode ser gravado em centenas de comprimentos de onda, ou um cromatograma em muitos tempos de eluição, por exemplo. Assim, é possível obter inúmeras informações de cada amostra (Ferreira, 2015).

As análises multivariadas exploratórias, como análise de componentes principais (PCA) e análise hierárquica de agrupamento (HCA), são utilizadas para processar um grande número de dados, auxiliando na interpretação dos resultados. Na análise de alimentos, essas técnicas são geralmente usadas para classificar e avaliar a autenticidade das amostras através de suas características químicas (Shi e Yu, 2017).

Na química analítica quantitativa dificilmente é possível estimar diretamente a concentração de uma ou mais espécies químicas (compostos, radicais) e para contornar esta situação, constrói-se um modelo de calibração, onde se determina a relação existente entre um sinal instrumental e a concentração. A calibração multivariada é uma das mais eficientes associações de métodos estatísticos com dados químicos. Os principais métodos utilizados são as regressões em componentes principais (PCR) e as regressões por quadrados mínimos parciais (PLS) (Neto, Scarminio e Bruns, 2006).

A utilização dos métodos quimiométricos, Figura 1, compreende várias etapas dependes da disposição dos dados experimentais coletados em forma de uma matriz, **X** (i x j), onde i representa cada linha e refere-se a medida de uma amostra, como por exemplo um espectro, e j representa cada coluna da matriz, e diz sobre a variável, como por exemplo, a absorbância em vários comprimentos de onda. Inicia-se pelo pré-processamentos de dados como centrar na média, alisamento, derivação, correção da linha base entre outros. Segue-se com o emprego de métodos exploratórios multivariados conforme a dimensionalidade dos dados. E, posteriormente tem-se a separação das amostras no conjunto de calibração (utilizado para ajustar o modelo) e no conjunto de previsão (utilizado para verificação da capacidade de generalização do modelo), seno esse um passo essencial anterior a aplicação de modelos multivariados de classificação ou regressão (Bona, Março e Valderrama, 2018).

Figura 1- Fluxograma simplificado para aplicação de métodos quimiométricos.



Fonte: Bona, Março e Valderrama (2018).

3.3.1 Pré-tratamentos

Após os dados experimentais serem organizados em uma matriz **X** (amostras nas linhas e variáveis nas colunas) é preciso realizar o pré-tratamento de dados. É fundamental escolher o pré-tratamento adequado, pois o mesmo tem como função diminuir as alterações no sinal que não são desejáveis. Porém, o pré-tratamento pode alterar de maneira prejudicial o resultado final se não for bem executado. Existem dois tipos de pré-tratamentos, a transformação que é aplicado à amostra nas linhas da matriz **X**; e o pré-processamento, que é aplicado às variáveis nas colunas da matriz **X** (Bona, Março e Valderrama, 2018).

3.3.1.1 Derivadas

Um erro instrumental ou de amostragem pode causar um deslocamento constante no espectro e para corrigir este deslocamento é utilizada a primeira derivada do espectro. A primeira derivada de uma constante é zero, então o espectro terá como

resultado zero de absorbância. A segunda derivada é utilizada caso o espectro mostre uma inclinação na linha de base devido ao decrescimento do número de onda. Estes pré-tratamentos com derivadas são bastante utilizados quando se observam espectros de refletância difusa, onde problemas de deslocamento e inclinação da linha de base acontecem com frequência (Ferreira, 2015).

O método de Savitzky-Golay é o mais utilizado para o cálculo das derivadas espectrais. É um filtro de média móvel que faz ajuste de um polinômio de grau *n*, por mínimos quadrados, aos (2m + 1) pontos da janela móvel e assim determinar o valor do polinômio no ponto central. Então, a primeira derivada de *x* em relação a *k* é dada pela Equação (1) (Ferreira, 2015).

$$\frac{dx}{dk} = a_1 + 2a_2k + \cdots + na_nk^{n-1} \quad (1)$$

O uso de derivadas é o método mais adequado para corrigir linhas de bases dos espectros, pois seu cálculo não determina o uso subjetivo das funções paralelas e nem inclui alta variância nos dados. Tem como desvantagem o decréscimo progressivo da razão sinal/ruído durante o cálculo das derivadas, chegando a constituir resultados inaceitáveis, mas caso a razão não seja elevada, não implicará em um problema (Ferreira, 2015).

3.3.1.2 Correção Multiplicativa de Espalhamento – MSC

A transformação MSC se aplica aos espectros de infravermelho, em espectros Raman e na região do UV-VIS. Os efeitos de espalhamento aditivos e multiplicativos na absorbância são corrigidos pelo MSC. Esses efeitos podem ser resultados de fenômenos físicos: como alteração do caminho ótico, sensibilidade do detector e amplificador; alterações na pressão e temperatura; divergência na granulometria das amostras (Ferreira, 2015).

A grande vantagem da utilização do MSC em relação a derivadas, é que o primeiro remove efeitos multiplicativos, preserva a forma original do espectro, possibilitando melhor interpretação dos resultados, mesmo não havendo correção na inclinação da linha de base, pois a transformação MSC usa a projeção dos espectros no espectro médio, que mantém a mesma orientação (Ferreira, 2015).

3.3.1.3 Centrar os Dados na Média

Centrar os dados na média é um pré-processamento utilizado nas variáveis, em cada coluna da matriz de dados. Esse método é realizado, através da diferença do valor médio de cada coluna da matriz de dados e de cada um dos valores da respectiva coluna Equação (2). Este pré-processamento realiza uma translação de eixos para o valor médio de cada um dos eixos, assim a estrutura dos dados mantém-se preservada (Ferreira, 2015).

$$x_{ij(cm)} = x_{ij} - \bar{x}_j \quad (2)$$

3.3.2 Análise de Componentes Principais – PCA

A análise de componentes principais (PCA) é uma técnica que diminui a dimensionalidade dos dados mantendo a maior parte da variação observada (Lancelot *et al.*, 2017). A PCA permite a visualização e interpretação das relações entre as variáveis e amostras. Por meio desta técnica, é possível observar as amostras que possuem comportamentos atípico, devido a projeção de dados em um espaço de menor dimensão (Ferreira, 2015).

A compressão de dados é resultado da combinação linear das variáveis originais, agrupando informações similares. As componentes principais (PC) são apresentadas como um novo conjunto de eixos onde as amostras serão projetadas. As PC são ortogonais, o que indica que a informação que contém em uma delas não está presente na outra. A PC1 é descrita pela direção que define a máxima variância dos dados originais, a PC2 está relacionada à direção de máxima variância dos dados ortogonal a PC1, assim como a PC3, será ortogonal à PC2, e assim por diante (Bona, Março e Valderrama, 2018).

A PCA envolve uma transformação matemática abstrata da matriz de dados original, *e.g.* os espectros NIR, conforme descrito na Equação (3)

$$\mathbf{X} = \mathbf{TP} + \mathbf{E}. \quad (3)$$

onde **T**, é a matriz de *scores* e possui a mesma quantidade de linhas que a matriz de dados original (**X**). Já **P**, é a matriz de *loadings* e possui a mesma quantidade de colunas que a matriz de dados original. A quantidade de colunas na matriz **T** é igual a quantidade de linhas na matriz **P** e corresponde ao número de PC escolhidas. **E** é uma matriz de resíduos, e possui as mesmas dimensões de **X**. Essa transformação geralmente é realizada através do método de decomposição em valores singulares (SVD) (Brereton, 2018).

3.3.3 Análise Hierárquica de Agrupamento – HCA

Este método tem como objetivo principal agregar amostras as mais semelhantes entre si formando um mesmo grupo. A finalidade da HCA é potencializar a homogeneidade dentro dos grupos e a heterogeneidade entre grupos (Ferreira, 2015). O resultado é geralmente apresentado em um dendrograma, um gráfico que apresenta a organização das amostras e suas relações (Granato *et al.*, 2018). Para aplicação do HCA é necessário definir uma métrica de distância e um método de agrupamento.

Segundo Brereton (2018), a distância Euclidiana no plano multidimensional (amostras *k* e *l*) é dada pela Equação (4),

$$d_{kl}^2 = (\mathbf{x}_k - \mathbf{x}_l)(\mathbf{x}_k - \mathbf{x}_l)^T \quad (4)$$

onde $\mathbf{x}_k$ e $\mathbf{x}_l$ são os vetores que representam, respectivamente, as amostras *k* e *l*. Quanto menor a distância Euclidiana, mais semelhantes são as amostras.

O método de Ward para o agrupamento de amostras aplica a soma quadrática das distâncias. Nesse método, a cada fase, dois grupos com menor acréscimo na soma quadrática total dentro do grupo são juntados, podendo ser denominado como método de variância mínima. Realizando o cálculo da soma dos quadrados das distâncias do centroide médio de cada grupo, obtém-se a distância entre os agrupamentos (Ferreira, 2015).

3.3.4 K-means

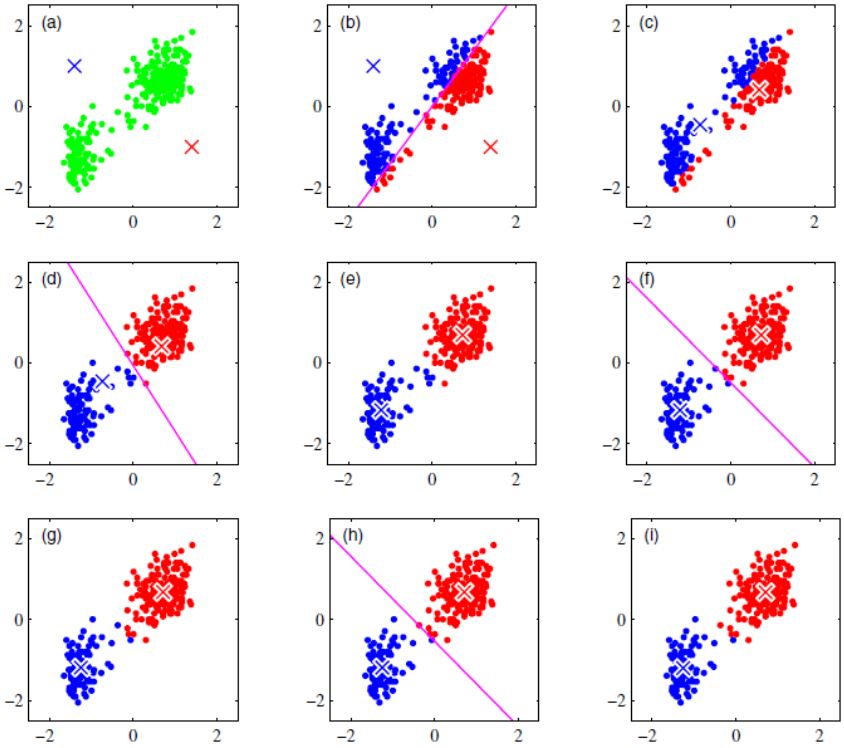
Este algoritmo tem como objetivo particionar o conjunto de dados em *k clusters*. *Cluster* é um grupo de pontos de dados onde as distâncias entre pontos são menores do que as distâncias para pontos fora do *cluster* (Bishop, 2006).

O algoritmo *k-means* é considerado uma técnica simples e eficiente, por ser um bom método de agrupamento, onde é capaz de classificar muitos dados numéricos de alta dimensão. Os dados que são agrupados pelo algoritmo *k-means* em um mesmo *cluster* (grupo) possuem alta similaridade (Yu *et al.*, 2018).

Na implementação direta do algoritmo *k-means*, em cada fase é preciso calcular a distância Euclidiana entre cada vetor protótipo e cada ponto de dados para realizar as médias de *cluster*, que é repetida até que seja clara a separação de *clusters* (Bishop, 2006).

O funcionamento do algoritmo *k-means* é apresentado na Figura 2 para K = 2, ou seja, dois *clusters*. Na Figura 2(a), os pontos verdes representam um conjunto de dados em um espaço Euclidiano de 2 dimensões. As cruzes vermelhas e azuis são as escolhas iniciais aleatórias para os centros dos possíveis grupos. Em 2(b) cada ponto de dados é atribuído ao *cluster* vermelho ou ao *cluster* azul, conforme a maior proximidade. Para 2(c), cada centro do *cluster* é recalculado, obtendo a média dos pontos atribuídos do próprio *cluster. De* 2(d) até 2(i) estão representadas etapas iterativas até o algoritmo convergir (Bishop, 2006).

Figura 2- Representação do funcionamento do algoritmo *k-means.*



Fonte: Bishop (2006).

3.3.5 PLS-DA - Análise Discriminante pelo Método de Quadrados Mínimos Parciais

A PLS-DA é um método quimiométrico supervisionado de classificação de padrões fundamentado na regressão por mínimos quadrados parciais (PLS). Essa técnica, como a regressão PLS apresenta uma matriz **X** que está relacionada a uma matriz **Y**, porém para PLS a matriz **Y** contém os valores da propriedade de interesse, enquanto a PLS-DA esta matriz contém informações sobre a classe de amostra. Cada classe é codificada por zero ou um, determinando se pertence ou não à classe. Na PLS-DA, as PCs do PCA que eram ortogonais, agora são conhecidas como variáveis latentes (LVs) e não são ortogonais (Bona, Março e Valderrama, 2018; Marquetti *et al.*, 2016).

3.3.6 Figuras de Mérito

Segundo Ferreira (2015), a exatidão de um modelo de regressão pode ser avaliada usando o erro médio quadrático de calibração (RMSEC) e de previsão (RMSEP) conforme as Equações (7) e (8),

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I - A}} \quad (7)$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I}} \quad (8)$$

onde y_i é o valor de referência, $\hat{y}_i$ é o valor previsto para a amostra *i*, I é o número de amostras utilizadas na validação cruzada e *A* é o número de variáveis latentes mais um, quando os dados são centrados na média.

Para modelos de classificação, além do RMSEC e RMSEP, é possível utilizar a área abaixo da curva (AUC) de característica de operação do receptor (ROC). Nessa curva é representada a taxa de verdadeiro positivo em relação à taxa de falso positivo para um limiar de decisão variável. A precisão é dada pela proporção de verdadeiros positivos e verdadeiros negativos entre o número total de casos examinados. A sensibilidade é a capacidade de classificar adequadamente as amostras previstas para estarem em uma classe como amostras pertencentes a essa classe. Já a especificidade é a capacidade de classificar as amostras previstas para estarem nas demais classes como amostras não pertencentes a classe em questão (Bona, Março e Valderrama, 2018). A sensibilidade e seletividade são definidas nas Equações (9) e (10), respectivamente (Brereton, 2018)

$$Sensibilidade = \frac{TP}{(TP + FN)} \quad (9)$$

$$Seletividade = \frac{TN}{(FP + TN)} \quad (10)$$

onde TP é a taxa de verdadeiro positivo; FP é a taxa de falso positivo, TN é a taxa de verdadeiro negativo e FN é a taxa de falso negativo.

4 MATERIAL E MÉTODOS

4.1 Material

As amostras de farinha de trigo foram cedidas pelo programa de Cereais de Inverno do Instituto Agronômico do Paraná (IAPAR-Londrina), onde os grãos foram moídos em moinho experimental e a textura do trigo fui atribuída em função da farinha de quebra. Foi coletado um total de 180 amostras de farinha de trigo de diferentes genótipos e provenientes de duas safras conforme descrito na Tabela 1.

Tabela 1. Genótipos, número total de amostras de cada genótipo e quantidade de amostra de cada safra.

Genótipo	Textura*	Total de Amostras	Locais de cultivo	Safra 2013	Safra 2014
LD122105	Macia	21	12	6	15
LD122206	Macia	28	15	11	17
LD121102	Dura	24	12	9	15
LD132210	Dura	26	14	8	18
LD131102	Dura	27	14	9	18
LD141103	Muito dura	18	12	0	18
LD141202	Semidura	18	13	0	18
LD142114	Dura	18	12	0	18

* Em função da farinha de quebra.

Fonte: Autoria própria.

Os genótipos foram cultivados em 17 municípios de 4 estados do Brasil. No estado do Paraná, os genótipos de trigo foram cultivados nos municípios de Cambará, Campo Mourão, Cascavel, Cruzmaltina, Guarapuava, Irati, Londrina, Mauá da Serra, Palotina, Pato Branco, Ponta Grossa e Warta. Em Santa Catarina foi cultivado em Campo Êre e Campos Novos. Em Mato Grosso do Sul foi cultivado em Maracajú e em São Paulo foi cultivado em Itaberá.

4.2 Métodos

4.2.1 Moagem experimental para extração da farinha

A moagem experimental foi realizada no moinho Chopin no IAPAR-Londrina, modelo CD1, após acondicionamento de 500g de trigo durante 16h para atingir a umidade de 15,5%, conforme método 26-10 descrito em AACC (1995). Os produtos de moagem obtidos são: farinha de quebra (FAQ), farinha de redução (FAR) e taxa de extração. FAQ e FAR foram misturadas para compor as amostras analisadas.

4.2.2 Coleta dos espectros em infravermelho próximo (NIRS).

Foram obtidos espectros nas regiões NIR das amostras de farinha de trigo em um espectrofotômetro NIRSystems 6500 (FossTecator AB, Höganäs, Suécia) por reflectância, no IAPAR-Londrina. As leituras foram feitas em temperatura ambiente, na faixa de comprimento de onda de 1100 a 2500 nm com 2 nm de resolução. Para aquisição dos espectros foi utilizado o software WinISI III versão 1,50e (FossNIRSystems/Tecator Infrasoft International, LLC, Silver Spring, MD, USA).

4.2.3 Pré-tratamento

As rotinas de cálculo foram realizadas com o software MATLAB (R2008b, The Mathworks, Inc., Natick, EUA). Após a aquisição dos espectros NIRS, os mesmos foram pré-tratados: correção do espalhamento multiplicativo (MSC) e segunda derivada (algoritmo de Savitzky-Golay, polinômio de quarto grau e janela de 25 pontos). Após o pré-tratamento foram aplicadas análises multivariadas lineares exploratórias e de agrupamento: PCA, HCA e *k-means*. Também foi aplicado PLS-DA para classificar as amostras de acordo com o genótipo do trigo.

4.2.3.1 Análise exploratória e de agrupamento

A PCA foi realizada na matriz de espectros (180 x 700) corrigidos pelo MSC e centrada na média. A matriz de covariância foi decomposta em autovalores e

autovetores usando o método de decomposição em valores singulares (SVD). Para confirmar os grupos sugeridos pela PCA, foi realizada uma análise hierárquica de agrupamento (HCA) usando os espectros médios de cada genótipo, o método de Ward e distância Euclidiana (Ferreira, 2015). Para avaliar o efeito responsável pela formação dos grupos, as amostras foram separadas em três grupos pelo método não supervisionado do *k-means* (Bishop, 2006), usando os espectros NIR como entrada.

4.2.3.2 Classificação dos genótipos usando o PLS-DA

Antes da construção dos modelos PLS-DA, as amostras foram separadas em um conjunto de calibração (67%) e previsão (33%) usando o algoritmo de Kennard-Stone (Westad e Marini, 2015).

Os modelos PLS-DA foram obtidos pelo algoritmo SIMPLS (Ferreira, 2015), usando como entrada os espectros NIR pré-tratados tanto com o MSC quanto MSC + 2ª derivada. O número de variáveis latentes de cada modelo foi selecionado com base nas figuras de mérito obtidas por meio de validação cruzada *leave-one-out*.

Ao estabelecer a quantidade de variáveis latentes para os espectros corrigidos pelo MSC e para os espectros corrigidos pelo MSC + 2ª derivada, o melhor pré-tratamento foi definido de acordo com as figuras de mérito de calibração e validação: RMSE, AUC e PCC (porcentagem de classificação correta).

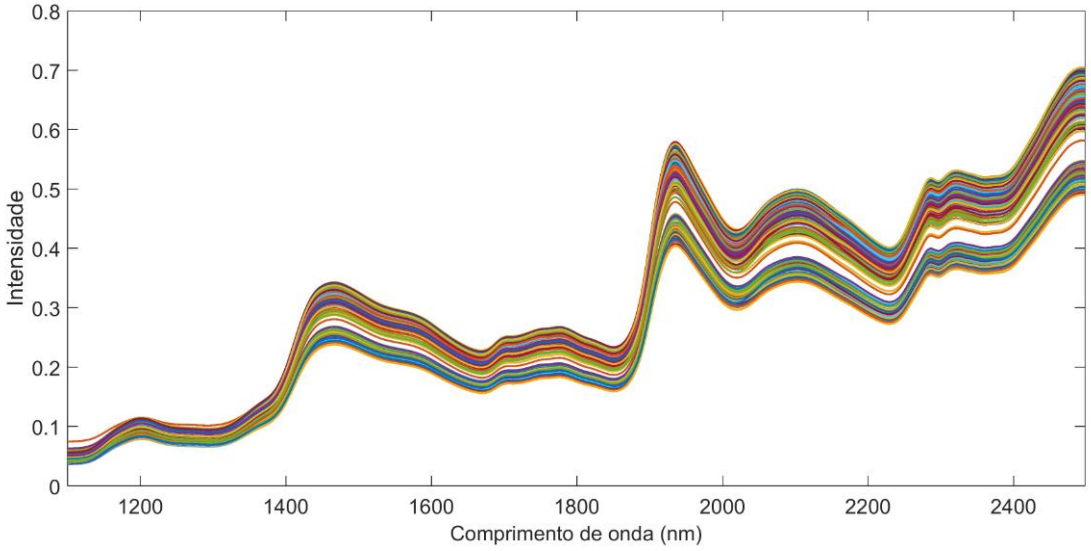
Após a definição do melhor pré-tratamento, foi avaliada a sensibilidade e seletividade do melhor modelo tanto para calibração quanto para previsão. Além disso, foram analisados os *scores* e *loadings* do melhor modelo para justificar a separação das classes pelo PLS-DA.

5 RESULTADOS E DISCUSSÃO

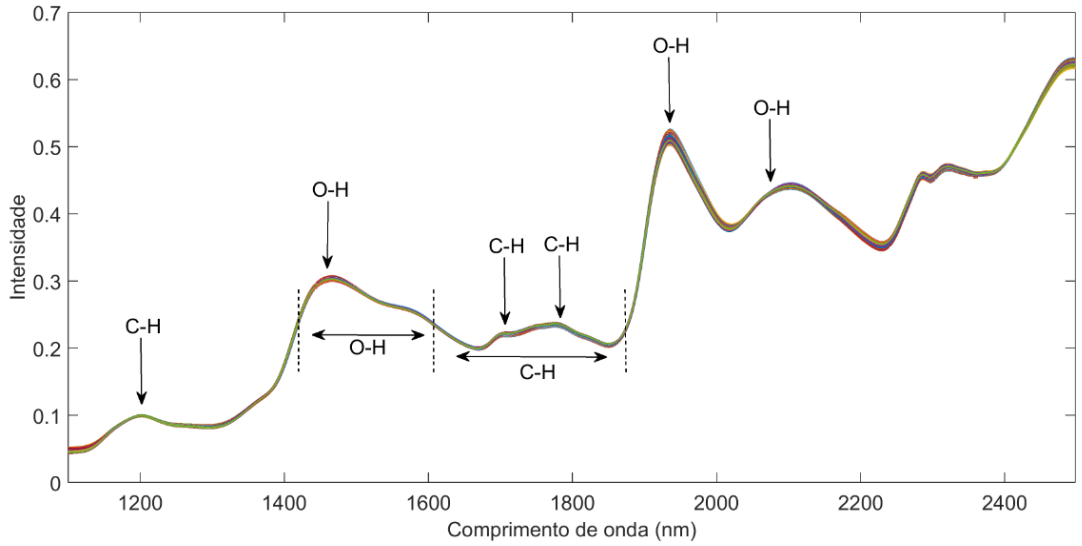
5.1 Pré-tratamento dos espectros

Os espectros NIR originais e pré-tratados, MSC e segunda derivada, de todas das amostras de farinha trigo estão apresentados na Figura 3.

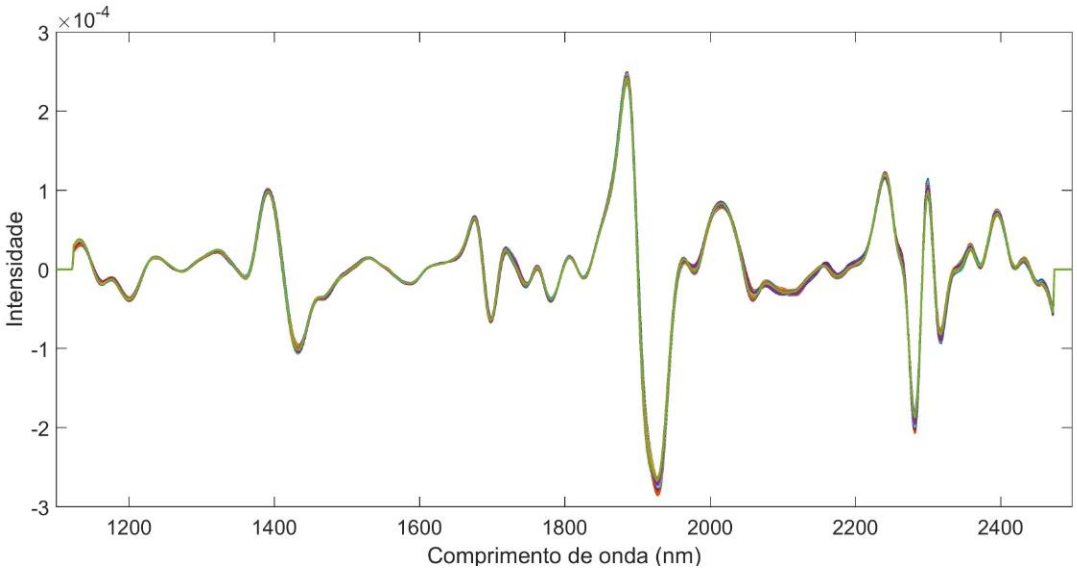
Figura 3- Espectros NIR das amostras de farinha de trigo. (a) originais, (b) pré-tratados com MSC e (c) pré-tratados com MSC + segunda derivada.



(a)



(b)



(c)

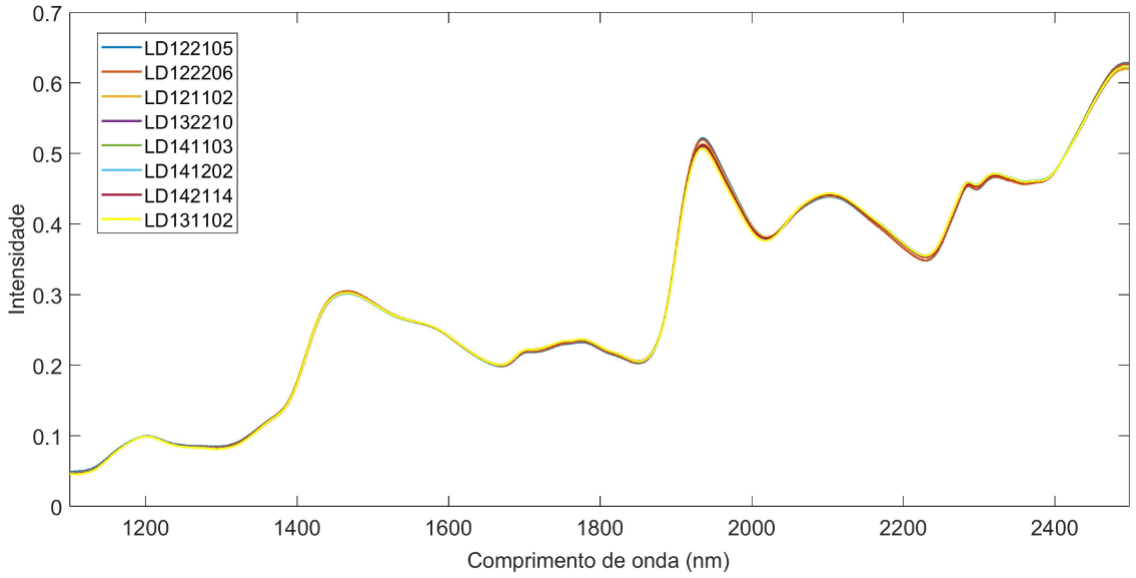
Fonte: Autoria própria.

Os espectros obtidos apresentam características semelhantes àquelas apresentadas em um trabalho que investigou a capacidade do método NIR em descrever mudanças físicas e químicas ocorridas durante a aglomeração úmida da farinha de trigo (Ait Kaddour e Cuq, 2009).

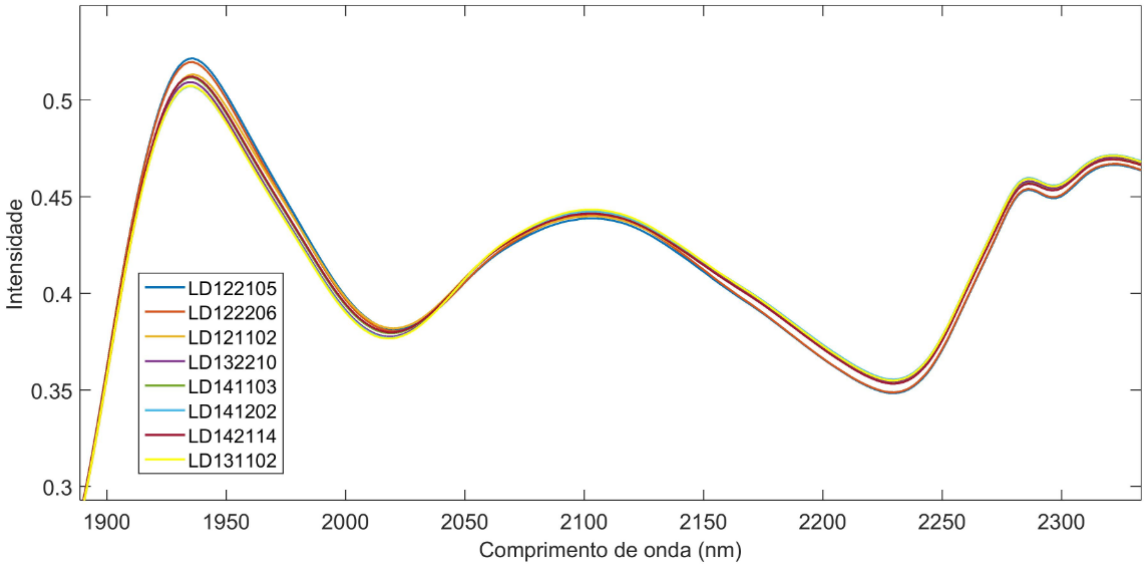
Foram observadas vibrações O-H nos comprimentos de onda de 1460 nm, 1934 nm e 2076 nm e vibrações C-H nos comprimentos de onda de 1202 nm, 1705 nm e 1781 nm. Duas bandas largas centradas em 1460 nm e 1934 nm e diferentes bandas pequenas centradas em 1202 nm, 1705 nm, 1781 nm e 2076 nm. Em estudo que traçou a origem de trigo e farinha chilenos através do NIR e quimiometria, foram observadas vibrações correspondentes aos comprimentos de onda de 1422 a 1608 nm que estão associadas à ligação O–H e uma banda na região de 1608 a 1878 nm que corresponde à ligação C–H (González-Martín *et al.*, 2014).

Os espectros médios das farinhas dos genótipos de trigo, pré-tratados com MSC estão apresentados na Figura 4.

Figura 4 - Espectros médios das amostras de farinha de trigo em relação aos genótipos: (a) espectro completo e (b) destaque para a região de 1900 a 2400nm.



(a)



(b)

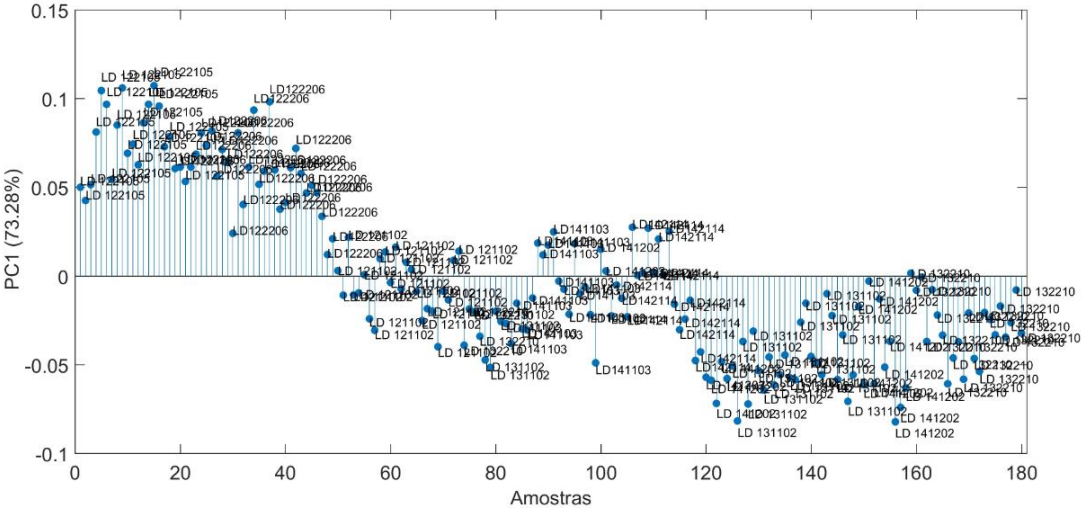
Fonte: Autoria própria.

Os espectros médios das 180 amostras de farinha de trigo apresentam semelhanças evidentes, mas foi possível observar, Figura 4(b), intensidades diferentes entre alguns genótipos nas regiões de 1920 a 1960 nm e 2150 a 2350 nm do espectro NIR.

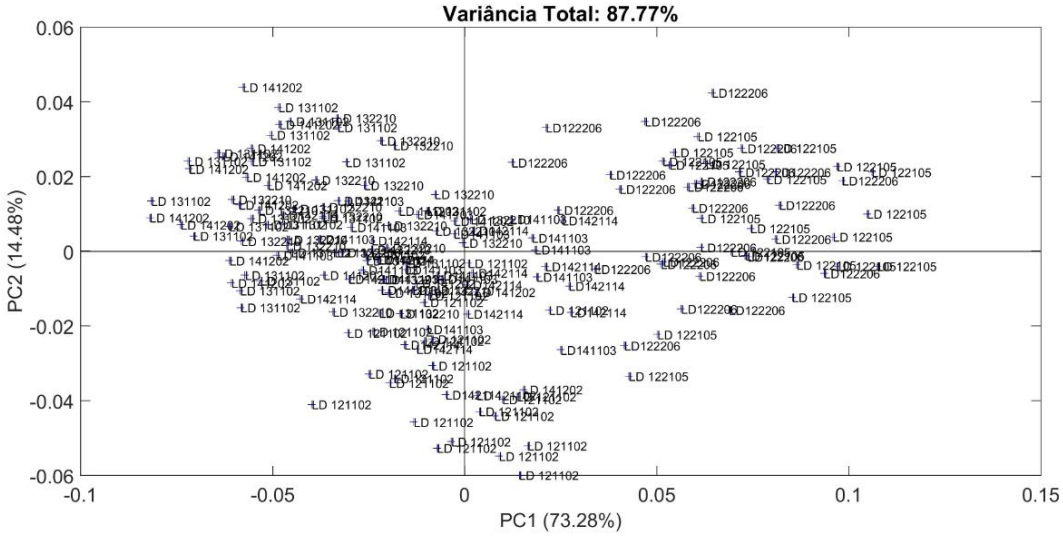
5.2 Análise de Componentes Principais – PCA

O gráfico de *scores* da PCA realizada nos espectros na região do infravermelho próximo, pré-tratadas com MSC está apresentado na Figura 5.

Figura 5 – Gráficos dos *scores* da PCA realizada nos espectros na região do infravermelho próximo, pré-tratadas com MSC:(a) PC1; (b) PC1 e PC2.



(a)



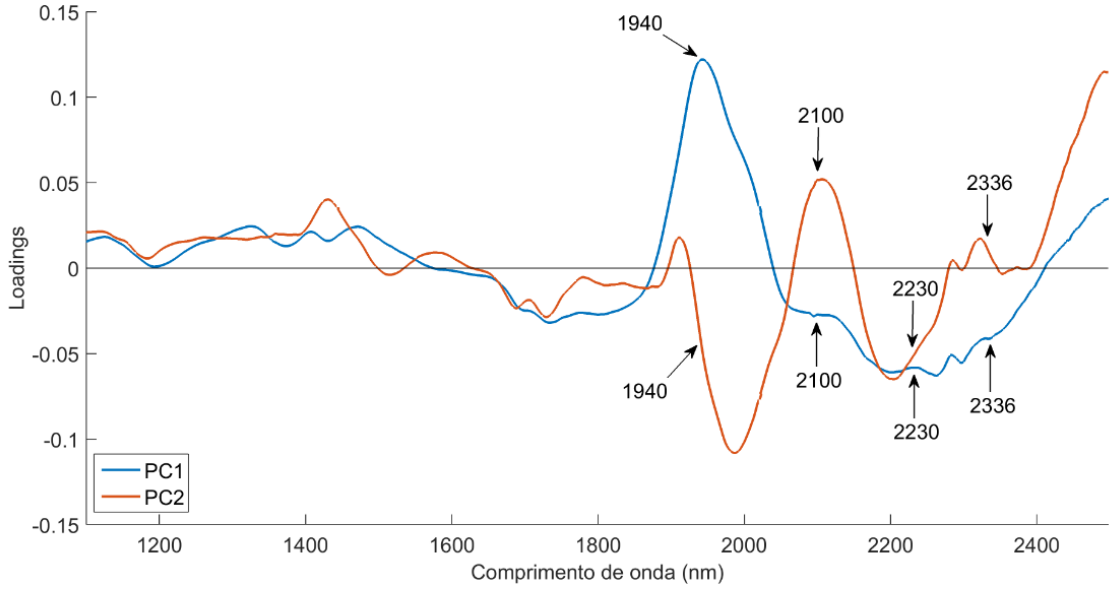
(b)

Fonte: Autoria própria.

A PCA, Figura 5 sugere a presença de três grupos, os genótipos LD122105 e LD122206 (quadrante positivo de PC1), os genótipos LD141103, LD142114 e LD121102 (distribuídos nos quadrantes positivo e negativo de PC1 e quadrante negativo da PC2) e os genótipos LD132210, LD131102 e LD141202 (quadrante negativo de PC1).

Por meio dos gráficos de *scores* da Figura 5, nota-se que as amostras dos genótipos LD122105 e LD122206 apresentam maior similaridade e foram agrupadas no quadrante positivo da PC1. Observando os espectros da Figura 4(b) e os *loadings* da PCA para PC1 e PC2 (Figura 6), nota-se que estes mesmos genótipos se separam dos demais nas bandas de 1940 nm, 2230nm e 2340nm.

Figura 6- Gráfico de *loadings* da PCA para PC1 e PC2.



Fonte: Autoria própria.

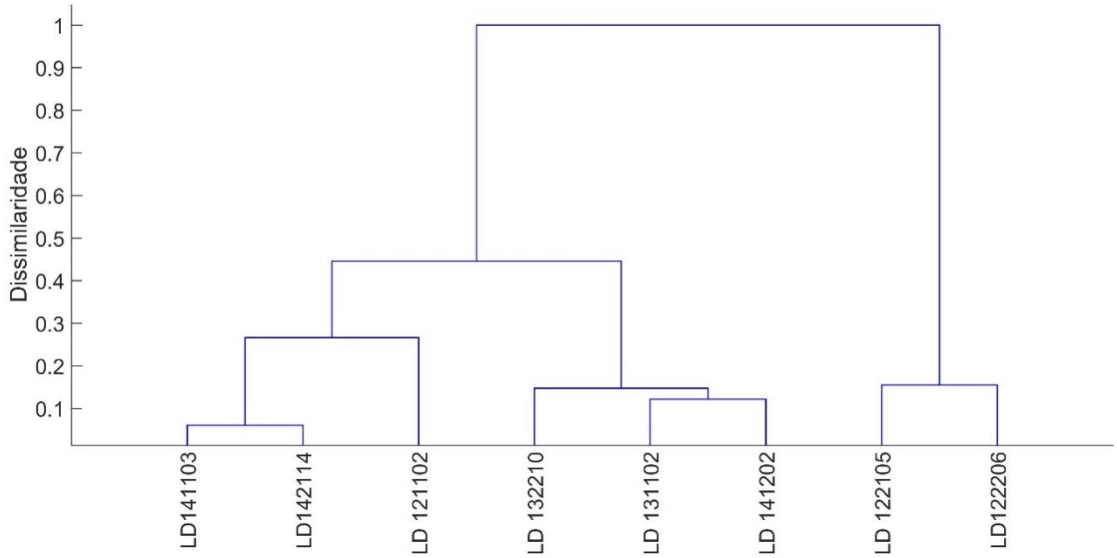
Segundo Burns & Ciurczak (2007), em 1940 nm encontra-se o ponto máximo da banda da umidade da farinha de trigo e esta não é sobreposta pelas bandas de outros constituintes da farinha. Em 2100 nm o amido tem uma relevante banda de absorção. A região de 2230 nm corresponde a uma banda de absorção do glúten de trigo, que representam as proteínas das farinhas. Em 2336 nm observa-se uma banda discreta característica da celulose, que é um dos principais componentes do farelo que se sobrepõe a banda do amido. Assim, as medidas NIR podem ser associadas ao teor de farelo da farinha e, consequentemente correlacionadas com o teor de cinzas.

Com isso, podemos observar que os genótipos LD122105 e LD122206 (textura macia) se separam das demais por terem intensidades diferentes nas bandas características de proteínas, umidade e farelo. No trigo de textura macia são encontradas altas concentrações de proteínas ricas em aminoácido triptofano, enquanto que em trigo de textura dura estas proteínas estão em baixas concentrações (Pasha, Anjum e Morris 2001). Na banda característica da umidade da farinha, os

genótipos LD122105 e LD122206 apresentaram maior intensidade, podendo ter mais umidade que as demais classes. Porém, na banda característica do farelo, esses genótipos apresentaram menor intensidade que as demais classes, já foi relatado na literatura que a dureza do grão tem correlação positiva com teor de cinzas na farinha (Cox *et al.*, 2001).

Na Figura 7 está representado o dendrograma obtido na HCA. Este dendrograma confirma os mesmos grupos indicados pela PCA. O grupo dos genótipos LD141103 (muito dura), LD142114 (dura) e LD121102 (dura) apresenta uma similaridade de 73,35% e uma semelhança de 93,91% entre os genótipos LD141103 e LD142114. Já o grupo dos genótipos LD132210 (dura), LD131102 (dura) e LD141202 (semidura) apresentou uma similaridade de 85,23% e o grupo dos genótipos LD122105 (macia) e LD122206 (macia) uma similaridade de 84,47%. A HCA também confirma que os genótipos LD122105 e LD122206 apresentam um comportamento mais distinto em relação aos demais genótipos analisados.

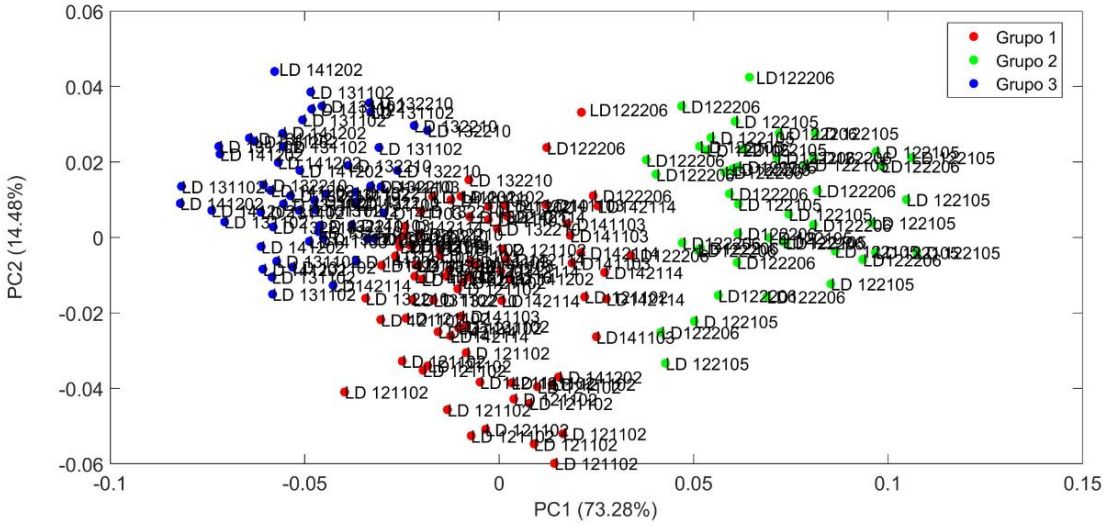
Figura 7 – Dendrograma para os espectros médios de cada genótipo.



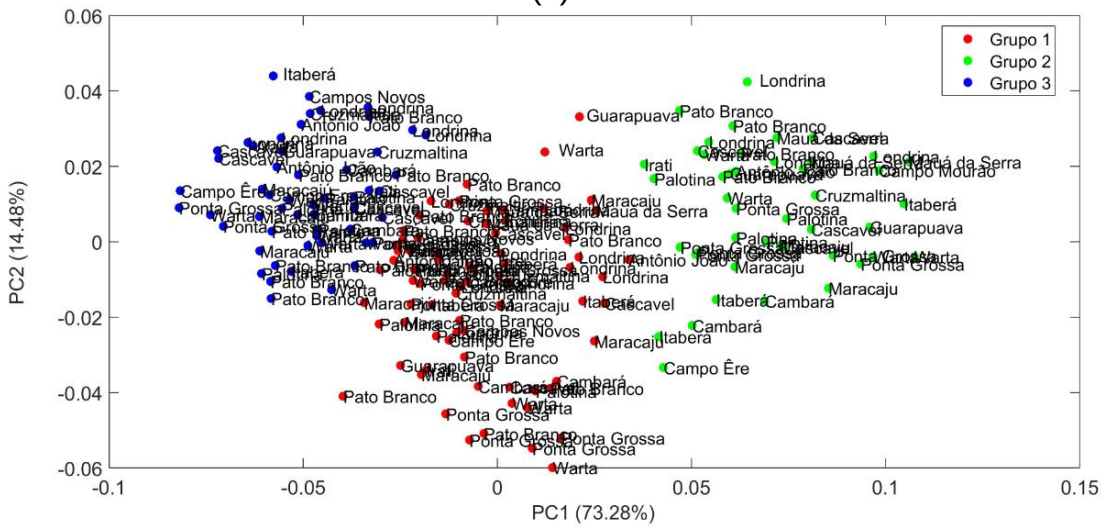
Fonte: Autoria própria.

O resultado do agrupamento do *k-means* para as amostras foi apresentado no eixo da PC1 e PC2 conforme a Figura 8. Além disso, as amostras foram associadas com o respectivo genótipo, Figura 8(a), cidade de cultivo, Figura 8(b), safra, Figura 8(c) e textura Figura 8(d).

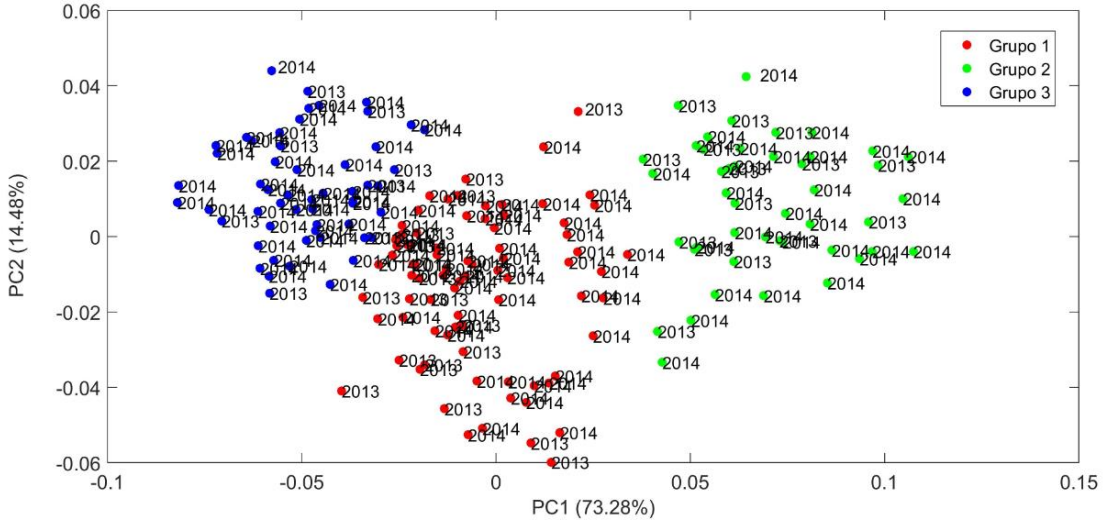
Figura 8 – Representação das amostras nos eixos da PC1 e PC2 de acordo com os grupos formados pelo *k-means* (a) Genótipo, (b) cidade de cultivo, (c) safra e (d) textura.



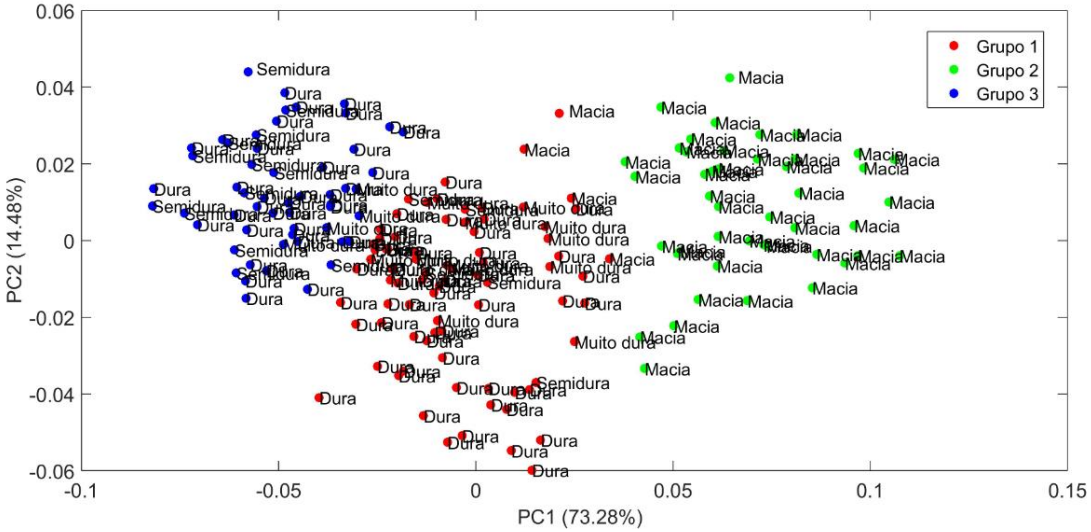
(a)



(b)



(c)



(d)

Fonte: Autoria própria.

A análise da Figura 8 confirma que os grupos são formados pelos genótipos como já indicados na PCA e HCA. A textura também influencia na formação dos grupos, o grupo 2 é formado exclusivamente por amostras de textura macia, apenas 4 amostras de textura macia foram classificadas como grupo 1. As amostras de textura semiduras em sua maioria estão no grupo 3, exceto 5 amostras que foram classificadas como pertencentes ao grupo 1. O local de cultivo e a safra não apresentam um efeito importante na formação desses grupos, evidenciando maior efeito das características genéticas na separação dos grupos em função dos espectros de NIR.

5.3 PLS-DA

Na Tabela 2 estão apresentadas as proporções por classe (genótipos) das amostras dos conjuntos de calibração e previsão separadas pelo algoritmo de Kennard-Stone.

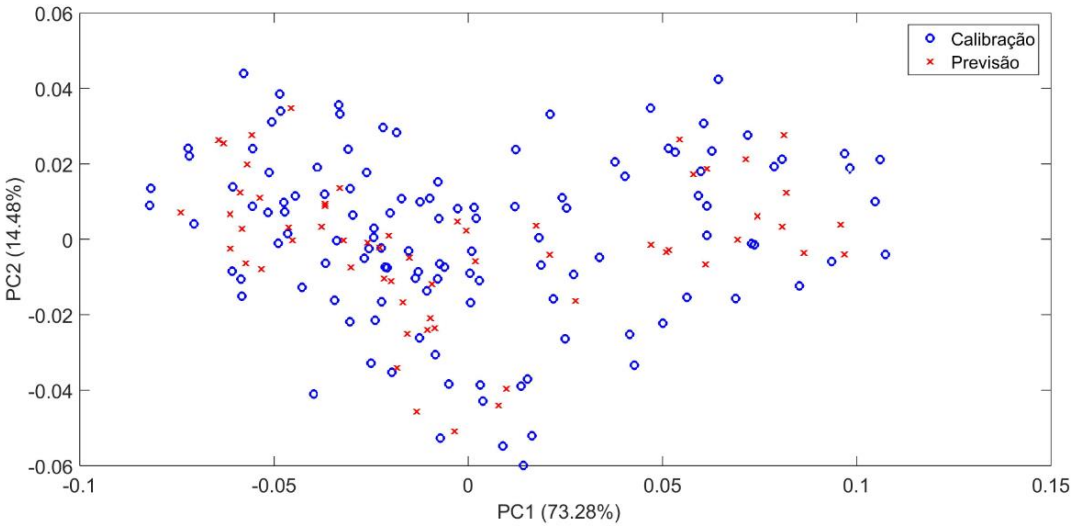
Tabela 2. Tabela dos percentuais de amostras para calibração e previsão de acordo com o Kennard-Stone.

Genótipo	Amostras		
	Calibração	Previsão	Total
LD122105	14 (66,7%)	7 (33,3%)	21 (100%)
LD122206	19 (67,9%)	9 (32,1%)	28 (100%)
LD121102	16 (66,7%)	8 (33,3%)	24 (100%)
LD132210	17 (65,4%)	9 (34,6%)	26 (100%)
LD131102	18 (66,7%)	9 (33,3%)	27 (100%)
LD141103	12 (66,7%)	6 (33,3%)	18 (100%)
LD141202	12 (66,7%)	6 (33,3%)	18 (100%)
LD142114	12 (66,7%)	6 (33,3%)	18 (100%)

Fonte: Autoria própria.

As amostras de calibração e previsão representadas nos eixos da PC1 e PC2 estão na Figura 9.

Figura 9 - Amostras de calibração e previsão nos eixos da PC1 e PC2.



Fonte: Autoria própria.

Modelos PLS-DA foram desenvolvidos para cada pré-tratamento dos espectros para discriminar as amostras de farinha de trigo por genótipo, conforme apresentado na Tabela 3.

Tabela 3 - Resultados PLS-DA para amostras de farinha de trigo por genótipo. O melhor resultado está destacado em negrito.

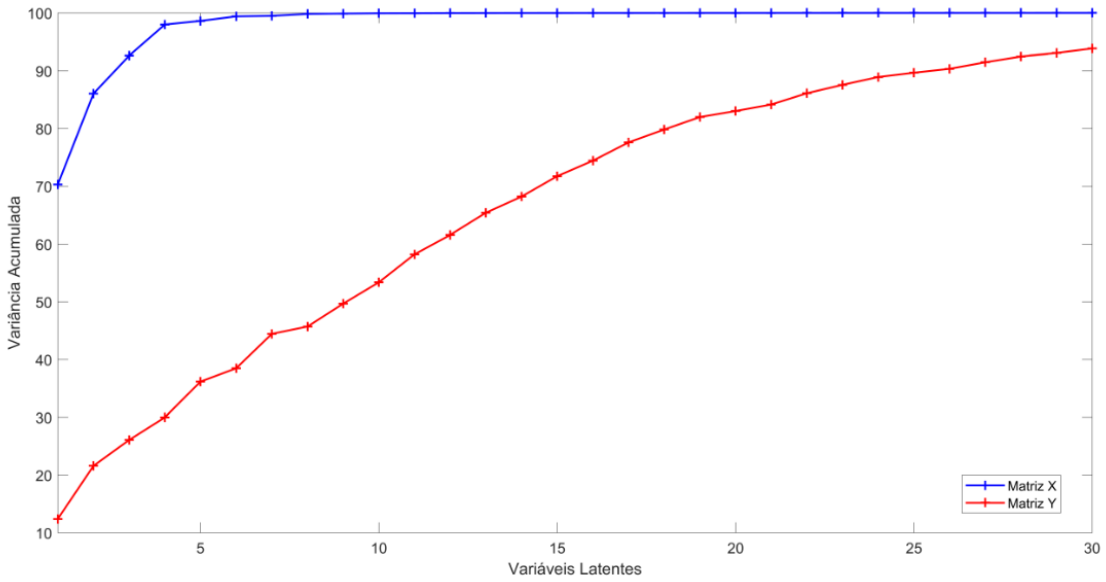
Pré-tratamento	LV	Calibração			Previsão		
		RMSE	AUC	PCC(%)	RMSE	AUC	PCC(%)
MSC	17	0,1563	0,9988	99,17	0,1682	0,9951	95,00
	24	0,1099	1,0000	100,00	0,1539	0,9977	95,00
	30	**0,0820**	**1,0000**	**100,00**	**0,1580**	**0,9973**	**96,67**
	35	0,0652	1,0000	100,00	0,1648	0,9961	96,67
MSC + 2ª derivada	15	0,1564	0,9963	95,04	0,2023	0,9756	88,14
	21	0,1234	0,9966	99,17	0,1828	0,9967	91,53
	31	0,0863	1,0000	100,00	0,1844	0,9883	93,22
	42	0,0594	1,0000	100,00	0,1824	0,9917	93,22

RMSE: raiz quadrada do erro quadrático médio; AUC: área abaixo da curva ROC; PCC: porcentagem de classificação correta.

Fonte: Autoria própria.

Ao analisar os resultados da Tabela 3, observa-se que o pré-tratamento MSC apresentou o melhor modelo PLS-DA, com 30 variáveis latentes que apresenta 99,99% da variância acumulada em **X** e 93,82% em **Y,** conforme Figura10, onde a variância capturada da matriz **X** aumentou rapidamente e para a matriz **Y** foram necessárias 30 LVs para aumentar a variância representada pelo modelo.

Figura 10 – Gráfico das variâncias na matriz **X** e **Y**



Fonte: Autoria própria.

Este modelo foi escolhido por apresentar a combinação dos menores valores de RMSE de calibração e previsão para cada classe e valores mais altos de

sensibilidade, especificidade, AUC, PCC tanto para calibração quanto para previsão, preferindo menor quantidade de variáveis latentes.

Os valores de sensibilidade e seletividade das classes (genótipos) do modelo escolhido estão apresentados na Tabela 4, apresentando valores médios de 95,83% para a sensibilidade e 99,53% para a seletividade e representa um modelo satisfatório dos espectros NIR.

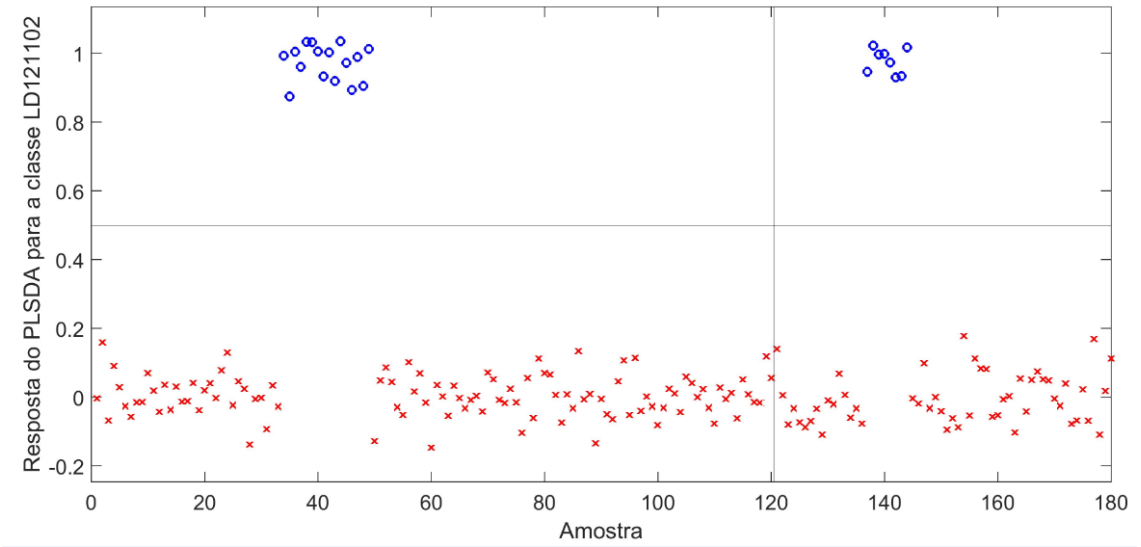
Tabela 4 – Melhor resultado da PLS-DA (MSC com 30 LV) e os valores de sensibilidade e seletividade da calibração e previsão do modelo.

Classe	Calibração		Previsão	
	Sensibilidade (%)	Seletividade (%)	Sensibilidade (%)	Seletividade (%)
LD122105	100,00	100,00	100,00	100,00
LD122206	100,00	100,00	100,00	100,00
LD121102	100,00	100,00	100,00	100,00
LD132210	100,00	100,00	100,00	100,00
LD131102	100,00	100,00	100,00	100,00
LD141103	100,00	100,00	83,33	100,00
LD141202	100,00	100,00	83,33	98,15
LD142114	100,00	100,00	100,00	98,15

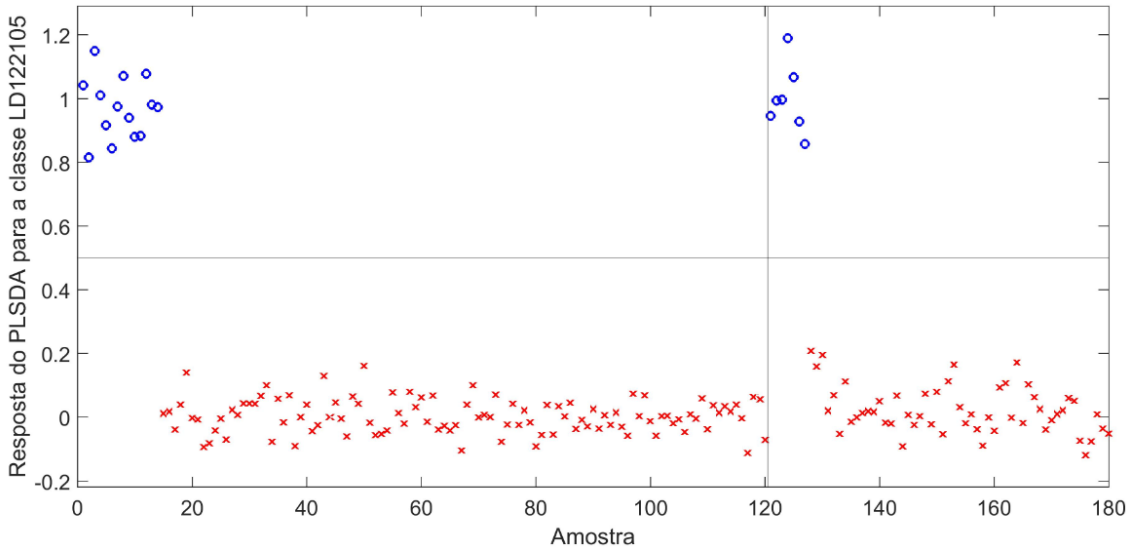
Fonte: Autoria própria.

As respostas dos modelos PLS-DA escolhidos para cada uma das classes (genótipos) estão apresentadas na Figura 11.

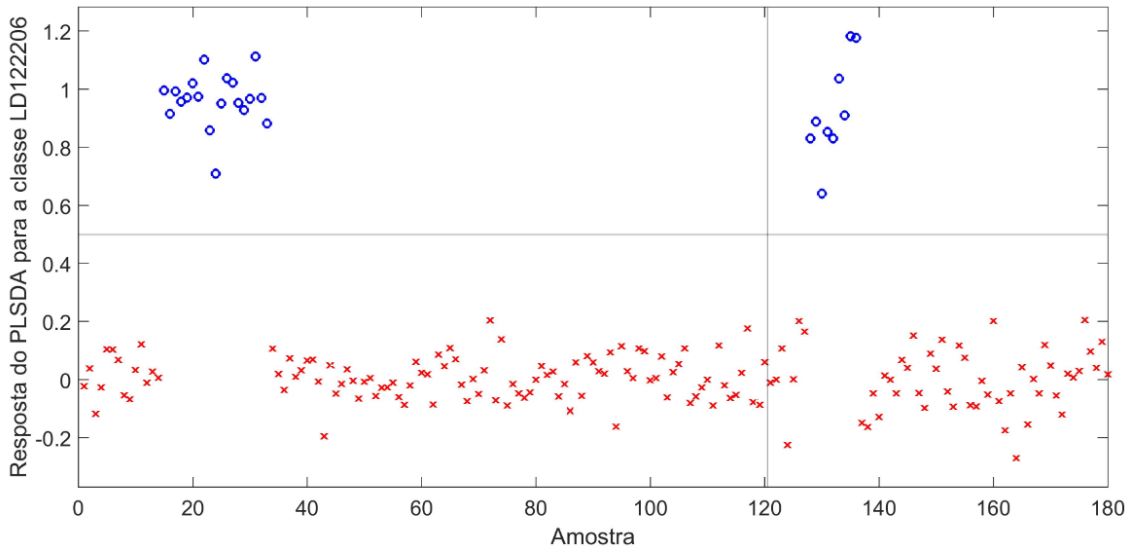
Figura 11 – Respostas do modelo PLS-DA para cada uma das classes (genótipos), onde a linha horizontal determina o limiar e a linha vertical delimita a calibração e a previsão.



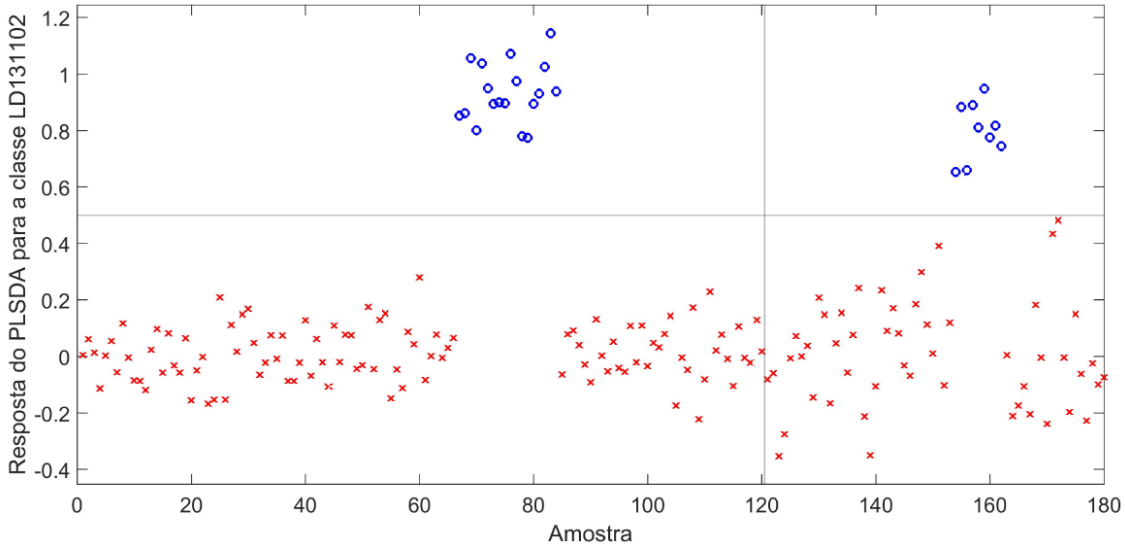
(a)



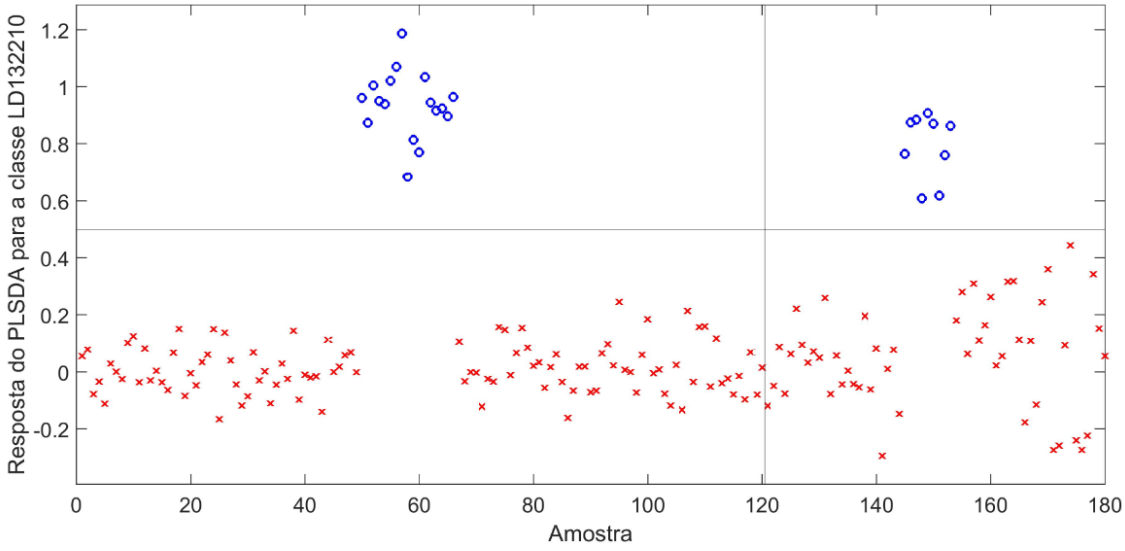
(b)



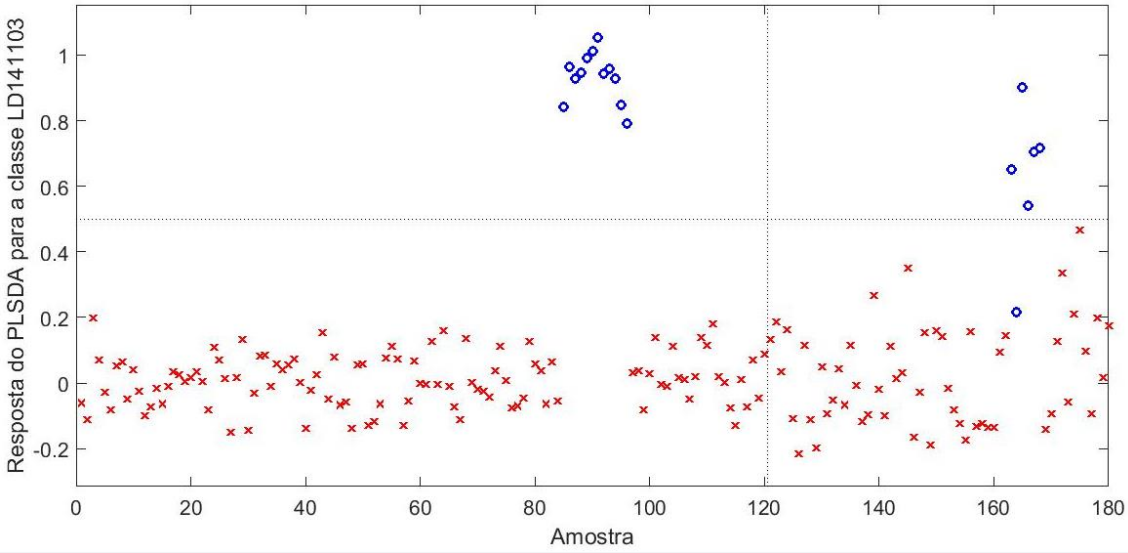
(c)


Resposta do PLSDA para a classe LD131102
Amostra

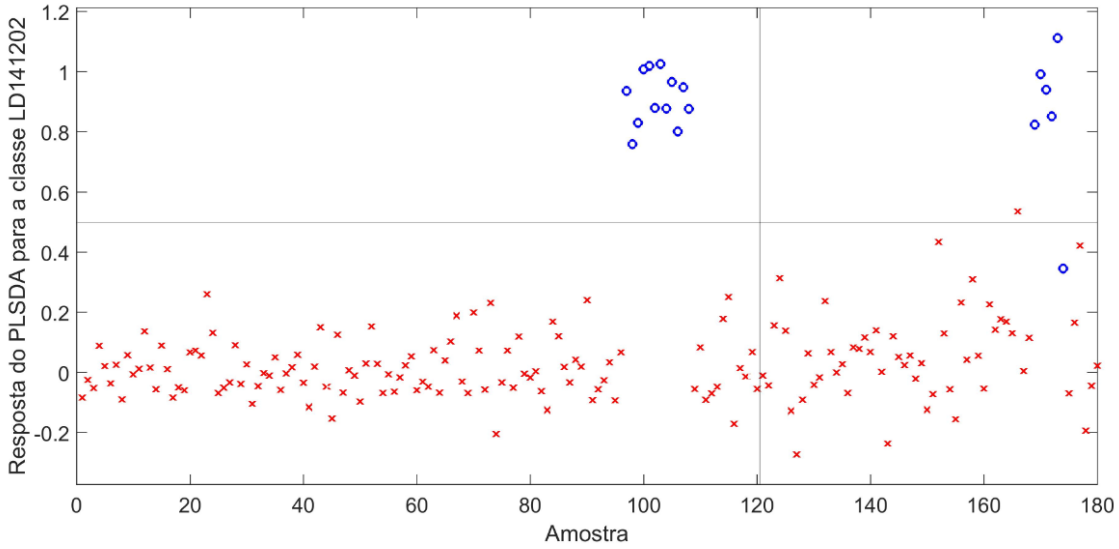
(d)


Resposta do PLSDA para a classe LD132210
Amostra

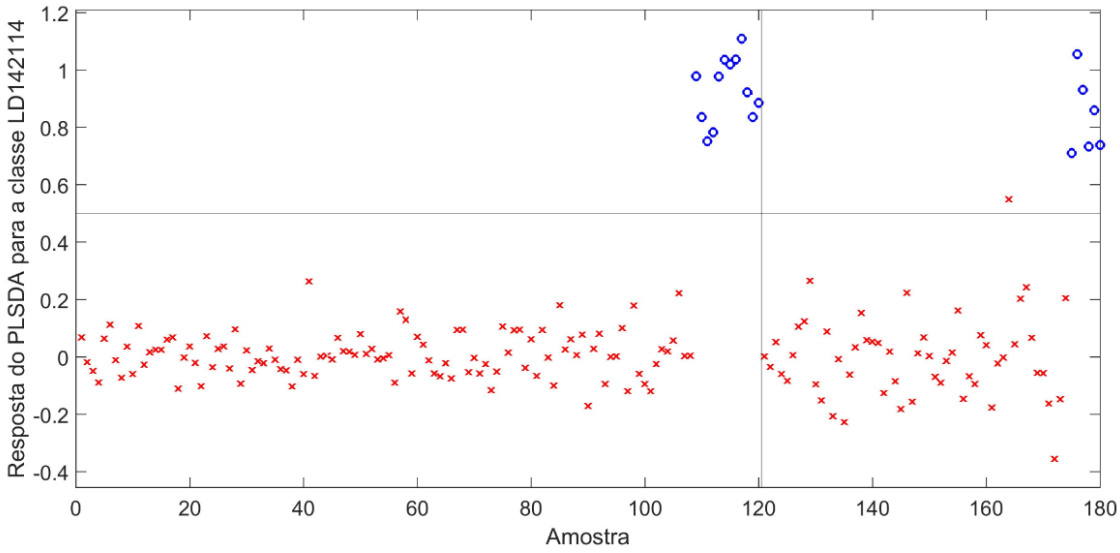
(e)


Resposta do PLSDA para a classe LD141103
Amostra

(f)



(g)



(h)

Fonte: Autoria própria.

Nas respostas da PLS-DA apresentadas na Figura 11(f) nota-se que há apenas uma amostra da classe LD141103 que é caracterizada como não sendo desta classe, indicando um falso negativo.

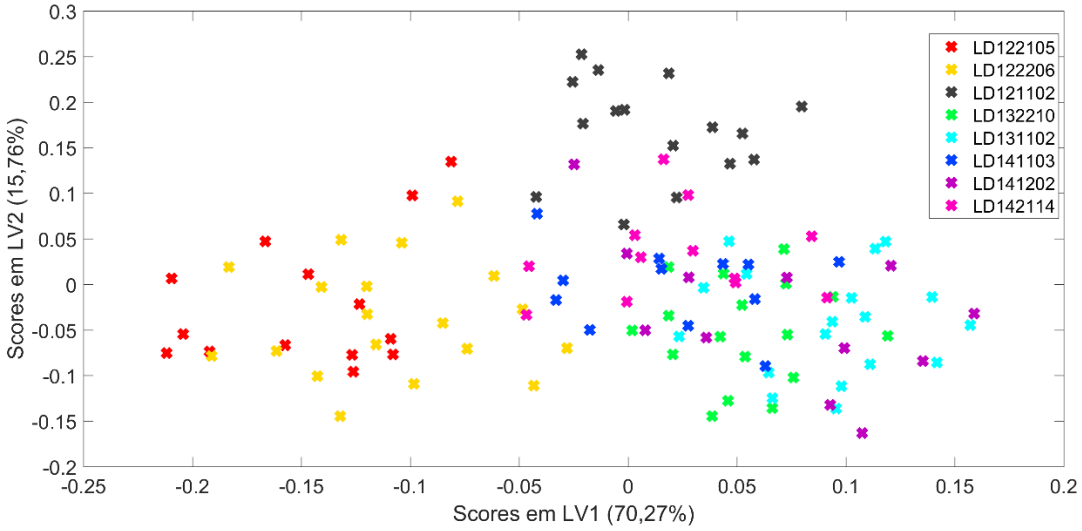
Entre as amostras do genótipo LD141202, conforme a Figura 11(g), uma se caracteriza como não sendo desta classe e uma que não tem as características do genótipo, portanto tem um falso negativo e um falso positivo. Na classe do genótipo LD142114, descrito na Figura 11(h) encontra-se uma amostra que não é da classe classificada como pertencente a esta classe de genótipo, ou seja, um falso positivo.

Cinco classes (genótipos) foram separadas totalmente das demais, com 100% de sensibilidade e 100% seletividade e três classes não foram totalmente separadas,

porém apresentam uma boa classificação. Com isso, o modelo PLS-DA obtido pode ser utilizado para a classificação das amostras de farinha de trigo por genótipos.

O gráfico de *scores* para o melhor modelo PLS-DA é apresentado na Figura 12, onde indica uma separação entre as amostras por genótipos.

Figura 12 – Gráfico de *scores* para o melhor modelo PLS-DA para todas as classes



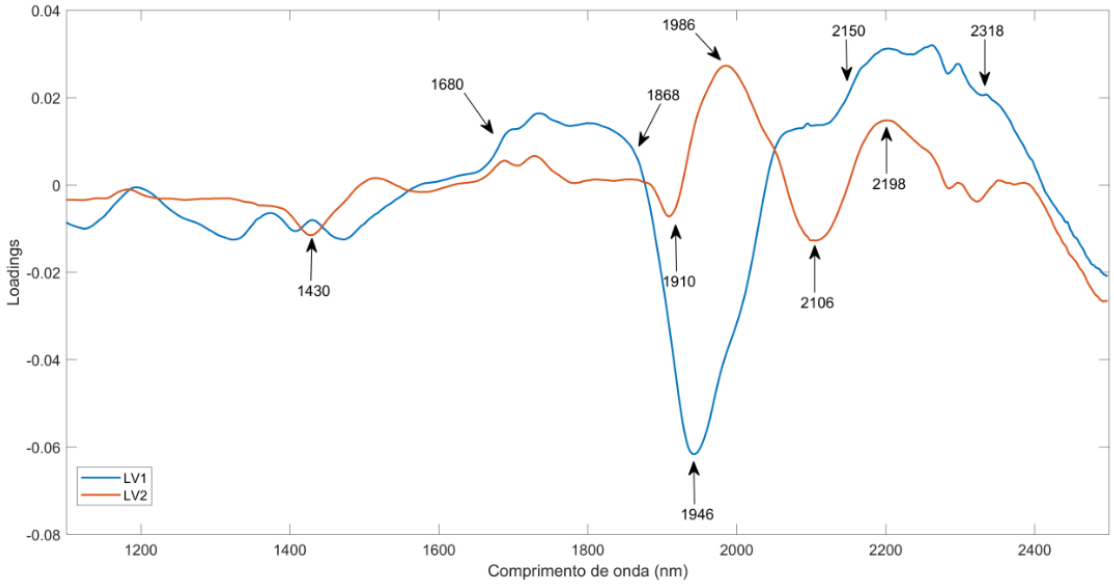
Fonte: Autoria própria.

As amostras LD122105 e LD122206 se separam claramente na região negativa de LV1, enquanto as amostras LD121102 foram discriminadas no quadrante positivo de LV2. As amostras LD132210 e LD131102 foram discriminadas no quadrante positivo de LV1 e no quadrante negativo de LV2.

Na Figura 13, o gráfico de *loadings* da PLS-DA é apresentado. No gráfico de *loadings* observa-se que o pico em 1946 nm contribuiu para a classificação das amostras das classes LD122105 e LD122206, devido aos valores negativos de LV1.

Para as amostras da classe LD121102 os picos 1986 e 2198 nm foram importantes para a distinção dessa classe, pois possuem valores positivos para LV2. As regiões 1680 a 1868 nm e 2150 e 2318 nm representam valores positivos para LV1 e os picos 1430, 1910 e 2106 nm representam os valores negativos para LV2 das classes LD132210 e LD131102.

Figura 13 – Gráfico de *loadings* da PLS-DA.



Fonte: Autoria própria.

6 CONCLUSÕES

Através da análise exploratória dos espectros NIR foi observada uma separação dos oito genótipos de trigo em três grupos. Esses genótipos se separaram principalmente nas bandas características da umidade, proteína e presença de farelo, apresentando variabilidade no teor destes componentes.

O emprego da técnica de espectroscopia de infravermelho próximo em conjunto com a quimiometria indicou que os grupos são segmentados principalmente pelas características genéticas das amostras de trigo. A textura também influenciou na formação dos grupos, o grupo formado exclusivamente por amostras de textura macia se separou das demais (semidura, dura e muito dura) em todas técnicas empregadas. Local de cultivo e a safra não apresentaram efeito no agrupamento das amostras.

A PLS-DA classificou assertivamente as amostras com valores médios de 95,83% para a sensibilidade e 99,53% para a seletividade, reafirmando o resultado da análise exploratória. Assim, conclui-se que foi possível classificar as amostras de trigo de diferentes safras e locais de cultivo através dos espectros NIR combinados com métodos quimiométricos.

7 REFERÊNCIAS

AHMAD, M. H.; NACHE, M.; WAFFENSCHMIDT, S.; HITZMANN, B. A fluorescence spectroscopic approach to predict analytical, rheological and baking parameters of wheat flours using chemometrics. **Journal of Food Engineering**, v. 182, p. 65–71, 2016.

AIT KADDOUR, A.; CUQ, B. In line monitoring of wet agglomeration of wheat flour using near infrared spectroscopy. **Powder Technology**, v. 190, n. 1–2, p. 10–18, 2009.

BARDINI, G.; BOUKID, F.; CARINI, E.; CURTI, E.; PIZZIGALLI, E.; VITTADINI, E. Enhancing dough-making rheological performance of wheat flour by transglutaminase and vital gluten supplementation. **LWT**, v. 91, p. 467–476, 1 maio 2018.

BENINCASA, P.; DOMINICI, F.; BOCCI, L.; GOVERNATORI, C.; PANFILI, I.; TOSTI, G.; TORRE, L.; PUGLIA, D. Relationships between wheat flour baking properties and tensile characteristics of derived thermoplastic films. **Industrial Crops and Products**, v. 100, p. 138–145, 2017.

BISHOP, C. M. **Pattern recognition and machine learning**. 1. ed. New York: Springer, 2006.

BONA, E.; MARÇO, P. H.; VALDERRAMA, P. Chemometrics Applied to Food Control. **Food Control and Biosecurity**, p. 105–133, 1 jan. 2018.

BRERETON, R. G. **Chemometrics: Data Driven Extraction for Science**. 2. ed. Hoboken, NJ, USA: John Wiley & Sons, 2018.

BURNS, D. A.; CIURCZAK, E. W. **Handbook of near-infrared analysis**. 3. ed. Boca Raton, FL: CRC Press: Taylor & Francis Group, 2007.

CHEN, J.; ZHU, S.; ZHAO, G. Rapid determination of total protein and wet gluten in commercial wheat flour using siSVR-NIR. **Food Chemistry**, v. 221, p. 1939–1946, 2017.

CHOY, A.; WALKER, C. K.; PANOZZO, J. F. An investigation of wheat milling yield based on grain hardness parameters Investigation of Wheat Milling Yield Based on Grain Hardness Parameters. n. November, 2015.

CORREIA, F. O.; SILVA, D. S.; COSTA, S. S. L.; SILVA, I. K. V.; SILVA, D. R. DA; ALVES, J. DO P. H.; GARCIA, C. A. B.; MARANHÃO, T. DE A.; PASSOS, E. A.; ARAUJO, R. G. O. Optimization of microwave digestion and inductively coupled

plasma-based methods to characterize cassava, corn and wheat flours using chemometrics. **Microchemical Journal**, v. 135, p. 190–198, 2017.

COX, T. S.; CLUSTER, D. M. R.; MARTIN, J. M.; FROHBERG, R. C.; MORRIS, C. F.; TALBERT, L. E.; GIROUX, M. J. Milling and Bread Baking Traits Associated with Puroindoline Sequence Type in Hard Red Spring Wheat. v. 41, n. February, p. 228–234, 2001.

DOBRASZCZYK, B. J.; MORGENSTERN, M. P. Rheology and the breadmaking process. **Journal of Cereal Science**, v. 38, n. 3, p. 229–245, 2003.

FAO. **No Title**. Disponível em: <http://www.fao.org/worldfoodsituation/csdb/en/>.

FERRÃO, M. F.; CARVALHO, C. W.; MÜLLER, E. I.; DAVANZO, C. U. Determinação simultânea dos teores de cinza e proteína em farinha de trigo empregando NIRR-PLS e DRIFT-PLS. **Ciência e Tecnologia de Alimentos**, v. 24, n. 3, p. 333–340, 2004.

FERREIRA, M. M. C. **Quimiometria: Conceitos, métodos e Aplicações**. Campinas, SP: Editora da Unicamp, 2015.

GONZÁLEZ-MARTÍN, M. I.; WELLS MONCADA, G.; GONZÁLEZ-PÉREZ, C.; ZAPATA SAN MARTÍN, N.; LÓPEZ-GONZÁLEZ, F.; LOBOS ORTEGA, I.; HERNÁNDEZ-HIERRO, J. M. Chilean flour and wheat grain: Tracing their origin using near infrared spectroscopy and chemometrics. **Food Chemistry**, v. 145, p. 802–806, 2014.

GRANATO, D.; SANTOS, J. S.; ESCHER, G. B.; FERREIRA, B. L.; MAGGIO, R. M. Trends in Food Science & Technology Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods : A critical perspective. **Trends in Food Science & Technology**, v. 72, n. December 2017, p. 83–90, 2018.

GUO, P.; YU, J.; COPELAND, L.; WANG, S.; WANG, S. Mechanisms of starch gelatinization during heating of wheat flour and its effect on in vitro starch digestibility. **Food Hydrocolloids**, v. 82, p. 370–378, 1 set. 2018.

HU, Y.; WANG, L.; LI, Z. Modi fi cation of protein structure and dough rheological properties of wheat fl our through superheated steam treatment. **Journal of Cereal Science**, v. 76, p. 222–228, 2017.

LANCELOT, E.; BERTRAND, D.; HANAFI, M.; JAILLAIS, B. Near-infrared hyperspectral imaging for following imbibition of single wheat kernel sections.

Vibrational Spectroscopy, v. 92, p. 46–53, 2017.

LI VIGNI, M.; BASCHIERI, C.; MARCHETTI, A.; COCCHI, M. RP-HPLC and chemometrics for wheat flour protein characterisation in an industrial bread-making process monitoring context. **Food Chemistry**, v. 139, n. 1–4, p. 553–562, 2013.

MARQUETTI, I.; LINK, J. V.; LEMES, A. L. G.; SCHOLZ, M. B. DOS S.; VALDERRAMA, P.; BONA, E. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. **Computers and Electronics in Agriculture**, v. 121, p. 313–319, 2016.

MISRA, N. N.; KAUR, S.; TIWARI, B. K.; KAUR, A.; SINGH, N.; CULLEN, P. J. Food Hydrocolloids Atmospheric pressure cold plasma (ACP) treatment of wheat fl our. **Food hydrocolloids**, v. 44, p. 115–121, 2015.

MORRIS, C. F. Puroindolines : the molecular genetic basis of wheat grain hardness. p. 633–647, 2002.

NETO, B. D. B.; SCARMINIO, I. S.; BRUNS, R. E. 25 Anos de quimiometria no Brasil. **Quimica Nova**, v. 29, n. 6, p. 1401–1406, 2006.

OSBORNE, B. G. Principles and practice of near-infrared (NIR) reflectance analysis. **Journal of Food Technology**, v. 16, p. 13–19, 1981.

PASHA, I.; ANJUM, F. M.; MORRIS, C. F. Grain Hardness : A Major Determinant of Wheat Quality. 2001.

RANZAN, C.; STROHM, A.; RANZAN, L.; TRIERWEILER, L. F.; HITZMANN, B.; TRIERWEILER, J. O. Wheat flour characterization using NIR and spectral filter based on ant colony optimization. **Chemometrics and Intelligent Laboratory Systems**, v. 132, p. 133–140, 2014.

SHI, H.; YU, P. Comparison of grating-based near-infrared (NIR) and Fourier transform mid-infrared (ATR-FT/MIR) spectroscopy based on spectral preprocessing and wavelength selection for the determination of crude protein and moisture content in wheat. **Food Control**, v. 82, p. 57–65, 2017.

VERDÚ, S.; IVORRA, E.; SÁNCHEZ, A. J.; BARAT, J. M.; GRAU, R. Study of high strength wheat flours considering their physicochemical and rheological characterisation as well as fermentation capacity using SW-NIR imaging. **Journal of Cereal Science**, v. 62, p. 31–37, 2015.

___. Spectral study of heat treatment process of wheat flour by VIS/SW-NIR image system. **Journal of Cereal Science**, v. 71, p. 99–107, 2016.

VERDÚ, S.; VÁSQUEZ, F.; GRAU, R.; IVORRA, E.; SÁNCHEZ, A. J.; BARAT,

J. M. Detection of adulterations with different grains in wheat products based on the hyperspectral image technique: The specific cases of flour and bread. **Food Control**, v. 62, p. 373–380, 2016.

WESTAD, F.; MARINI, F. Validation of chemometric models - A tutorial. **Analytica Chimica Acta**, v. 893, p. 14–24, 2015.

XING, J.; SYMONS, S.; HATCHER, D.; SHAHIN, M. Comparison of short-wavelength infrared (SWIR) hyperspectral imaging system with an FT-NIR spectrophotometer for predicting alpha-amylase activities in individual Canadian Western Red Spring (CWRS) wheat kernels. **Biosystems Engineering**, v. 108, n. 4, p. 303–310, 2011.

YU, S.; CHU, S.; WANG, C.; CHAN, Y. Two improved k- means algorithms. **Applied Soft Computing Journal**, v. 68, p. 747–755, 2018.

ZHANG, H.; ZHANG, W.; XU, C.; ZHOU, X. International Journal of Biological Macromolecules Studies on the rheological and gelatinization characteristics of waxy wheat flour. **International Journal of Biological Macromolecules**, v. 64, p. 123–129, 2014.

ZHAO, H.; GUO, B.; WEI, Y.; ZHANG, B. Near infrared reflectance spectroscopy for determination of the geographical origin of wheat. **Food Chemistry**, v. 138, n. 2–3, p. 1902–1907, 2013.

ZIEGLER, J. U.; LEITENBERGER, M.; LONGIN, C. F. H.; WÜRSCHUM, T.; CARLE, R.; SCHWEIGGERT, R. M. Near-infrared reflectance spectroscopy for the rapid discrimination of kernels and flours of different wheat species. **Journal of Food Composition and Analysis**, v. 51, p. 30–36, 2016.