UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Programa de Pós-Graduação em Tecnologia de Alimentos

IZABELE MARQUETTI

CLASSIFICAÇÃO DE GENÓTIPOS DE CAFÉ ARÁBICA USANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO

DISSERTAÇÃO

CAMPO MOURÃO

2014

IZABELE MARQUETTI

CLASSIFICAÇÃO DE GENÓTIPOS DE CAFÉ ARÁBICA USANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO

Dissertação apresentada ao programa de Pós Graduação em Tecnologia de Alimentos da Universidade Tecnológica Federal do Paraná, como parte dos requisitos para obtenção do título de mestre em Tecnologia de Alimentos.

CAMPO MOURÃO
2014

Dados Internacionais de Catalogação na Publicação

M357c Marquetti, Izabele

Classificação de genótipos de café arábica usando espectroscopia de infravermelho próximo / Izabele Marquetti – 2014.

80 f. : il. ; 30 cm.

Orientador: Evandro Bona

Co-orientadora: Patricia Valderrama

Dissertação (Mestrado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Tecnologia de Alimentos. Medianeira, 2014.

Inclui bibliografias.

1. Café – cultivo. 2. Interação genótipo-ambiente. 3. Alimentos – Dissertações. I. Bona, Evandro, orient. II.Valderrama, Patricia, co-orient. III. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Tecnologia de Alimentos. IV. Título.

CDD: 664

Biblioteca Câmpus Medianeira

Marci Lucia Nicodem Fischborn 9/1219

TERMO DE APROVAÇÃO

CLASSIFICAÇÃO DE VARIETAIS DE CAFÉ ARÁBICA USANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO

POR

IZABELE MARQUETTI

Essa dissertação foi apresentada às quatorze horas, do dia nove de junho de dois mil e quatorze, 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, 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. Luiz Henry Monken e Silva (Membro Externo – UniCesumar)

Profa. Dra. Maria Brigida dos Santos Scholz (Membro Externo – IAPAR)

Orientador

Professor Dr. Evandro Bona

Co-orientadora

Professora Dra. Patrícia Valderrama

AGRADECIMENTOS

Aos professores Dr. Evandro Bona e Dra. Patrícia Valderrama, pela oportunidade de desenvolvimento do projeto, por toda a orientação e dedicação.

Ao doutorando Jade Varaschim Link pela ajuda durante o projeto.

Aos alunos de Iniciação Científica André Luis Guimarães Lemes, Vinícius Arca, Lucas Sabino e Rhayanna Gonçalves por todo o auxílio durante a pesquisa.

À Dra. Maria Brígida dos Santos Scholz do Instituto Agronômico do Paraná pelo fornecimento das amostras de café e pela valiosa contribuição na discussão dos resultados.

Ao Dr. Marcelo Caldeira Viegas, pela colaboração, e ao pessoal da Café Iguaçu pela parceria ao projeto disponibilizando o equipamento NIR para as análises.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação Araucária, pelo suporte financeiro ao projeto.

À Coordenação de Aperfeiçoamento de Pessoal em Nível Superior (CAPES), pela concessão da bolsa de mestrado.

RESUMO GERAL

MARQUETTI, Izabele. Classificação de genótipos de café arábica usando espectroscopia de infravermelho próximo. 2014. 79 f. Dissertação – Programa de Pós-Graduação em Tecnologia de Alimentos, Universidade Tecnológica Federal do Paraná. Campo Mourão, 2014.

As condições ambientais do cultivo do café, como clima, tipo de solo e altitude, associadas a práticas agrícolas, são responsáveis pela composição química final do grão. Além disso, o genótipo cultivado também influencia diretamente nas características essenciais da bebida, aumentando o seu valor agregado. Portanto, comprovações da origem geográfica e genotípica da genótipo do café devem ser realizadas utilizando métodos confiáveis. A espectroscopia no infravermelho próximo (NIRS), na região de 1100 a 2498 nm, foi utilizada na análise de genótipos de café arábica, cultivadas em diferentes cidades do estado do Paraná, Brasil. Como primeira aproximação, os métodos lineares, análise de componentes principais (ACP) e mínimos quadrados parciais com análise discriminante (PLS-DA), foram utilizados para a interpretação dos dados devido à complexidade e elevada quantidade de informação contida nos espectros. Os modelos PLS-DA obtidos para a classificação geográfica apresentaram uma sensibilidade média de 93,75% e uma especifidade de 100%. Já para a classificação dos genótipos a performance do PLS-DA foi de 93,75% para sensibilidade e 97,13% para a especificidade. Na tentativa de melhorar a performance e a confiabilidade de classificação foram desenvolvidos modelos de dois estágios. Tanto os *scores* da ACP como as variáveis latentes do PLS-DA foram alimentados em dois tipos diferentes de redes neurais artificiais, o perceptron de múltiplas camadas (MLP) e a rede de funções de base radial (RBF) que são modelos inerentemente não-lineares. Os respectivos parâmetros de arquitetura dessas redes foram otimizados através do método de busca direta simplex sequencial. Os modelos de dois estágios, linear com PLS-DA e não-linear com RBF, foram capazes de classificar geograficamente e genotipicamente com 100% de seletividade e especificidade todas as amostras de treinamento e de teste. As variáveis latentes do PLS-DA por serem determinadas levando-se em consideração a resposta desejada contêm mais informação que os *scores* da ACP. Já a rede RBF, por possuir um número menor de parâmetros livres e uma estrututa mais simples quando comparada à MLP, possui um treinamento mais rápido e covergente. Quando comparados com os resultados obtidos na espectroscopia de infravermelho médio (FTIR), os modelos obtidos usando os espectros NIRS apresentaram uma performance melhor e mais confiável. Estes resultados indicam que os espectros NIRS contêm informações importantes que aliadas a métodos adequados de reconhecimento de padrões resultam em uma classificação eficiente de amostras de café arábica verde por genótipo e local de cultivo. Além disso, uma análise dos *loadings* das variáveis latentes do PLS-DA permite associar quais bandas são características em cada classe. Essa informação pode ser correlacionada com a composição química das amostras fornecendo, assim, dados preliminares para avaliar o efeito da região de cultivo e do tipo de genótipo selecionado nas carecterísticas químicas do grão de café verde.

Palavras-chave: Café verde. Modelos de dois estágios. Análise de Componentes Principais. PLS-DA. Redes Neurais Artificiais.

GENERAL ABSTRACT

MARQUETTI, Izabele. Coffee arabica genotype classification using near infrared spectroscopy. 2014. 79 f. Dissertação – Programa de Pós-Graduação em Tecnologia de Alimentos, Universidade Federal Tecnológica do Paraná. Campo Mourão, 2014.

The environmental conditions in coffee cultivation, such as climate, soil type and altitude, associated with agronomic practices, are responsible for influence the final chemical composition of the bean. They directly influence the essential features of the beverage, increasing its aggregate price. Proof of geographic and genotypic origin of the coffee genotypes must be done using reliable methods. Thus, the near infrared spectroscopy (NIRS), in the 1100 to 2498nm range, was used for analyze different coffee genotypes that were cultivated in different cities (Brazil - Paraná State). As first approach linear methods, principal components analysis (PCA) and partial least squares with discriminant analysis (PLS-DA), were used for data interpretation due to the high complexity and amount of information contained in the spectra. The obtained PLS-DA models had an average sensitivity of 93.75% and a specificity of 100% for the geographical classification. While for genopyte classification, the PLS-DA performance was 93.75% for sensitivity and 97.13% for specificity. In an attempt to improve the performance and reliability of the developed classifiers, both the PCA scores and the PLS-DA latent variables were fed into two artificial neural networks, the multilayer perceptron (MLP) and radial basis function network (RBF), that are nonlinear models. The architecture parameters of these networks were optimized using the sequential simplex method. The two-stage models, linear with PLS-DA and nonlinear with RBF, were able to classify geographically and genotypically with 100% of selectivity and specificity all the training and test samples. The latent variables of the PLS-DA are determined by taking into account the desired response, so it contains more information than the scores of the PCA. While the RBF network, by having fewer free parameters and a simpler architecture compared to the MLP, has a faster and covergente training. The spectra analysis in near-infrared region showed better results than mid-infrared spectra. These results indicate that NIRS spectra contain important information that, combined with appropriate methods of pattern recognition, allow the classification of green arabica coffee samples by genotype and growing region. Besides, the PLS-DA loadings analysis allows associating which NIRS bands are specific of each class. This information can be correlated with the samples chemical composition, providing preliminary data to evaluate the effect of growing region and genotype in the selected green coffee chemical composition.

Keywords: Green Coffee. Two-stage models. Principal Component Analysis. PLS-DA. Artificial Neural Networks.

LISTA DE FIGURAS

Capítulo 1

Capítulo 2

LISTA DE TABELAS

Capítulo 1

Capítulo 2

LISTA DE ABREVIATURAS

ACG – Ácidos Clorogênicos

ACP – Análise de Componentes Principais

AE – Autoescalonamento

CB – *City Block*

COR – Correlação

COS – Cosseno

CP – Cornélio Procópio

DE – Distância euclidiana

DP – Dados Puros

FL – Função Logística

FTIR – Espectrocopia de Infravermelho Médio com Transformada de Fourier (do inglês, *Fourier transform Mid-infrared spectroscopy*)

G – Função Gaussiana

IAPAR – Instituto Agronômico do Paraná

IUPAC – *International Union of Pure and Applied Chemistry*

L – Função Linear

LD – Londrina

LV – Variáveis Latentes (do inglês, *Latent Variables)*

MD – Mandaguari

MLP – Perceptron de Múltiplas Camadas (do inglês, *Multi Layer Perceptron*)

MM – Normalização Minimax (mínimo = -1 e máximo = +1)

MQ – Função Multiquádrica

MQI – Função Multiquádrica inversa

MSC – Correção do espalhamento multiplicativo (do inglês, *Multiplicative Scatter Correction*)

MSE – Erro Quadrático Médio (do inglês, *Mean Square Error*)

NIRS – Espectroscopia de Infravermelho Próximo (do inglês, *Near-Infrared Spectroscopy*)

PC – Componente Principal (do inglês, *Principal Component*)

PLS-DA – Mínimos Quadrados Parciais com Análise Discriminante (do inglês, *Partial Least Squares with Discriminant Analysis*)

PV – Paranavaí

RBF – Rede de Funções de Base Radial (do inglês, *Radial Basis Function Network*)

RNA – Redes Neurais Artificiais

THS – Função Tangente Hiperbólica Sigmóide

VU – Vetor Unitário

SUMÁRIO

APRESENTAÇÃO

Esta dissertação é composta por dois capítulos que foram elaborados na forma de artigo para publicação e estão apresentados nas normas da revista para onde serão encaminhados.

Capítulo 1 – Autores: Izabele Marquetti, Jade Varaschim Link, André Luis Guimarães Lemes, Maria Brígida dos Santos Scholz, Patrícia Valderrama, Evandro Bona. Título: Classificação de genótipos de café arábica utilizando espectroscopia no infravermelho próximo e mínimos quadrados parciais com análise discriminante. Periódico: Journal of Near Infrared Spectroscopy. *(Já submetido)*

Capítulo 2 – Autores: Izabele Marquetti, André Luis Guimarães Lemes, Evandro Bona, Maria Brígida dos Santos Scholz e Patrícia Valderrama. Título: Classificação de genótipos de café arábica utilizando espectroscopia de infravermelho próximo e modelos de dois estágios. Periódico: Food Chemistry.

CAPÍTULO 1:

Classificação de genótipos de café arábica utilizando espectroscopia no infravermelho próximo e mínimos quadrados parciais com análise discriminante

1 CLASSIFICAÇÃO DE GENÓTIPOS DE CAFÉ ARÁBICA UTILIZANDO ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO E MÍNIMOS QUADRADOS PARCIAIS COM ANÁLISE DISCRIMINANTE

Resumo

As condições ambientais do cultivo do café, como clima, tipo de solo e altitude, associadas a práticas agrícolas, são responsáveis pela composição química final do grão. Além disso, o genótipo cultivado também influencia diretamente nas características essenciais da bebida, aumentando o seu valor agregado. Portanto, a comprovação da origem geográfica e genotípica do café deve ser realizada utilizando métodos confiáveis. A espectroscopia no infravermelho próximo (NIRS), na região de 1100 a 2498 nm, foi utilizada na análise de genótipos de café, cultivadas em diferentes cidades do estado do Paraná, Brasil. O método quimiométrico mínimos quadrados parciais com análise discriminante (PLS-DA) foi utilizado para a interpretação dos dados devido à complexidade e elevada quantidade de informação contida nos espectros. Dois pré-processamentos, correção do espalhamento multiplicativo (MSC) e segunda derivada de *Savitzky-Golay*, foram testados a fim de avaliar qual fornece a maior porcentagem de classificação correta. Os melhores modelos obtidos possibilitaram uma classificação correta total de 94,4% das amostras de validação tanto por origem geográfica e quanto por genotípica. Estes resultados indicam que os espectros NIRS podem ser utilizados para a previsão de amostras de café arábica por genótipo e local de cultivo.

Palavras-chave: café verde; origem geográfica, origem genotípica, espectroscopia no infravermelho próximo, PLS-DA

1.1 INTRODUÇÃO

O café pertence ao gênero *Coffea* e à família Rubiaceae. Este gênero possui cerca de 100 espécies, mas apenas duas são comercializadas, a *Coffea arabica* e a *Coffea canephora*.[1,2] O consumo do café tem aumentado continuamente, devido a fatores como melhoria na qualidade da bebida, melhorias nas práticas agrícolas, sua associação com benefícios à saúde e disponibilidade de novos produtos. A qualidade da bebida está associada à seleção adequada dos genótipos de café, melhorando seu aroma e sabor.[3] O café arábica produz uma bebida de

melhor qualidade, com aroma intenso, menor teor de cafeína e menos amargor e, consequentemente, tem maior valor agregado.[4,5]

Existem vários genótipos de café arábica disponíveis, grande parte delas foram obtidas por melhoramento genético. O desenvolvimento destes genótipos modernos busca obter grãos mais adaptados a várias condições climáticas, de solo e também resistentes a doenças e pragas, aumentando assim a produtividade e melhorando a qualidade.[6] A variabilidade genética existente entre os genótipos da espécie *C. arabica* em interação com o ambiente interfere quantitativa e qualitativamente nos componentes químicos e nas características físico-químicas dos grãos de café.[7,8]

Apesar da qualidade do café como bebida estar relacionada com a composição química do café torrado, esta depende da composição do café verde, já que os compostos do café verde reagem entre si em todos os estágios da torrefação, gerando bebidas diversificadas.[9] Cafés de melhor qualidade estão relacionados com o aumento de teores de sacarose, lipídeos, aminoácidos e trigonelina; e a redução de ACG e cafeína, que são responsáveis por aumentar o amargor do café. [5,10] A possibilidade de associar a composição química e a qualidade ao local de cultivo tornou-se uma maneira de agregar valor em mercados altamente competitivos.[11,12] Para garantir ao consumidor a origem geográfica e genética do café, métodos analíticos rápidos e eficientes que permitam atingir estes propósitos são cada vez mais requeridos.

Para discriminar os genótipos de café utilizadas para o preparo da bebida faz-se o uso de técnicas analíticas como análises cromatográficas, microscopia eletrônica de varredura e espectroscopia no infravermelho médio.[3,13] Muitas dessas análises são demoradas, pois necessitam de preparo de amostras, possuem um custo elevado e geram muitos resíduos. Uma alternativa é a espectroscopia no infravermelho próximo (NIRS) cujas vantagens são a rapidez, preparo mínimo da amostra e a possibilidade de análises simultâneas.[14,15] Por ser uma técnica com complexidade de interpretação e elevada quantidade de informação, proveniente de sobretons e bandas de combinação, é necessária a utilização de métodos estatísticos multivariados para auxiliar na sua interpretação.[16]

Os sinais obtidos no infravermelho próximo são devido a sobretons e bandas de combinação de vibrações moleculares fundamentais, principalmente estiramentos e deformações angulares, referentes às ligações C=O, C-H, C-N, C-O, N-H, NO_2 e O-H.[17] A Tabela 1 mostra algumas bandas de NIRS de compostos orgânicos na região de comprimento de onda utilizada neste estudo.

Tabela 1 - Comprimentos de onda de algumas bandas de NIRS de compostos orgânicos.[17]

Região	Faixa de comprimento de onda (nm)	Modos vibracionais
1	1100 – 1225	2º sobretom de C-H
2	1300 – 1420	1º sobretom de bandas de combinação de C-H
3	1400 – 1600	1º sobretom de N-H e O-H
4	1620 – 1800	1º sobretom de C-H
5	1900 – 2000	2º sobretom de C=O, 1º sobretom de C=O e bandas de combinação de O-H
6	2000 – 2200	Bandas de combinação de N-H e O-H
7	2200 – 2460	Bandas de combinação de C-H

Esta técnica já foi utilizada para discriminação de mistura de cafés arábica e robusta, obtendo resultados satisfatórios.[15,19,20] No entanto, são escassas as pesquisas que utilizam a técnica NIRS para discriminação de genótipos de café arábica por genótipo e região de cultivo.

Este trabalho teve como objetivo realizar uma segmentação geográfica e genotípica de cafés cultivados no Brasil. Para este propósito, foram desenvolvidos modelos PLS-DA a partir dos espectros obtidos por NIRS.

1.2 MATERIAIS E MÉTODOS

1.2.1 AMOSTRAS DE GENÓTIPOS DE CAFÉ

Foram analisados quatro genótipos de *C. arabica* desenvolvidas pelo Instituto Agronômico do Paraná (IAPAR): IPR 99, IPR 105, IPR 106 e Iapar 59. O genótipo Iapar 59 foi lançada em 1994, originada do cruzamento entre *C. arabica*, "Villa Sarchi 971/10" e o "Híbrido de Timor 832/2", assim como a IPR 99, apresenta resistência aos tipos de ferrugem conhecidos.[21] O genótipo IPR 105 é derivado do genótipo Catuaí; o genótipo IPR 106 originou-se do genótipo Icatu, ambos igualmente resistentes a ferrugens em diferentes níveis. Dentre estes genótipos, apenas o genótipo Iapar 59 e a IPR 99 foram disponibilizados para os cafeicultores.[22] Logo, encontrar semelhanças entre os genótipos IPR 105 e IPR 106 com os genótipos já disponíveis comercialmente (IPR 99 e Iapar 59) pode ser uma contribuição para o lançamento de novos genótipos para o mercado consumidor. Assim, torna-se importante obter ferramentas eficazes de análise que ajudem a encontrar essas semelhanças, principalmente em relação à composição química.

Foram utilizadas 18 amostras de cafés cultivadas em quatro locais: Cornélio Procópio (CP), Paranavaí (PV), Mandaguari (MD) e Londrina (LD), todos no estado do Paraná, Brasil. Foi utilizada uma amostra de cada genótipo por cidade, com exceção das amostras do genótipo IA 59 cultivadas em Paranavaí e Cornélio Procópio, onde foram utilizadas duas amostras por cidade.

As amostras de Paranavaí e Cornélio Procópio foram colhidas na safra de 2008 e as de Londrina e Mandaguari foram colhidas na safra de 2010. As condições climáticas, bem como a latitude, longitude e altitude são mostradas na Tabela 2.[23]

Após a colheita, as amostras foram enviadas a estação experimental do IAPAR em Londrina, onde foram colocadas em caixas de madeira com uma malha de fundo e movidas oito vezes por dia até obter uma umidade dos grãos de 11-12%. Em seguida, as amostras foram beneficiadas, removendo a casca e o pergaminho.[24] Os grãos verdes beneficiados foram moídos (0,5 mm) e armazenados em um freezer a -18ºC, para serem analisados posteriormente.

Tabela 2 - Condições climáticas das cidades.[23]

Cidade	Latitude	Longitude	Altitude	Temperatura média anual
Mandaguari	23°32'52”S	51°40’15”W	650 m	20-21°C
Londrina	23°18'36”S	51°09’56”W	585 m	21-22°C
Cornélio Procópio	23°10'51"S	50°38’48"W	658 m	21-22°C
Paranavaí	23°04'22"S	52°27’55"W	470 m	22-23°C

1.2.2 ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO

Os espectros de café verde foram obtidos em um espectrofotômetro de infravermelho próximo NIRSystem 5000-M (Foss Tecator AB, Höganäs, Suécia). As leituras foram feitas em temperatura ambiente (23^{o}C), na faixa de comprimento de onda de 1100 a 2498nm em intervalos de 2nm. Para cada amostra de café foram realizadas 5 repetições, obtendo-se assim um total de 90 espectros. O *software* WinISI III versão 1.50e (Foss NIRSystems/Tecator Infrasoft International, LLC, Silver Spring, MD, USA) foi utilizado para aquisição dos espectros. A absorbância foi obtida como logarítmo decimal do inverso da transmitância, log (1/T).

1.2.3 PROCESSAMENTO DOS DADOS

Para reduzir fontes de variação que não carregam informações relevantes durante a calibração do modelo multivariado, dois pré-tratamentos foram utilizados nos espectros: correção do espalhamento multiplicativo (MSC) e segunda derivada de *Savitzky-Golay*.

A correção do espalhamento multiplicativo (MSC) corrige simultaneamente os efeitos aditivos e multiplicativos do espalhamento de luz, gerados por diferenças na granulometria, morfologia e orientação das partículas. Para realizar a correção é utilizada uma regressão linear das variáveis espectrais versus o espectro médio.[25,26]

A segunda derivada por meio do algoritmo de *Savitzky-Golay* remove problemas devido a mudanças de inclinação entre as amostras.[27] Um polinômio de 2^{o} grau e 7 pontos de janela foram utilizados.

1.2.3.1 Mínimos Quadrados Parciais com Análise Discriminante

O PLS-DA é um método quimiométrico supervisionado, ou seja, utiliza a resposta desejada para as amostras de treinamento na decomposição dos dados em *scores* e *loadings*.[28] Foi desenvolvido a partir dos algoritmos do PLS (*Partial Least Squares*) utilizados para a calibração multivariada, mas, é aplicado para a classificação.[29] No PLS-DA, assim como no PLS, é estabelecida uma relação linear entre a variável dependente (**Y**) e a variável independente (**X**). Tanto o PLS quanto o PLS-DA são baseados no método da análise de componentes principais (ACP). A matriz **X** é decomposta no produto de duas matrizes, *scores* e *loadings*, como no ACP. A diferença é que no PLS e no PLS-DA ocorre uma leve rotação no eixo das componentes principais objetivando buscar a máxima covariância de **X** com **Y** e os componentes principais passam a ser chamados de variáveis latentes (LV).[30] No PLS, **Y** contém os valores de uma propriedade de interesse, enquanto que no PLS-DA a matriz **Y** contém informações acerca das classes das amostras. O número de colunas é igual ao número de classes, ou seja, cada classe tem uma coluna em **Y**. Em cada classe é assumido o valor de 0 ou 1, dependendo se ele pertence ou não à classe representada por aquela coluna.[31]

O modelo consiste de duas etapas: a calibração ou treinamento, em que as características dos dados são investigadas a fim de encontrar um modelo para seu comportamento; e a validação ou teste, em que algumas amostras que não participaram da calibração são utilizadas para avaliar a qualidade do modelo construído.[32]

A performance do modelo de classificação pode ser avaliada utilizando alguns parâmetros como sensibilidade e especificidade. A sensibilidade é a habilidade do modelo de classificar corretamente as amostras, relacionando as amostras previstas como sendo da classe

com as amostras presentes de fato na classe. Enquanto a especificidade relaciona as amostras previstas como não sendo da classe com as amostras reais que não são da classe.[33]

Também é possível calcular um valor limite (*threshold)* que separa as classes. Desta maneira, minimiza-se o número de falsos positivos/negativos para a validação dos dados.[33] O valor do *threshold* corresponde ao encontro das curvas de probabilidade *a posteriori* determinadas através do teorema de Bayes (1).[34]

$$p(C_k|y)= \frac{p(y|C_k)p(C_k)}{p(y)} \qquad \text{(Eq. 1)}$$

Na equação (1) $p(y|C_k)$ é a probabilidade condicional calculada pela distribuição Gaussiana, $p(C_k)$ é a probalilidade *a priori* e $p(y)$ é uma constante de normalização.

O cálculo da exatidão do conjunto de calibração é avaliado utilizando o erro quadrático médio de calibração (RMSEC), enquanto que a exatidão do conjunto de previsão é avaliada pelo erro quadrático médio de previsão (RMSEP), conforme as equações (2) e (3). Onde n é o número de amostras e *v* é o número de variáveis latentes +1, para dados centrados na média.[35]

$$RMSEC= \sqrt{\frac{\sum \left(\hat{y}_p - y_r\right)^2}{n-v}} \qquad \text{(Eq. 2)}$$

$$RMSEP= \sqrt{\frac{\sum \left(\hat{y}_p - y_r\right)^2}{n}} \qquad \text{(Eq. 3)}$$

Modelos PLS-DA foram construídos para diferenciar as amostras de café arábica com relação aos genótipos e aos locais de cultivo. Foram utilizadas 72 amostras de calibração e 18 amostras de validação (uma repetição de cada um dos cafés estudados). Os dados foram centrados na média.

Todas as análises quimiométricas dos espectros NIRS foram realizadas no MATLAB R2008b (The MathWorks Inc., Natick, USA).

1.3 RESULTADOS E DISCUSSÕES

A Figura 1-(a) mostra os espectros originais obtidos por NIRS das 18 amostras de café, com 5 repetições cada, perfazendo um total de 90 espectros. A Figura 1-(b) mostra os espectros tratados com MSC para minimizar os efeitos do espalhamento de luz. Os espectros

tratados pela segunda derivada, utilizando o algoritmo de *Savitzky-Golay* são mostrados na Figura 1-(c) e aqueles tratados com a segunda derivada após o tratamento com MSC podem ser visualizados na Figura 1-(d).

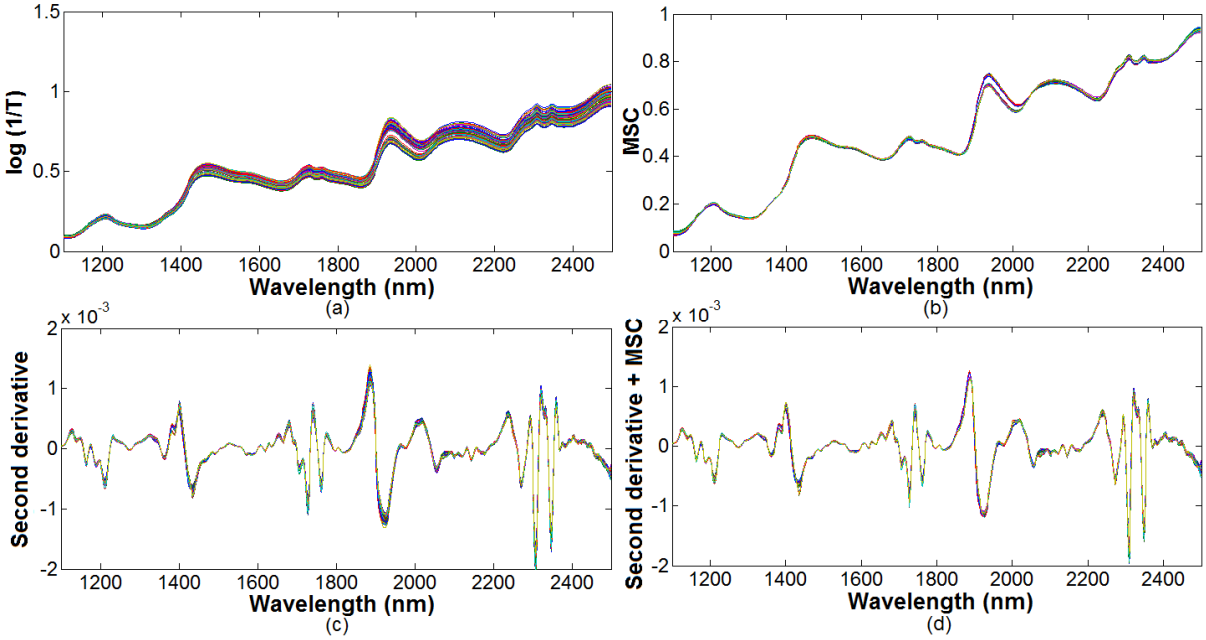


Figura 1 - (a) Espectro das amostras de café; (b) espectros após aplicação de MSC; (c) espectros após aplicação de segunda derivada; (d) espectros com aplicação de MSC e segunda derivada conjuntamente.

1.3.1 ORIGEM GEOGRÁFICA

Um modelo supervisionado PLS-DA foi desenvolvido para cada um dos três conjuntos de espectros (Figura 1-(b), (c), (d)), com o objetivo de discriminar as amostras de café por origem geográfica. Os dados foram centrados na média e o melhor processamento, bem como o número de variáveis latentes, foi escolhido baseado nos menores valores de RMSEC e RMSEP por classe e maiores valores de sensibilidade e especificidade. A Tabela 3 contém o melhor modelo para cada pré-tratamento utilizado.

Tabela 3. Seleção do melhor modelo supervisionado PLS-DA.

Pré-processamento	LV	Classe	RMSEC	RMSEP	Calibração		Validação	
					Sens.	Espec.	Sens.	Espec.
MSC+2ª derivada	6	CP	0,1598	0,1853	1,000	1,000	1,000	1,000
		PV	0,1184	0,1734	1,000	1,000	1,000	1,000
		MD	0,2059	0,2186	1,000	1,000	0,750	1,000
		LD	0,0919	0,1024	1,000	1,000	1,000	1,000
2ª derivada	5	CP	0,2043	0,2860	1,000	1,000	1,000	0,923
		PV	0,1842	0,2459	1,000	1,000	0,800	1,000
		MD	0,2109	0,1894	0,938	0,982	1,000	1,000
		LD	0,1172	0,1192	1,000	1,000	1,000	1,000
MSC	4	CP	0,3360	0,3544	0,800	0,846	1,000	0,692
		PV	0,3290	0,3664	0,950	0,827	0,800	0,615
		MD	0,2371	0,2400	0,938	0,946	0,750	1,000
		LD	0,1645	0,1734	1,000	1,000	1,000	1,000

Como pode ser observado na Tabela 3, o melhor modelo PLS-DA foi o tratado com MSC e 2ª derivada conjuntamente, com 6 variáveis latentes. Os resultados apresentados a seguir são todos para este modelo que apresentou uma variância acumulada de 93,53% em **X** e 87,90% em **Y**. A Tabela 4 apresenta a variância acumulada para as seis primeiras variáveis latentes.

Tabela 4. Porcentagens de variância explicada pelo modelo PLS-DA, espectros tratados com MSC e 2ª derivada conjuntamente.

LV	Bloco X		Bloco Y	
	LV	Variância acumulada	LV	Variância acumulada
1	66,46	66,46	29,36	29,36
2	11,65	78,10	16,47	45,82
3	7,95	86,06	10,92	56,75
4	3,48	89,53	17,49	74,24
5	3,13	92,66	6,13	80,37
6	0,87	93,53	7,52	87,90

Analisando os valores dos resíduos e do *leverage*, nenhuma amostra foi considerada como *outlier*, já que nenhuma apresentou simultaneamente altos valores de *leverage* e altos valores de resíduos, como pode ser visto na Figura 2.

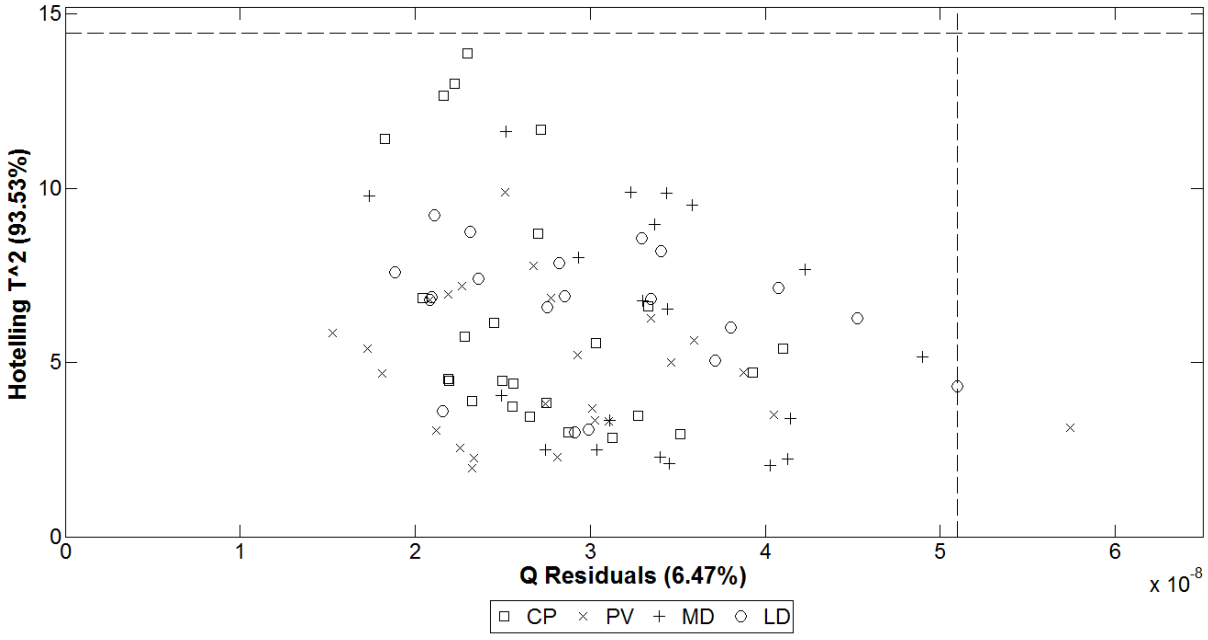


Figura 2. Resíduos espectrais (Q) versus leverage (T^2 de Hotelling) para modelo PLS-DA, com espectros tratados com MSC e 2ª derivada conjuntamente.

Os gráficos dos *scores* para este modelo PLS-DA estão apresentados na Figura 3-(a) e (b), como pode ser visto houve uma separação entre as amostras por cidades, indicando que as amostras de Mandaguari foram discriminadas pela parte positiva da LV 1 e negativa da LV 2, as amostras de Londrina, pela parte positiva da LV 1 e positiva da LV 2. Enquanto que as amostras de Paranavaí, foram discriminadas pela parte negativa da LV 1 e positiva da LV 3; e as de Cornélio Procópio, pela parte negativa da LV 1 e da LV 3.

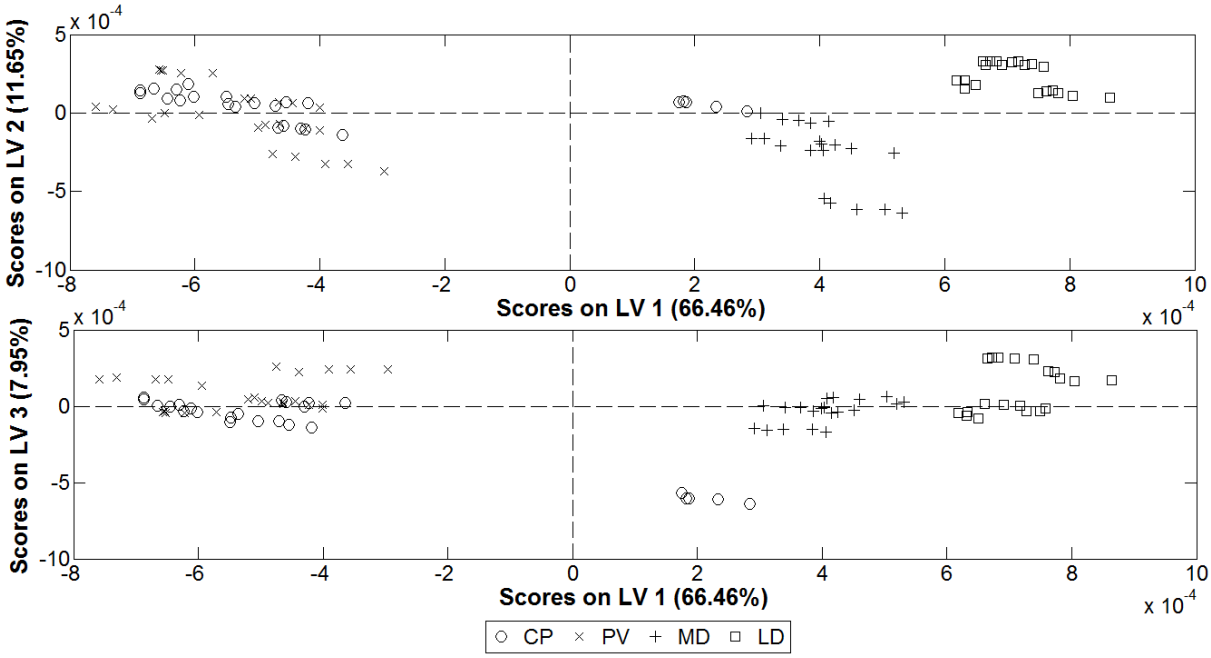


Figura 3. Scores do modelo PLS-DA, com espectros tratados com MSC e 2ª derivada conjuntamente, (a) LV 1 versus LV 2 e (b) LV 1 versus LV 3.

O gráfico dos *loadings* do PLS-DA, apresentado na Figura 4, mostra os picos de NIRS mais intensos, com valores maiores que ± 0,1, que contribuíram significativamente para a separação entre as classes.

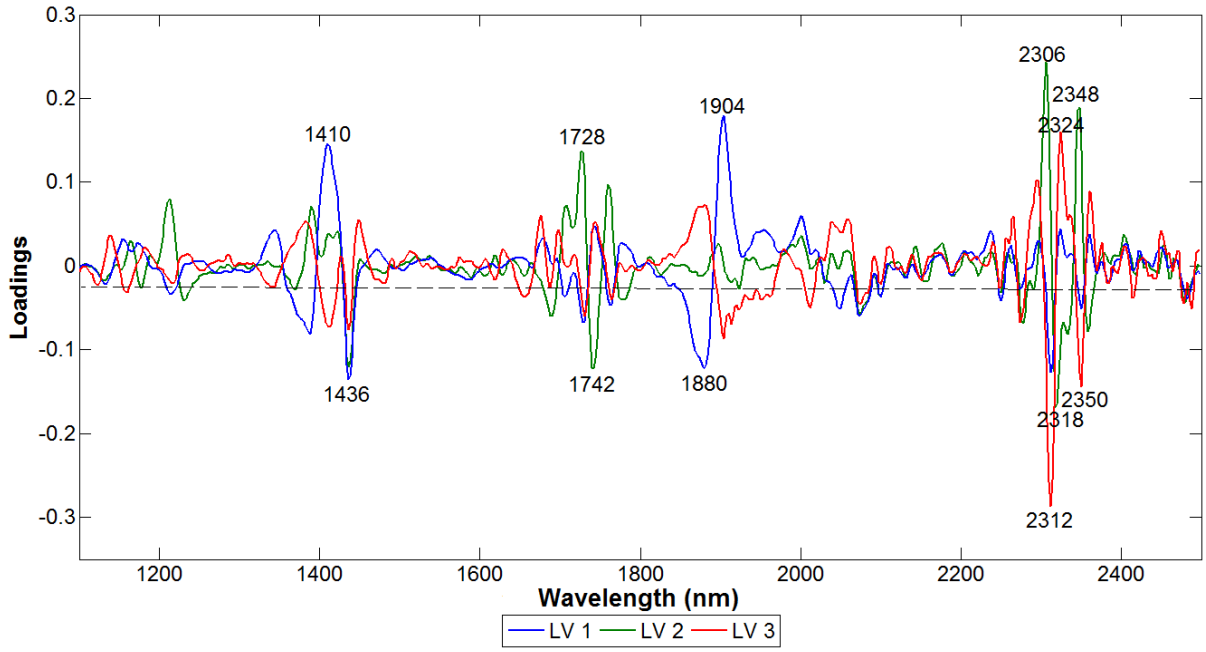


Figura 4. Gráfico dos loadings das LV 1, 2 e 3 versus comprimento de onda (variáveis) para o modelo PLS-DA, com espectros tratados com MSC e 2ª derivada conjuntamente.

Analisando a Figura 4 juntamente com a Tabela 1, nota-se que os picos de 1410, 1742, 1904, e 2318nm contribuíram para a classificação das amostras de Mandaguari, já que possuem *loadings* positivos para LV 1 ou negativos para LV 2. O pico de 1410nm está relacionado com bandas atribuídas aos lipídeos, água e aos carboidratos; o de 1742nm, à cafeína; o de 1904, aos ácidos clorogênicos, aos lipídeos, à água e aos carboidratos; e o de 2318nm, à cafeína, proteínas, aminoácidos, lipídeos e aos açúcares.[9]

Na classificação das amostras de Londrina, destacaram-se os picos de 1410, 1728, 1904, 2306 e 2348nm, pois possuem *loadings* positivos para LV 1 ou LV 2. O pico de 1728nm está relacionado com bandas atribuídas à cafeína, proteínas e aminoácidos; o de 2306nm, à cafeína, proteínas, aminoácidos, lipídeos e aos açúcares; e o de 2348nm, à cafeína, proteínas, aminoácidos e lipídeos.[9]

Para as amostras de Paranavaí, os picos de 1436, 1880, 2324nm foram os que contribuíram para a classificação das mesmas, pois possuem *loadings* negativos para LV 1 ou positivos para LV 3. O pico de 1436nm está relacionado com bandas atribuídas aos ácidos clorogênicos, cafeína, lipídeos, água, açúcares e aos carboidratos; e o pico de 2324nm, aos ácidos clorogênicos, proteínas, aminoácidos, lipídeos, açúcares e aos carboidratos.[9]

Já para as amostras de Cornélio Procópio, foram os picos 1436, 1880, 2312 e 2350nm, pois possuem *loadings* negativos para LV 1 ou para LV 3. Sendo que o pico de 2350nm está relacionado com bandas referentes aos ácidos clorogênicos, proteínas, aminoácidos, lipídeos e aos carboidratos.[9]

As respostas previstas para as amostras de calibração e de validação deste modelo PLS-DA estão apresentados por classe na Figura 5. Os valores de *threshold* foram de 0,4964, 0,5320, 0,4432 e 0,3667 para as classes CP, PV, MD e LD, respectivamente, e estão representados no gráfico pela linha tracejada em vermelho. Apenas uma amostra de Mandaguari foi classificada erroneamente como sendo de Londrina, estes resultados indicam uma classificação correta total das amostras de 94,4%.

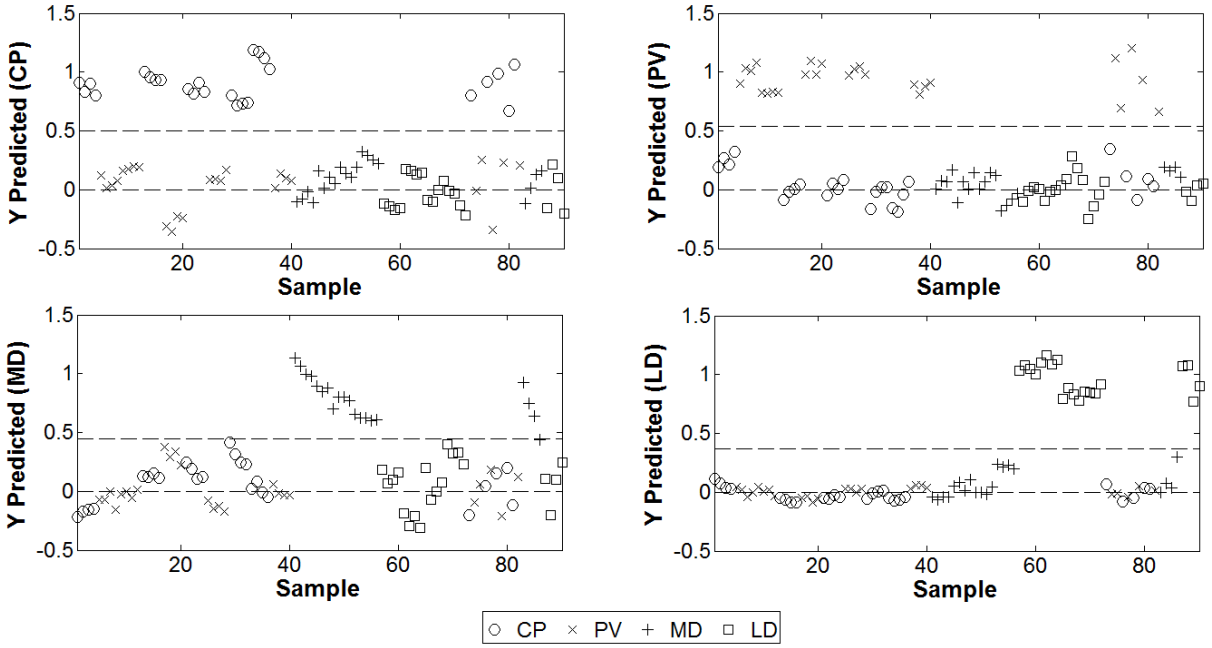


Figura 5. Amostras classificadas pelo modelo PLS-DA, espectros tratados com MSC e 2ª derivada conjuntamente. (a) Classe 1 – CP, (b) classe 2 – PV, (c) classe 3 – MD e (d) classe 4 – LD.

Os valores de sensibilidade e especificidade obtidos por classe para este modelo (Tabela 3) foram próximos ou iguais a 1, indicando que o modelo pode ser utilizado para a classificação das amostras por origem geográfica.

A discriminação entre as amostras de Londrina e Mandaguari em todos os tratamentos talvez esteja relacionada às condições ambientais e climáticas, já que Londrina tem menor altitude e clima mais quente que Mandaguari (Tabela 2), o que faz com que os grãos se desenvolvam mais rapidamente.[36] Resultados semelhantes foram obtidos em outros estudos nesta mesma região.[37] A separação entre as amostras de Cornélio Procópio e Paranavaí que foram os extremos de longitude analisados foi menos evidente.

1.3.2 ORIGEM GENOTÍPICA

Um modelo supervisionado PLS-DA foi desenvolvido para cada um dos três conjuntos de espectros (Figura 1-(b), (c), (d)), com o objetivo de discriminar as amostras de

café por origem genotípica. Os dados foram centrados na média e o melhor processamento, bem como o número de variáveis latentes, foi escolhido baseado nos menores valores de RMSEC e RMSEP e maiores valores de sensibilidade e especificidade. A Tabela 5 contém o melhor modelo para cada pré-tratamento utilizado.

Tabela 5. Seleção do melhor modelo supervisionado PLS-DA.

Pré-processamento	LV	Classe	RMSEC	RMSEP	Calibração		Validação	
					Sens.	Espec.	Sens.	Espec.
MSC+2ª derivada	6	IPR 105	0,2315	0,2858	1,000	0,964	1,000	0,929
		IPR 106	0,2539	0,2607	0,750	0,946	0,750	1,000
		IPR 99	0,2536	0,2525	1,000	0,964	1,000	1,000
		IA 59	0,1439	0,1585	1,000	1,000	1,000	1,000
2ª derivada	6	IPR 105	0,2923	0,2907	0,938	0,875	0,750	0,786
		IPR 106	0,2872	0,2969	0,750	0,946	0,750	1,000
		IPR 99	0,2407	0,2523	1,000	0,946	1,000	1,000
		IA 59	0,1727	0,2079	1,000	1,000	1,000	1,000
MSC	4	IPR 105	0,4136	0,4171	0,750	0,357	0,500	0,429
		IPR 106	0,4057	0,4024	0,625	0,536	0,750	0,357
		IPR 99	0,3144	0,3316	0,750	0,982	0,750	0,929
		IA 59	0,3576	0,3443	0,708	0,833	0,667	1,000

Como pode ser observado na Tabela 5, o melhor modelo PLS-DA foi o tratado com MSC e 2ª derivada conjuntamente, com 6 variáveis latentes. Os resultados apresentados a seguir são todos para este modelo que apresentou uma variância acumulada de 92,50% em **X** e 72,58% em **Y**. A Tabela 6 apresenta a variância acumulada para as seis primeiras variáveis latentes.

Tabela 6. Porcentagens de variância explicada pelo modelo PLS-DA, espectros tratados com MSC e 2ª derivada conjuntamente.

LV	Bloco X		Bloco Y	
	LV	Variância acumulada	LV	Variância acumulada
1	32,22	32,22	14,20	14,20
2	47,84	80,06	5,14	19,34
3	2,85	82,91	17,49	36,83
4	2,74	85,65	13,06	49,89
5	2,24	87,90	14,31	64,20
6	4,60	92,50	8,38	72,58

Analisando os valores dos resíduos e do *leverage*, nenhuma amostra foi considerada como *outlier*, já que nenhuma apresentou simultaneamente altos valores de *leverage* e altos valores de resíduos (Figura 6).

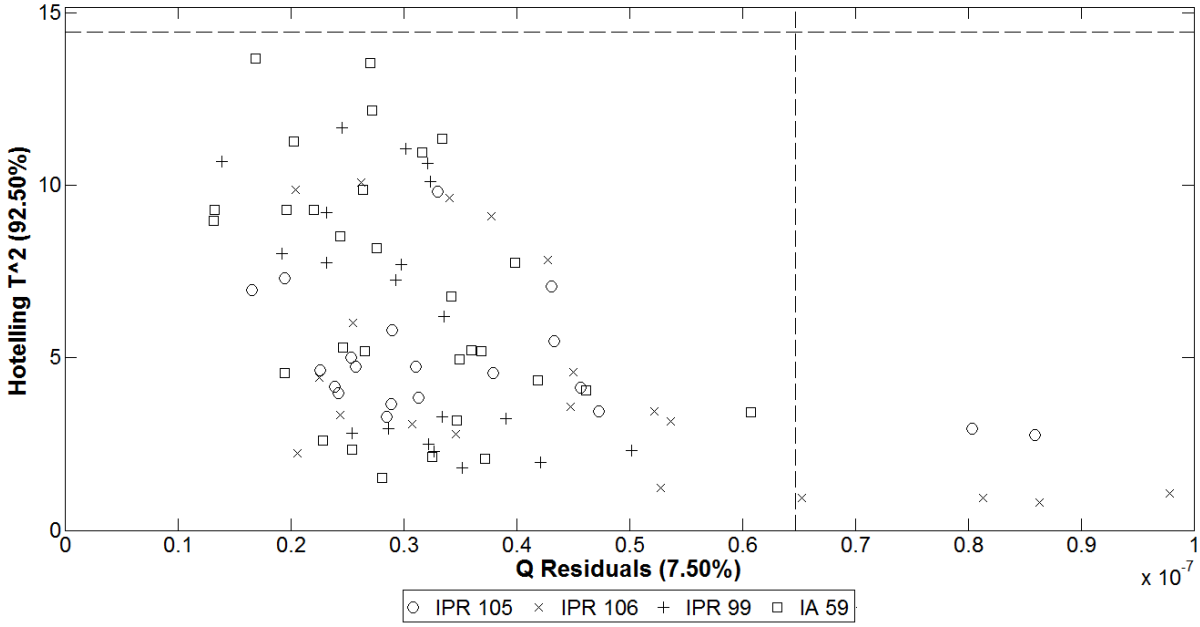


Figura 6. Resíduos espectrais (Q) versus leverage (T^2 de Hotelling) para modelo PLS-DA, com espectros tratados com MSC e 2ª derivada conjuntamente.

Os gráficos dos *scores* para este modelo PLS-DA estão apresentados na Figura 7-(a) e (b), como pode ser visto houve uma discreta separação entre as amostras por genótipos, indicando que as amostras IPR 105 e IPR 106 foram discriminadas pela parte negativa da LV 3, as amostras IPR 99, pela parte negativa da LV 1 e da LV 2; e as amostras IA 59, pela parte positiva da LV 3. As amostras IPR 105 e IPR 106 não apresentaram uma boa separação entre si.

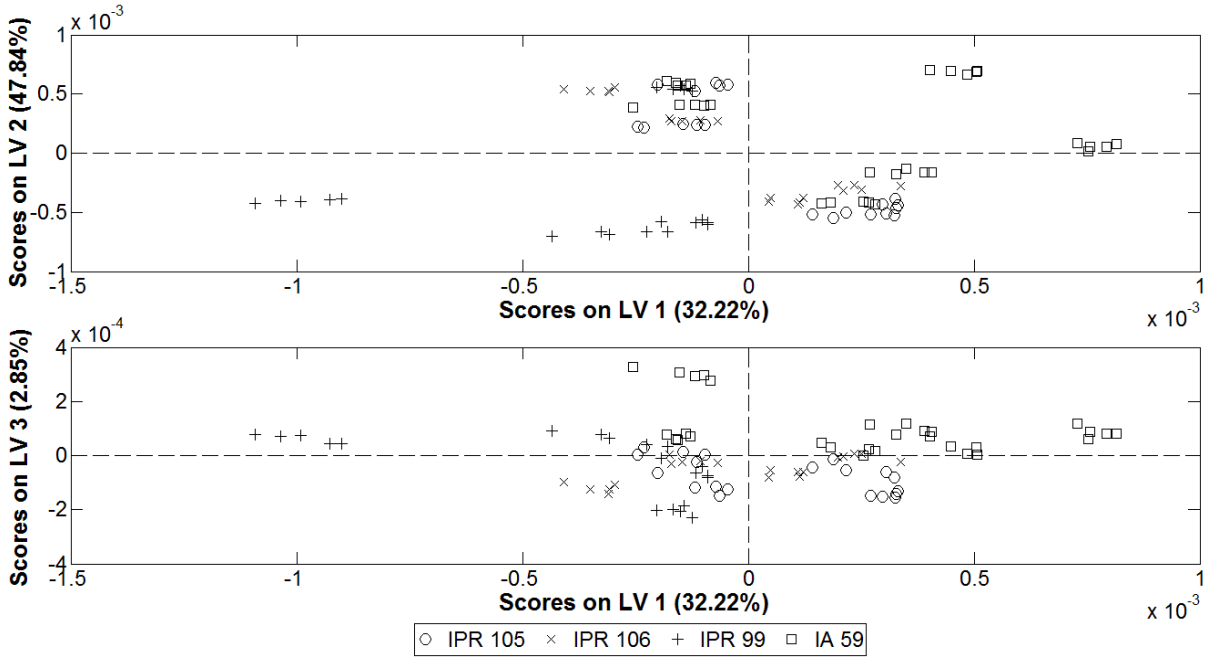


Figura 7. Scores do modelo PLS-DA, com espectros tratados com MSC e 2ª derivada conjuntamente. (a) LV 1 versus LV 2 e (b) LV 1 versus LV 3.

O gráfico dos *loadings* do PLS-DA, apresentado na Figura 15, mostra os picos de NIRS mais intensos, com valores maiores que ± 0,1, que contribuíram significativamente para a separação entre as classes.

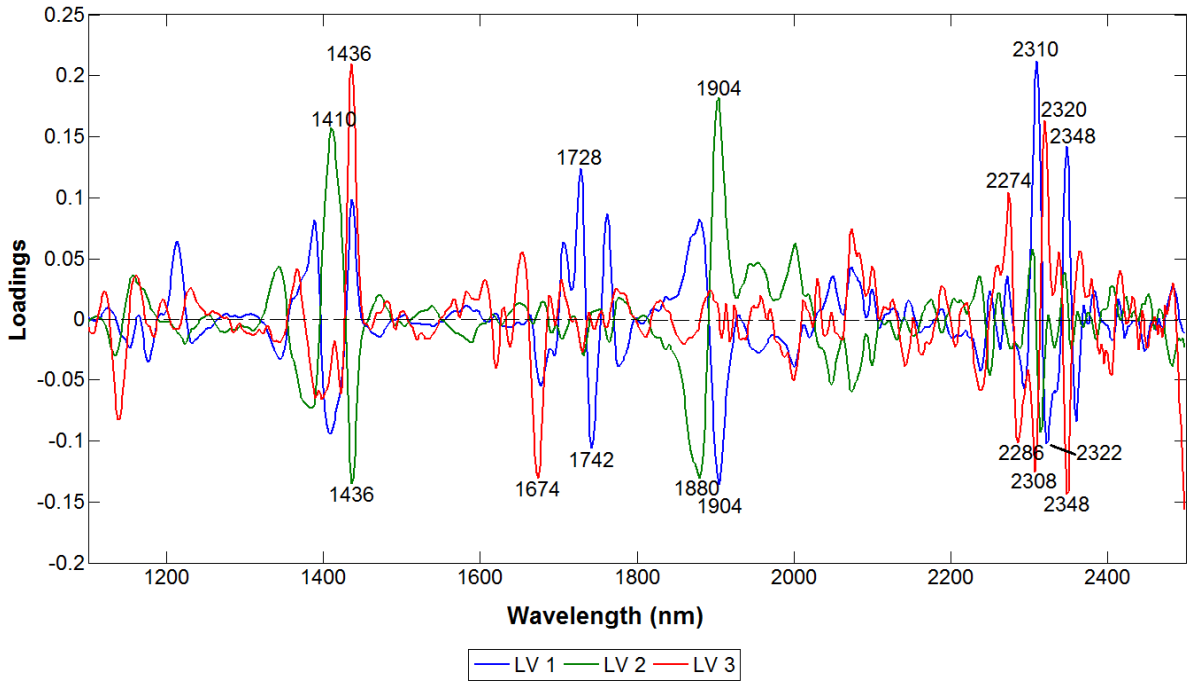


Figura 8. Gráfico dos loadings das LV 1, 2 e 3 versus comprimento de onda (variáveis) para o modelo PLS-DA, com espectros tratados com MSC e 2ª derivada conjuntamente.

Analisando a Figura 8 juntamente com a Tabela 1, nota-se que os picos de 1674, 2286, 2308, e 2348nm contribuíram para a classificação das amostras dos genótipos IPR 105 e IPR 106, já que possuem *loadings* negativos para LV 3. O pico de 1674nm está relacionado com bandas atribuídas à cafeína e aos ácidos clorogênicos; o de 2286nm, à cafeína, proteínas e aminoácidos; o de 2308nm, à cafeína, proteínas, aminoácidos, lipídeos e açúcares; e o de 2348nm, à cafeína, proteínas, aminoácidos e lipídeos.[9]

Na classificação das amostras do genótipo IPR 99, destacaram-se os picos de 1436, 1742, 1880, 1904, e 2322nm, pois possuem *loadings* negativos para LV 1 ou LV 2. O pico de 1436nm está relacionado com bandas atribuídas aos ácidos clorogênicos, cafeína, lipídeos, água, açúcares e carboidratos; de 1742nm, à cafeína; o de 1904, aos ácidos clorogênicos, lipídeos, água e carboidratos; e o de 2322nm, à cafeína, proteínas, aminoácidos, lipídeos e açúcares.[9]

Já para as amostras do genótipo IA 59, foram os picos 1436, 2274 e 2320nm, pois possuem *loadings* positivos para LV 3. Sendo que o pico de 2274nm está relacionado com bandas referentes à trigonelina, cafeína, proteínas, aminoácidos e carboidratos; e o 2320nm, à cafeína, proteínas, aminoácidos, lipídeos e açúcares.[9]

Com base nos resultados obtidos, percebe-se que os genótipos que já estão disponíveis para os cafeicultores, a IPR 99 e a IA 59, apresentaram bandas referentes aos carboidratos nas variáveis responsáveis por suas diferenciações. Enquanto que nenhuma banda de carboidrato é encontrada em destaque para os demais genótipos. Além disso, apenas o genótipo IA 59 tem banda de trigonelina nas variáveis responsáveis por suas diferenciações.

Os resultados para a validação deste modelo PLS-DA estão apresentados por classe na Figura 9. Os valores de *threshold* foram de 0,4859, 0,3368, 0,3662 e 0,4984 para as classes IPR 105, IPR 106, IPR 99 e IA 59, respectivamente, e estão representados no gráfico pela linha tracejada superior. Apenas uma amostra de validação do genótipo IPR 106 foi classificada erroneamente como sendo da IPR 99. Estes resultados indicam uma classificação correta total das amostras de validação de 94,4%, que é um valor muito bom para amostras que apresentam uma base genética muito semelhante.

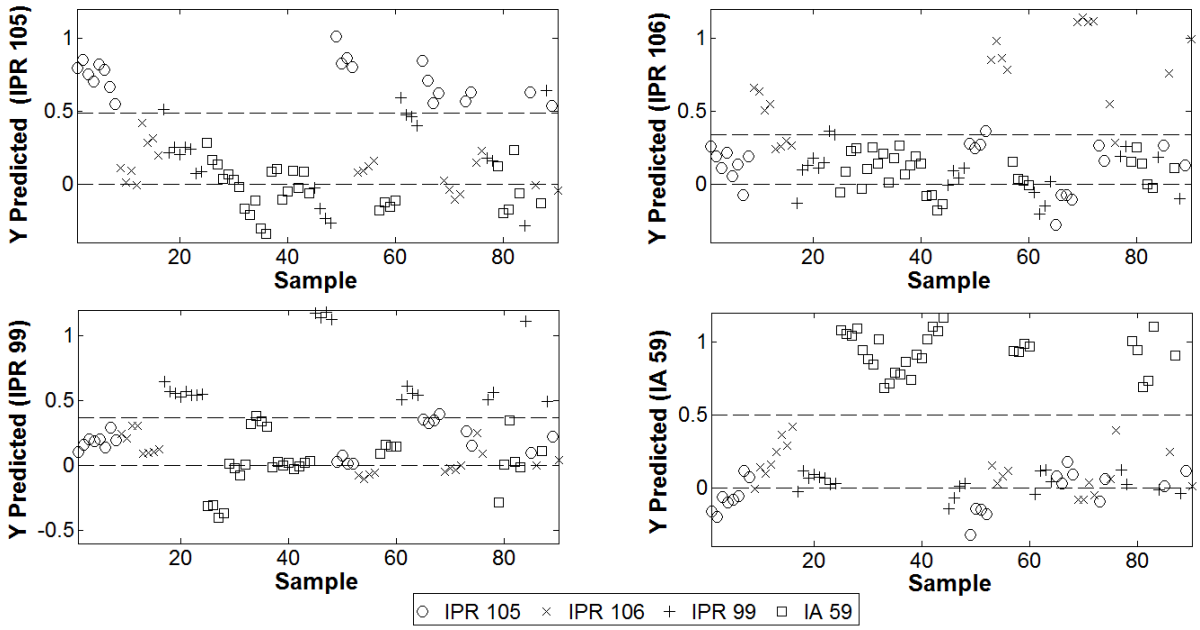


Figura 9. Amostras classificadas pelo modelo PLS-DA, espectros tratados com MSC e 2ª derivada conjuntamente. (a) Classe 1 – IPR 105, (b) classe 2 – IPR 106, (c) classe 3 – IPR 99 e (d) classe 4 – IA 59.

A alta porcentagem de classificação correta das amostras, bem como os resultados satisfatórios de RMSEC, RMSEP, sensibilidade e especificidade indicam que o modelo PLS-DA obtido pode ser utilizado para classificar genótipos de café arábica.

O perfil espectral obtido no NIR é altamente influenciado pelas condições ambientais.[38] Estes efeitos estão presentes quando se pretende discriminar os genótipos provenientes destes quatro locais. Provavelmente, as diferenças espectrais causadas pelas interferências ambientais foram maiores que aquelas atribuídas às diferenças genéticas entre os genótipos, o que reduziu os valores de sensibilidade e especificidade para a classificação genotípica, quando comparado à geográfica. Além disso, os genótipos modernos, como os analisados neste estudo, apresentam estreita base genética, ou seja, composição genética similar.[39,6]

1.4 CONCLUSÃO

O pré-processamento que resultou nos melhores modelos PLS-DA foi a correção do espalhamento multiplicativo e a segunda derivada conjuntamente, comprovando que os espectros NIRS sofrem influência do espalhamento de luz e da mudança de inclinação das amostras.

Tanto geograficamente quanto genotípicamente, o método PLS-DA foi capaz de classificar as amostras de café arábica. Em ambos os casos, apenas uma amostra de validação foi classificada erroneamente, obtendo-se uma classificação correta total de 94,4%.

Entretanto, a sensibilidade e a especificidade das amostras de calibração e validação, fatores que avaliam o desempenho do modelo, foram maiores na discriminação por local de cultivo, sugerindo um modelo com melhor performance.

O estudo mostra que, para a classificação geográfica, os fatores ambientais, como: condições climáticas, altitude, tipo de solo, são responsáveis por variações na composição química do grão, influenciando no sabor e no aroma do café como bebida. Na discriminação das amostras de Mandaguari destacaram-se picos com presença de lipídeos, que são responsáveis por reter o aroma no café; as de Londrina por picos com presença de cafeína; e as de Paranavaí e Cornélio Procópio, picos com ácidos clorogênicos. Sendo que, cafés de melhor qualidade estão relacionado com maiores teores de trigonelina, sacarose, lipídeos e aminoácidos; e menores teores de cafeína e ácidos clorogênicos.

A alta porcentagem de classificação correta obtida no modelo PLS-DA por origem genotípica mostra os benefícios da utilização das informações da classe das amostras na construção do modelo, evidenciando as vantagens de um método supervisionado para classificação. Os resultados obtidos indicam que mesmo os genótipos tendo composição genética similiar, os espectros NIRS contêm informações importantes para a discriminação por genótipo. Percebeu-se que há diferença entre os genótipos de café já disponibilizadas para os cafeicultores pelo IAPAR, IPR 99 e IA 59, quando comparados às demais, IPR 105 e IPR 106. Sendo que o genótipo IA 59 destacou-se pela presença de bandas com trigonelina nas variáveis responsáveis por suas diferenciações, sugerindo um café de melhor qualidade.

1.5 REFERÊNCIAS

1. J. Berthaud, and A. Charrier, *Genetic Resources of Coffea*. Elsevier Aplied Science, London, UK (1988).

2. R. J. Clarke, and O. G. Vitzthum, *Coffee Recent Developments*. Blackwell Science Ltda, Oxford, UK (2001).

3. A. Farah, "Coffee as speciality and functional beverage", in *Speciality and functional beverages*, Ed by P. Paquin. CRC press, Cambrige, USA (2009).

4. P. Lashermes, and F. Anthony, "Genome mapping and molecular breeding in plants", in *Technical crops*, Ed by C. Kole. Springer Berlin Heidelberg, Berlin, DE, p. 109 (2007).

5. C.-L. Ky, J. Louarn, S. Dussert, B. Guyot, S. Hamon, and N. M., "Caffeine, trigonelline, chlorogenic acids and sucrose diversity in wild *Coffea arabica* L. and *C. canephora* P. accessions", *Food Chemistry* **75**, 223 (2001)

6. T. Sera, "Coffee genetic breeding at IAPAR", *Crop Breeding and Applied Biotechnology* **1**, 179 (2001).

7. M. R. Malta, and S. J. d. R. Chagas, "Evaluation of non-volatile compounds in different cultivars of coffee cultivated in southern Minas Gerais", *Acta Scientiarum. Agronomy* **31**, 57 (2009).

8. M. B. S. Scholz, V. R. G. Figueiredo, J. V. N. Silva, and C. S. G. Kitzberger, "Características físico-químicas de grãos verdes e torrados de cultivares de café (Coffea arábica L.) do IAPAR", *Coffee Science* **6**, 245 (2011).

9. J. S. Ribeiro, M. M. C. Ferreira, and T. J. G. Salva, "Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy", *Talanta* **83**, 1352 (2011).

10. A. Stalmach, W. Mullen, C. Nagai, and A. Crozier, "On-line HPLC analysis of the antioxidant activity of phenolic compounds in brewed, paper-filtered coffee", *Braz. J. Plant Physiol.* **18**, 253 (2006).

11. A. Haiduc, C. Gancel, and V. Leloup, "NIR-based Determination of differences in green coffee chemical composition due to geographical", (2010). Available at http://www.heliospir.net/medias/upload/8eme_heliospir_Haiduc.pdf.

12. R. Teuber, "Geographical indications of origin as a tool of product differentiation: the case of coffee", *Journal of International Food & Agribusiness Marketing* **22**, 277 (2010).

13. R. D. M. C. Amboni, A. Francisco, and E. Teixeira, "Utilização de microscopia eletrônica de varredura para detecção de fraudes em café torrado e moído", *Ciência e Tecnologia de Alimentos* **19**, 311 (1999).

14. K. M. Tavares, R. G. F. A. Pereira, C. A. Antônio Nunes, A. C. M. Pinheiro, M. P. Rodarte, and M. C. Guerreiro, "Mid-infrared spectroscopy and sensory analysis applied to detection of adulteration in roasted coffee by addition of coffee husks", *Química Nova* **35**, 1164 (2012).

15. I. Esteban-Díez, J. M. González-Sáiz, C. Sáenz-González, and C. Pizarro, "Coffee varietal differentiation based on near infrared spectroscopy", *Talanta* **71**, 221 (2007).

16. P. Valderrama, *Avaliação de figuras de mérito em calibração multivariada na determinação de parâmetros de controle de qualidade em indústria alcooleira por espectroscopia no infravermelho próximo.* Unicamp, Campinas, BR (2005).

17. J. Workman, and L. Weyer, *Practical Guide to Interpretive Near-Infrared Spectroscopy.* CRC Press, (2007).

18. Metrohm NIRSystems, *A guide to near-infrared spectroscopic analysis of industrial manufacturing processes.* Metrohm AG, Herisau, CH (2013).

19. L. Alessandrini, S. Romani, G. Pinnavaia, and M. D. Rosa, "Near infrared spectroscopy: An analytical tool to predict coffee roasting degree", *Analytica Chimica Acta* **625**, 95 (2008).

20. J. R. Santos, M. C. Sarraguça, A. O. S. S. Rangel, and J. A. Lopes, "Evaluation of green coffee beans quality using near infrared spectroscopy: a quantitative approach", *Food Chemistry* **135**, 1828 (2012).

21. T. Sera, L. H. Shigueoka, G. H. Sera, J. A. Azevedo, F. G. Carvalho, and E. Andreazi, "Nova seleção da cultivar de café Iapar 59 com grãos mais graúdos", in *Simpósio de Pesquisa dos Cafés do Brasil*, Araxá, BR, (2011).

22. G. H. Sera, T. Sera, I. C. d. B. Fonseca, and D. S. Ito, "Resistance to leaf rust in coffee cultivars", *Coffee Science* **5**, 59 (2010).

23. P. H. Caramori, J. H. Caviglione, M. S. Wrege, S. L. Gonçalves, R. T. Faria, A. F. Androcioli, T. Sera, J. C. D. Chaves, and M. S. Koguishi, "Climatic risk zoning for coffee (Coffea arabica L.) in Paraná state, Brazil", *Revista Brasileira de Agrometeorologia* **9**, 486 (2001).

24. Brasil, Regulamento técnico de identidade e de qualidade para a classificação do café beneficiado e de grão verde. 2003, Instrução Normativa n° 8 (2011).

25. J. Huang, S. Romero-Torres, and M. Moshgbar, Practical Considerations in data pre-treatment for NIR and Raman spectroscopy. *American Pharmaceutical Review*, (2010).

26. T. Isaksson, and T. Næs, "The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy", Applied Spectroscopy **42**, 1273 (1988).

27. A. Savitzky, and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures", *Analytical Chemistry* **36**, 1627 (1964).

28. M. Barker, and W. Rayens, "Partial least squares for discrimination", *Journal of Chemometrics* **17**, 166 (2003).

29. S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis", *Chemometrics and Intelligent Laboratory Systems* **2**, 37 (1987).

30. M. Bassbasi, M. De Luca, G. Ioele, A. Oussama, and G. Ragno, "Prediction of the geographical origin of butters by partial least square discriminant analysis (PLS-DA) applied to infrared spectroscopy (FTIR) data", *Journal of Food Composition and Analysis*, (2014).

31. S. Masoum, D. J. R. Bouveresse, J. Vercauteren, M. Jalali-Heravi, and D. N. Rutledge, "Discrimination of wines based on 2D NMR spectra using learning vector quantization neural networks and partial least squares discriminant analysis", *Analytica Chimica Acta* **558**, 144 (2006).

32. P. Geladi, and B. R. Kowalski, "Partial least-squares regression: a tutorial", *Analytica Chimica Acta* **185**, 1 (1986).

33. M. R. Almeida, C. H. V. Fidelis, L. E. S. Barata, and R. J. Poppi, "Classification of Amazonian rose wood essential oil by Raman spectroscopy and PLS-DA with reliability estimation", *Talanta* **117**, 305 (2013).

34. C.M. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, US, (2006).

35. R. S. N. Paganotti, *Desenvolvimento de métodos analíticos para a análise de própolis utilizando técnicas espectrométricas e análise multivariada*. UFMG, Belo Horizonte, BR (2013).

36. P. Vaast, B. Bertrand, J.-J. Perriot, B. Guyot, and M. Génard, "Fruit thinning and shade improve bean characteristics and beverage quality of coffee (Coffea arabica L.) under optimal conditions", *Journal of the Science of Food and Agriculture* **86**, 197 (2006).

37. R. Garrett, E. M. Schmidt, L. F. P. Pereira, C. S. G. Kitzberger, M. B. S. Scholz, M. N. Eberlin, and C. M. Rezende, "Discrimination of arabica coffee cultivars by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry and chemometrics", *LWT - Food Science and Technology* **50**, 496 (2013).

38. H. Posada, M. Ferrand, F. Davrieux, and B. Bertrand, "*Near infrared spectral signature and their stability across environments*", in *22nd International Conference on Coffee Science*. Campinas, BR (2009).

39. D. Steiger, C. Nagai, P. Moore, C. Morden, R. Osgood, and R. Ming, "AFLP analysis of genetic diversity within and among *Coffea arabica* cultivars", *Theoretical and Applied Genetics* **105**, 209 (2002).

CAPÍTULO 2:

Classificação de genótipos de café arábica utilizando espectroscopia de infravermelho próximo e modelos de dois estágios

2 CLASSIFICAÇÃO DE GENÓTIPOS DE CAFÉ ARÁBICA UTILIZANDO ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO E MODELOS DE DOIS ESTÁGIOS

Resumo

A qualidade do café depende das condições ambientais do seu cultivo, fatores como clima, tipo de solo e altitude, associados a práticas agrícolas, influenciam diretamente na composição química final do grão. Este estudo desenvolveu modelos de dois estágios para comprovar a origem geográfica e genotípica de grãos verdes de café arábica. Para o primeiro estágio os métodos lineares mínimos quadrados parciais com análise discriminante (PLS-DA) e análise de componentes principais (ACP) foram testados. Como segundo estágio, duas redes neurais, modelos não-linares, o perceptron de múltiplas camadas (MLP) e a rede de função de base radial (RBF) foram avaliadas. Amostras de quatro genótipos, cultivadas em diferentes cidades do estado do Paraná, Brasil, foram analisados utilizando a espectroscopia no infravermelho próximo (NIRS), na região de 1100 a 2498 nm. Após a aquisição dos espectros três métodos de tratamento dos mesmos foram empregados, a correção do espalhamento multiplicativo (MSC), a segunda derivada (2ªd) e a coobinção de ambos (MSC+2ªd). Os melhores modelos foram os obtidos com os espectros tratados usando MSC e 2ª derivada, o PLS-DA como primeiro estágio seguido por uma rede RBF. Obteve-se assim, uma sensibilidade e especificidade de 100% para as amostras de treinamento e de teste tanto por origem geográfica e quanto por genotípica. Os espectros na região do infravermelho próximo apresentaram uma melhor separação das classes quando comparados com aqueles obtidos no infravermelho médio (FTIR). Estes resultados indicam que os espectros NIRS, aliados a técnicas adequadas de reconhecimento de padrões, podem ser utilizados como uma técnica rápida e eficiente para a classificação de amostras de café arábica por genótipo e local de cultivo.

Palavras-chave: correção do espalhamento multiplicativo, análise de componentes principais, PLS-DA, redes neurais artificiais, simplex sequencial.

2.1 INTRODUÇÃO

Os grãos de café da espécie *Coffea arabica* são conhecidos devido ao seu aroma e doçura intensos, menor amargor e melhor sabor, produzindo uma bebida de melhor qualidade quando comparado ao café robusta, sendo mais apreciado pelo consumidor e com maior valor agregado (Lashermes & Anthony, 2007; Ky, Louarn, Dussert, Guyot, Hamon & Noirot, 2001). Além da seleção adequada dos genótipos de café, as condições ambientais de cultivo, como clima, tipo de solo e altitude; e as práticas agrícolas influenciam diretamente na qualidade da bebida, pois são responsáveis pela composição química final do grão (Farah, 2009; Farah & Donangelo, 2006). Cafés de qualidade superior estão relacionados com teores maiores de sacarose, lipídeos, aminoácidos e trigonelina; e menores de ácidos clorogênicos e cafeína, que são responsáveis por aumentar o amargor do café (Ky, Louarn, Dussert, Guyot, Hamon & Noirot, 2001; Stalmach, Mullen, Nagai, & Crozier, 2006). Associar a composição química e a qualidade ao local de cultivo é uma forma de agregar valor ao café em mercados altamente competitivos (Haiduc, Gancel & Leloup, 2010; Teuber, 2010). Para garantir ao consumidor a origem geográfica e genética do café, métodos analíticos rápidos e eficientes são cada vez mais requeridos. Uma técnica muito utilizada em análises de café é a espectroscopia no infravermelho próximo (NIRS, do inglês *Near-Infrared Spectroscopy*) que dispensa o preparo da amostra e permite análises simultâneas (Esteban-Díez, González-Sáiz, Sáenz-González & Pizarro, 2007).

Conseguir classificar corretamente os genótipos de café arábica por genótipo e origem geográfica é uma tarefa complexa devido ao número elevado de variáveis independentes. Por isso, é necessária a utilização de ferramentas de reconhecimento de padrão. A análise de componentes principais (ACP) é muito utilizada para este fim, pois reduz a dimensionalidade dos dados agrupando as informações altamente correlacionadas. Entretanto, devido aos dados serem descritos por combinações lineares, sistemas não lineares não são bem representados. Além disso, a qualidade do resultado pode ser influenciada por amostras discrepantes (Haykin, 2001). Outro método é o mínimos quadrados parciais com análise discriminante (PLS-DA, do inglês *Partial Least Squares*) que se destaca por ser supervisionado, ou seja, utiliza as informações prévias das amostras na descomposição dos dados (Barker & Rayens, 2003). Ainda assim, existe a

posssibilidade destes métodos não apresentarem resultados satisfatórios, se o número de componentes significativos for alto é difícil extrair informações úteis dos dados (Haykin, 2001).

As redes neurais artificiais (RNAs) são um conjunto de métodos matemáticos que vem sendo utilizados para classificação e reconhecimento de padrões. São ferramentas computacionais não lineares capazes de modelar funções extremamente complexas, onde o conhecimento é adquirido pelo treinamento por um processo de aprendizagem (Graupe, 2007; Marini, 2009). São utilizadas para mapear os dados de entrada (*inputs*) em dados desejáveis de saída (*outputs*) e são implementadas utilizando componentes eletrônicos ou por simulação de programação em um computador digital (Haykin, 2001; Priddy & Keller, 2005; Marini, Bucci, Magrì & Magrì, 2008). São aplicadas em diferentes áreas incluindo autentificação de alimentos, análise sensorial e mapeamento de preferência do consumidor (Marini, 2009).

Este estudo teve como objetivo a classificação geográfica e genotípica de grãos verdes de café arábica. Para isto, espectros NIRS foram analisados em modelos de dois estágios, primeiro utilizando um método linear, ACP ou PLS-DA, e em seguida empregando um método não-linear baseado em redes neurais artificiais do tipo perceptron de múltiplas camadas (MLP) ou redes de função de base radial (RBF).

2.2 MATERIAIS E MÉTODOS

Todos os pré-processamentos dos espectros, a ACP, o PLS-DA, as redes neurais artificiais e o método de otimização simplex sequencial foram realizados no *software* MATLAB R2008b (The MathWorks Inc., Natick, USA).

2.2.1 AMOSTRAS DE GENÓTIPOS DE CAFÉ

Foram analisados quatro genótipos de *Coffea arabica* desenvolvidas pelo Instituto Agronômico do Paraná (IAPAR): IPR 99, IPR 105, IPR 106 e Iapar 59 (IA 59). O genótipo IA 59 foi lançada em 1994, originada do cruzamento entre *Coffea arabica*, “Villa Sarchi 971/10” e o “Híbrido de Timor 832/2”, assim como o IPR 99, apresenta resistência aos tipos de ferrugem conhecidos (Sera, Shigueoka, Sera,

Azevedo, Carvalho & Andreazi, 2011). O genótipo IPR 105 é derivado do genótipo Catuaí; o genótipo IPR 106 originou-se do genótipo Icatu, ambos igualmente resistentes a ferrugens em diferentes níveis. Dentre estes genótipos, apenas o genótipo Iapar 59 e a IPR 99 foram disponibilizados para os cafeicultores (Sera, Sera, Fonseca & Ito, 2010).

Foram utilizadas 18 amostras de cafés cultivadas em quatro locais: Cornélio Procópio (CP), Paranavaí (PV), Mandaguari (MD) e Londrina (LD), todos no estado do Paraná, Brasil. Foi utilizada uma amostra de cada genótipo por cidade, com exceção das amostras do genótipo IA 59 cultivadas em Paranavaí e Cornélio Procópio, onde foram utilizadas duas amostras por cidade. A Figura 1 apresenta a localização das cidades onde os cafés foram cultivados.

Figura 1 - Mapa da distância entre as cidades dos cafés estudados.
Fonte: Google maps, 2013.

As amostras de Paranavaí e Cornélio Procópio foram colhidas na safra de 2008 e as de Londrina e Mandaguari foram colhidas na safra de 2010. As condições climáticas, bem como a latitude, longitude e altitude são mostradas na Tabela 1 (Caramori et al., 2001).

Após a colheita, as amostras foram enviadas a estação experimental do IAPAR em Londrina, onde foram colocadas em caixas de madeira com uma malha de fundo e movidas oito vezes por dia até obter uma umidade dos grãos de 11-12%. Em seguida, as amostras foram beneficiadas, removendo a casca e o pergaminho (Brasil, 2011). Os grãos verdes beneficiados foram moídos (0,5 mm) e armazenados em um freezer a -18ºC, para serem analisados posteriormente.

Tabela 1 - Condições climáticas das cidades.

Cidade	Latitude	Longitude	Altitude	Temperatura média anual
Mandaguari	23°32'52"S	51°40'15"W	650 m	20-21°C
Londrina	23°18'36"S	51°09'56"W	585 m	21-22°C
Cornélio Procópio	23°10'51"S	50°38'48"W	658 m	21-2 2°C
Paranavaí	23°04'22"S	52°27'55"W	470 m	22-23°C

2.2.2 ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO

Os espectros de café verde foram obtidos em um espectrofotômetro de infravermelho próximo NIRSystem 5000-M (Foss Tecator AB, Höganäs, Suécia). As leituras foram feitas em temperatura ambiente (23^{o}C), na faixa de comprimento de onda de 1100 a 2498nm em intervalos de 2nm. Para cada amostra de café foram realizadas 5 repetições, obtendo-se assim um total de 90 espectros. O *software* WinISI III versão 1.50e (Foss NIRSystems/Tecator Infrasoft International, LLC, Silver Spring, MD, USA) foi utilizado para aquisição dos espectros. A absorbância foi obtida como logarítmo decimal do inverso da transmitância, log (1/T). Para o treinamento das redes neurais artificiais, foram utilizados 72 espectros tratados (80%) como amostras de treinamento e 18 espectros como amostras de teste (20%), uma repetição, escolhida aleatoriamente, de cada um dos cafés estudados.

2.2.3 PROCESSAMENTO DE DADOS

Duas transformações foram realizadas na matriz de dados dos espectros originais: a correção do espalhamento multiplicativo (MSC, do inglês *Multiplicative Scatter Correction*) (Isaksson & Næs, 1988) e a segunda derivada por meio do algoritmo de *Savitzky-Golay* (Savitzky & Golay, 1964) (7 pontos de janela e polinômio de 2^{o} grau).

A MSC utiliza uma regressão linear das variáveis espectrais versus o espectro médio para corrigir simultaneamente os efeitos aditivos e multiplicativos do espalhamento de luz. Já a segunda derivada remove eventuais problemas devido a mudanças de inclinação entre as amostras (Isaksson & Næs, 1988).

2.2.4 MODELO DE DOIS ESTÁGIOS

2.2.4.1 Primeiro Estágio Linear

Após estes pré-processamentos, como primeiro estágio do modelo de classificação foram empregados a ACP (Wold, Esbensen & Geladi, 1987) e o PLS-DA (Barker & Rayens, 2003), para avaliar qual o melhor modelo preliminar para fornecer as entradas para as redes neurais.

A ACP é um método não supervisionado capaz de reduzir a dimensionalidade dos dados ao agrupar as informações altamente correlacionadas em um novo sistema de eixos, examinar possíveis agrupamentos das amostras de acordo com sua origem genética e geográfica e identificar possíveis *outliers* (Wold, 1987). Esta análise transforma matematicamente os dados espectrais em componentes ortogonais, chamadas componentes principais, cujas combinações lineares mantêm as informações dos dados originais. Com as componentes principais é possível obter novos conjuntos de dados, chamados *scores* e *loadings*. Os *scores* são as projeções das amostras nos novos eixos. E os *loadings* possuem informação do peso de cada variável original na composição dos novos eixos (Matos, Pereira-Filho, Poppi & Arruda, 2003; Valderrama, 2005; Geladi & Kowalski, 1986).

O PLS-DA é um método muito utilizado para a classificação que, por ser supervisionado, utiliza a resposta desejada para cada amostra de treinamento na decomposição dos dados em *scores* e *loadings* (Barker & Rayens, 2003). Neste método é estabelecida uma relação linear entre a variável dependente (**Y**) e a variável independente (**X**). A matriz **X** é decomposta no produto de duas matrizes, *scores* e *loadings,* assim como no ACP. A diferença entre os dois métodos é que no PLS-DA ocorre uma leve rotação no eixo das componentes principais buscando a máxima covariância de **X** com **Y** e os componentes principais passam a ser chamados de variáveis latentes (LV) (Bassbasi, Luca, Ioele, Oussama & Ragno, 2014).

A quantidade de componentes principais empregadas, bem como a quantidade de variáveis latentes, foi um dos parâmetros otimizados através do método simplex sequencial.

2.2.4.2 Normalização

Antes dos vetores de entrada (componentes principais ou variáveis latentes escolhidas) serem alimentados na rede neural, eles foram pré-processados a fim de evitar um erro de *overflow* ou para previnir que as funções de ativação dos neurônios artificiais sejam saturadas (Haykin, 2001). Para isso, foram utilizados os pré-processamentos máximo e mínimo (minimax), que transforma os valores para uma escala de -1 a +1; autoescalonamento, em que os dados são centrados na média e divididos pela variância, assim após a transformação cada variável vai ter média nula e variância igual a 1, limitanto os dados a uma faixa de - 3 a +3; ou vetor unitário, em que o vetor de dados é dividido por sua norma euclidiana, após a transformação a norma de cada variável é 1 (Pérez-Magariño, Ortega-Heras, González-San José & Boger, 2004).

2.2.4.3 Segundo Estágio Não-Linear

Para o segundo estágio do modelo de classificação foram testadas as redes neurais artificiais MLP e RBF, que são modelos não-lineares. Como entrada dessas redes foram utilizados os *scores* da ACP ou as variáveis latentes do PLS-DA, devidamente normalizados.

2.2.4.3.1 Perceptron de Múltiplas Camadas

As redes neurais do tipo Perceptron de Múltiplas Camadas (MLP - *Multi Layer Perceptron*) são amplamente empregadas para classificação de padrões (Bona, Silva, Borsato & Bassoli, 2011; Borsato, Pina, Spacino, Scholz, & Androcioli, 2011; Galão, Borsato, Pinto, Visentainer, & Carrão-Panizzi, 2011; Link, Lemes, Sato, Scholz & Bona, 2012). Esta rede é constituída por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Na camada de entrada há um neurônio para cada componente principal ou variável independente utilizada (Haykin, 2001). A camada oculta é responsável por processar a informação recebida da camada de entrada, separando padrões através da formação de fronteiras de decisão, a quantidade de neurônios nesta camada depende da complexidade do problema e foi definida por otimização (Debska & Guzowska-Swider, 2011). A

camada de saída possui quatro neurônios, um para cada região ou genótipo de café. Deste modo, o vetor de resposta apresenta dimensão igual a 4, para uma amostra pertencente a classe k o k-ésimo valor é igual a 1 e todos os outros são zerados. Na Figura 2, está uma representação de um MLP com duas camadas ocultas.

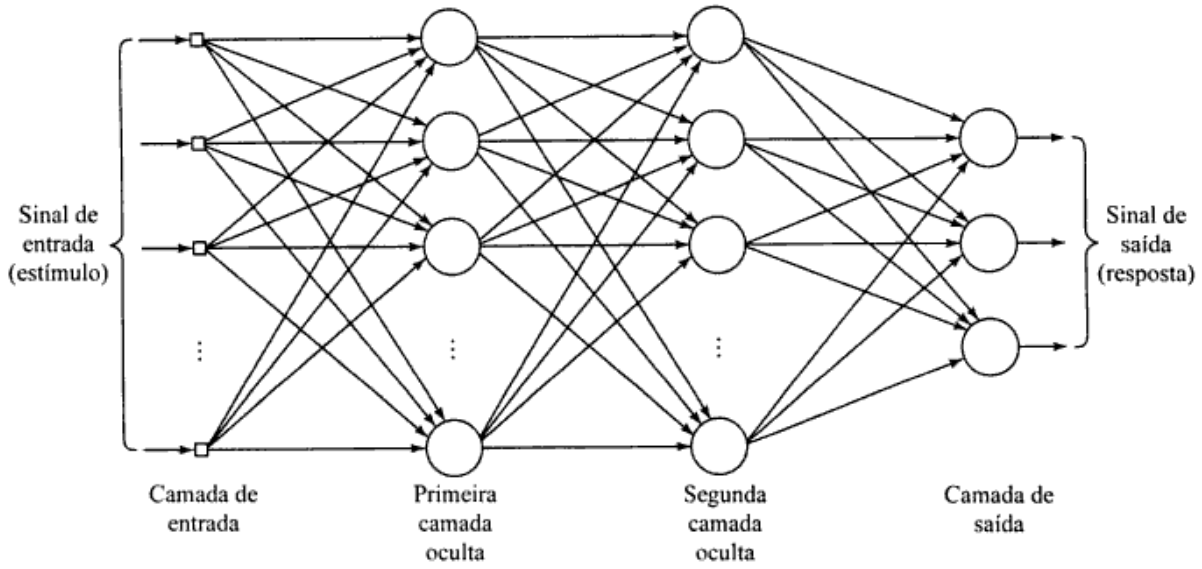


Figura 2 - Perceptron de múltiplas camadas com duas camadas ocultas.
Fonte: Haykin, 2001.

Em cada um dos neurônios das camadas da rede é realizada uma soma ponderada pelos pesos sinápticos dos sinais dos neurônios da camada anterior. A esta soma, chamada de campo local induzido (1), é aplicada uma função de ativação não linear (2) que produz a saída do neurônio (Bishop, 2006; Haykin, 2001).

$$v_i^l = \sum_{j=0}^{m_i} w_{ij}^l y_j^{l-1} \qquad i=1,2,\ldots N^l;\ l=1,2,\ldots,L \tag{1}$$

$$y_i^l = \varphi\left(v_i^l\right) \qquad i=1,2,\ldots N^l;\ l=1,2,\ldots,L \tag{2}$$

Onde i é o neurônio na camada l, w_{ij}^l é o peso sináptico associado ao sinal de entrada y_j^{l-1} do neurônio j da camada anterior (l-1), m_i é o número de entradas associadas ao neurônio i. Para $j=0$, $y_0^{l-1}=1$ e w_{i0}^l é chamado de termo de polarização, ou *bias* que, dependendo se ele é positivo ou negativo, aumenta ou diminui a entrada líquida da função de ativação.

Para a rede MLP desenvolvida foram testadas as funções de ativação do tipo logística (3) ou tangente hiperbólica sigmoide (4) para a camada oculta de neurônios artificiais (Haykin, 2001).

$$\varphi = \frac{1}{1+e^{-\upsilon}} \tag{3}$$

$$\varphi = \frac{2}{1+e^{-2\upsilon}} - 1 \tag{4}$$

O algoritmo de Levenberg-Marquardt foi utilizado para realizar o processo de aprendizagem das redes MLP construídas, tendo a função de modificar ordenadamente os pesos sinápticos da rede. Neste algoritmo há um parâmetro que regula o tamanho do passo das correções do peso, propondo uma solução de compromisso entre o algoritmo do gradiente descendente (retropropagação) e o método de Gauss-Newton (Bishop, 2006).

O processo de aprendizagem ocorre até que os pesos sinápticos e os níveis de *bias* se estabilizem e o erro médio quadrático regularizado (MSE, do inglês *Mean Square Error*), uma função dos parâmetros livres do sistema, tenha convergido a um valor mínimo (Bona, Silva, Borsato & Bassoli, 2011; Haykin, 2001), conforme equação (5).

$$\text{MSE} = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2 + \frac{\lambda}{2}\|w\|^2 \tag{5}$$

onde $\|w\|^2 = w^T w = w_0^2+w_1^2+\ldots+w_M^2$, λ é o parâmetro de regularização que controla a importância relativa do termo de regularização comparado com o termo da soma dos quadrados do erro (Bishop, 2006).

2.2.4.3.2 Rede de Função de Base Radial

Uma rede de função de base radial consiste de uma camada de entrada, uma de saída e uma única camada oculta. Essas três camadas apresentam funções totalmente diferentes. A camada de entrada é formada por nós de fonte, que são unidades sensoriais, conectando a rede ao seu ambiente. A camada oculta realiza

uma transformação não-linear do espaço de entrada para o espaço oculto, sendo que este normalmente é de alta dimensionalidade, é constituída de funções de base radial. A camada de saída é uma combinação linear das funções de base radial e fornece a resposta da rede ao sinal de ativação aplicado à camada de entrada (Haykin, 2001; Buhmann & Ablowitz, 2003).

Existe um neurônio na camada de entrada para cada componente principal ou variável latente utilizada. Os neurônios da camada oculta são representados por φ_k funções de base radial, tanto a quantidade de bases radiais como a quantidade de componentes principais ou variáveis latentes utilizadas na camada de entrada, foram parâmetros determinados pelo simplex sequencial. O número de neurônios ocultos associados a cada classe depende da complexidade dos padrões a serem separados (Tudu, Jana, Metla, Ghosh, Bhattacharyya & Bandyopadhyay, 2009).

A rede RBF possui arquitetura mais simples, quando comparada à MLP, consistindo de duas camadas de pesos, sendo que a primeira possui os parâmetros das funções de base radial e a segunda cria combinações lineares das funções a fim de gerar uma saída. A Figura 3 representa uma rede de função de base radial (Haykin, 2001).

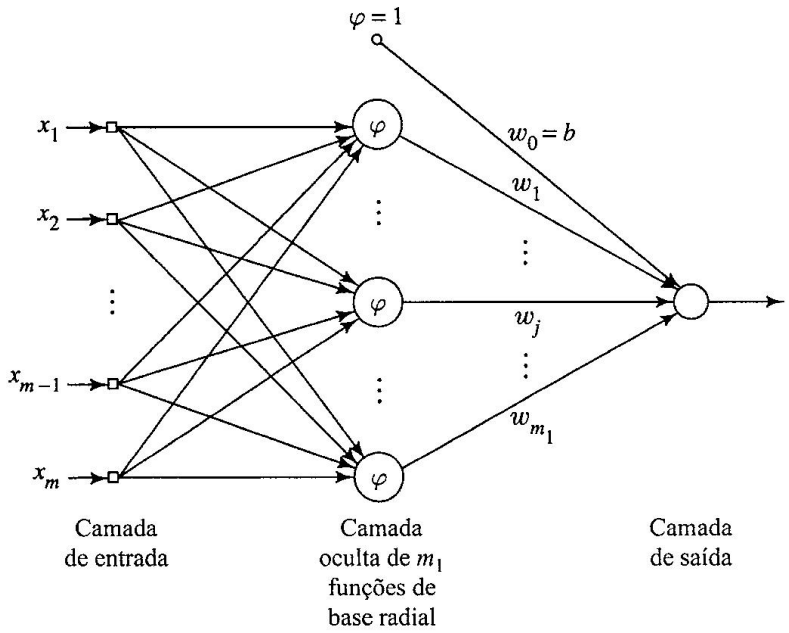


Figura 3 - Representação de uma rede de função de base radial (RBF).
Fonte: Haykin, 2001.

Foram utilizadas neste estudo três tipos de função de base radial, as multiquádricas (6), multiquádricas inversas (7) e as funções Gaussianas (8) (Haykin, 2001). A função mais adequada foi determinada pela otimização simplex sequencial.

$$\varphi(r) = (r^2+\sigma^2)^{\frac{1}{2}} \tag{6}$$

$$\varphi(r) = \frac{1}{(r^2+\sigma^2)^{\frac{1}{2}}} \tag{7}$$

$$\varphi(r) = e^{\left(-\frac{r^2}{2\sigma^2}\right)} \tag{8}$$

Onde r é a distância euclideana entre o centro da base e o vetor da amostra e σ é a largura da base, um parâmetro que controla a suavidade da função de interpolação.

O método de aprendizagem realizado foi híbrido, ou seja, em dois estágios. Primeiramente, a aprendizagem é auto-organizada, estimando as localizações adequadas para os centros das funções de base radial na camada oculta. Por fim, faz-se uma aprendizagem supervisionada de rápida convergência, estimando a largura das bases (dispersão) e os pesos lineares da camada de saída (Bishop, 2006; Haykin, 2001).

O processo de aprendizagem auto-organizada foi realizado neste estudo utilizando um método de agrupamento que separa o conjunto de dados em subgrupos, o mais homogêneo possível. Para isso, foi utilizado o algoritmo de *K-means,* que posiciona os centros das funções de base radial apenas nas regiões densamente povoadas do espaço multidimensional da entrada (Bishop, 2006).

A largura inicial das bases foi determinada de acordo com a equação (9) proposta por Haykin (2001).

$$\sigma = \frac{d_{max}}{\sqrt{2m_1}} \tag{9}$$

Onde m_1 é o número de centros e d_{max} é a distância máxima entre os centros escolhidos. Utilizando a equação (9) evita-se que as funções de base radial sejam pontiagudas ou planas demais. Depois, este valor foi otimizado com o método

de busca direta Quase-Newton, a fim de minimizar o erro médio quadrático para as amostras de treinamento (Beveridge & Schechter, 1987).

Para a determinação dos pesos lineares da camada de saída da rede, foi utilizado um procedimento direto que calcula a pseudo-inversa da matriz ϕ regularizada, conforme equação (10) (Bishop, 2006).

$$w = \left(\lambda I+\phi^{T}\phi\right)^{-1}\phi^{T}t \tag{10}$$

Na equação (10), ϕ^{T} é a matriz transposta de ϕ, que é definida por

$$\phi = \varphi_{k}\left(\left\|x_j-t_k\right\|^{2},\sigma\right) \qquad j=1,2,\ldots N;\ k=1,2,\ldots,m_1 \tag{11}$$

onde x_j representa o j-ésimo vetor de entrada da amostra; k, a k-ésima função de base radial e t_k é o centro da base *k* (Tudu, Jana, Metla, Ghosh, Bhattacharyya & Bandyopadhyay, 2009). O parâmetro de regularização também foi determinado pela otimização simplex sequencial.

2.2.5 Otimização simplex da arquitetura de rede

Em ambas as redes, MLP e RBF, alguns parâmetros (Tabela 2) foram otimizados utilizando o método simplex sequencial a fim de maximizar a porcentagem de classificação correta e reduzir o erro quadrado médio para o conjunto de amostras de validação com o menor modelo possível (Bona, Silva, Borsato & Bassoli, 2011).

Tabela 2- Parâmetros otimizados através do simplex sequencial.

Rede	Parâmetros	Variação
Perceptron de Múltiplas Camadas	Quantidade de neurônios na 1ª camada oculta	2 a 15
	Quantidade de neurônios na 2ª camada oculta	0 a 5
	Função de pré-processamento das entradas	Dados Puros, Minimax, Autoescalonamento ou Vetor Unitário
	Função de ativação utilizada na camada oculta	Logística ou Tangente hiperbólica sigmóide
	Função de ativação utilizada na camada de saída	Logística, Tangente hiperbólica sigmoide ou Linear
	Quantidade de variáveis independentes (entradas)	3 a 20
	Parâmetro de regularização ln (1/λ)	0 a 10
Redes de Função de Base Radial	Quantidade de bases radiais	2 a 15
	Função de pré-processamento das entradas	Dados Puros, Minimax, Autoescalonamento ou Vetor Unitário
	Função de distância para o algoritmo K-*means*	Distância euclidiana, *City Block*, Cosseno ou Correlação
	Parâmetro de regularização ln (1/λ)	0 a 10
	Tipo de base radial	Gaussiana, Multiquadrática ou Multiquadrática inversa
	Quantidade de variáveis independentes (entradas)	3 a 20

A otimização pelo princípio do simplex básico consiste em deslocar uma figura regular, como um triângulo equilátero, sobre uma superfície, quando duas variáveis estão sendo consideradas (Splendley, Himsworth & Hext, 1962). Ela ocorre até que o valor do erro médio quadrático varie no valor estabelecido de 0,001 ou pela avaliação gráfica que auxilia na visualização da otimização, é representada como uma suavização na variação das respostas e variáveis independentes. O algoritmo utilizado neste estudo está descrito em Gao & Han (2012), Link, Lemes, Marquetti, Scholz & Bona, (2014) e Link, Lemes, Sato, Scholz & Bona, (2012), e encontra-se em maiores detalhes no Apêndice B.

A performance do modelo de classificação foi avaliada utilizando um valor limite (*threshold)* que separa as classes. Assim, minimiza-se o número de falsos positivos/negativos para a validação dos dados (Almeida, Fidelis, Barata & Poppi, 2013). O valor do *threshold* corresponde ao encontro das curvas de probabilidade (Figura 4) *a posteriori* encontradas utilizando o teorema de Bayes (Bishop, 2006).

$$p(C_k|y) = \frac{p(y|C_k)p(C_k)}{p(y)} \quad (12)$$

Onde $p(y|C_k)$ é a probabilidade condicional calculada pela distribuição Gaussiana, $p(C_k)$ é a probalilidade *a priori* e $p(y)$ é a constante de normalização.

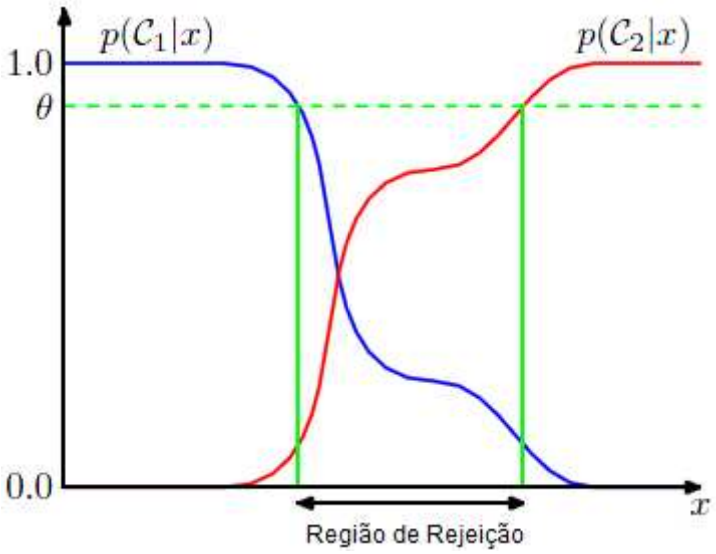


Figura 4 - Curvas de probabilidade *a posteriori*.

Desta maneira, amostras localizadas na região de rejeição devem ter sua classificação avaliada com cuidado.

2.3 RESULTADOS E DISCUSSÕES

A Figura 5-(a) mostra os espectros originais obtidos por NIRS das 18 amostras de café, com 5 repetições cada, perfazendo um total de 90 espectros. A Figura 5-(b) mostra os espectros tratados com MSC para minimizar os efeitos do espalhamento de luz. Os espectros tratados pela segunda derivada, utilizando o algoritmo de *Savitzky-Golay* (Savitzky & Golay, 1964) são mostrados na Figura 5-(c)

e aqueles tratados com a segunda derivada após o tratamento com MSC podem ser visualizados na Figura 5-(d).

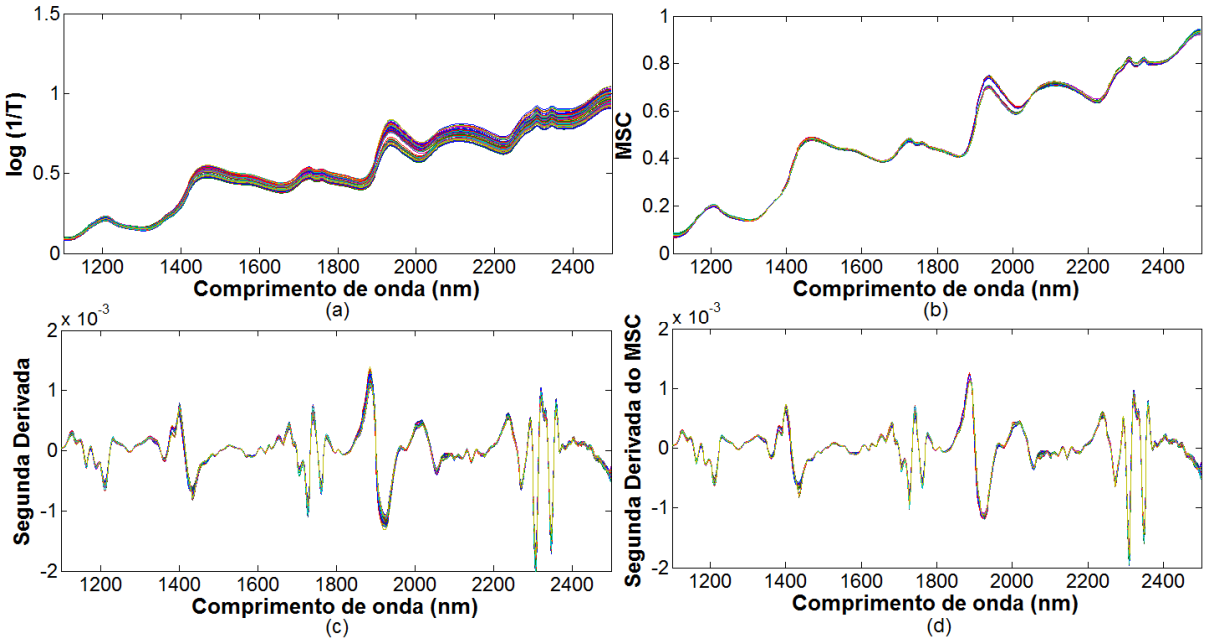


Figura 5 - (a) Espectro das amostras de café; (b) espectros após aplicação de MSC; (c) espectros após aplicação de segunda derivada; (d) espectros com aplicação de MSC e segunda derivada conjuntamente.

Após a realização da otimização dos parâmetros selecionados para a rede MLP (Tabela 2), mil redes foram criadas para cada tratamento dos espectros apresentados na Figura 5-(b), (c), (d) e para cada processamento, ACP e PLS-DA. A melhor rede de cada tipo foi escolhida de acordo com o menor erro médio quadrático e a maior porcentagem de classificação correta para as amostras de teste. Os resultados obtidos para as redes perceptrons de múltiplas camadas propostas para a classificação por origem geográfica e genotípica podem ser visualizados nas Tabelas 3 e 4.

Tabela 3 - Resultados obtidos para os perceptron de múltiplas camadas propostos para a classificação geográfica de café arábica.

	Processamento					
	ACP			PLS-DA		
	Rede A	Rede B	Rede C	Rede D	Rede E	Rede F
Tratamento dos Espectros	MSC	MSC + 2ª derivada	2ª derivada	MSC	MSC + 2ª derivada	2ª derivada
Neurônios (camada oculta)[a]	2	2	3	2	2	2
Função de pré-processamento das entradas[b]	AE	AE	AE	AE	AE	AE
Função de ativação (camada oculta)[c]	THS	FL	THS	FL	FL	THS
Função de ativação (saída)[c]	FL	FL	L	FL	FL	FL
Componentes/Variáveis Latentes	5	5	6	3	4	6
Variância acumulada (**X**)	99,70	93,06	95,30	98,13	89,53	94,61
Variância acumulada (**Y**)	-	-	-	53,01	74,24	86,36
Parâmetro de regularização	0,9994	0,9999	0,9999	1,0000	0,9990	0,9990
Parâmetros livres (pesos)	24	24	37	20	22	26
Desempenho das melhores redes obtidas						
Erro quadrático médio (treinamento)	0,0002	0,0001	$<10^{-4}$	$<10^{-4}$	0,0019	0,0001
Erro quadrático médio (teste)	0,0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0,0025	0,0002
Abordagem Bayesiana						
Classe – CP						
Threshold	0,1846	0,1699	0,0496	0,0100	0,4948	0,3799
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – PV						
Threshold	0,1596	0,4189	0,1527	0,0100	0,4949	0,1484
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – MD						
Threshold	0,1700	0,1399	0,0989	0,0100	0,2430	0,2639
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – LD						
Threshold	0,1296	0,1196	0,1686	0,0100	0,5906	0,1999
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

[a] Todas as redes otimizadas apresentaram zero neurônios na segunda camada oculta.
[b] Função de pré-processamento: MM (Minimax) e AE (Autoescalonamento).
[c] Função de ativação: FL (Função Logística), THS (Tangente Hiperbólica Sigmoide) e L (Linear).

Analisando a Tabela 3, percebe-se que todas as redes MLP classificaram corretamente 100% das amostras de treinamento e teste por origem geográfica. Entretanto, o perceptron de múltiplas camadas que obteve o menor erro médio quadrático com a menor quantidade de parâmetros livres foi a Rede D, em que os espectros foram tratados com MSC e aplicado PLS-DA. Os parâmetros de arquitetura de rede são dependentes do problema, assim a aplicação da otimização simplex possibilita uma escolha automatizada do melhor conjunto de valores para cada caso (Link, Lemes, Marquetti, Scholz & Bona, 2014). No geral, observa-se que a quantidade de componentes principais ou variáveis latentes resultou em variâncias acumuladas maiores que 90% porém a quantidade de parâmetros livres foi menor quando utilizado o PLS-DA no primeiro estágio. Os valores do parâmetro de regularização foram elevados, mesmo para um baixo número de neurônios ocultos, indicando a necessidade de uma suavização do mapeamento. O número de pesos sinápticos em todas as redes foi menor que o número de exemplos utilizados no treinamento da rede neural (72 exemplos), indicando haver graus de liberdade suficiente para que a aprendizagem seja considerada segura sem a ocorrência de sobreajuste.

Tabela 4 - Resultados obtidos para os perceptron de múltiplas camadas propostos para a classificação genotípica de café arábica.

	Processamento					
	ACP			PLS-DA		
	Rede A	Rede B	Rede C	Rede D	Rede E	Rede F
Tratamento dos Espectros	MSC	MSC + 2ª derivada	2ª derivada	MSC	MSC + 2ª derivada	2ª derivada
Neurônios (camada oculta)[a]	[4 4]	3	2	3	[2 3]	3
Função de pré-processamento das entradas[b]	AE	AE	AE	AE	AE	AE
Função de ativação (camada oculta)[c]	THS	THS	THS	THS	THS	FL
Função de ativação (saída)[c]	THS	L	FL	FL	L	FL
Componentes/Variáveis Latentes	4	8	10	7	5	4
Variância acumulada (**X**)	99,42	95,40	97,24	99,87	87,90	90,69
Variância acumulada (**Y**)	-	-	-	50,76	64,20	41,39
Parâmetro de regularização	0,9980	0,9984	0,9924	0,9990	0,9986	0,9997
Parâmetros livres (pesos)	60	43	34	40	37	31
Desempenho das melhores redes obtidas						
Erro médio quadrático (treinamento)	0,0167	0,0009	0,0040	0,0003	0,0016	0,0367
Erro médio quadrático (teste)	0,0363	0,0006	0,0035	0,0090	0,0042	0,0554
Abordagem Bayesiana						
Classe – IPR 105						
Threshold	0,3919	0,4187	0,3379	0,4620	0,3404	0,4102
Sensibilidade (Treinamento)	0,9375	1,0000	1,0000	1,0000	1,0000	0,9375
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	0,9821
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	0,7500
Especificidade (Teste)	0,8571	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – IPR 106						
Threshold	0,4656	0,3311	0,3959	0,4596	0,3500	0,4030
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	0,9643	1,0000	1,0000	1,0000	1,0000	0,9821
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	0,7500
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – IPR 99						
Threshold	0,7243	0,4778	0,5407	0,1997	0,2884	0,2685
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	0,7500
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	0,9464
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	0,7500
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – IA 59						
Threshold	0,7456	0,4987	0,4789	0,0900	0,4613	0,4409
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	0,8333
Especificidade (Teste)	1,0000	1,0000	1,0000	0,9167	1,0000	1,0000

[a] Para redes com duas camadas ocultas a quantidade de neurônios está em sequência
[b] Função de pré-processamento: MM (Minimax) e AE (Autoescalonamento).
[c] Função de ativação: FL (Função Logística), THS (Tangente Hiperbólica Sigmoide) e L (Linear)

Para a classificação genotípica (Tabela 4), três redes classificaram corretamente 100% das amostras de treinamento e teste, as Redes B, D e E. Porém, todas as demais redes apresentaram porcentagem de classificação elevada, mesmo se tratando de genótipos com estreita base genética, ou seja, composição genética similar (Sera, 2001). A Rede C, de maneira geral, apresentou desempenho satisfatório, já que teve elevada porcentagem de classificação com menor quantidade de parâmetros livres quando comparada às demais.

Nesta classificação, todas as melhores redes foram as com função de ativação da camada oculta do tipo tangente hiperbólica sigmóide, indicando ser uma função de ativação adequada. Já para a classificação geográfica, destacaram-se as redes com função de ativação da camada oculta do tipo logística. Com relação à normalização, tanto na classificação geográfica quanto na genotípica, o método autoescalonamento apresentou os melhores resultados. A média das componentes principais ou variáveis latentes foi de 6 para ambas as classificações.

As redes treinadas para a classificação geográfica, no entanto, apresentaram menor número de pesos sinápticos e variáveis de entrada que as para a classificação genotípica, isto ocorre devido às amostras serem da mesma espécie de café, sendo um problema mais difícil de ser modelado. Em Link et al. (2012) esse mesmo comportamento também foi observado. Também foram observados elevados valores do parâmetro de regularização e modelos neurais menores quando usado o PLS-DA como primeiro estágio.

Após a realização da otimização dos parâmetros selecionados para as redes RBF (Tabela 2), mil redes foram criadas para cada tratamento dos espectros apresentados na Figura 5-(b), (c), (d) e para cada processamento, ACP e PLS-DA. A melhor rede de cada tipo foi escolhida de acordo com o menor erro quadrado médio e maior porcentagem de classificação para as amostras de teste. Os resultados obtidos para as redes de base radial propostas para a classificação por origem geográfica e genotípica podem ser visualizados nas Tabelas 5 e 6.

Tabela 5 - Resultados obtidos para as redes de base radial propostas para a classificação geográfica de café arábica.

	Processamento					
	ACP			PLS-DA		
	Rede A	Rede B	Rede C	Rede D	Rede E	Rede F
Tratamento dos Espectros	MSC	MSC + 2[a] derivada	2[a] derivada	MSC	MSC + 2[a] derivada	2[a] derivada
Número de bases radiais	4	6	5	5	4	4
Função de distância para o K-médias[a]	DE	CB	COS	DE	DE	DE
Tipo de base radial[b]	G	G	G	G	G	G
Componentes/Variáveis Latentes	3	9	9	7	9	6
Variância acumulada (**X**)	98,33	95,91	96,98	99,85	95,40	94,61
Variância acumulada (**Y**)	-	-	-	86,77	96,25	86,36
Função de normalização[c]	MM	VU	MM	MM	AE	AE
Parâmetro de regularização	0,0724	0,0002	0,0022	$<10^{-4}$	0,0118	$<10^{-4}$
Largura das bases de radiais	0,5665	0,2669	1,1019	0,9746	4,9731	2,5569
Parâmetros livres (pesos)	20	28	24	24	20	20
Desempenho das melhores redes obtidas						
Erro médio quadrático	0,0261	0,0285	0,0214	0,0090	0,0075	0,0196
Erro médio quadrático	0,0253	0,0262	0,0250	0,0118	0,0145	0,0208
Abordagem Bayesiana						
Classe – CP						
Threshold	0,4211	0,4059	0,5041	0,3677	0,4463	0,5254
Sensibilidade (Treinamento)	0,9000	0,9500	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	0,9808	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – PV						
Threshold	0,3965	0,4458	0,4704	0,4410	0,4008	0,4527
Sensibilidade (Treinamento)	1,0000	0,9500	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – MD						
Threshold	0,4092	0,4345	0,4495	0,4545	0,5374	0,4877
Sensibilidade (Treinamento)	0,9375	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	0,9821	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – LD						
Threshold	0,4440	0,4280	0,3597	0,3581	0,3104	0,3655
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	0,9821	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

[a] Função de distância para o K-médias: DE (Distância euclidiana), CB (*City Block*), COS (Cosseno) e COR (Correlação).
[b] Tipo de base radial: G (Gaussiana), MQ (Multiquadrática), MQI (Multiquadrática inversa).
[c] Função de pré-processamento: DP (Dados Puros), MM (Minimax), AE (Autoescalonamento) e VU (Vetor Unitário).

Para a classificação geográfica do café arábica (Tabela 5), quatro redes RBF classificaram corretamente 100% das amostras de treinamento e teste: as Redes B, D, E e F. Porém a Rede E, treinada com os *scores* do PLS-DA e espectros tratados com MSC e segunda derivada conjuntamente, destacou-se por apresentar menores valores do erro médio quadrático para as amostras de treinamento e teste e menor número de parâmetros livres, indicando ser um modelo mais simples. Mais uma vez a otimização simplex se mostrou eficaz na escolha dos parâmetros de arquitetura de rede. Também observou-se uma redução, em alguns casos, da quantidade de entradas e bases radiais quando feita a comparação entre PLS-DA e ACP. Diferentemente da rede MLP, para a RBF os parâmetros de regularização são pequenos, pois nesse tipo de rede a largura da base já é responsável pela suavização do mapeamento.

Todas as melhores redes para esta classificação foram do tipo Gaussiana, provando que este tipo de base radial é o mais adequado para a segmentação geográfica.

Tabela 6 - Resultados obtidos para as redes de base radial propostas para a classificação genotípica de café arábica.

	Processamento					
	ACP			PLS-DA		
	Rede A	Rede B	Rede C	Rede D	Rede E	Rede F
Tratamento dos Espectros	MSC	MSC + 2ª derivada	2ª derivada	MSC	MSC + 2ª derivada	2ª derivada
Número de bases radiais	11	8	9	9	4	7
Função de distância para o K-médias[a]	COS	COS	COS	COS	COS	CB
Tipo de base radial[b]	MQ	MQ	MQ	G	G	G
Componentes/Variáveis Latentes	17	13	19	13	12	10
Variância acumulada (**X**)	100,00	96,97	98,51	99,99	96,67	96,73
Variância acumulada (**Y**)	-	-	-	89,22	96,10	91,41
Função de normalização[c]	AE	VU	VU	VU	VU	AE
Parâmetro de regularização	0,0630	$<10^{-4}$	0,0004	0,0004	$<10^{-4}$	0,0032
Largura das bases radiais	1,3109	0,2755	0,4387	0,4202	0,5989	4,2084
Parâmetros livres (pesos)	48	36	40	40	20	32
Desempenho das melhores redes obtidas						
Erro médio quadrático	0,0206	0,0423	0,0356	0,0154	0,0072	0,0184
Erro médio quadrático	0,0275	0,0435	0,0463	0,0192	0,0183	0,0309
Abordagem Bayesiana						
Classe – IPR 105						
Threshold	0,4687	0,5124	0,4586	0,4049	0,3821	0,4210
Sensibilidade (Treinamento)	1,0000	0,9375	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	0,9821	0,9821	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – IPR 106						
Threshold	0,4724	0,4570	0,4383	0,3229	0,4704	0,5251
Sensibilidade (Treinamento)	1,0000	1,0000	0,8750	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	0,9643	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	0,7500	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	0,9286	1,0000	1,0000	1,0000
Classe – IPR 99						
Threshold	0,4387	0,4516	0,4415	0,4190	0,4091	0,4603
Sensibilidade (Treinamento)	1,0000	0,9375	0,9375	1,0000	1,0000	1,0000
Especificidade (Treinamento)	0,9821	0,9643	0,9821	0,9643	1,0000	1,0000
Sensibilidade (Teste)	0,7500	0,7500	0,7500	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Classe – IA 59						
Threshold	0,4640	0,4400	0,4706	0,5412	0,4932	0,5104
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

[a] Função de distância para o K-médias: DE (Distância euclidiana), CB (*City Block*), COS (Cosseno) e COR (Correlação).
[b] Tipo de base radial: G (Gaussiana), MQ (Multiquadrática), MQI (Multiquadrática inversa).
[c] Função de pré-processamento: DP (Dados Puros), MM (Minimax), AE (Autoescalonamento) e VU (Vetor Unitário).

Do mesmo modo, para segmentação genotípica (Tabela 6), quatro redes RBF classificaram corretamente 100% das amostras de treinamento e teste: as Redes A, D, E e F. Entretanto a Rede E, treinada com os *scores* do PLS-DA e espectros tratados com MSC e segunda derivada conjuntamente, possui uma quantidade reduzida de parâmetros livres quando comparada às demais. Além disso, apresentou menores valores do erro médio quadrático para as amostras de treinamento e teste.

As melhores redes treinadas com os *scores* da ACP foram, para a classificação genotípica, do tipo multiquádrica, enquanto que as do PLS-DA foram do tipo Gaussiana. Dentre as funções de distância para o algoritmo *K-means*, a função cosseno apresenta uma superioridade em relação às demais funções testadas.

As RBF desenvolvidas tanto com ACP quanto com o PLS-DA apresentaram performance semelhante ao MLP proposto para a classificação geográfica e genotípica, ambas tiveram 100% de classificação das amostras de teste para a melhor rede. De acordo com Haykin (2001), toda rede MLP apresenta uma RBF com desempenho equivalente. No entanto, as redes RBF treinadas com o PLS-DA, utilizando os espectros tratados com MSC e 2^{a} derivada conjuntamente, podem ser consideradas como melhor opção para ambas as classificação, por possuírem um número menor de parâmetros quando comparadas às demais, indicando uma estrutura mais simples e um treinamento mais rápida.

A ACP é um método muito utilizado nas redes neurais artificiais, a fim de reduzir a dimensionalidade dos dados. Entretanto, os melhores resultados obtidos com o PLS-DA mostram as vantagens da utilização de um método supervisionado linear, que utiliza as informações das classes para construção do modelo, como *input* das redes neurais. Atualmente, são escassos os estudos que utilizam o PLS-DA e as redes neurais artificias, Ciosek, Brzozka, Wroblewski, Martinelli, Di Natale & D'Amico (2005) compararam a ACP e o PLS-DA para a utilização em RNA e obtiveram 100% de classificação nas redes que utilizaram como entrada duas matrizes do PLS-DA, uma *scores* e outra de **Y** previsto. Portanto, a utilização do PLS-DA deve ser considerada como opção no estudo das RNA.

A Figura 6 apresenta as respostas geradas por classe pela melhor rede RBF (Rede E) para as amostras classificadas por origem geográfica. Os valores de *threshold Bayesiano*, valor limite que separa as classes, foram de 0,4463, 0,4008,

0,5374 e 0,3104 para as classes CP, PV, MD e LD, respectivamente, e estão representados na figura pela linha horizontal tracejada.

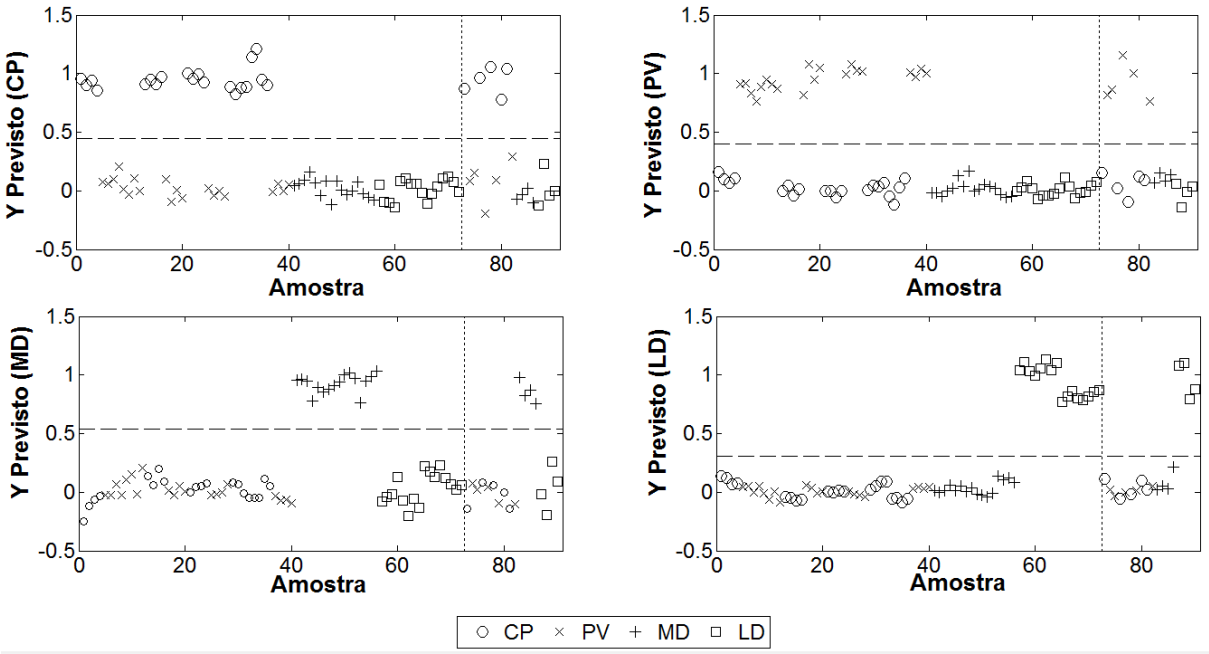


Figura 6- Resposta da rede RBF para classificação geográfica, espectros tratados com MSC e 2ª derivada, PLS-DA como primeiro estágio. A linha pontilhada vertical separa as amostras de treinamento daquelas utilizados para o teste.

É possível verificar que nenhuma amostra da classe analisada ficou abaixo da linha do *threshold*, indicando que todas as classes obtiveram sensibilidade (fator que relaciona as amostras previstas como sendo da classe com as amostras presentes de fato na classe) igual a 1, e nenhuma das demais classes ficou acima da linha do *threshold*, indicando que todas as classes obtiveram especificidade (fator que relaciona as amostras previstas como não sendo da classe com as amostras reais que não são da classe) igual a 1. Já no modelo PLS-DA, uma amostra de teste da classe MD foi classificada erroneamente, obtendo uma sensibilidade para esta classe de 0,750.

Além disso, a rede E não apresentou nenhuma amostra na região de rejeição para todas as classes estudadas, quando analisadas as curvas de probabilidade *a posteriori*, diferentemente das demais redes RBF para a classificação geográfica (Figura 7). Isto mostra que as amostras ficaram distantes do *threshold*, representado pela linha tracejada, o que garante uma maior confiabilidade no modelo.

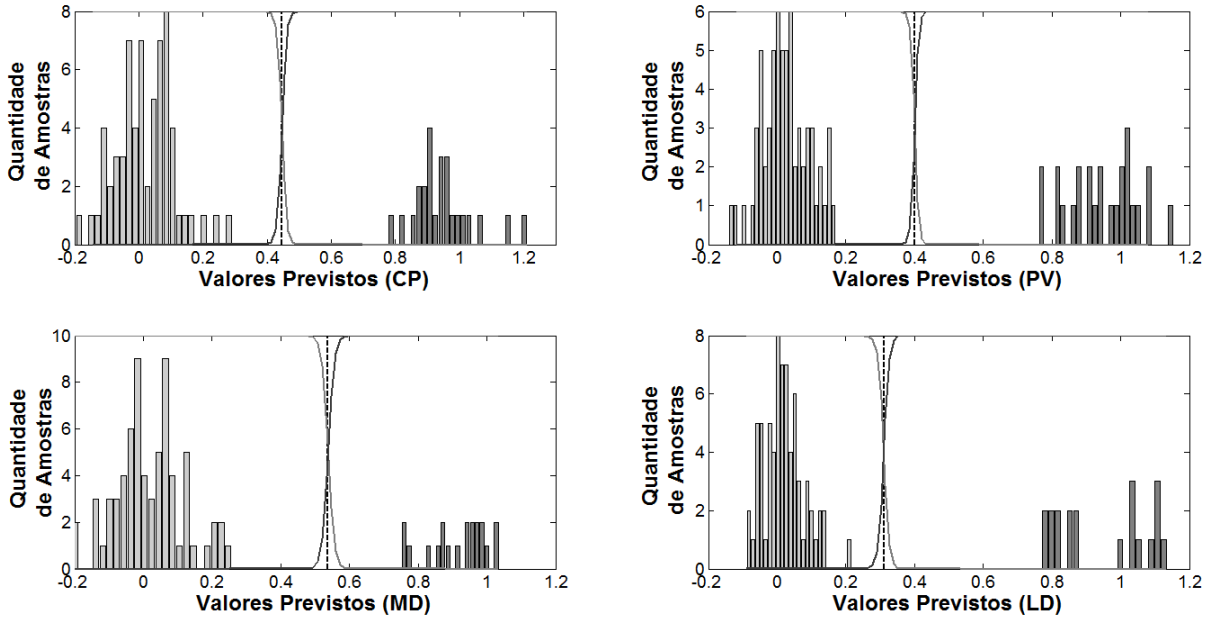


Figura 7 - Curva de probabilidade *a posteriori* por classe para a classificação geográfica.

A Figura 8 apresenta as respostas geradas por classe pela melhor rede RBF (Rede E) para as amostras classificadas por origem genotípica. Os valores de *threshold* foram de 0,3821, 0,5251, 0,4603 e 0,5104 para as classes IPR 105, IPR 106, IPR 99 e IA 59, respectivamente, e estão representados na figura pela linha tracejada. Da mesma maneira, nenhuma amostra da classe analisada ficou abaixo da linha do *threshold*, indicando que todas as classes obtiveram sensibilidade igual a 1, e nenhuma das demais classes ficou acima da linha do *threshold*, indicando que todas as classes obtiveram especificidade igual a 1. Enquanto que, no modelo PLS-DA, apesar de que apenas uma amostra de teste da classe IPR 106 tenha sido classificada erroneamente, o modelo obtido apresentou especificidade abaixo de 1 para classes IPR 105, IPR 106 e IPR 99 durante o treinamento e para a classe IPR 105 no teste.

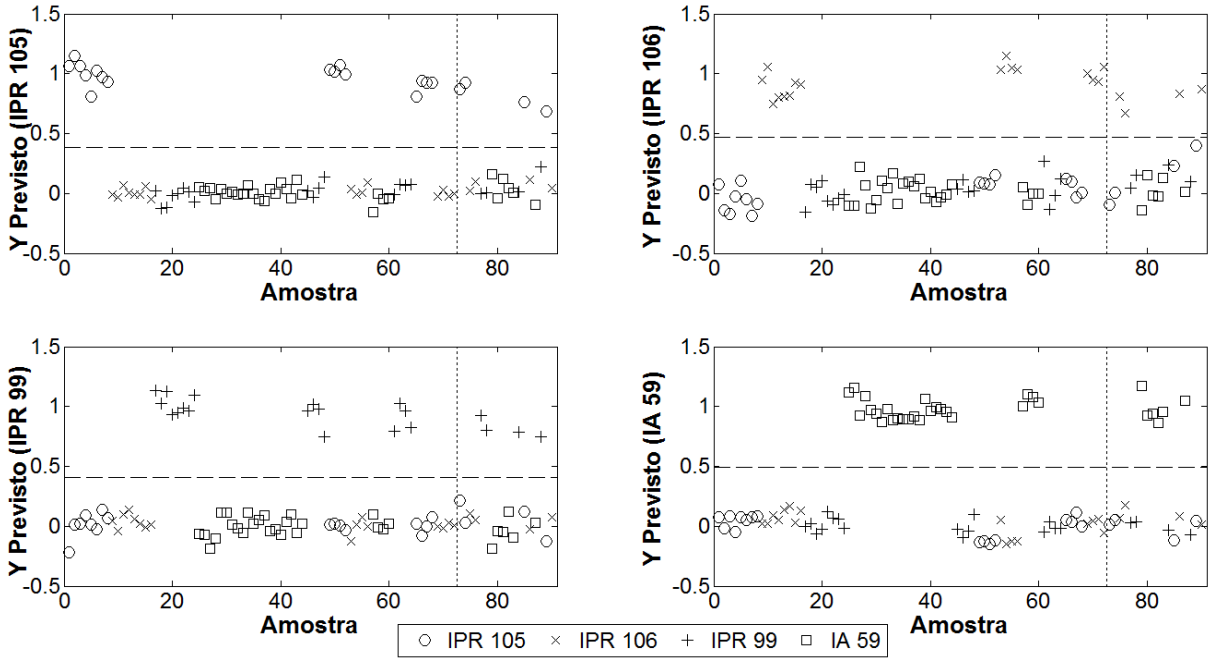


Figura 8 - Resposta da rede RBF para classificação genotípica, espectros tratados com MSC e 2ª derivada, PLS-DA como primeiro estágio. A linha pontilhada vertical separa as amostras de treinamento daquelas utilizados para o teste.

Da mesma maneira, para a classificação genotípica, a rede E não apresentou nenhuma amostra na região de rejeição para todas as classes estudadas, de acordo com as curvas de probabilidade *a posteriori*, diferentemente das demais redes RBF (Figura 9).

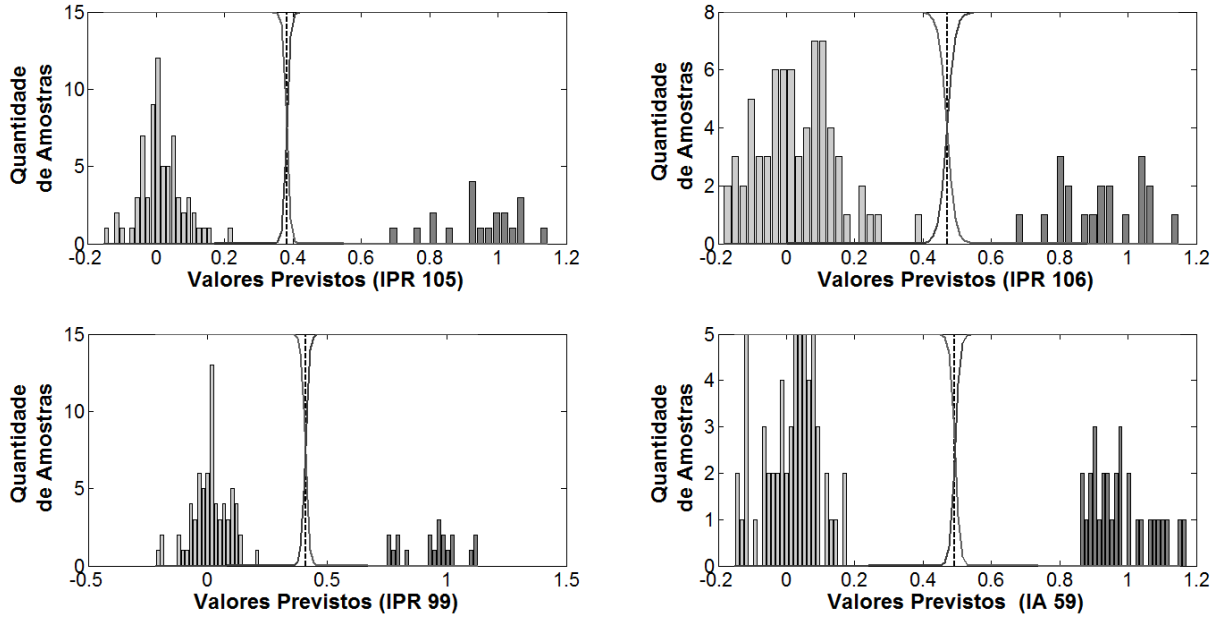


Figura 9 - Curva de probabilidade *a posteriori* por classe para a classificação genotípica.

No estudo realizado utilizando apenas PLS-DA para a classificação, os modelos classificaram corretamente 94,4% das amostras por genótipo e região de cultivo. Com as redes neurais foi possível aumentar a porcentagem de classificação e melhorar os parâmetros que avaliam o desempenho do modelo: a seletividade e a especificidade. Em ambas as melhores redes MLP e RBF, treinadas com o PLS-DA, utilizando os espectros tratados com MSC e 2ª derivada conjuntamente, os valores de seletividade e especificidade foram 1 para todas as classes.

2.3.1 NIRS VERSUS FTIR

As redes MLP e RBF também foram aplicadas em trabalhor prévios às mesmas amostras, entretanto, analisadas com a espectroscopia de infravermelho médio com transformada de Fourier (FTIR, do inglês *Fourier transform infrared spectroscopy*), utilizando apenas a ACP como processamento (Link, Lemes, Marquetti, Scholz & Bona, 2014; Link, Lemes, Sato, Scholz & Bona, 2012). A Tabela 7 mostra a porcentagem de classificação correta para as amostras de teste, analisadas por NIRS e FTIR.

Tabela 7 – Porcentagem de classificação correta das melhores redes MLP e RBF por origem geográfica e genotípica, comparação entre espectros analisados por NIRS e FTIR, ambos processados com ACP.

Espectros	FTIR				NIRS			
	Geográfica	MSE	Genotípica	MSE	Geográfica	MSE	Genotípica	MSE
MLP	100%[a]	0,0353	77,78%[b]	0,0986	100%[c]	0,0001	100%[c]	0,0030
RBF	100%[a]	0,0433	94,44%[b]	0,0840	100%[c]	0,0262	100%[c]	0,0491

[a] Espectros originais.
[b] 1ª derivada dos espectros.
[c] MSC + 2ª derivada.

As amostras analizadas por NIRS apresentaram menores valores de erro médio quadrático e conseguiram classificar corretamente 100% das amostras de teste por genótipo e origem geográfica para ambas as redes, MLP e RBF. Enquanto que, as analisadas por FTIR, conseguiram 100% de classificação correta apenas na classificação geográfica e na classificação genotípica, a rede MLP apresentou um número muito elevado de pesos sinápticos, indicando que a rede não aprendeu de maneira confiável (Link, Lemes, Sato, Scholz & Bona, 2012).

Os resultados obtidos indicam que a espectroscopia no infravermelho próximo é uma técnica adequada para ser empregada na identifiação geográfica e genotípica de café verde, podendo ser utilizada como uma análise alternativa na indústria. Por necessitar de um preparo mínimo das amostras, a NIRS descarta possíveis erros experimentais que possam ocorrer durante as análises, fazendo com que o resultado seja mais preciso. Além disso, possui vantagem de ser uma técnica rápida e que posssibilita a reutilização da amostra após a análise.

2.4 CONCLUSÃO

As redes de base radial e os perceptrons de múltiplas camadas otimizados foram capazes de classificar corretamente todas as amostras de café arábica geograficamente e genotipicamente. Entretanto, as redes de base radial se sobressaíram sobre as demais, principalmente as treinadas com as variáveis latentes do PLS-DA, que tiveram um desempenho satisfatório, de baixo valor do erro médio quadrático e reduzido número de parâmetros livres nas melhores redes. A maior quantidade de informação presente no modelo PLS-DA faz com que a rede precise de uma menor quantidade de pesos sinápticos para realizar a classificação, evidenciando assim, os benefícios de se utilizar um método supervisionado como entrada da rede. O simplex sequencial utilizado provou ser uma eficiente metodologia para a determinação dos parâmetros ótimos, maximizando a performance e e minimizando o tamanho das redes.

Os espectros NIRS, processados com ACP, apresentaram melhores resultados de classificação para ambas as redes, MLP e RBF, ao serem comparados com os espectros do FTIR. As redes MLP treinadas com o FTIR, não conseguiram classificar corretamente as amostras por genótipo. Já as que utilizaram os espectros NIRS, tiveram 100% de classificação e um número menor de pesos sinápticos que o número de exemplos utilizados no treinamento da rede (72 exemplos), problema encontrado no estudo realizado com o FTIR. Sendo assim, a espectroscopia no infravermelho próximo pode ser utilizada como uma técnica alternativa na análise de café, podendo ser realizada de maneira mais eficiente e rápida, evitando a destruição das amostras.

2.5 REFERÊNCIAS

Almeida, M.R., Fidelis, C.H.V., Barata, L.E.S., & Poppi, R.J. (2013). Classification of Amazonian rose wood essential oil by Raman spectroscopy and PLS-DA with reliability estimation. *Talanta,* 117, 305-311.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics,* 17, 166-173.

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, 43, 3-31.

Bassbasi, M., De Luca, M., Ioele, G., Oussama, A. & Ragno, G. (2014). Prediction of the geographical origin of butters by partial least square discriminant analysis (PLS-DA) applied to infrared spectroscopy (FTIR) data. *Journal of Food Composition and Analysis*.

Beveridge, G.S.G., & Schechter, R.S. (1987). *Optimization theory and pratice.* Tokyo: Mac Graw-Hill & Sons.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bona, E., Silva, R.S.S.F., Borsato, D., & Bassoli, D. G. (2011). Optimized Neural Network for Instant Coffee Classification through an Electronic Nose. *International Journal of Food Engineering,* 7.

Borsato, D., Pina, M.V.R., Spacino, K.R., Scholz, M.B.S., & Androcioli, A.F. (2011). Application of artificial neural networks in the geographical identification of coffee samples. *Eur Food Res Technol,* 233, 533-543.

Buhmann, M.D., & Ablowitz, M.J. (2003). *Radial basis functions: Theory and implementations.* United Kingdom: Cambridge University Press.

Brasil, Regulamento técnico de identidade e de qualidade para a classificação do café beneficiado e de grão verde, Instrução Normativa no 8 de 11 jun. 2003, Ministério da Agricultura Pecuária e Abastecimento, 2011. URL http://www.claspar.pr.gov.br/arquivos/File/pdf/cafebenef008_03.pdf .

Caramori, P.H. et al. (2001). Climatic risk zoning for coffee (Coffea arabica L.) in Paraná state, Brazil. *Revista Brasileira de Agrometeorologia*, 9, 486-494.

Ciosek, P., Brzozka, Z., Wroblewski, W., Martinelli, E., Di Natale, C., & D'Amico, A. (2005). Direct and two-stage data analysis procedures based on PCA, PLS-DA and ANN for ISE-based electronic tongue - Effect of supervised feature extraction. *Talanta,* 67, 590-596.

Debska, B., & Guzowska-Swider, B. (2011). Application of artificial neural network in food classification. *Analytica Chimica Acta*, 705, 283-291.

Esteban-Díez, I., González-Sáiz, J. M., Sáenz-González, C., & Pizarro, C. (2007). Coffee cultivar differentiation based on near infrared spectroscopy. *Talanta*, 71, 221-229.

Farah, A. (2009). Coffee as speciality and functional beverage. *Functional and speciality beverage technology*, 370-395.

Farah, A., & Donangelo, C.M. (2006). Phenolic compounds in coffee. *Braz. J. Plant Physiol.*, 18, 23-36.

Galão, O. F., Borsato, D., Pinto, J.P., Visentainer, J.V., & Carrão-Panizzi M.C. (2011). Artificial neural networks in the classification and identification of soybean cultivars by planting region. *J. Braz. Chem. Soc.*, 22, 142-147.

Gao, F., & Han, L. (2012). Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Comput. Optim. Appl.*, 51, 259-277.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta,* 185, 1.

Graupe, D. (2007). *Principles of Artificial Neural Networks: Advanced Series on Circuits and Systems*. (2nd ed).Chicago: World Scientific.

Google Maps, 2013. URL: https://maps.google.com.br/.

Haiduc, A., Gancel, C., & Leloup, V. NIR-based Determination of differences in green coffee chemical composition due to geographical, 2010. URL http://www.heliospir.net/medias/upload/8eme_heliospir_Haiduc.pdf. Accessed 16/02/2014.

Haykin, S. (2001). *Redes Neurais: Princípios e Práticas* (2nd ed.). Porto Alegre:Bookman.

Isaksson, T., & Næs, T. (1988). The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. *Applied Spectroscopy*, 42, 1273-1287.

Ky, C-L., Louarn, J., Dussert, S., Guyot, B., Hamon, S., & Noirot, M. (2001). Caffeine, trigonelline, chlorogenic acids and sucrose diversity in wild *Coffea arabica* L. and *C. canephora* P. accessions. *Food Chemistry*, 75, 223-230.

Lashermes, P., & Anthony, F. (2007). Genome mapping and molecular breeding in plants. In C. Kole (Ed), *Technical crops.* Berlin: Springer Berlin Heidelberg.

Link, J.V., Lemes, A.L.G., Sato, H.P. Scholz, M.B.S., & Bona, E. (2012). Optimized multilayer perceptron for the geographical and genotypic classification of four genotypes of arabica coffee. *Revista Brasileira de Pesquisa em Alimentos* (REBRAPA), 3, 72-81.

Link, J.V., Lemes, A.L.G., Marquetti, I., Scholz, M.B.S., & Bona, E. (2014). Geographical and genotypic classification of arábica coffee using Fourier transform infrared spectroscopy and radial-basis function networks. *Chemometrics and Intelligent Labotory Systems*, 135, 150-156.

Marini, F. (2009). Artificial Neural Networks in foodstuff analyses: Trends and perspectives - A review. *Analytica Chimica Acta*, 635, 121-131.

Marini, F., Bucci, R., Magrì, A.L., & Magrì, A.D. (2008). Artificial neural networks in chemometrics: History, examples and perspectives. *Microchemical Journal*, 88, 178-185.

Matos, G.D. Pereira-Filho, E.R., Poppi, R.J. & Arruda, M.A.Z. (2003). Análise exploratória em química analítica com emprego de quimiometria: PCA e PCA de imagens. *Revista Analytica* 3*,* 38-50.

Pérez-Magariño, S., Ortega-Heras, M., González-San José, M. L., & Boger, Z. (2004). Comparative study of artificial neural network and multivariate methods to classify Spanish DO rose wines. *Talanta*, 62, 983-990.

Priddy, K.L., & Keller, P.E. (2005). *Artificial Neural Network: an introduction.* Washington: SPIE.

Savitzky, A., & Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36, 1627-1639.

Sera, T. (2001). Coffee genetic breeding at IAPAR. *Crop Breeding and Applied Biotechnology,* 1, 179-199.

Sera, G.H., Sera, T., Fonseca, I.C.d.B., & Ito, D.S. (2010). Resistance to leaf rust in coffee cultivars. *Coffee Science,* 5, 59-66.

Sera, T., Shigueoka, L.H., Sera, G.H., Azevedo, J.A., Carvalho, F.G., & Andreazi, E. (2011). Nova seleção da cultivar de café Iapar 59 com grãos mais graúdos. In *Simpósio de Pesquisa dos Cafés do Brasil*, Araxá, BR, (2011).

Splendley, W., Himsworth, F.R., & Hext, G.R. (1962). Sequential application of simplex designs in optimization and evolutionary operation. *Technometrics*, 4, 441-461.

Stalmach, A., Mullen, W., Nagai, C., & Crozier, A. (2006). On-line HPLC analysis of the antioxidant activity of phenolic compounds in brewed, paper-filtered coffee. *Braz. J. Plant Physiol*, 18, 253-262.

Teuber, R. (2010). Geographical indications of origin as a tool of product differentiation: the case of coffee. *Journal of International Food & Agribusiness Marketing*, 22, 277-298.

Tudu, B., Jana, A., Metla, A., Ghosh, D., Bhattacharyya, N., & Bandyopadhyay, R. (2009). Electronic nose for black tea quality evaluation by an incremental RBF network. *Sensors and Actuators B: Chemical, 138*, 90-96.

Valderrama, P. (2005). *Avaliação de figuras de mérito em calibração multivariada na determinação de parâmetros de controle de qualidade em indústria alcooleira por espectroscopia no infravermelho próximo.* Campinas: Unicamp.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2, 37-52.

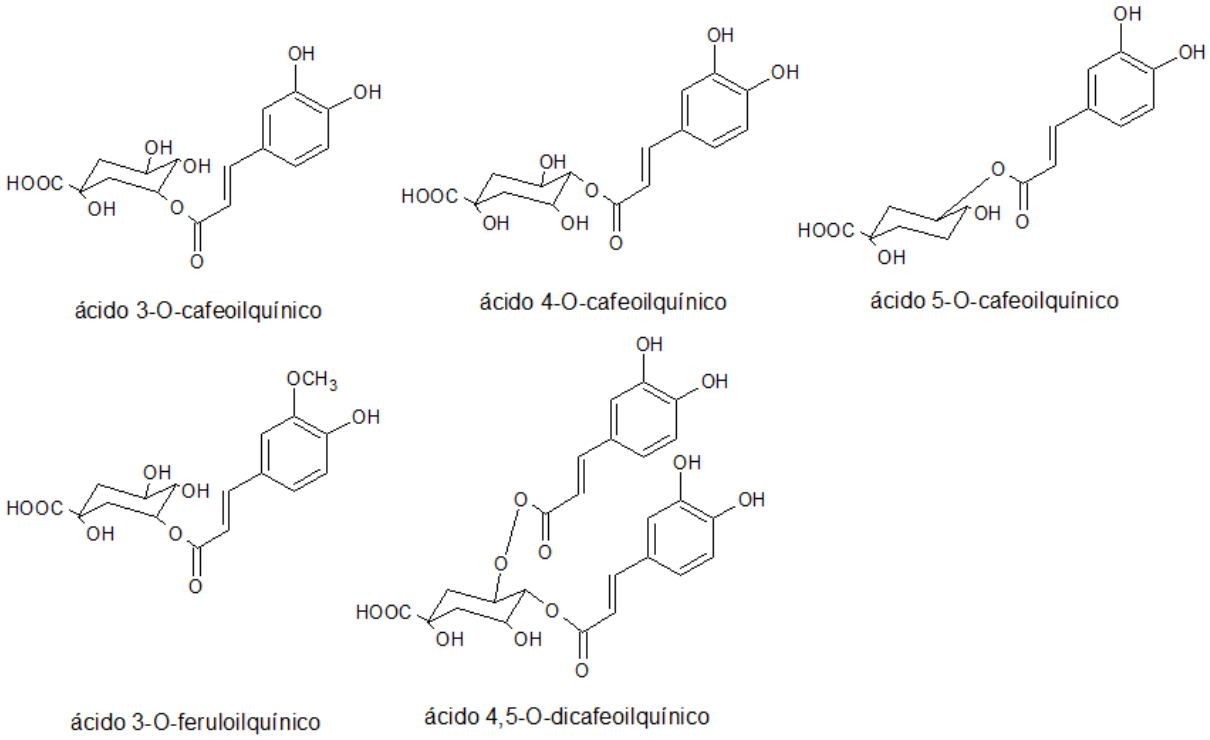
3 APÊNDICE A

3.1 COMPOSIÇÃO QUÍMICA DO CAFÉ

Grande parte do aroma de alimentos ocorre devido a componentes voláteis. O aroma do café é formado basicamento por hidrocarbonetos, alcoóis, aldeídos, éteres, cetonas, furanos, pirróis, pirazinas, ácidos orgânicos, compostos fenólicos (Freitas & Mosca, 1999).

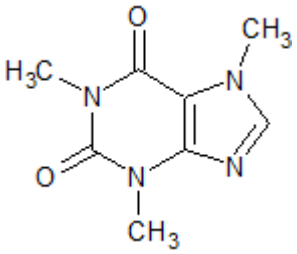
Os compostos fenólicos são metábolitos secundários que representam um papel importante na interação das plantas com o ambiente, protegendo contra a radiação ultravioleta e ataque de patógenos (Farah & Donangelo, 2006; Manach, Scalbert, Morand, Remesy & Jimenez, (2004).

Ácidos clorogênicos (ACG) e seus isômeros são os principais componentes fenólicos dos grãos de café verde, com teores de até 14% em massa. Estes compostos variam com a espécie e genótipo, grau de maturação, práticas agrícolas, clima e solo. Seus isômeros podem ser subdivididos em grupos: ácidos cafeoilquínico, ácidos dicafeoilquínico, ácidos feruloilquínico, ácidos p-coumaroilquínico e ácidos cafeoilferuloilquínicos, (Clifford, 1979) (Figura 1). Geralmente, o *Coffea arabica* apresenta teores totais de ACG menores que o *Coffea canephora* (Farah & Donangelo, 2006).



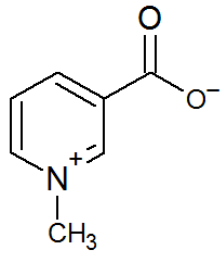
Estruturas dos principais ácidos clorogênicos encontrados no café.
Fonte: Campa, Chrestin, Kochko, Bertrand, Leroy & Noirot, 2001

A cafeína é um alcalóide pertencente ao grupo das xantinas, um tipo de purina, que é um composto orgânico heterocíclico com base nitrogenada, cujo nome segundo a IUPAC (*International Union of Pure and Applied Chemistry*) é 1,3,7-trimetil- 1*H*-purino- 2,6(3*H*,7*H*)-diona (Figura 2), que pode ser isolada de aproximadamente 60 plantas. Em sua forma pura, é um pó cristalino branco, inodoro, e que contribui para o sabor amargo. Em doses moderadas, possui efeito estimulante leve com pequeno risco de efeitos nocivos. Entretanto, seu consumo excessivo está associado a doenças, principalmente cardíacas
(Watson, 2003; Chou & Benowitz, 1994).



Estrutura da cafeína.
Fonte: Watson, 2003

A trigonelina, N-metil-nicotínico, é um alcalóide natural encontrado no café em concentrações relativamente elevadas. Com a torrefação, a trigonelina é transformada em diversos compostos voláteis, como alquilpiridina e niacina, sendo a niacina encontrado no café como ácido nicotínico e nicotinamida (Damodaran, Parkin & Fennema, 2008; Clarke & Macrae, 1985). Os teores de trigonelina no café são altamente dependentes dos genótipos, sendo maiores no *Coffea arabica* do que no *Coffea robusta* (Campa, Ballester, Doulbeau, Dussert, Hamon & Noirot, 2004). A Figura 3 ilustra a estrutura química da trigonelina, ácido nicotínico e nicotinamida.



Trigonelina

Estrutura da trigonelina.
Fonte: Damodaran, Parkin & Fennema, 2008.

3.2 REFERÊNCIAS

M. Costa Freitas, and A. I. Mosca, "Coffee geographic origin — an aid to coffee differentiation", *Food Research International* **32**, 565 (1999).

A. Farah, and C. M. Donangelo, "Phenolic compounds in coffee", *Braz. J. Plant Physiol.* **18**, 23 (2006).

C. Manach, A. Scalbert, C. Morand, C. Remesy, and L. Jimenez, "Polyphenols: food sources and bioavailability", *American Journal of Clinical Nutrition* **79**, 727 (2004).

M. N. Clifford, "Chlorogenic acids—Their complex nature and routine determination in coffee beans", *Food Chemistry* **4**, 63 (1979).

J. Watson, "Caffeine" in *Encyclopedia of Food Sciences and Nutrition (Second Edition)*, Ed by B. Caballero. Academic Press, Oxford, UK, p. 745 (2003).

T. M. Chou, and N. L. Benowitz, "Caffeine and coffee: effects on health and cardiovascular disease", *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology* **109**, 173 (1994).

S. Damodaran, K. L. Parkin, and O. R. Fennema, *Fennema's Food Chemistry.* CRC Press, (2008).

R. J. Clarke, and R. Macrae, in ***Coffee***, Elsevier Applied Science Publishers Ltd, **(1985)** Vol. 3.

C. Campa, J. F. Ballester, S. Doulbeau, S. Dussert, S. Hamon, and M. Noirot, "Trigonelline and sucrose diversity in wild Coffea species", *Food Chemistry* **88**, 39 (2004).

4 APÊNDICE B

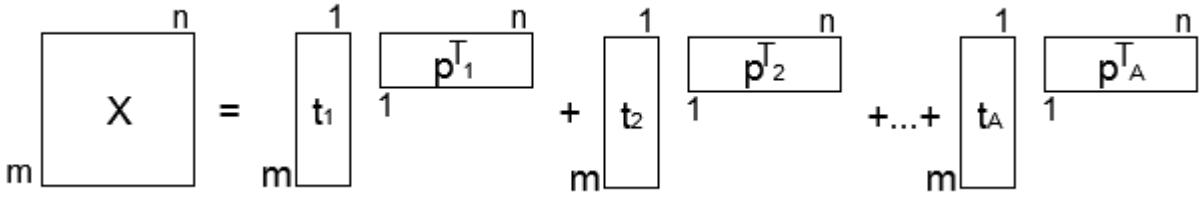
4.1 SEGMENTAÇÃO DO CAFÉ ARÁBICA VERDE USANDO ACP

Depois de realizados estes pré-tratamentos foi utilizada a análise de componentes principais (ACP), um método não supervisionado capaz de reduzir a dimensionalidade dos dados ao agrupar as informações altamente correlacionadas em um novo sistema de eixos; examinar possíveis agrupamentos das amostras de acordo com sua origem genética e geográfica e identificar possíveis *outliers* (Wold, 1987). O método descarta combinações lineares que possuem variâncias pequenas e as responsáveis pela descrição de ruídos instrumentais, retendo apenas termos que têm variâncias significativas (Haykin, 1999).

Esta análise transforma matematicamente os dados espectrais em componentes ortogonais, cujas combinações lineares mantêm as informações dos ados originais. Essas novas variáveis são denominadas componentes principais ou fatores (Bishop, 1995).

Com as componentes principais é possível obter novos conjuntos de dados, chamados *scores* e *loadings*. Os *scores* são as projeções das amostras nos novos eixos. E os *loadings* possuem informação do peso de cada variável original na composição dos novos eixos (Matos, Pereira-Filho, Poppi & Arruda, 2003; Valderrama, 2005; Geladi & Kowalski, 1986).

. A Figura abaixo mostra a decomposição de uma matriz de dados X de dimensão (m x n) em vetores de *scores* (t) e *loadings* (p) (Valderrama, 2005):



Decomposição em A componentes principais por ACP.

Nota: o conteúdo deste apêndice refere-se a um material suplementar retirado do artigo PLS-DA submetido ao JNIRS, mas que será utilizado futuramente em outro artigo.

Com o gráfico dos *scores* dos componentes principais consegue-se observar as semelhanças e diferenças entre as amostras; e a importância das variáveis para o modelo podem ser analisadas no gráfico dos *loadings* (Almeida, Fidelis, Barata & Poppi, 2003).

Entretanto, devem-se levar em conta algumas desvantagens da análise. Os dados são descritos por combinações lineares, deste modo, sistemas não lineares não são bem representados. É preciso avaliar a qualidade do resultado que pode ser influenciada por amostras que deveriam ser desconsideradas. Após a análise, o número de componentes significativos pode ser grande, dificultando a extração de informações úteis, a partir dos dados gerados (Haykin, 2001).

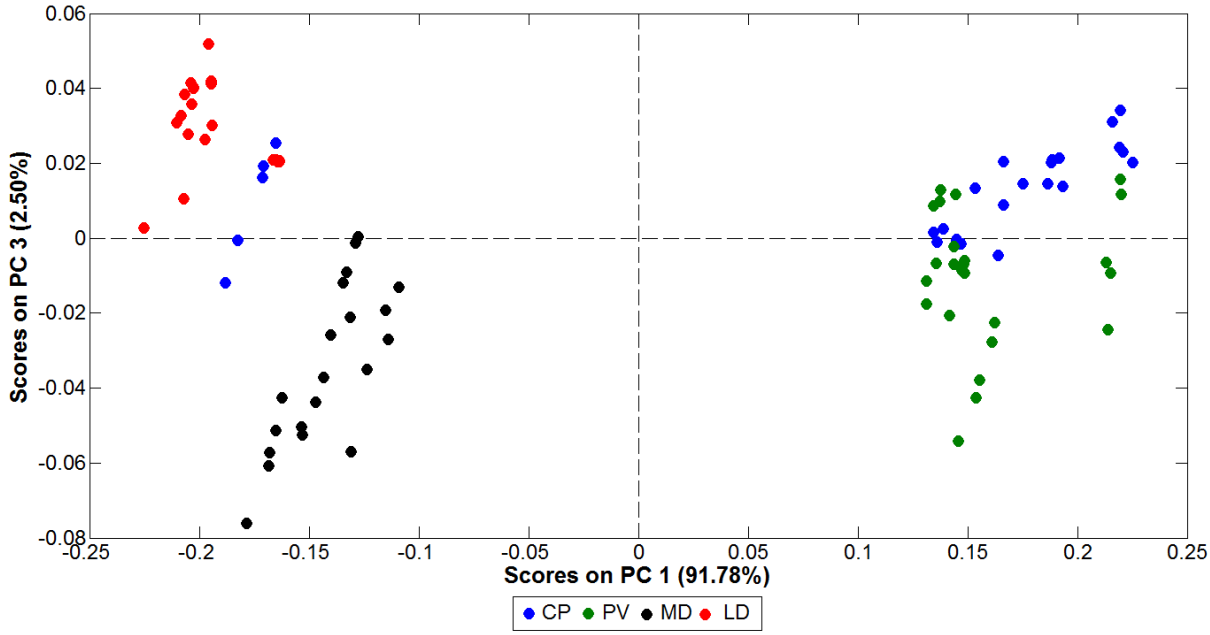
Na ACP realizada, os dados foram centrados na média, de modo que uma média das intensidades para cada comprimento de onda é calculada e, em seguida, subtrai-se cada intensidade do respectivo valor médio (Valderrama, 2005). Para verificar a existência de *outliers*, foram utilizados os valores dos resíduos espectrais (Q) e o *leverage* (T^2 de Hotelling) (Wise, Gallagher, Bro, Shaver, Winding & Koch, 2006).

Origem Geográfica

Para realizar uma análise exploratória do conjunto de dados, primeiramente os dados foram centrados na média e foi feita a ACP a fim de verificar uma possível separação entre amostras de café por origem geográfica, utilizando os espectros apresentados na Figura 6-(b), (c), (d).

O número de componentes principais foi escolhido de acordo com a variância explicada. Componentes que explicam uma quantidade elevada de variância devem ser considerados no modelo, enquanto que os que explicam pouca variância consistem de ruídos (Almeida, Fidelis, Barata & Poppi, 2003).

Para os espectros tratados com MSC, três componentes principais explicaram 98,3% da informação contida nos dados originais. A Figura 7 mostra os gráficos dos *scores* para o modelo do ACP tratados com MSC construído para as amostras de café. Pode ser observada a discriminação entre as cidades onde as amostras foram cultivadas. O componente principal 1 (PC 1) explicou 91,78%, o PC 2 explicou 4,06% e o PC 3, 2,50%. Apenas uma amostra de Cornélio Procópio e suas repetições ficaram separadas das demais.



***Scores* do modelo ACP com espectros após aplicação de MSC.**

Três componentes principais explicaram aproximadamente 89% das informações contidas nos dados para os espectros tratados com a segunda derivada de *Savitzky-Golay*, e 88,27% para os tratados com MSC e segunda derivada conjuntamente. Entretanto, as amostras não ficaram tão bem agrupadas como nas tratadas com MSC.

Analisando os três tratamentos utilizados, os espectros tratados com MSC apresentaram melhor discriminação em relação aos demais. Portanto, todos os resultados apresentados a seguir foram feitos utilizando este modelo ACP. As amostras de Mandaguari foram discriminadas pela parte negativa do PC 1 e do PC 3; as de Londrina, pela parte negativa do PC 1 e positiva do PC 3; e as de Paranavaí e Cornélio Procópio pela parte positiva do PC 1 (Figura 7).

Desta maneira, os *loadings* do ACP, do melhor pré-tratamento, espectro após aplicação de MSC, foram estudados a fim de identificar quais variáveis seriam responsáveis pela separação dos grupos de café por origem geográfica.

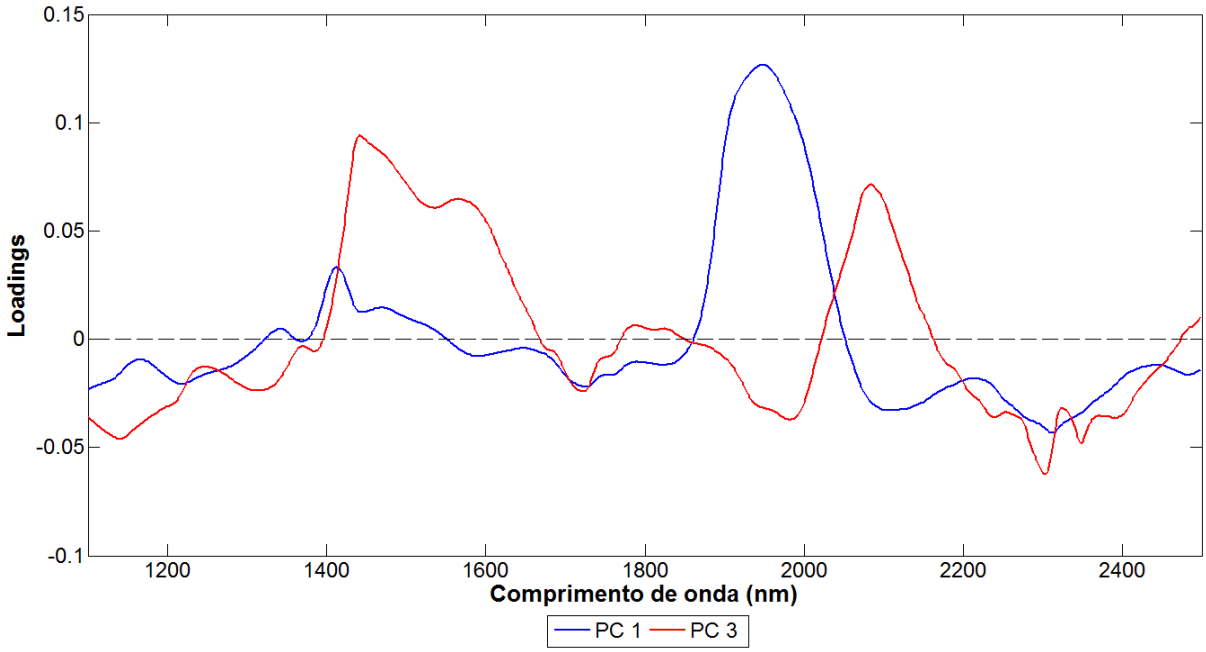


Gráfico dos *loadings* das PC 1 e 3 *versus* comprimento de onda (variáveis) para o modelo ACP dos espectros após aplicação de MSC.

Analisando a Figura acima juntamente com a Tabela das bandas NIRS, verifica-se que as regiões **1**, **2**, **4**, **5**, **6** e **7** contribuem significativamente para a separação das amostras de Mandaguari, pois apresentam valores de *loadings* negativos. A região **1** possui bandas que estão presentes na trigonelina, cafeína, proteínas e aminoácidos, lipídeos, açúcares e carboidratos. A região **2** possui bandas atribuídas à cafeína, aos lipídeos e aos açúcares. A região **4** possui bandas tipicamente encontradas na trigonelina, cafeína, ácidos clorogênicos, proteínas e aminoácidos, lipídeos e açúcares. A região **5** possui bandas que estão presentes na cafeína, ácidos clorogênicos, proteínas e aminoácidos, lipídeos, água e carboidratos. E as regiões **6** e **7** são de bandas atribuidas a todos os compostos citados acima (Ribeiro, Ferreira & Salva, 2011)

As região de *loadings* negativos para PC 1 e positivos para PC 3 contribuem para a discriminação das amostras de Londrina. A faixa de *loadings* negativos para PC 1 compreendem as regiões **1**, **2**, **4**, **6**, e **7**, conforme citado anteriormente. A faixa de *loadings* positivos para PC 3 compreendem as regiões **3** e **6**. Sendo que a região **3** é de bandas encontradas na cafeína, ácidos clorogênicos, proteínas e aminoácidos, lipídeos, água e carboidratos (Ribeiro, Ferreira & Salva, 2011).

As regiões de *loadings* positivos para PC 1 são referentes à Paranavaí e Cornélio Procópio, compreendendo as regiões **3** e **5**. É possível verificar que essa não é uma região onde a trigonelina está presente, entretanto, isso não significa que

as amostras de Paranavaí e Cornélio Procópio não tenham trigonelina em sua composição, mas sugere que a quantidade presente é muito pequena em relação às demais cidades.

Origem Genotípica

Para realizar uma análise exploratória do conjunto de dados, primeiramente foi feita a ACP a fim de verificar uma possível separação entre amostras de café por origem genotípica, utilizando os espectros apresentados na Figura 6-(b), (c), (d). Os dados foram centrados na média.

Analisando as amostras pela origem genotípica, o modelo ACP não foi suficiente para realizar uma boa discriminação entre as amostras em todos os pré-tratamentos utilizados, provavelmente devido à alta variabilidade no perfil espectral de cada genótipo decorrente dos efeitos ambientais.

Conclusão

O estudo mostra que, para a classificação geográfica, os fatores ambientais, como: condições climáticas, altitude, tipo de solo, são responsáveis por variações na composição química do grão, influenciando no sabor e no aroma do café como bebida. Para o modelo ACP, as bandas de NIRS responsáveis pela discriminação das amostras de Paranavaí e Cornélio Procópio foram aquelas com pouca quantidade de trigonelina, enquanto que as bandas que separaram as amostras de Londrina e Mandaguari foram as que continham quantidades elevadas de cafeína e lipídeos. Sendo que, cafés de melhor qualidade estão relacionado com maiores teores de trigonelina, sacarose, lipídeos e aminoácidos; e menores teores de cafeína e ACG.

4.2 REFERÊNCIAS

Wold, K. Esbensen, and P. Geladi, "Principal component analysis", *Chemometrics and Intelligent Laboratory Systems* **2**, 37 (1987).

S. Haykin, *Neural Networks: A Comprehensive Foundation.* MacMillan Publishing Company, (1999).

C. M. Bishop, *Neural Networks for Pattern Recognition.* Oxford University, Oxford, UK (1995).

G. D. Matos, E. R. Pereira-Filho, R. J. Poppi, and M. A. Z. Arruda, "Análise exploratória em química analítica com emprego de quimiometria: PCA e PCA de imagens", *Revista Analytica* **3***,* 38 (2003).

P. Valderrama, *Avaliação de figuras de mérito em calibração multivariada na determinação de parâmetros de controle de qualidade em indústria alcooleira por espectroscopia no infravermelho próximo.* Unicamp, Campinas, BR (2005).

P. Geladi, and B. R. Kowalski, "Partial least-squares regression: a tutorial", *Analytica Chimica Acta* **185**, 1 (1986).

M. R. Almeida, C. H. V. Fidelis, L. E. S. Barata, and R. J. Poppi, "Classification of Amazonian rose wood essential oil by Raman spectroscopy and PLS-DA with reliability estimation", *Talanta* **117**, 305 (2013).

B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Winding, R. S. Koch, *PLS-Toolbox and Solo, v 6.5*. Eigenvector Research Inc., (2006)

J. S. Ribeiro, M. M. C. Ferreira, and T. J. G. Salva, "Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy", *Talanta* **83**, 1352 (2011).

5 APÊNDICE C

5.1 OTIMIZAÇÃO SIMPLEX

O princípio do simplex básico utilizado consiste em deslocar uma figura regular, como um triângulo equilátero sobre uma superfície, ao se considerar duas variáveis (Splendley, Himsworth & Hext, 1962).

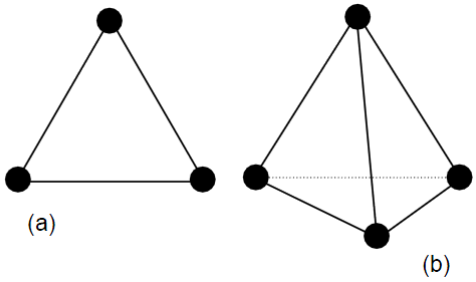


Figura 1 - Interpretação geométrica do simplex para (a) duas variáveis (b) três variáveis.

O método do simplex modificado altera o tamanho e a forma do simplex básico, adaptando-se melhor à superfície de resposta. Este método é o mais utilizado para resolver o problema de otimização sem restrições (Nelder & Mead, 1965). Um método em que se inicia a otimização (maximização ou minimização) atribuindo-se limites inferiores (L_i) e superiores (U_i) para cada fator que será controlado foi desenvolvido posteriormente (Pires, Borsato & Silva, 1998). Uma técnica que permite calcular as demais coordenadas do simplex inicial segundo as expressões (1) e (2) foi apresentada, onde *n* é o número de variáveis (contínuas ou qualitativas) e *t* a distância entre dois vértices (geralmente tomada como 1) (Splendley, Himsworth & Hext, 1962).

$$p = \frac{t}{n\sqrt{2}}\left(\sqrt{n+1}+n-1\right) \qquad (1)$$

$$q = \frac{t}{n\sqrt{2}}\left(\sqrt{n+1}-1\right) \qquad \text{.....} \qquad (2)$$

Ao simplex denominado supermodificado, foram adicionados os limites inferiores e superiores dos fatores utilizados nas expressões (3) e (4) (Nakai Koide & Eugester, 1984).

$$m_1 = L_i + p(U_i - L_i) \quad (3)$$

$$m_2 = L_i + q(U_i - L_i) \quad (4)$$

As coordenadas dos vértices de um simplex regular são representadas pela matriz **M**, onde as colunas representam os componentes dos vértices, numerados de 1 até n+1 e as linhas representam as coordenadas, i = 1 até n (Himmelblau, 1972).

$$M = \begin{bmatrix} L_i & m_1 & m_2 & m_2 \\ L_i & m_2 & m_1 & m_2 \\ \cdots & \cdots & \cdots & \cdots \\ L_i & m_2 & m_2 & m_1 \end{bmatrix} \text{Matriz n x n+1}$$

O algoritmo simplex é muito utilizado para encontrar o mínimo de uma função. Com as respostas obtidas em cada interação, os vértices do simplex são ordenados de acordo com seus valores em **B** (melhor), **N** (intermediários) e **W** (pior) (Gao e Han, 2012).

$$\mathbf{B} \leq \mathbf{N} \leq \cdots \leq \mathbf{W}$$

O novo simplex é determinado rejeitando-se o vértice correspondente à pior resposta e substituindo-se esse vértice por uma operação. O algoritmo utiliza as operações: a reflexão, a expansão, a contração externa e interna e o encolhimento sendo cada uma delas está associada a um parâmetro escalar: α (reflexão), β (expansão), γ (contração externa e interna) e δ (encolhimento). Os valores destes parâmetros devem satisfazer α > 0, β > 1, 0 < γ < 1, e 0 < δ < 1 (Gao e Han, 2012). No simplex estes parâmetros foram calculados adaptativamente as n dimensões do problema de acordo com as expressões (5, 6, 7 e 8),

$$\alpha = 1 \quad (5)$$

$$\beta = 1 + \frac{2}{n} \quad (6)$$

$$\gamma = 0{,}75 - \frac{1}{2n} \tag{7}$$

$$\delta = 1 - \frac{1}{n} \tag{8}$$

A Figura 2 representa a direção do movimento de reflexão,determinada pelo centroide ($\overline{P}$) formado pelos pontos remanescentes (Neto, Scarminio & Bruns, 2010).

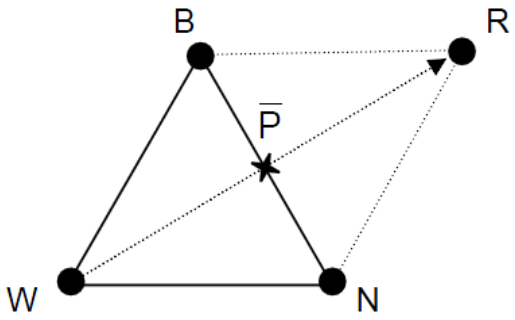


Figura 2 - Movimento de reflexão em um simplex para duas variáveis.

Os passos realizados pelo simplex para a realização da otimização são (Neto, Scarminio & Bruns, 2010; Gao e Han, 2012):

- Ordenar: Avaliar a função (rede neural artificial) nos n+1 vértices calculados pelo simplex inicial e classificar os vértices em **B** (melhor), **N** (intermediários) e **W** (pior).
- Reflexão: Calcular o ponto de reflexão (R):

$$R = \overline{P} + \alpha(\overline{P} - W) \tag{9}$$

- Avaliar **R**: Se **B** $\leq$ **R** $\leq$ **N**, substituir **W** por **R**.
- Expansão: Se **R**<**B**, calcular o ponto de expansão (**S**):

$$S = \overline{P} + \beta(R - W) \tag{10}$$

e avaliar **S** . Se **S**<**R**, substituir **W** por **S**, caso contrário substituir **W** por **R**.

- Contração externa: Se **N**$\leq$**R**<**W**, calcular o ponto de contração externa (**U**):

$$U = \bar{P} + \gamma(R - \bar{P}) \quad (11)$$

e avaliar U. Se $U \leq R$, substitua W por U, caso contrário, prosseguir para o passo 6.

- Contração interna: Se **R≥W**, calcular o ponto de contração interna (**T**):

$$T = \bar{P} - \gamma(R - \bar{P}) \quad (12)$$

e avaliar **T**: Se **T<W**, substituir **W** com **T**, caso contrário, prosseguir para o passo seguinte.

- Encolhimento: Para $2 \leq i \leq$ **W**, definir:

$$E_i = B + \delta(R_i - B) \quad (13)$$

A otimização ocorre até que o valor do erro quadrado médio varie apenas dentro da tolerância estabelecida que de 0,001 ou pela avaliação gráfica, que auxilia na visualização da otimização, representada como um abrandamento na variação das respostas e variáveis independentes.

5.2 REFERÊNCIAS

Gao, F.; Han, L. (2012). Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Comput Optim Appl***,** 51, 259–277.

Himmelblau, D. M. (1972). *Applied Nonlinear Programming*. McGraw-Hill Book Company, New York, 498p.

Nakai, S.; Koide, R.; Eugester, K. A. (1984). A new mapping super-simplex optimization for food products and process development. *J. Food Sci.,*49, 1143-1148.

Nelder, J. A.; Mead, R. (1965). A. simplex method for function minimization. *Computer J,* 7, 308-312.

Neto, B. D. B.; Scarminio, I. S.; Bruns, R. E. (2010). *Como fazer experimentos. Pesquisa e desenvolvimento na ciência e na indústria.* (4th ed.). Porto Alegre: Bookman, 413p.

Pires, M. V. P.; Borsato, D.; Silva, R. S. F. (1998). Desenvolvimento de aplicativo para microinformática visando a otimização de sistemas alimentares, 16, 1998, Rio de Janeiro. Anais ... SBCTA. *In: Congresso Brasileiro de Ciência e Tecnologia de Alimentos***,** 3,1565-1568.

Splendley, W.; Himsworth, F. R.; Hext, G. R. (1962). Sequential application of simplex designs in optimization and evolutionary operation. *Technometrics*, 4, 441-461.