UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO DE ENGENHARIA DE ALIMENTOS

ANDRÉ LUIS GUIMARÃES LEMES

APLICAÇÃO DE MODELOS DE DOIS ESTÁGIOS EM PROBLEMAS DE CLASSIFICAÇÃO DE ALTA COMPLEXIDADE: SEGMENTAÇÃO GEOGRÁFICA E GENOTÍPICA DE CAFÉ ARÁBICA

TRABALHO DE CONCLUSÃO DE CURSO

CAMPO MOURÃO
2014

ANDRÉ LUIS GUIMARÃES LEMES

APLICAÇÃO DE MODELOS DE DOIS ESTÁGIOS EM PROBLEMAS DE CLASSIFICAÇÃO DE ALTA COMPLEXIDADE: SEGMENTAÇÃO GEOGRÁFICA E GENOTÍPICA DE CAFÉ ARÁBICA

Trabalho de conclusão de curso de graduação, apresentado à disciplina de Trabalho de Conclusão de Curso II, do Curso Superior de Engenharia de Alimentos do Departamento Acadêmico de Alimentos, da Universidade Tecnológica Federal do Paraná – UTFPR, Câmpus Campo Mourão, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Alimentos.

CAMPO MOURÃO

2014

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Departamento Acadêmico de Alimentos

TERMO DE APROVAÇÃO

APLICAÇÃO DE MODELOS DE DOIS ESTÁGIOS EM PROBLEMAS DE CLASSIFICAÇÃO DE ALTA COMPLEXIDADE: SEGMENTAÇÃO GEOGRÁFICA E GENOTÍPICA DE CAFÉ ARÁBICA

Por

ANDRÉ LUIS GUIMARÃES LEMES

Esse trabalho de conclusão de curso foi apresentado às 15 horas e 30 minutos do dia 04 de agosto de 2014, como requisito parcial para a obtenção do Título de Bacharel em Engenharia de Alimentos, Departamento Acadêmico de Alimentos, da Universidade Tecnológica Federal do Paraná. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Prof. Dr. Evandro Bona (Orientador – DALIM – UTFPR-CM)

_____
Prof. Dra. Karla Silva (DALIM – UTFPR-CM)

_____
Prof. Dr. Paulo Henrique Março (DALIM – UTFPR-CM)

A folha de aprovação assinada encontra-se na Coordenação do Programa

AGRADECIMENTOS

Reverencio o Professor Dr. Evandro Bona pela sua dedicação, pela orientação e pela amizade durante os mais de quatro anos de trabalho.

À Professora Dra. Patrícia Valderrama, por todo o apoio durante o projeto.

Ao doutorando e amigo Jade Varaschim Link por todo o auxílio durante o projeto.

À mestre Izabele Marquetti pela ajuda durante a pesquisa.

A Dra. Maria Brígida dos Santos Scholz, pela colaboração, e ao pessoal do Instituto Agronômico do Paraná - Londrina (IAPAR) pelo fornecimento das amostras de café.

Ao professor Dr. Dionísio Borsato, a técnica de laboratório Msc. Ivanira Moreira e a todos do Departamento de Química da Universidade Estadual de Londrina (UEL) pelo apoio nas análises no equipamento FTIR.

Aos amigos Gustavo Yasuo Figueiredo Makimori, Rodrigo Mochi Guazelli, Oswaldo Takeshi Koike e João Henrique Mallmann, por todo apoio durante o curso.

E agradeço principalmente aos meus pais pelo apoio e compreensão, pois sem eles, nada disso seria possível.

RESUMO

LEMES, André Luis Guimarães. **Aplicação de modelos de dois estágios em problemas de classificação de alta complexidade: segmentação geográfica e genotípica de café arábica.** 2014. 59 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Alimentos, Universidade Tecnológica Federal do Paraná. Campo Mourão, 2014.

Atualmente o Brasil é o maior produtor mundial de café, sendo responsável por 33,6% da produção mundial. O café pertence ao gênero *Coffea*, e família *Rubiacea*. Suas espécies arábica e canéfora (robusta) têm grande importância econômica mundial, sendo a arábica responsável por 90% da produção. Além da espécie, o genótipo do café também influencia na qualidade da bebida. O objetivo deste projeto foi desenvolver uma metodologia capaz de discriminar os diferentes genótipos de café arábica cultivados no Brasil e identificar sua região de origem. Setenta e quatro amostras de grãos verdes de 20 genótipos do café arábica, cultivados nas cidades de Mandaguari, Londrina, Paranavaí e Cornélio Procópio foram fornecidos pelo IAPAR (Londrina – PR). Foram obtidos espectros das amostras por espectroscopia de infravermelho com transformada de Fourier (FTIR). Após a realização dos pré-tratamentos dos dados, foram criados modelos de dois estágios: um estágio linear e outro não linear. No primeiro estágio do modelo de classificação foram empregados a análise de componentes principais (ACP) e o método de mínimos quadrados parciais com análise discriminante (PLS-DA) com o objetivo de reduzir a dimensionalidade dos dados. Com a realização do PLS-DA, também foi possível realizar a classificação das amostras, proporcionando uma posterior comparação entre o modelo linear e o modelo de dois estágios. Na criação do segundo estágio do modelo, foi utilizada uma rede neural artificial denominada de rede de funções de base radial de regularização (RBF de regularização). Na etapa de construção das redes neurais, uma série de parâmetros deveriam ser escolhidos, e para isto utilizou-se método simplex sequencial para otimização dos mesmos. Na classificação geográfica, o melhor modelo foi o PLS-DA utilizando a faixa espectral entre 750 e 3750 cm^{-1} com os dados puros, pois além de classificar corretamente 100% das amostra, teve melhor performance calculada através dos limiares estabelecidos pelo teorema de Bayes. Na classificação genotípica, o melhor modelo encontrado foi o modelo de dois estágios que utilizou a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados e PLS-DA como primeiro estágio. Esse modelo foi capaz de classificar corretamente 89,04% das amostras de teste, e obteve melhor performance calculada através do teorema de Bayes. Mesmo realizando uma classificação geográfica correta de 100% das amostras, a performance de Bayes mostrou que os modelos ainda devem ser modificados na tentativa de encontrar melhores resultados de sensibilidade e especificidade e diminuir o número de amostras na região de rejeição.

Palavras-chave: FTIR. ACP. PLS-DA. Redes Neurais Artificiais.

ABSTRACT

LEMES, André Luis Guimarães. **Application of two-stage models in high complexity pattern recognition problems: Geographical and genotypic segmentation of green arabica coffee.** 2014. 59 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Alimentos, Universidade Tecnológica Federal do Paraná. Campo Mourão, 2014.

Currently, Brazil is the largest producer of coffee, accounting for 33.6% of world production. The coffee belongs to the *Coffea* genus, from *Rubiaceae* family. The arabica and canephora (robust) species have great global economic importance, being the arabica responsible for 90% of production. In addition to the species, the coffee genotype also influences the quality of the beverage. The objective of this project was to develop a methodology to discriminate the different genotypes of arabica coffee, and also identify the cultivation region. Seventy-four samples of green beans of 20 genotypes of arabica coffee, grown in the cities of Mandaguari, Londrina, Paranavaí and Cornélio Procópio were provided by IAPAR (Londrina-PR). Spectra of samples were obtained by infrared spectroscopy with Fourier transform (FTIR). So, two-stage models were created using a first linear stage and a second nonlinear one. For the linear stage it was used the principal component analysis (PCA) and partial least squares method with discriminant analysis (PLS-DA). With PLS-DA, it was also possible to perform the classification of samples, providing a further comparison between the linear model and the two-stage model. For the second stage of the model it was used a regularized radial basis functions artificial neural network (RBF-R). In neural networks construction several parameters should be optimized and, in this work the sequential simplex method was used for this purpose. For geographical classification, the best model was the PLS-DA using the raw spectra in the range of 750 and 3750 cm^{-1}. The obtained model classify correctly 100% of the samples and, had better performance confirmed by the thresholds established by Bayes' theorem. In genotypic classification, the best model found was the two-stage one using the first derivative of spectra in the range between 800 and 1900 cm^{-1} and PLS-DA as first stage. This model was able to correctly classify 89.04% of test specimens, and obtained better performance based on Bayes theorem. Even performing a 100% correct geographical classification of samples, Bayes' inference showed that the models should still be modified in an attempt to find better results for sensitivity and specificity, and decrease the number of samples in the rejection region.

Keywords: FTIR. PCA. PLS-DA. Artificial Neural Networks.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Atualmente o Brasil é o maior produtor mundial de café, sendo responsável por 33,6% da produção mundial, volume equivalente à soma da produção dos outros três maiores países produtores. O consumo interno de café no Brasil, no período de novembro de 2012 e outubro de 2013 foi de 20,08 milhões de sacas, mantendo-se quase estável em relação ao período anterior correspondente (ABIC, 2013). Em relação à exportação, no período entre agosto de 2012 e julho de 2013, foram exportadas pouco mais de 30 milhões de sacas totalizando uma receita de aproximadamente US$ 5,8 bilhões (CECAFÉ, 2013).

O grão de café é produzido a partir do fruto do cafeeiro, um pequeno arbusto pertencente ao gênero *Coffea*, da família *Rubiaceae*. Suas espécies arábica e canéfora (robusta) têm grande importância econômica mundial, sendo que a espécie arábica representa cerca de 90% da produção mundial de café, e a canéfora cerca de 9%. O café arábica tem um valor comercial maior, pois possui um sabor melhor que o robusta (KEMSLEY; RUAULT; WILSON,1995).

De acordo com Ferreira et al. (2013), o genótipo do café influencia na qualidade da bebida. Entre as cultivares comerciais do cafeeiro *Coffea* arábica disponíveis para o plantio, a cultivar Bourbon apresenta elevado potencial quanto à qualidade de bebida e é altamente valorizada no mercado de cafés especiais por possuir características sensoriais diferenciadas.

A análise de amostras por espectroscopia de infravermelho com transformada de Fourier (FTIR) é uma técnica que fornece impressões digitais bioquímicas de amostras de forma rápida e não destrutiva. Nos últimos anos, esta técnica foi utilizada para analisar e autenticar misturas de café (WANG et al., 2009). A leitura do FTIR é realizada na faixa de 4000 a 400 cm^{-1} com dois feixes de radiação, onde um permanece fixo e o outro se move. Com a variação das distâncias percorridas pelos dois feixes, obtêm-se o chamado interferograma que é uma sequência de interferências que provocam variações na intensidade de radiação recebida pelo detector. A transformação de Fourier em posições sucessivas do espelho dá origem

ao espectro completo de infravermelho. Como a técnica permite uma alta resolução do espectro e utiliza uma grande faixa de comprimento de onda, pode se obter uma quantidade enorme de variáveis (SILVERSTEIN; WEBSTER; KIEMLE, 2007).

Quando se dispõem de muitas variáveis faz-se necessário o uso de técnicas multifatoriais para extrair informações contida nos dados. A maneira mais natural para solucionar problemas de reconhecimento de padrões é a forma estatística, que reconhece a natureza probabilística tanto dos dados a serem processados, quanto da maneira que os resultados devem ser expressados (BISHOP, 2006). As redes neurais artificiais (RNA) são um conjunto de técnicas baseadas em princípios estatísticos que é amplamente aplicada para o reconhecimento de padrões e classificação (HAYKIN, 2001). Uma das principais causas desse sucesso é a sua capacidade de aproximação universal (MELEIRO; VON ZUBEN; MACIEL FILHO, 2009).

As RNA são modelos para o processamento de informação compostas por um conjunto de elementos conectados, unidades ou nós de processamento simples, cuja funcionalidade foi baseada no funcionamento de um neurônio do córtex cerebral. A habilidade de processamento da rede está na memória armazenada nas conexões dessas unidades, chamadas de pesos, que são adquiridas por um processo chamado de aprendizagem que faz uso de padrões de treinamento (GURNEY, 1997). No entanto, devido à sua estrutura genérica, os modelos neurais geralmente requerem a estimativa de um grande número de parâmetros (MELEIRO; VON ZUBEN; MACIEL FILHO, 2009).

Para superar esse obstáculo é possível o emprego de um método de otimização automatizado. A otimização de sistemas é um processo de ajuste para os parâmetros que o influenciam na tentativa de produzir o melhor resultado. O sucesso de um método de otimização depende da sua eficácia para encontrar o ótimo corretamente. Uma técnica muito utilizada é o simplex sequencial, que é uma figura que se desloca sobre uma superfície, de modo a evitar regiões de resposta não satisfatória. No espaço n-dimensional o simplex é um poliedro com faces planas contendo n+1 vértices, onde n é o número de variáveis independentes. Uma vez nas vizinhanças do ótimo, o simplex pode sofrer contração com o objetivo de determinar uma posição mais precisa (BONA et al., 2000).

Em reconhecimento estatístico de padrões, é necessária uma seleção prévia de características, onde um espaço de dados é transformado em um espaço de características. Ou seja, o conjunto de dados sofre uma redução de dimensionalidade. Essa transformação é projetada de modo que o conjunto de dados possa ser representado por um número reduzido de características efetivas mantendo a maior parte da informação intrínseca dos dados (HAYKIN, 2001).

Um modo de reduzir a dimensão dos dados é a análise de componentes principais (ACP) que consiste em uma combinação linear das variáveis originais formando componentes principais (CP) ortogonais. Esta transformação é definida de modo que a primeira CP armazene a maior variância possível, e cada componente seguinte armazene também a maior variância possível (BONA et al., 2012). Com o gráfico das CP é possível, também, visualizar a distribuição das amostras e identificação de possíveis pontos anormais (*outliers*) (WOLD; ESBENSEN; GELADI, 1987)

Outra maneira de realizar a redução da dimensionalidade dos dados é o método PLS-DA (Partial Least Squares Discriminant Analysis), ou Análise Discriminante por Mínimos Quadrados Parciais. O PLS-DA é um método de reconhecimento de padrões supervisionado, ou seja, leva em consideração a matriz de respostas pré-definida. Este método maximiza a distância entre as classes pré-definidas ao invés de explicar as variações dentro de um conjunto de dados, dando origem às variáveis latentes (VL) (WONG et al., 2013).

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 redes neurais artificiais, do tipo rede de funções de base radial de regularização (RBF) que são modelos inerentemente não-lineares (MARQUETTI, 2014).

2. OBJETIVOS

2.1. OBJETIVO GERAL

O objetivo deste projeto foi desenvolver uma metodologia capaz de discriminar diferentes genótipos de café arábica cultivados no Brasil e identificar sua região de origem. Para esse fim, os espectros obtidos no FTIR foram analisados através do emprego de modelos de dois estágios.

2.2. OBJETIVOS ESPECÍFICOS

- Coletar, registrar e armazenar as amostras dos genótipos que serão fornecidas pelo Instituto Agronômico do Paraná (IAPAR, Londrina – PR);
- Obter os espectros infravermelhos no equipamento de FTIR e realizar os pré-processamentos necessários (correção de linha de base, suavização, etc.);
- Definir a melhor faixa de trabalho na região do infravermelho médio;
- Testar as diferentes formas de apresentação dos espectros (espectro puro, primeira derivada, segunda derivada, etc.);
- Verificar qual o melhor método para ser utilizado como primeiro estágio do modelo: ACP ou PLS-DA;
- Realizar o treinamento e definir a melhor arquitetura dos modelos de dois estágios, para a identificação da região de origem das amostras de café arábica;
- Realizar o treinamento e definir a melhor arquitetura para a diferenciação dos genótipos de café arábica dos modelos de dois estágios;

- Comparar a capacidade de classificar corretamente as amostras para o modelo linear (PLS-DA) e para o modelo de dois estágios, e escolher a melhor opção para o problema proposto.

3. METODOLOGIA

3.1. GENÓTIPOS DE CAFÉ ARÁBICA

Setenta e quatro amostras de grãos verdes de 20 genótipos do café arábica foram fornecidos pelo IAPAR (Londrina – PR). Os grãos foram secos, moídos, peneirados, embalados e mantidos congelados até a realização das análises. As amostras são das safras de 2009 e 2010, e foram cultivadas nas seguintes cidades: Mandaguari, Londrina, Paranavaí e Cornélio Procópio. A Tabela 1 mostra os genótipos utilizados e seus respectivos códigos. A relação das amostras está disposta na Tabela 1-B, no Apêndice B.

Tabela 1: Genótipos de café analisados com seus respectivos códigos.

Código	Genótipo	Código	Genótipo	Código	Genótipo	Código	Genótipo
IP097	IPR 97	IP102	IPR 102	IP107	IPR 107	IA059	IAPAR 59
IP098	IPR 98	IP103	IPR 103	IP108	IPR 108	IC001	Icatu
IP099	IPR 99	IP104	IPR 104	CT001	Catuaí	MN001	Mundo Novo
IP100	IPR 100	IP105	IPR 105	BB001	Bourbon	IE059	IA 59 enxertado
IP101	IPR 101	IP106	IPR 106	TU001	Tupi	IE105	IPR 105 enxertado

3.2. ESPECTROSCOPIA DE INFRAVERMELHO (FTIR)

Para preparar as pastilhas foram adicionados em torno de 100 mg de KBr seco (SIGMA-ALDRICH - padrão cromatográfico) e aproximadamente 1 mg de

amostra finamente moída. A mistura foi então prensada em uma prensa hidráulica (Bovenau, P15 ST) usando um molde (ICL, ICL's Macro/Micro KBr die) usando 7 toneladas de pressão produzindo, assim, uma pastilha transparente. Antes da análise de cada amostra o FTIR (Shimadzu, IR Affinity-1) foi programado para realizar um espectro de background do ar, sendo o mesmo utilizado para descontar a influência dos componentes do ar na amostra. A pastilha foi então posicionada no feixe do instrumento e os espectros foram obtidos na faixa de 4000 a 400 cm^{-1}. Foram realizadas 5 repetições (pastilhas) para cada amostra e foi usada uma apodização do tipo Happ-Genzel com 32 varreduras acumuladas para formar o espectro final. Para este trabalho foi considerado o uso da região entre 3750 a 750 cm^{-1} do espectro, desconsiderando assim os ruídos presentes além desta região. Também foi testada a região entre 1900 e 800 cm^{-1} conforme recomendação da literatura consultada (WANG et al., 2009; BRIANDET; KEMSLEY; WILSON, 1996).

3.3. PRÉ-TRATAMENTO

Após obtenção dos espectros foi realizado um pré-processamento que consiste em várias etapas. Primeiramente realizou-se os ajustes necessários ao espectro (correção de linha de base, suavização, etc.). Após a normalização, realizou-se uma análise de componentes principais com o objetivo de identificar possíveis *outliers*.

Posteriormente, realizou-se a análise de componentes independentes (ICA), para extrair a influência do CO_2 no espectro. A ICA é uma técnica de separação que tem sido desenvolvida com o objetivo de extrair os sinais puros subjacentes a partir de um conjunto de sinais misturados com proporções desconhecidas (HYVÄRINEN; OJA, 2000; BOUVERESSE; BENABID; RUTLEDGE, 2007; VALDERRAMA et al., 2011). A ideia principal da ICA é encontrar uma transformação matemática dos dados em uma combinação linear de componentes estatisticamente independentes (PARASTAR; JALALI-HERAV; TAULER, 2012). Posteriormente foram realizadas a primeira e a segunda derivada de cada espectro através do algoritmo de *Savitzky-Golay*, usando 7 pontos de janela e um polinômio de 2^o grau, com o objetivo de

remover ruídos e acentuar as diferenças entre as amostras (SAVITZKY; GOLAY, 1964; WANG et al., 2009).

3.4. MODELO DE DOIS ESTÁGIOS

Após a realização dos pré-tratamentos dos dados, foram criados modelos de dois estágios. Estes modelos foram constituídos de um estágio linear e outro não linear (CIOSEK et al., 2005).

3.4.1. Primeiro estágio (linear)

No primeiro estágio do modelo de classificação foram empregados a ACP (WOLD; ESBENSEN; GELADI, 1987) e o PLS-DA (BARKER; RAYENS, 2003) em todos os bancos de dados (espectros puros, primeira derivada e segunda derivada, utilizando para cada um a região entre 1900 a 800 cm^{-1} e 3750 a 750 cm^{-1}).

A ACP é um método não supervisionado que agrupa informações altamente correlacionadas em um novo sistema de eixos, proporcionando assim uma redução da dimensionalidade dos dados. Com a realização da ACP, pode-se examinar possíveis agrupamentos das amostras e identificar possíveis *outliers*. Esta análise transforma matematicamente os dados espectrais em componentes ortogonais, chamadas componentes principais (CP), cujas combinações lineares mantêm as informações dos dados originais. Na ACP a matriz de dados é decomposta em dois 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 (MARQUETTI, 2014).

O PLS-DA é um método supervisionado muito utilizado para a classificação de padrões. Ou seja, utiliza a resposta desejada para cada amostra de treinamento na decomposição dos dados em *scores* e *loadings*. 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 na 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 (VL) (MARQUETTI, 2014).

O objetivo principal da aplicação da ACP e do PLS-DA foi reduzir a dimensionalidade dos dados, sendo que as componentes principais e as variáveis latentes encontradas em cada análise foram utilizadas como variáveis de entrada no segundo estágio dos modelos. Com a realização do PLS-DA, também foi possível realizar a classificação das amostras, proporcionando uma posterior comparação entre o modelo linear e o modelo de dois estágios. Para a classificação das amostras utilizando o PLS-DA, o numero de VL foi determinado através da análise dos valores de porcentagem de classificação correta e erro quadrado médio (EQM) tanto para as amostras de treinamento e teste.

3.4.2. Normalização

Antes de serem alimentados no segundo estágio do modelo os vetores de entrada (componentes principais ou variáveis latentes) foram normalizados (HAYKIN, 2001). Essa etapa é necessária para que as funções de ativação dos neurônios artificiais não fossem facilmente saturadas ou ocorresse um erro de *overflow*, número grande demais para ser representado de maneira binária e ser manipulado pelo processador do computador. Os métodos de normalização utilizados foram: máximo e mínimo (minimax), transformação para uma escala entre -1 e 1 e autoescalonamento, vetor de entrada com média zero e variância unitária (PÉREZ-MAGARIÑO *et al.*, 2004).

3.4.3. Segundo estágio (não-linear)

Na criação do segundo estágio do modelo, foi utilizada uma rede neural artificial denominada de rede de funções de base radial de regularização (RBF de regularização). Esse tipo de rede baseia-se no método da interpolação exata e na teoria da regularização, que envolve a adição de uma função de penalidade que pune mapeamentos que não são suaves (BISHOP, 2006). A arquitetura de uma rede neural RBF, mostrada na Figura 1, envolve três camadas com papéis totalmente diferentes entre si. Os neurônios de entrada conectam a rede ao seu ambiente. A segunda camada, a única camada oculta da rede, aplica uma transformação não linear do espaço de entrada para um espaço oculto, também conhecido como espaço de características. Essas unidades ocultas fornecem um conjunto de funções radiais que constituem uma base arbitrária para os padrões de entrada. A camada de saída faz uma combinação linear das bases radiais, fornecendo a resposta da rede ao padrão de ativação aplicado à camada de entrada (HAYKIN, 2001).

O número de neurônios da camada de entrada é determinado pela quantidade de CP ou VL utilizadas. Na camada oculta, de acordo com a metodologia da interpolação exata, cada amostra de treinamento é utilizada como uma base radial (BISHOP, 2006). O número de neurônios da camada de saída é definido de acordo com o número de classes existentes e como é feita a classificação. Neste caso, utilizou-se 20 neurônios na camada de saída para classificação de genótipos e 4 neurônios para a classificação geográfica, sendo que a resposta de um deles é igual a 1 e as outras são iguais a 0, indicando assim, uma das classes existentes (HAYKIN, 2001).

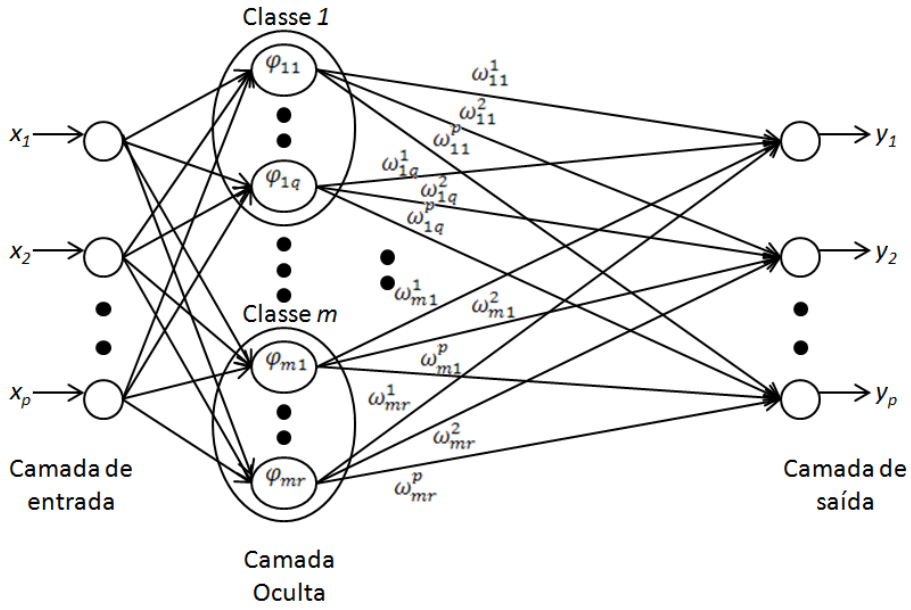

Figura 1: Representação de uma rede de função de base radial.

Há uma grande classe de funções de base radial que são de particular interesse no estudo de redes RBF (HAYKIN, 2001). Dentre elas destacam-se as multiquádrica (1), multiquádrica inversa (2) e função gaussiana (3).

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

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

$$\varphi(r) = exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{3}$$

onde r é a distância euclidiana entre o centro da base e o padrão de entrada e σ é o raio da base radial que controla a suavidade da função de interpolação.

Na RBF de regularização, cada amostra de treinamento é utilizada como uma base radial, e o valor do y_p (resultado previsto) é dado pela equação (4),

$$y_p = \Phi W \tag{4}$$

sendo a matriz Φ o valor de cada função de base radial para cada uma das amostras avaliadas. A matriz W é calculada pela equação (5),

$$W = (\Phi + \lambda I)^{-1} y \tag{5}$$

sendo que Φ é obtida com as amostras de treinamento, λ é o parâmetro de regularização, I é uma matriz identidade e y é a resposta observada das amostras de treinamento. O parâmetro de regularização tem a função de suavizar o mapeamento evitando que a interpolação seja exata e assim aumentando a capacidade de generalização do modelo para as amostras de teste.

3.5. OTIMIZAÇÃO DOS PARÂMETROS DE REDE

Na etapa de construção das redes neurais, uma série de parâmetros devem ser escolhidos para que o modelo criado seja o melhor possível para a classificação dos dados. Neste trabalho, os parâmetros otimizados foram:

- Quantidade de CP ou VL utilizadas (3 a 100);
- Função de pré-processamento das entradas (minimax ou autoescalonamento);
- Função de base radial (multiquádrica, multiquádrica inversa ou função gaussiana);
- Raio da base radial(σ) (1 a 14)
- Parâmetro de regularização (λ) (0,001 a 5).

A fim de maximizar a porcentagem de classificação correta e reduzir o erro quadrado médio com o menor modelo possível foi realizada uma otimização multiobjetivo dos parâmetros empregando as funções de desejabilidade (BONA et al., 2011).

O método simplex sequencial utilizado consiste em uma figura de n+1 vértices, que se altera em tamanho e forma adaptando-se melhor ao espaço de resposta, onde *n* é o número de variáveis. A otimização é iniciada atribuindo-se

limites inferiores (L_i) e superiores (U_i) para cada fator que será controlado. As coordenadas do simplex inicial são calculadas utilizando as equações (6, 7, 8 e 9), onde *t* é a distância entre dois vértices (geralmente tomada como 1), e são formadas conforme a 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* (LINK, 2013).

$$p = \frac{t}{n\sqrt{2}}\left(\sqrt{n+1}+n-1\right) \quad (6)$$

$$q = \frac{t}{n\sqrt{2}}\left(\sqrt{n+1}-1\right) \quad (7)$$

$$m_1 = L_i + p\left(U_i - L_i\right) \quad (8)$$

$$m_2 = L_i + q\left(U_i - L_i\right) \quad (9)$$

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

Com as respostas obtidas em cada iteração, os vértices do simplex são ordenados de acordo com seus valores em **B** (melhor), **N** (intermediários) e **W** (pior) (GAO; HAN, 2012).

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: reflexão, expansão, contração externa e interna e encolhimento. Cada

uma delas está associada a um parâmetro de escala: 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; HAN, 2012). No simplex estes parâmetros foram calculados adaptativamente para as *n* dimensões do problema de acordo com as expressões (10, 11, 12 e 13).

$$\alpha = 1 \tag{10}$$

$$\beta = 1 + \frac{2}{n} \tag{11}$$

$$\gamma = 0{,}75 - \frac{1}{2n} \tag{12}$$

$$\delta = 1 - \frac{1}{n} \tag{13}$$

Uma descrição detalhada do algoritmo pode ser encontrada em Gao e Han (2012).

A otimização segue através de uma sequencia das operações citadas até que o valor da resposta varie apenas dentro da tolerância estabelecida que foi de 0,001 ou pela visualização gráfica que também pode ser utilizada como um critério de parada da otimização.

3.6. AVALIAÇÃO DA DESEMPENHO DO MODELO

A performance do modelo de classificação foi avaliada utilizando um valor limite (*threshold* ou *limiar)* que separa as classes. Assim, minimiza-se o número de

falsos positivos/negativos para a validação dos dados (ALMEIDA et al., 2013). O valor do *threshold* corresponde ao encontro das curvas de probabilidade *a posteriori* (Figura 2) encontradas utilizando o teorema de Bayes (BISHOP, 2006),

$$p(C_k|y) = \frac{p(y|C_k)p(C_k)}{p(y)} \tag{21}$$

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.

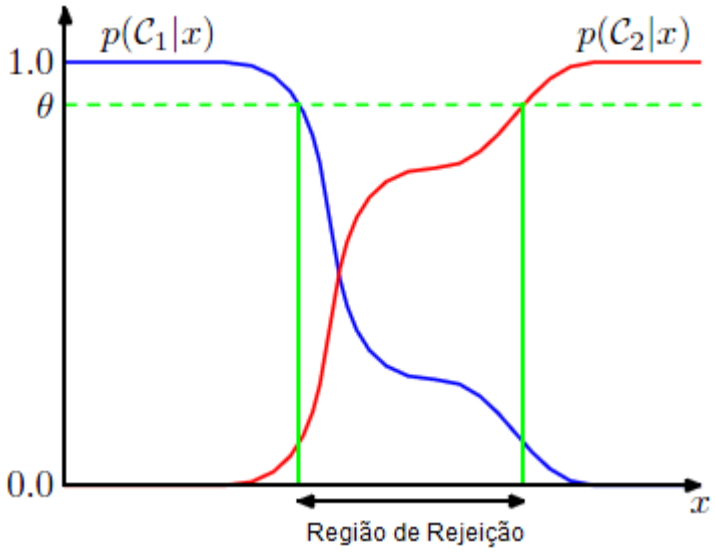

Figura 2: Curvas de probabilidade *a posteriori*.

Desta maneira, amostras localizadas na região de rejeição devem ter sua classificação avaliada com cuidado. Pois, na região de rejeição, as probabilidades da amostra ser ou não da classe, são próximas.

3.7. IMPLEMENTAÇÃO COMPUTACIONAL

Todas as análises matemáticas e/ou estatísticas, assim como a ACP, PLS-DA, ICA, otimização e as redes neurais artificiais foram realizadas no software MATLAB R2008b (The MathWorks Inc., Natick, USA).

4. RESULTADOS E DISCUSSÕES

4.1. PRÉ-TRATAMENTOS

Após a realização da espectroscopia de infravermelho e os pré-tratamentos de suavização e correção de linha de base, foram obtidos os espectros dispostos na Figura 3.

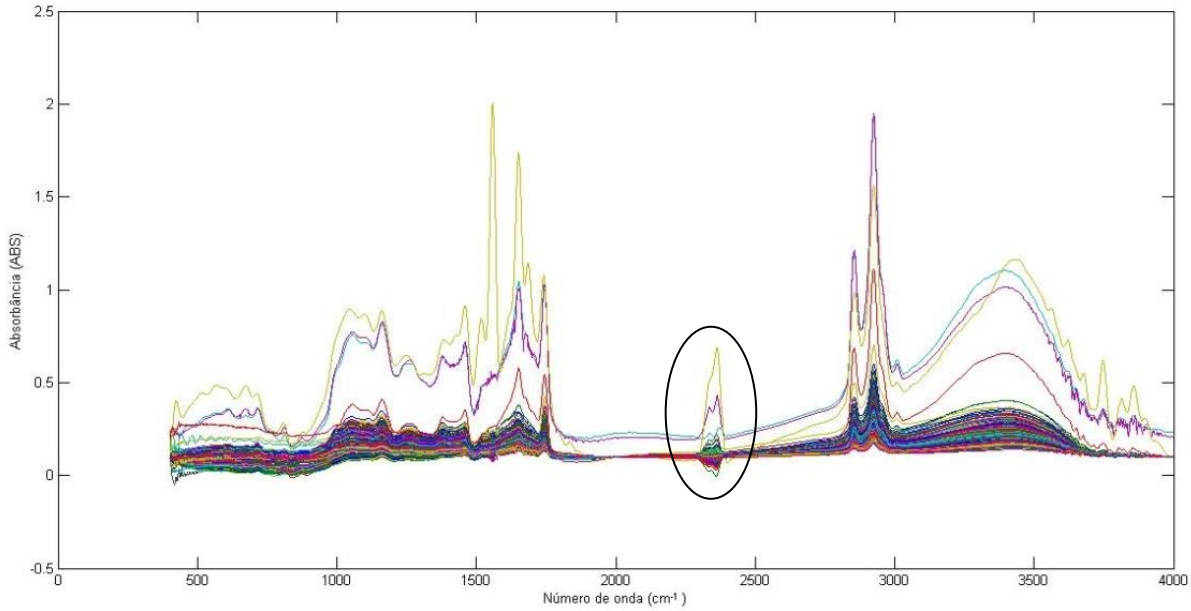

Figura 3: Espectros das amostras de café, a banda do CO_2 está destacada.

Usando uma ACP preliminar, foi possível visualizar alguns dados com comportamento distinto das demais (*outliers*), ou seja, amostras distantes da nuvem de pontos, conforme destacado na Figura 4.

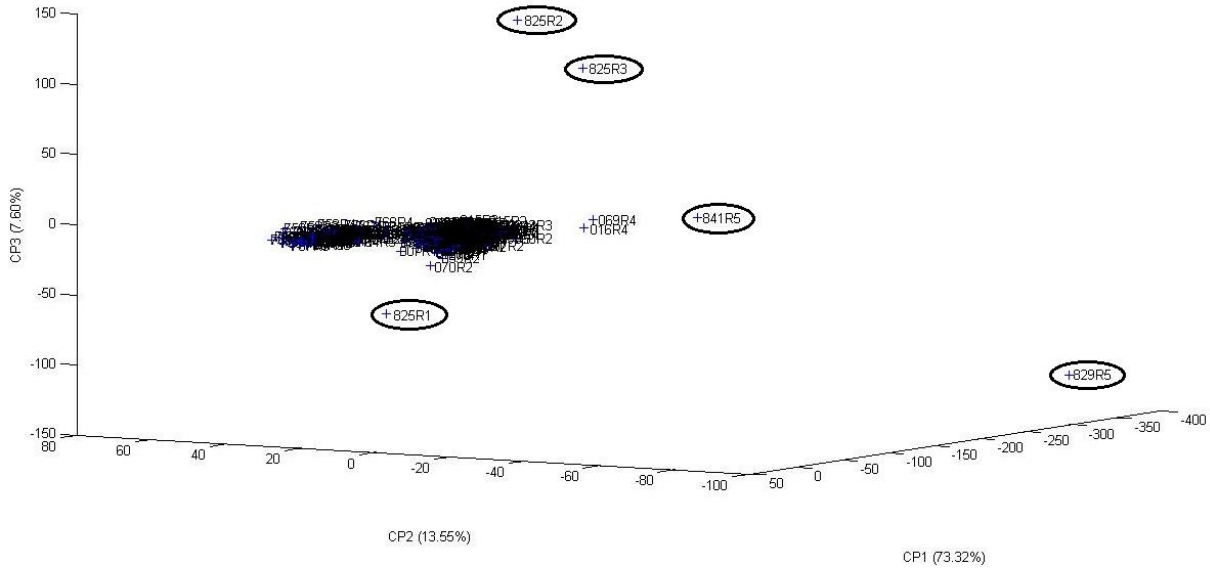

Figura 4: Outliers identificados através da ACP com uma variância acumulada de 94,47%.

Após a remoção dessas cinco amostras discrepantes, o banco de dados passou a possuir 364 amostras, sendo que 291 amostras (80%) foram utilizadas para o treinamento da rede neural e 73 amostras (20%) foram utilizadas para teste e avaliação da capacidade de generalização do modelo. Para realizar a seleção das amostras de teste, uma das repetições de cada amostra foi escolhida aleatoriamente.

Posteriormente foi realizada a ICA para remover a interferência do CO_2 nas amostras, e então foram obtidos os espectros da Figura 5.

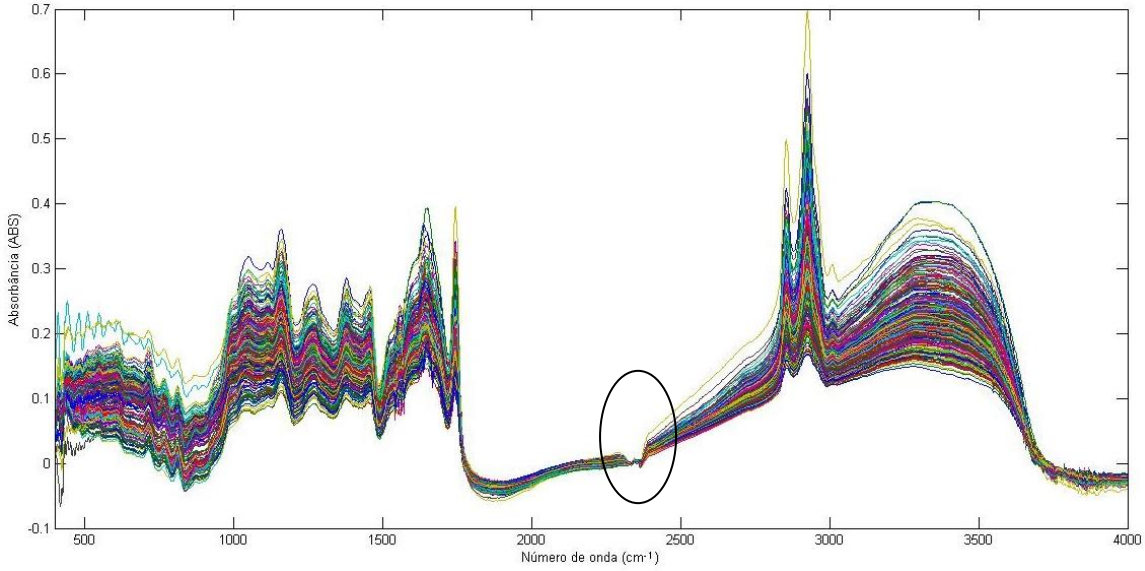

Figura 5: Espectros das amostras de café após a remoção de outliers e realização da ICA, o destaque mostra a eliminação da banda do CO_2.

Como pode ser observado na Figura 5, a ICA foi capaz de remover completamente a interferência de CO_2 nas amostras, sendo que estava compreendida na região entre 2300 a 2400 cm^{-1} dos espectros. Esses dados pré-tratados foram utilizados nos modelos de classificação.

4.2. CLASSIFICAÇÃO GEOGRÁFICA

Os resultados da classificação geográfica realizada através do PLS-DA estão dispostos na Tabela 2.

Tabela 2: Resultados da classificação geográfica através do PLS-DA.

Parâmetros do PLS-DA	**Faixa do espectro utilizada (cm^{-1})**					
	800 - 1900			**750 - 3750**		
Tratamento dos espectros	Dados Puros	1ª Derivada	2ª Derivada	Dados Puros	1ª Derivada	2ª Derivada
Número de LV's	27	29	30	42	34	33
Desempenho do PLS-DA para os dados de treinamento						
Erro quadrado médio	0,0303	0,0236	0,0227	0,0172	0,0172	0,0148
% de classificação correta	100	100	100	100	100	100
Desempenho do PLS-DA para os dados de teste						
Erro quadrado médio	0,0399	0,0383	0,0476	0,0394	0,0422	0,0591
% de classificação correta	100	100	97,2603	100	98,6301	95,8904

Como se pode observar na Tabela 2, o PLS-DA foi capaz de classificar corretamente 100% das amostras (treinamento e teste).

Os resultados da classificação geográfica realizada através dos modelos de dois estágios estão dispostos na Tabela 3.

Tabela 3: Resultados de classificação geográfica, obtidos através do modelo de dois estágios.

Primeiro Estágio	**ACP**						**PLS-DA**					
Faixa do espectro utilizada (cm^{-1})	**800 - 1900**			**750 - 3750**			**800 - 1900**			**750 - 3750**		
Modelo	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Tratamento dos espectros	Dados Puros	1ª Derivada	2ª Derivada	Dados Puros	1ª Derivada	2ª Derivada	Dados Puros	1ª Derivada	2ª Derivada	Dados Puros	1ª Derivada	2ª Derivada
Número de CP/VL	81	46	93	77	54	98	44	31	16	51	31	37
Função da base radial[a]	G	G	G	G	G	G	MQ	MQ	G	G	G	G
Normalização[b]	AE	AE	AE	MM	AE	AE	AE	AE	AE	AE	AE	AE
Parâmetro de regularização	0,002	0,6847	0,2141	0,002	0,0113	0,0529	2,0735	1,9919	0,002	0,002	0,0241	0,0342
Dispersão	11,0486	9,1075	13,6456	9,9384	12,1176	13,7651	14	14	2,7241	9,8037	6,5784	8,9707
Desempenho do modelo para os dados de treinamento												
Erro quadrado médio	$3,63 \times 10^{-5}$	0,0328	0,0228	0,0145	0,0045	0,0115	0,0264	0,0296	$9,29 \times 10^{-5}$	$4,46 \times 10^{-5}$	0,002	0,0041
% de classificação correta	100	99,3127	100	100	100	100	100	100	100	100	100	100
Desempenho do modelo para os dados de teste												
Erro quadrado médio	0,032	0,145	0,0996	0,0449	0,1064	0,1092	0,0405	0,0438	0,0337	0,0384	0,0339	0,0504
% de classificação correta	100	76,7123	82,1918	100	82,1918	75,3425	100	100	97,2603	100	100	95,8904

[a] Função de base radial: G (Gaussiana), MQ (Multiquadrática), MQI (Multiquadrática inversa).
[b] Normalização: MM (Minimax) e AE (Autoescalonamento).

Como pode ser observado na Tabela 3, os melhores modelos encontrados foram o **A**, **D**, **G**, **H**, **J** e **K**, pois foram capazes de classificar corretamente 100% das amostras de treinamento e teste.

Também pode ser observado na Tabela 3, que a quantidade de VL utilizadas nos modelos que utilizaram como primeiro estágio o PLS-DA foi menor que a quantidade de CP nos modelos que utilizaram como primeiro estágio a ACP. Isso demonstra que as VL carregam mais variância relacionada à diferenciação das classes que as CP.

Ao comparar a porcentagem de classificação correta e o EQM da classificação realizada pelo PLS-DA (Tabela 2) e pelo modelo de dois estágios (Tabela 3), verifica-se que além dos melhores modelos classificarem corretamente 100% as amostras, o valor de EQM dos melhores modelos foram bem parecidos.

Para avaliar melhor a performance dos modelos que classificaram corretamente 100% das amostras, foi realizada a análise de performance através do teorema de Bayes. Os resultados da performance de cada modelo encontrados através do teorema de Bayes estão dispostos nas Tabelas 4 (PLS-DA) e 5 (modelos de dois estágios).

Tabela 4: Performance do PLS-DA calculada através do teorema de Bayes.

Faixa do espectro (cm^{-1})	**800 - 1900**	**800 - 1900**	**750 - 3750**
Tratamento dos espectros	**Dados Puros**	**1ª Derivada**	**Dados Puros**
Classe: Paranavaí			
Limiar	0,4425	0,4314	0,4281
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	0,9953	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	1,0000
Classe: Cornélio Procópio			
Limiar	0,3865	0,3791	0,3943
Sensibilidade (Treinamento)	0,9500	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000
Sensibilidade (Teste)	0,9000	1,0000	1,0000
Especificidade (Teste)	1,0000	0,9841	1,0000
Classe: Mandaguari			
Limiar	0,4248	0,4472	0,4721
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	0,9310
Especificidade (Teste)	0,9318	0,9773	0,9773
Classe: Londrina			
Limiar	0,3908	0,3822	0,3970
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000
Especificidade (Teste)	0,9831	0,9831	1,0000
Média ponderada pela quantidade de amostras			
Sensibilidade (Treinamento)	0,9931	1,0000	1,0000
Especificidade (Treinamento)	1,0000	0,9989	1,0000
Sensibilidade (Teste)	0,9863	1,0000	0,9726
Especificidade (Teste)	0,9817	0,9863	0,9954

Observando os resultados da Tabela 4, nota-se que em todos os modelos ocorreram erros de classificação das amostras, porém o melhor modelo encontrado foi o modelo que utilizou a faixa espectral entre 750 e 3750 cm^{-1} com os dados puros. Este modelo atingiu os níveis máximos de sensibilidade e especificidade para as classes Paranavaí, Cornélio Procópio e Londrina, e apenas na classe Mandaguari

ocorram erros de classificação. Além disso, uma quantidade menor de amostras ficou localizada na região de rejeição (Figura 6).

Tabela 5: Performance dos melhores modelos de dois estágios, calculada através do teorema de Bayes.

	Modelo					
	A	**D**	**G**	**H**	**J**	**K**
Primeiro estágio	ACP	ACP	PLS-DA	PLS-DA	PLS-DA	PLS-DA
Faixa do espectro (cm^{-1})	800 - 1900	750 - 3750	800 - 1900	800 - 1900	750 - 3750	750 - 3750
Tratamento dos espectros	Dados Puros	Dados Puros	Dados Puros	1ª Derivada	Dados Puros	1ª Derivada
Classe: Paranavaí						
Limiar	0,1710	0,3781	0,4789	0,4646	0,2279	0,4052
Sensibilidade (Treinamento)	1,0000	1,0000	0,9875	0,9875	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	0,9953	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,8679	1,0000	1,0000	1,0000	0,9075	0,9811
Classe: Cornélio Procópio						
Limiar	0,1916	0,3471	0,4481	0,4340	0,2130	0,3645
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	0,9750	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	0,9000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,9206	0,9683	0,9841	0,9841	0,8730	1,0000
Classe: Mandaguari						
Limiar	0,2419	0,4311	0,4749	0,4645	0,2605	0,4610
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	0,9943	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	0,9310	0,9655	0,9655	1,0000	0,8621
Especificidade (Teste)	0,8636	0,9773	1,0000	0,9773	0,8409	1,0000
Classe: Londrina						
Limiar	0,1622	0,3621	0,4292	0,4154	0,2073	0,4103
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)	0,9153	1,0000	1,0000	1,0000	0,9322	1,0000
Média ponderada pela quantidade de amostras						
Sensibilidade (Treinamento)	1,0000	1,0000	0,9966	0,9931	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	0,9989	0,9989	1,0000	1,0000
Sensibilidade (Teste)	1,0000	0,9589	0,9863	0,9863	1,0000	0,9452
Especificidade (Teste)	0,8950	0,9863	0,9954	0,9909	0,8908	0,9954

Observando os resultados da Tabela 5, também ocorreram erros de classificação das amostras, porém o melhor modelo encontrado foi o modelo **D**, que utilizou a faixa espectral entre 750 e 3750 cm^{-1} com os dados puros e a ACP como primeiro estágio. Pois, além de obter bons valores das médias ponderadas de sensibilidade e especificidade, uma quantidade menor de amostras ficou localizada na região de rejeição (Figura 7).

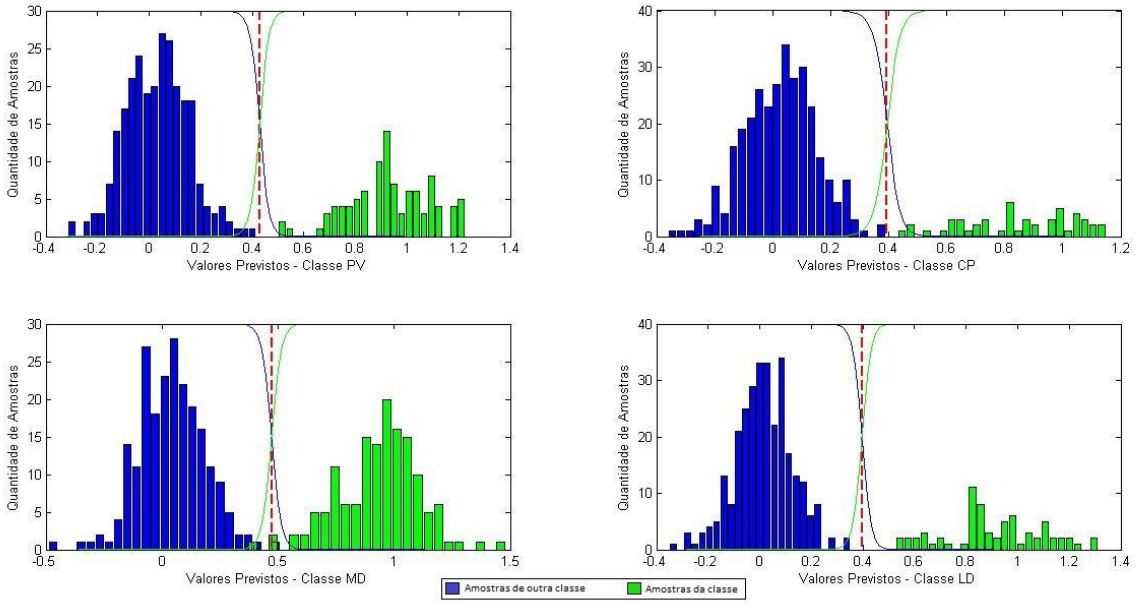

Figura 6: Curva de probabilidade a posteriori por classe para a classificação geográfica do PLS-DA utilizando a faixa espectral entre 750 e 3750 cm-1 com os dados puros.

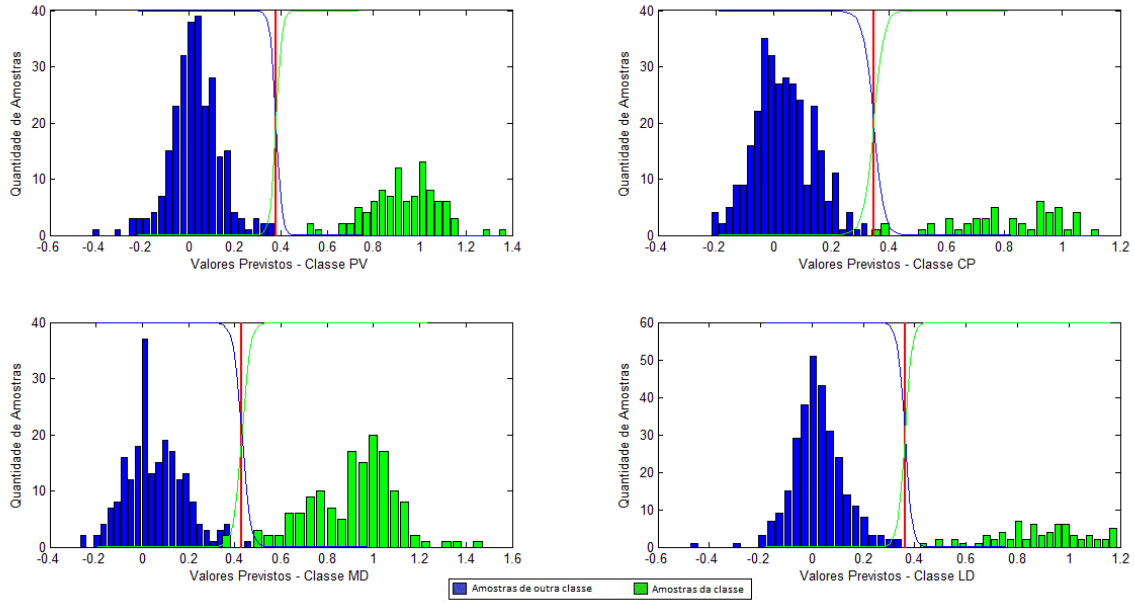

Figura 7: Curva de probabilidade a posteriori por classe para a classificação geográfica do modelo de dois estágios D, utilizando a faixa espectral entre 750 e 3750 cm^{-1} com os dados puros e ACP como primeiro estágio.

Quando se compara os dados de performance calculados através do teorema de Bayes, para a classificação com o PLS-DA (Tabela 4) e com o modelo de dois estágios (Tabela 5), verifica-se que não existe grandes diferenças entre os dois modelos. Além disso, a análise dos gráficos das Figuras 6 e 7 indica que a quantidade de amostras na área de confusão é baixa para ambos os modelos.

As respostas fornecidas pelo PLS-DA utilizando a faixa espectral entre 750 e 3750 cm^{-1} com os dados puros, e as respostas fornecidas pelo modelo de dois estágios D, foram plotadas nas Figuras 8 e 9 respectivamente, para melhor análise.

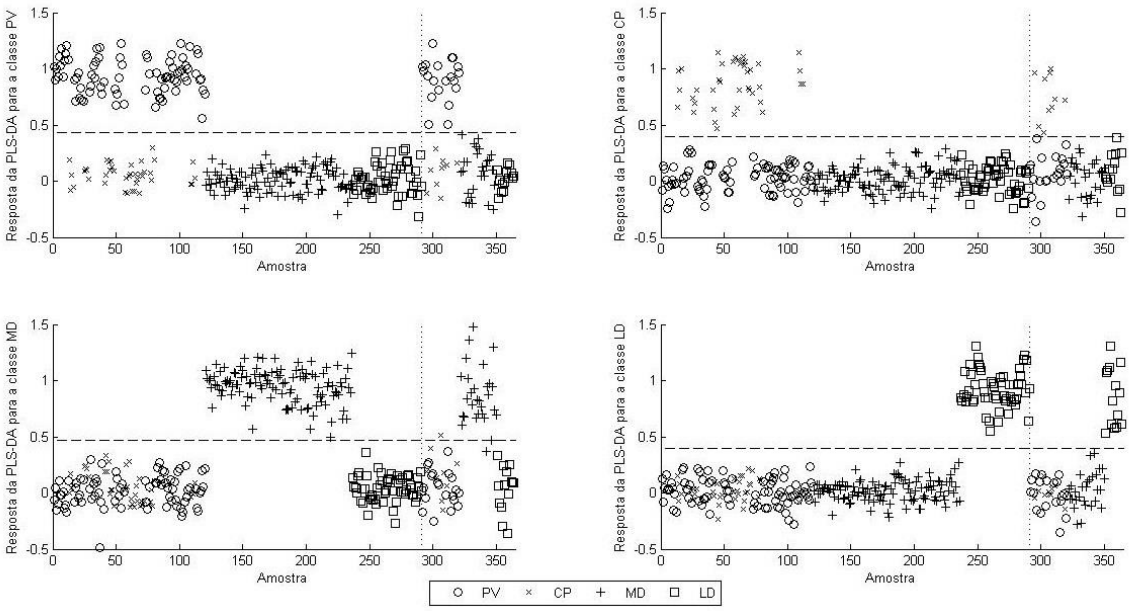

Figura 8: Resposta do PLS-DA para classificação geográfica, utilizando a faixa espectral entre 750 e 3750 cm^{-1} com os dados puros. A linha pontilhada vertical separa as amostras de treinamento daquelas utilizados para o teste.

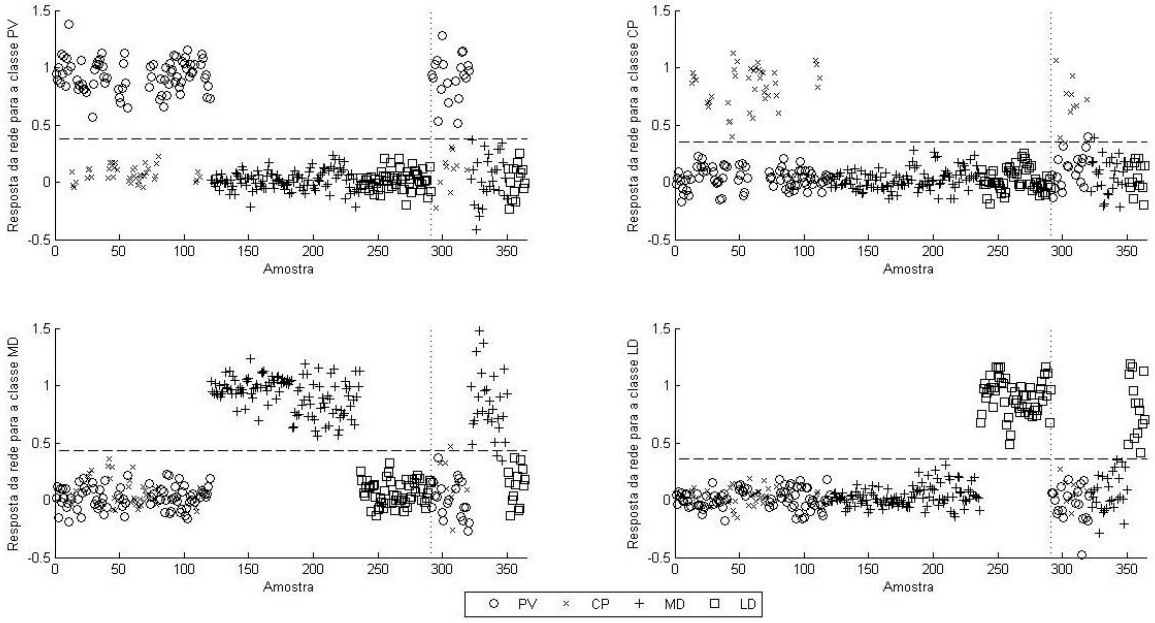

Figura 9: Resposta do modelo de dois estágios D, utilizando a faixa espectral entre 750 e 3750 cm^{-1} com os dados puros e ACP como primeiro estágio. A linha pontilhada vertical separa as amostras de treinamento daquelas utilizados para o teste.

Como a performance do modelo de dois estágios foi bem parecida com o PLS-DA não justifica-se a utilização do primeiro para a classificação geográfica do café arábica. Vale salientar que ainda serão testadas algumas alterações no segundo estágio, redes neurais, na tentativa de melhor os resultados.

4.3. CLASSIFICAÇÃO GENOTÍPICA

Os resultados da classificação genotípica realizada através do PLS-DA estão dispostos na Tabela 6.

Tabela 6: Resultados da classificação genotípica através do PLS-DA

Parâmetros do PLS-DA	**Faixa do espectro utilizada (cm^{-1})**					
	800 - 1900			**750 - 3750**		
Tratamento dos espectros	Dados Puros	1ª Derivada	2ª Derivada	Dados Puros	1ª Derivada	2ª Derivada
Número de LV's	52	92	70	53	58	59
Desempenho do PLS-DA para os dados de treinamento						
Erro quadrado médio	0,0228	0,014	0,0172	0,0217	0,0181	0,0167
% de classificação correta	94,1581	99,3127	97,5945	96,2199	97,9381	97,5945
Desempenho do PLS-DA para os dados de teste						
Erro quadrado médio	0,0339	0,0483	0,0526	0,0369	0,039	0,0484
% de classificação correta	78,0822	71,2329	57,5342	71,2329	69,863	50,6849

Como se pode observar na Tabela 6, o PLS-DA não foi capaz de realizar uma boa classificação genotípica das amostras, sendo o modelo que utiliza a faixa do espectro entre 800 e 1900 cm^{-1} com os dados puros (modelo X), foi capaz de classificar 78,8% das amostras de teste corretamente.

Os resultados da classificação geográfica realizada através dos modelos de dois estágios estão dispostos na Tabela 7.

Tabela 7: Resultados de classificação genotípica, obtidos através do modelo de dois estágios.

	ACP						**PLS-DA**					
	Faixa do espectro utilizada (cm^{-1})											
	800 - 1900			**750 - 3750**			**800 - 1900**			**750 - 3750**		
Modelo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Tratamento dos espectros	Dados Puros	1ª Derivada	2ª Derivada	Dados Puros	1ª Derivada	2ª Derivada	Dados Puros	1ª Derivada	2ª Derivada	Dados Puros	1ª Derivada	2ª Derivada
Número de CP/VL	45	97	51	57	98	94	43	40	62	53	51	45
Função da base radial[a]	MQ	G	MQ	G	G	G	G	G	G	G	G	G
Normalização[b]	AE	AE	AE	AE	AE	AE	AE	MM	AE	AE	AE	AE
Parâmetro de regularização	0,001	0,9209	4,5377	0,0205	0,6788	0,0038	0,013	0,0471	0,4376	1,0707	0,0325	1,395
Dispersão	3,7925	4,0705	8,0816	5,0189	5,396	3,9621	7,0193	2,2576	4,4514	4,9102	3,6204	4,5829
Desempenho do modelo para os dados de treinamento												
Erro quadrado médio	2,33E-08	0,0103	0,0598	9,13E-05	0,0084	1,05E-06	4,76E-4	0,0057	0,0051	0,0121	5,71E-05	0,016
% de classificação correta	100	100	28,1787	100	100	100	100	99,6564	100	99,6564	100	99,3127
Desempenho do modelo para os dados de teste												
Erro quadrado médio	0,0247	0,05	0,1209	0,0346	0,0496	0,0499	0,022	0,024	0,0349	0,0313	0,0281	0,0382
% de classificação correta	89,0411	53,4247	12,3288	68,4932	50,6849	28,7671	91,7808	89,0411	64,3836	80,8219	79,4521	56,1644

[a] Função de base radial: G (Gaussiana), MQ (Multiquadrática), MQI (Multiquadrática inversa).
[b] Normalização: MM (Minimax) e AE (Autoescalonamento).

Como pode ser observado na Tabela 7, os melhores modelos encontrados foram o **1**, **7** e **8**, pois foram capazes de classificar corretamente 89,04%, 91,78% e 89,04%, respectivamente, as amostras de teste.

As médias ponderadas dos resultados da performance dos modelos **1**, **7**, **8**, e **X**, encontrados através do teorema de Bayes estão dispostos na Tabela 8.

Tabela 8: Médias ponderadas dos resultados de performance dos melhores modelos, calculados através do teorema de Bayes.

Modelo	**1**	**7**	**8**	**X**
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	0,8591
Especificidade (Treinamento)	1,0000	1,0000	0,9984	0,9958
Sensibilidade (Teste)	0,9863	0,9178	0,9315	0,6164
Especificidade (Teste)	0,4932	0,9481	0,9668	0,9870

De acordo com a Tabela 8, verificou-se que o modelo 8 obteve melhores valores de performance. O resultado completo de performance do modelo 8 está disposto na Tabela 9. Os resultados completos de performance dos modelos 1, 7 e X, estão dispostos nas Tabelas 1-A, 2-A e 3-A, respectivamente no Apêndice A.

Tabela 9: Performance do modelo 8, calculada através do teorema de Bayes.

Modelo: 8	**Faixa do espectro (cm^{-1}): 800 - 1900**		**1ª Derivada**		**PLS-DA**
Classes	**IP100**	**IP106**	**BB001**	**TU001**	**IC001**
Limiar	0,2288	0,2849	0,1390	0,2464	0,1381
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	0,9964	1,0000	0,9965	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,9286	0,9853	1,0000	0,9429	0,9722
Classes	**IP102**	**IP108**	**CT001**	**IP101**	**IP098**
Limiar	0,2398	0,1995	0,2534	0,2499	0,2460
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	0,9926	1,0000	1,0000
Sensibilidade (Teste)	1,0000	0,7500	1,0000	1,0000	1,0000
Especificidade (Teste)	1,0000	1,0000	0,9706	0,9853	0,9857
Classes	**IP104**	**IP097**	**IA059**	**IE105**	**IP103**
Limiar	0,1867	0,2702	0,2879	0,1962	0,1904
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	0,9962	0,9965	0,9964
Sensibilidade (Teste)	0,5000	1,0000	0,8571	1,0000	1,0000
Especificidade (Teste)	0,9275	0,8971	0,9394	0,9859	0,9857
Classes	**IP105**	**IP099**	**MN001**	**IE059**	**IP107**
Limiar	0,2845	0,2713	0,1496	0,1609	0,1756
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	0,9965	0,9964
Sensibilidade (Teste)	1,0000	0,8000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,9706	0,9559	0,9718	0,9718	0,9571

Para melhor análise dos dados, foram plotadas as curvas de probabilidade a posteriori da classificação realizada pelo modelo 8. As curvas de probabilidade a posteriori das classes BB001, CT001, IA059 e MN001 estão dispostas na Figura 10. As curvas de probabilidade a priori das demais classes estão dispostas nas Figuras 1-A, 2-A, 3-A e 4-A, no Apêndice A.

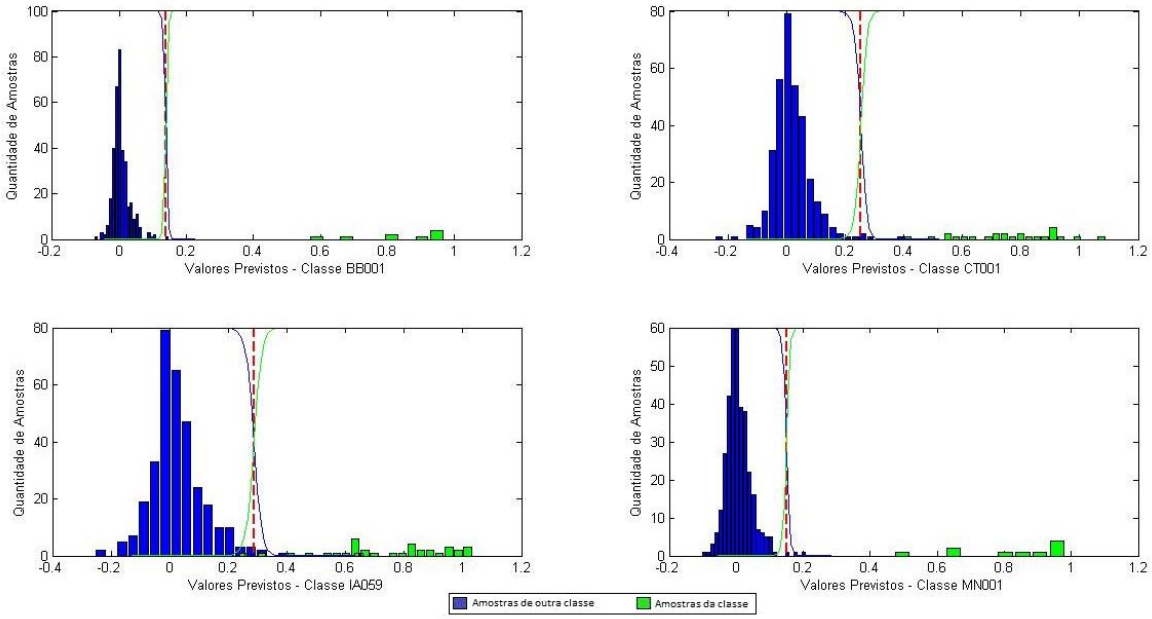

Figura 10: Curva de probabilidade a posteriori por classe para a classificação genotípica do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados e PLS-DA como primeiro estágio. Classes: BB001, CT001, IA059 e MN001.

Analisando a Figura 10, verifica-se que algumas amostras se apresentaram na região de confusão e outras classificadas incorretamente. A Figura 11 apresenta a resposta do modelo 8 para as classes BB001, CT001, IA059 e MN001. As respostas do modelo 8 para as demais classes estão dispostas nas Figuras 5-A, 6-A, 7-A e 8-A, no Apêndice A.

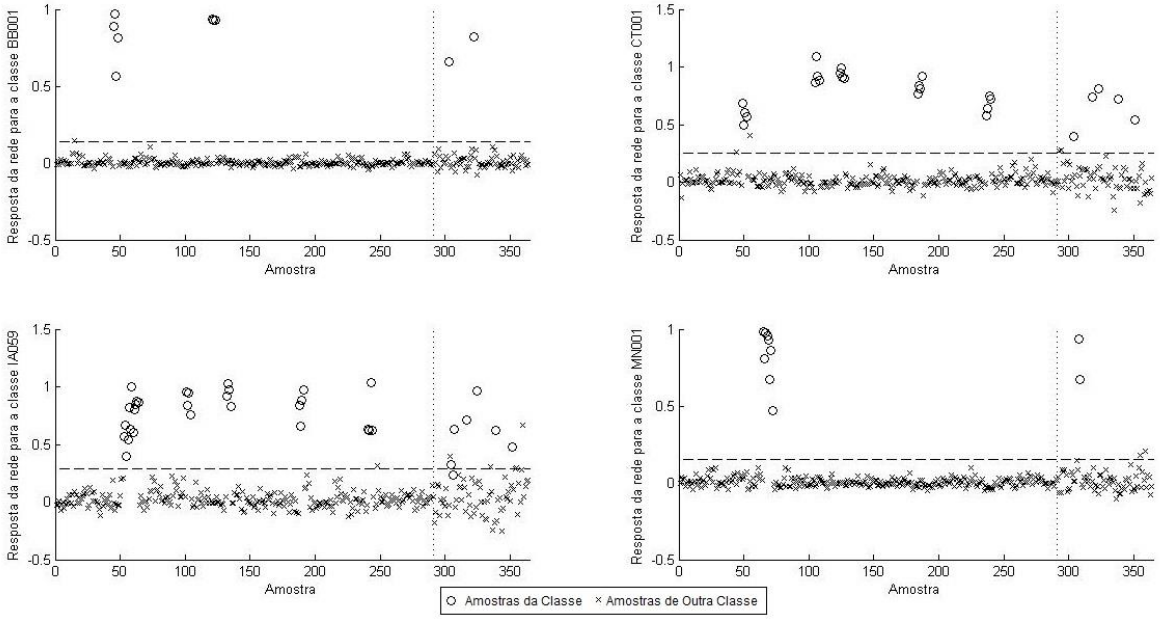

Figura 11: Resposta do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados puros e PLS-DA como primeiro estágio. Classes: BB001, CT001, IA059 e MN001. A linha pontilhada vertical separa as amostras de treinamento daquelas utilizados para o teste.

Mesmo classificando grande parte das amostras corretamente, o modelo criado ainda apresenta dificuldade na classificação genotípica das amostras. Vale salientar que ainda serão testadas algumas alterações no segundo estágio, redes neurais, na tentativa de melhor os resultados.

5. CONCLUSÃO

Alguns modelos criados com o objetivo de realizar a classificação geográfica das amostras foram capazes de classificar corretamente 100% das amostras. Porém, os modelos criados apresentaram valores de sensibilidade e especificidade inferiores ao máximo. Além disso, os resultados apresentaram amostras na região de confusão. Isso mostra que mesmo classificando corretamente as amostras, os modelos ainda devem ser modificados na tentativa de encontrar melhores resultados.

O PLS-DA proporcionou melhor classificação geográfica das amostras que os modelos de dois estágios. Portanto, não justifica-se a utilização dos modelos de dois estágios criados para a classificação geográfica destas amostras.

O PLS-DA e o modelo de dois estágios não foram capazes de realizar a classificação genotípica das amostras com êxito. Mostrando assim, a complexidade da classificação proposta.

Mesmo não classificando 100% das amostras corretamente na classificação genotípica, o modelo de dois estágios apresentou uma classificação muito superior ao PLS-DA. Isso mostra a necessidade de um modelo não linear nesse tipo de classificação.

6. REFERENCIAS

ABIC. **Indicadores da indústria de café no Brasil – 2013**. Disponível em: <http://www.abic.com.br/publique/cgi/cgilua.exe/sys/start.htm?sid=61#1910>. Acesso em: 19 de julho de 2014.

ALMEIDA, M. R.; FIDELIS, C. H. V.; BARATA, L. E. S.; POPPI, R. J. "Classification of Amazonian rose wood essential oil by Raman spectroscopy and PLS-DA with reliability estimation", **Talanta.** v.117, p.305, 2013.

BARKER, M.; RAYENS, W. Partial least squares for discrimination. **Journal of Chemometrics***,* 17, 166-173, 2003.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. Oxford: Oxford University: Springer. 738p., 2006.

BONA, E.; BORSATO, D. ; SILVA, R. S. F.; HERRERA, R. P. Aplicativo para otimização empregando o método simplex seqüencial. **Acta ScientiarumTechnology**, v.22, p. 1201-1206, 2000.

BONA, E.; SILVA, R. S. F.; BORSATO, D.; BASSOLI, D. G. Optimized Neural Network for Instant Coffee Classification through an Electronic Nose. **International Journal of Food Engineering**, v. 7, p. 6, 2011.

BONA, E.; SILVA, R. S. F.; BORSATO, D.; BASSOLI, D. G., Self-organizing maps as a chemometric tool for aromatic pattern recognition of soluble coffee. **ActaScientiarum Technology**.v.34, p. 111-119, 2012**.**

BOUVERESSE, Delphine. J.-R.; BENABID, Hamida; RUTLEDGE, Douglas N. Independent component analysis as a pretreatment method for parallel factor analysis to eliminate artefacts from multiway data. **Analytica Chimica Acta.** p. 589, 2007.

BRIANDET, R.; KEMSLEY, E. K.; WILSON, R. H. Approaches to adulteration detection in instant coffees using infrared spectroscopy and chemometrics. **Journal of the Science of Food and Agriculture**, v.71, p.359-366, 1996.

CECAFÉ. **Resumo das exportações de café – JULHO/2013**. Disponível em: **<**http://www.cecafe.com.br/Menu/dados/exportacoes/CECAF%C9%20-%20Resumo%20das%20Exporta%E7%F5es%20de%20Cafe%20JULHO%202013.pdf**>.** Acesso em: 18 de Agosto de 2013.

CIOSEK, P., BRZOZKA, Z., WROBLEWSKI, W., MARTINELLI, E., DI NATALE, C., D'AMICO, A. 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, 2005.

FERREIRA, André D et al . Desempenho agronômico de seleções de café Bourbon Vermelho e Bourbon Amarelo de diferentes origens. **Pesquisa Agropecuária Brasileira**, Brasília , v. 48, n. 4, Abr. 2013.

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

GURNEY, Kevin; **An Introduction to Neural Networks**. London: Routledge, 1997, 234p.

HAYKIN, S. **Redes Neurais**: Princípios e Prática. 2ª edição. Porto Alegre: Bookman, 2001. 900p.

HYVÄRINEN, A.; OJA, E. Independent component analysis: algorithms and applications. **Neural Networks**. v.13(4-5) p.411-430, 2000.

KEMSLEY, E. K.; RUAULT, S.; WILSON, R. H. Discrimination between Coffeaarabica and Coffeacanephora variant robusta beans using infrared spectroscopy. **Food Chemistry**. v.54, n.3, p. 321-326, 1995.

LINK, J. V. **Estudo dos Genótipos de Café Arábica Utilizando FTIR e Redes Neurais Artificiais**. Dissertação – Programa de Pós-Graduação em Tecnologia de Alimentos, Universidade Federal Tecnológica do Paraná. Campo Mourão, 2013.

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

MELEIRO, L. A. C.; VON ZUBEN, F. J.; MACIEL FILHO, R. Constructive learning neural network applied to identification and control of fuel-ethanol fermentation process**. Engineering Aplicationsof Artificial Intelligence**, v.22, p.201-215, 2009.

PARASTAR, H.; JALALI-HERAVI, M.; TAULER, R. Is independent component analysis appropriate for multivariate resolution in analytical chemistry?. **Trends in Analytical Chemistry,** v. 31. 2012.

PÉREZ-MAGARIÑO, S. et al. Comparative study of artificial neural network and multivariate methods to classify Spanish DO rose wines. **Talanta,** v. 62, n. 5, p. 983-990, 2004

SAVITZKY, A; GOLAY, M. J. E. "Smoothing and differentiation of data by simplified least squares procedures", **Analytical Chemistry***,* 38, p.1627-1639, 1964.

SILVERSTEIN, R. M.; WEBSTER, F. X.; KIEMLE, D. J. **Identificação espectrométrica de compostos orgânicos.** 7. ed. Rio de Janeiro, RJ: LTC, xiv, 490 p., 2007.

VALDERRAMA, P.; MARÇO, P. H.; LOCQUET, N.; AMMARI, F.; RUTLEDGE, D. N. A procedure to facilitate the choice of the number of factors in multi-way data analysis applied to the natural samples: Application to monitoring the thermal degradation of oils using front-face fluorescence spectroscopy. **Chemometrics and Intelligent Laboratory Systems,** p.106, v. 2, 2011.

WANG, J.; JUN, S.; BITTENBENDER; H. C.; GAUTZ, L.; LI, Q. X. Fourier Transform Infrared Spectroscopy for Kona Coffee Authentication. **Journal of Food Science**, v.74, p. 385-389, 2009.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, v.2, p.37-52, 1987.

WONG, K. H.; RAZMOVSKI-NAUMOVSKI, V.; LI, K. M.; LI, G. Q.; CHAN, K., Differentiation of Puerarialobata and Puerariathomsonii using partial least square discriminant analysis (PLS-DA), **Journal of Pharmaceutical and Biomedical Analysis,**v.84, p. 5-13, 2013.

7. APÊNDICE A

Tabela 1-A: Performance do modelo 1, calculada através do teorema de Bayes.

Modelo: 1 Faixa do espectro (cm^{-1}): 800 - 1900 Dados Puros ACP					
Classes	**IP100**	**IP106**	**BB001**	**TU001**	**IC001**
Limiar	0,0093	0,0091	0,0096	0,0094	0,0098
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,4143	0,3971	0,4789	0,5429	0,6111
Classes	**IP102**	**IP108**	**CT001**	**IP101**	**IP098**
Limiar	0,0096	0,0097	0,0091	0,0096	0,0098
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	0,8000	1,0000	1,0000
Especificidade (Teste)	0,4638	0,4928	0,5147	0,3676	0,5143
Classes	**IP104**	**IP097**	**IA059**	**IE105**	**IP103**
Limiar	0,0095	0,0094	0,0092	0,0095	0,0096
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,4783	0,4265	0,3333	0,5493	0,4714
Classes	**IP105**	**IP099**	**MN001**	**IE059**	**IP107**
Limiar	0,0094	0,0092	0,0097	0,0098	0,0094
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,5882	0,5000	0,6197	0,5915	0,4857

Tabela 2-A: Performance do modelo 7, calculada através do teorema de Bayes.

Modelo: 7	**Faixa do espectro (cm^{-1}): 800 - 1900**		**Dados Puros**		**PLS-DA**
Classes	**IP100**	**IP106**	**BB001**	**TU001**	**IC001**
Limiar	0,2049	0,3101	0,1083	0,2513	0,1404
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,9000	0,9706	0,9577	0,9000	0,9583
Classes	**IP102**	**IP108**	**CT001**	**IP101**	**IP098**
Limiar	0,2282	0,2164	0,2625	0,2494	0,2821
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	0,5000	0,8000	1,0000	1,0000
Especificidade (Teste)	0,9565	0,9855	0,9853	0,9853	1,0000
Classes	**IP104**	**IP097**	**IA059**	**IE105**	**IP103**
Limiar	0,1686	0,2696	0,2550	0,1560	0,1550
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	0,7500	0,8000	0,8571	1,0000	1,0000
Especificidade (Teste)	0,9420	0,8971	0,9394	1,0000	0,8714
Classes	**IP105**	**IP099**	**MN001**	**IE059**	**IP107**
Limiar	0,2514	0,2532	0,0956	0,1833	0,1618
Sensibilidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
Sensibilidade (Teste)	1,0000	1,0000	1,0000	1,0000	1,0000
Especificidade (Teste)	0,9706	0,9265	0,9155	0,9718	0,9286

Tabela 3-A: Performance do modelo X, calculada através do teorema de Bayes.

Modelo: X Faixa do espectro (cm-1): 800 - 1900 Dados Puros					
Classes	**IP100**	**IP106**	**BB001**	**TU001**	**IC001**
Limiar	0,3050	0,3701	0,2859	0,2905	0,2564
Sensibilidade (Treinamento)	0,7143	1,0000	1,0000	0,9167	1,0000
Especificidade (Treinamento)	0,9892	1,0000	0,9965	0,9928	1,0000
Sensibilidade (Teste)	0,3333	1,0000	1,0000	0,3333	1,0000
Especificidade (Teste)	1,0000	0,9853	1,0000	0,9571	1,0000
Classes	**IP102**	**IP108**	**CT001**	**IP101**	**IP098**
Limiar	0,3406	0,3287	0,3642	0,3608	0,3144
Sensibilidade (Treinamento)	1,0000	0,8125	0,9500	0,8500	1,0000
Especificidade (Treinamento)	1,0000	0,9891	0,9852	1,0000	1,0000
Sensibilidade (Teste)	0,7500	0,5000	0,8000	0,6000	0,6667
Especificidade (Teste)	0,9855	1,0000	0,9853	0,9853	0,9857
Classes	**IP104**	**IP097**	**IA059**	**IE105**	**IP103**
Limiar	0,3013	0,3411	0,3823	0,2390	0,2729
Sensibilidade (Treinamento)	0,8000	0,7000	0,7857	0,8750	0,5833
Especificidade (Treinamento)	0,9928	0,9963	1,0000	1,0000	0,9928
Sensibilidade (Teste)	0,2500	0,2000	0,7143	0,5000	0,6667
Especificidade (Teste)	0,9855	0,9265	0,9697	1,0000	1,0000
Classes	**IP105**	**IP099**	**MN001**	**IE059**	**IP107**
Limiar	0,3619	0,3483	0,2658	0,2694	0,2947
Sensibilidade (Treinamento)	0,9000	0,8000	1,0000	1,0000	0,8333
Especificidade (Treinamento)	0,9963	0,9926	1,0000	0,9965	0,9964
Sensibilidade (Teste)	0,8000	0,4000	1,0000	1,0000	0,3333
Especificidade (Teste)	0,9853	0,9853	1,0000	1,0000	1,0000

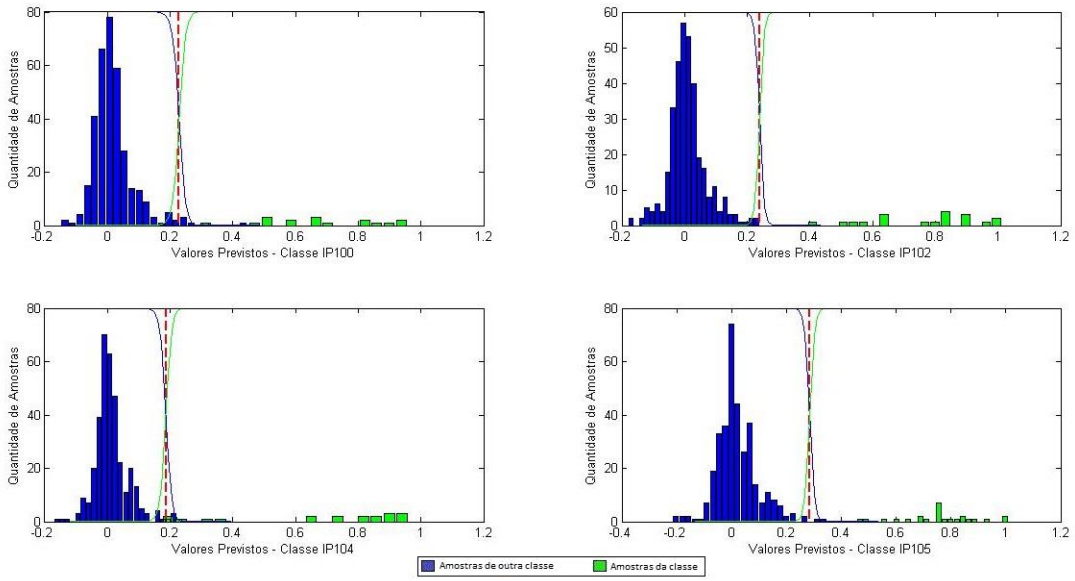

Figura 1-A: Curva de probabilidade a posteriori por classe para a classificação genotípica do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados e PLS-DA como primeiro estágio. Classes: IP100, IP102, IP104 e IP105.

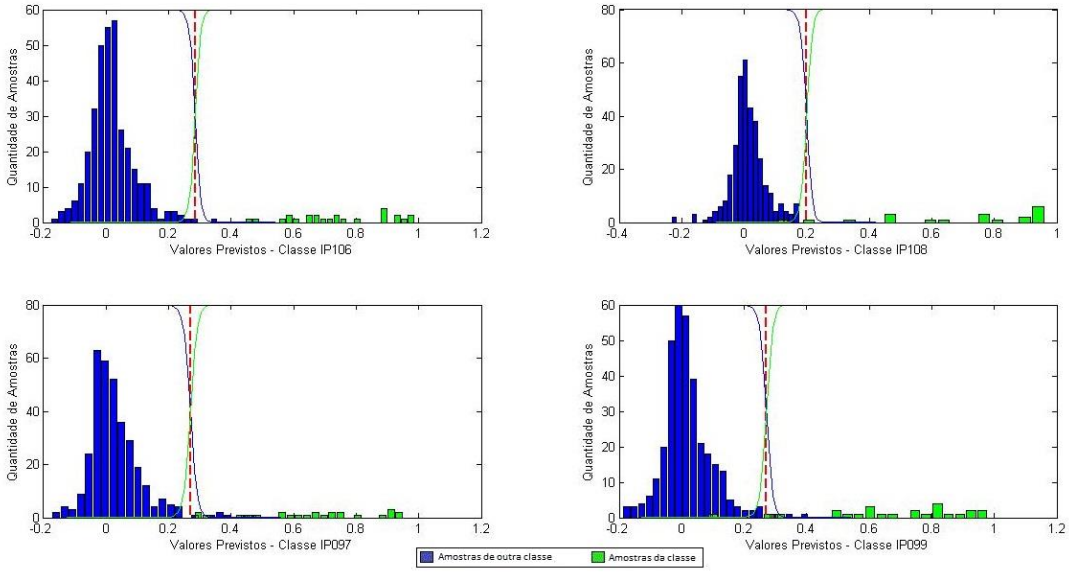

Figura 2-A: Curva de probabilidade a posteriori por classe para a classificação genotípica do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados e PLS-DA como primeiro estágio. Classes: IP106, IP108, IP097 e IP099.

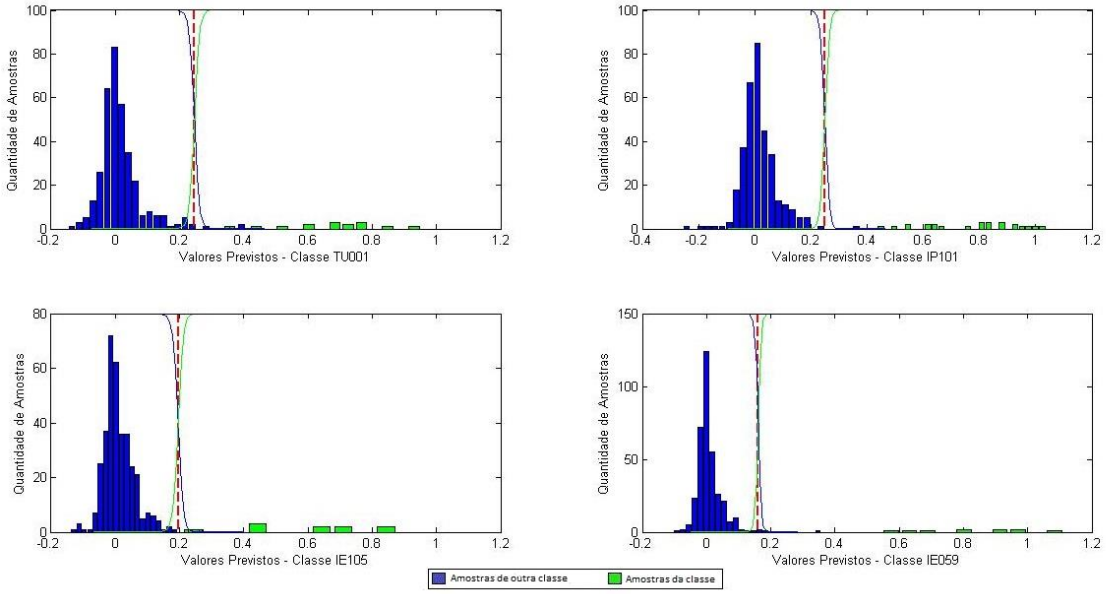

Figura 3-A: Curva de probabilidade a posteriori por classe para a classificação genotípica do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados e PLS-DA como primeiro estágio. Classes: TU001, IP101, IE105 e IE059.

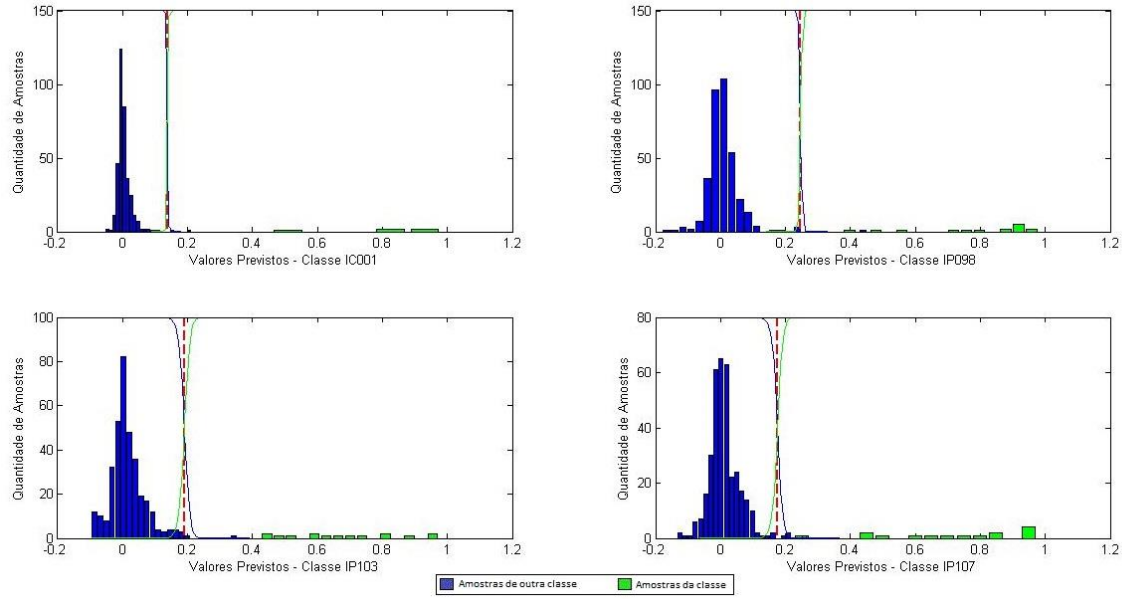

Figura 4-A: Curva de probabilidade a posteriori por classe para a classificação genotípica do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados e PLS-DA como primeiro estágio. Classes: IC001, IP098, IP103 e IP107.

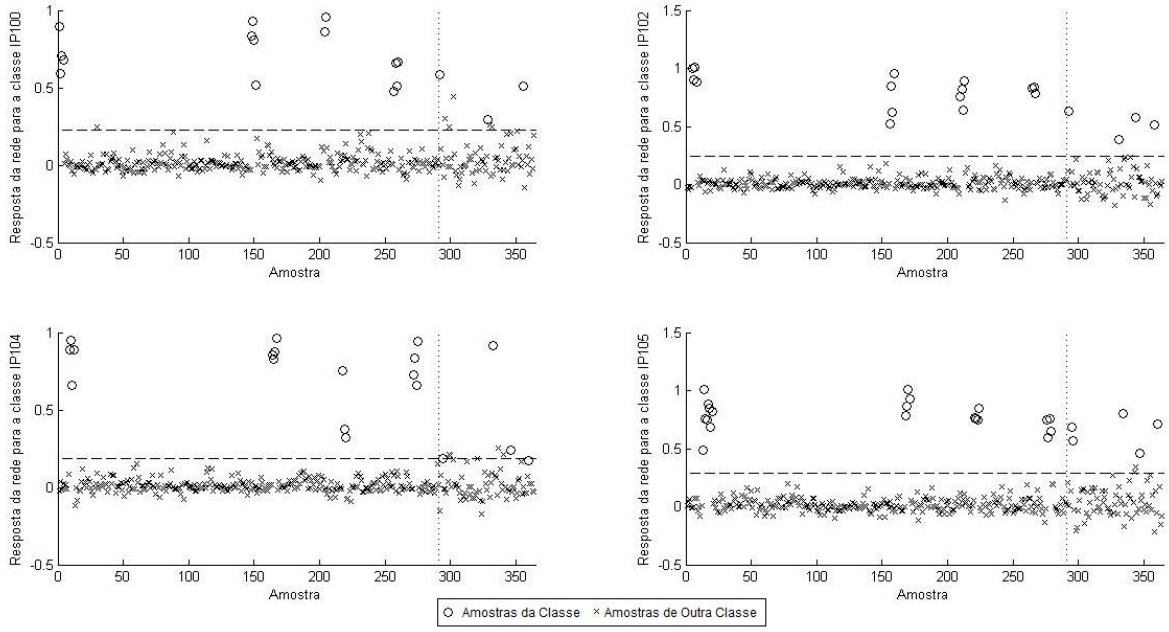

Figura 5-A: Resposta do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados puros e PLS-DA como primeiro estágio. Classes: IP100, IP102, IP104 e IP105. A linha pontilhada vertical separa as amostras de treinamento.

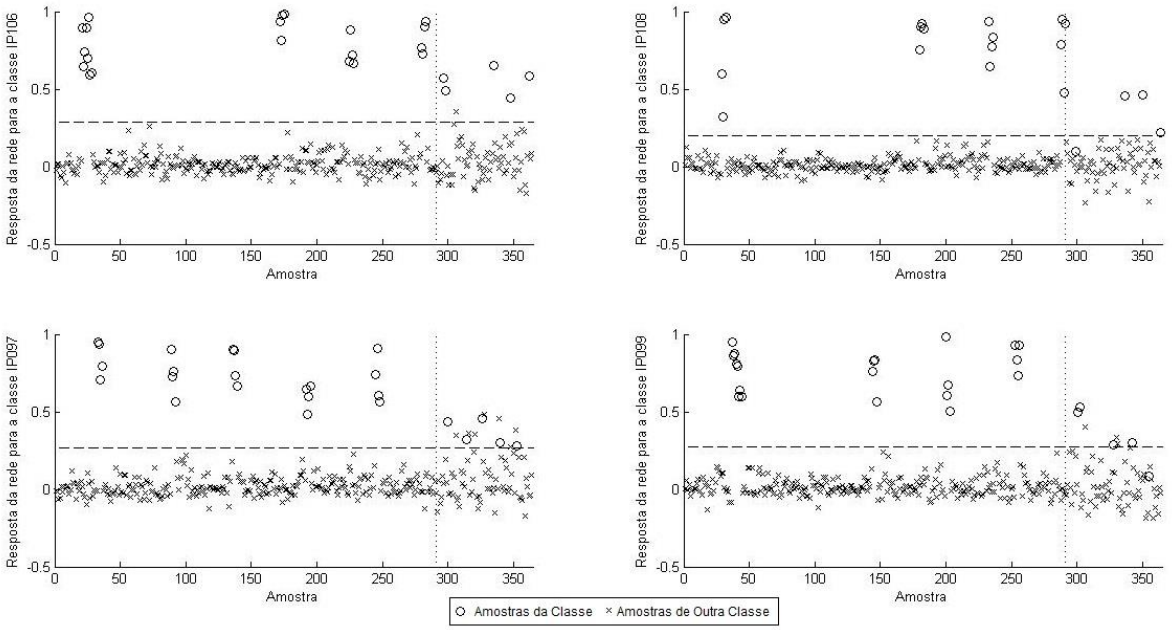

Figura 6-A: Resposta do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados puros e PLS-DA como primeiro estágio. Classes: IP106, IP108, IP097 e IP099. A linha pontilhada vertical separa as amostras de treinamento.

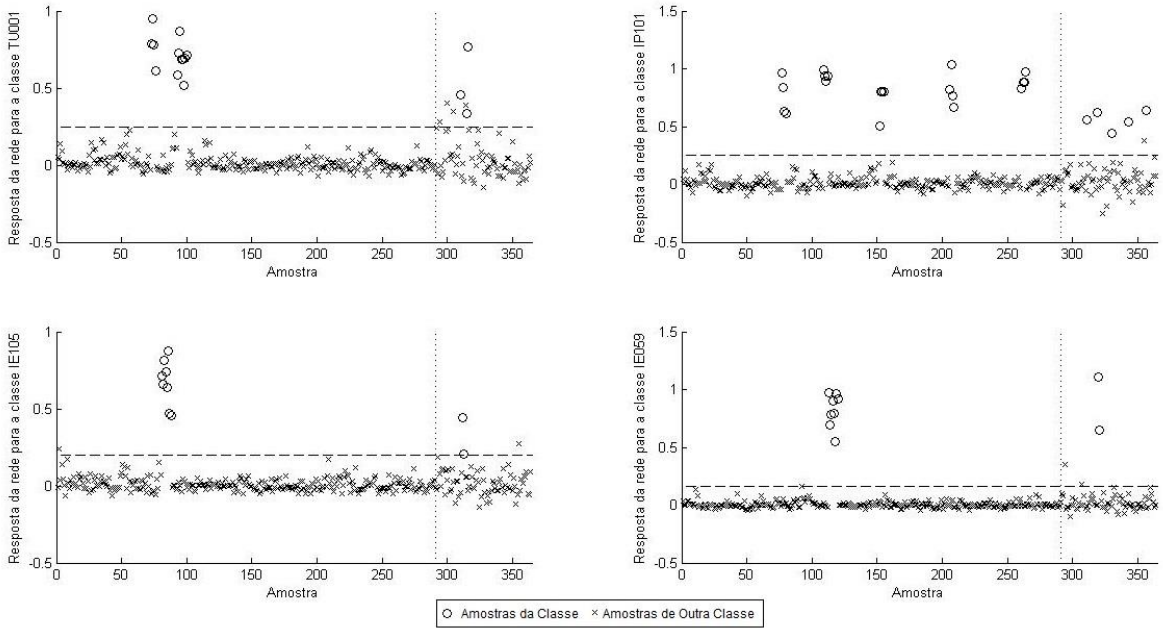

Figura 7-A: Resposta do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados puros e PLS-DA como primeiro estágio. Classes: TU001, IP101, IE105 e IE059. A linha pontilhada vertical separa as amostras de treinamento.

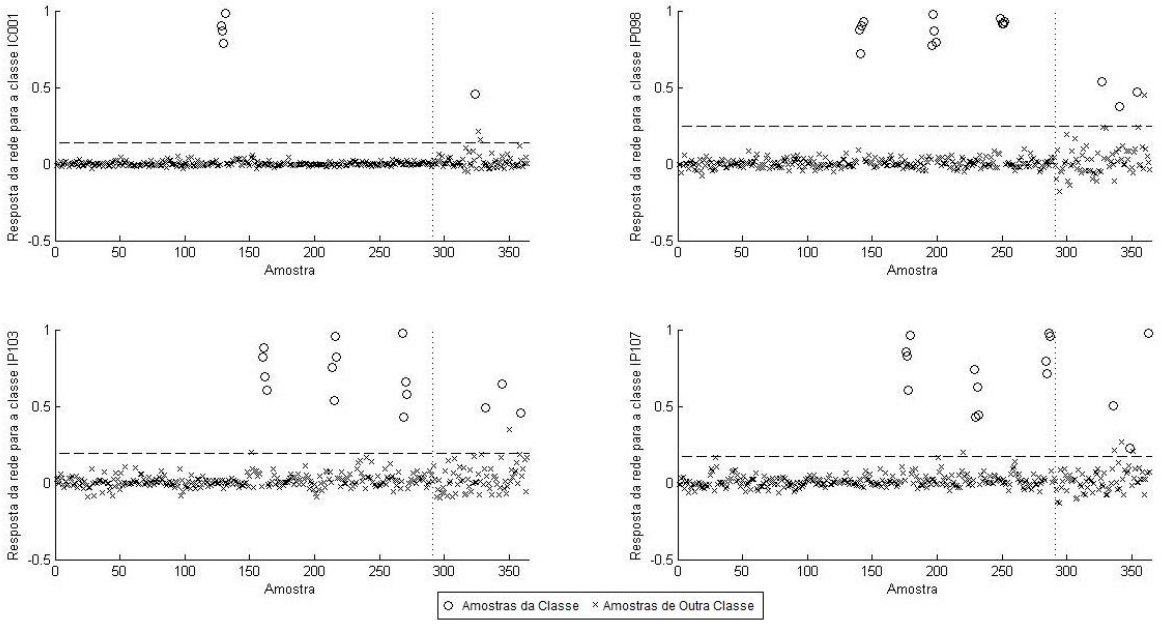

Figura 8-A: Resposta do modelo de dois estágios 8, utilizando a faixa espectral entre 800 e 1900 cm^{-1} com a primeira derivada dos dados puros e PLS-DA como primeiro estágio. Classes IC001, IP098, IP103 e IP107. A linha pontilhada vertical separa as amostras de treinamento.

8. APÊNDICE B

Tabela 1-B: Relação de amostras fornecidas pelo IAPAR - Londrina.

Genótipo	Ano	Local	N° de amostras
IP097	2009	Mandaguari	1
	2010	Paranavaí	2
		Mandaguari	1
		Londrina	1
IP098	2009	Mandaguari	1
	2010	Mandaguari	1
		Londrina	1
IP099	2009	Mandaguari	1
	2010	Paranavaí	1
		Cornélio Procópio	1
		Mandaguari	1
		Londrina	1
IP100	2009	Mandaguari	1
	2010	Paranavaí	1
		Mandaguari	1
		Londrina	1
IP101	2009	Mandaguari	1
	2010	Cornélio Procópio	2
		Mandaguari	1
		Londrina	1
IP102	2009	Mandaguari	1
	2010	Paranavaí	1
		Mandaguari	1
		Londrina	1
IP103	2009	Mandaguari	1
	2010	Mandaguari	1
		Londrina	1
IP104	2009	Mandaguari	1
	2010	Paranavaí	1
		Mandaguari	1
		Londrina	1
IE105	2010	Paranavaí	2
IP105	2009	Mandaguari	1
	2010	Paranavaí	1
		Cornélio Procópio	1
		Mandaguari	1
		Londrina	1
IP106	2009	Mandaguari	1
	2010	Paranavaí	1
		Cornélio Procópio	1
		Mandaguari	1
		Londrina	1
IP107	2009	Mandaguari	1
	2010	Mandaguari	1
		Londrina	1
IP108	2009	Mandaguari	1
	2010	Paranavaí	1
		Mandaguari	1
		Londrina	1
CT001	2009	Mandaguari	1
	2010	Paranavaí	2
		Mandaguari	1
		Londrina	1
BB001	2009	Mandaguari	1
	2010	Cornélio Procópio	1
TU001	2010	Paranavaí	3
IA059	2009	Mandaguari	1
	2010	Paranavaí	2
		Cornélio Procópio	2
		Mandaguari	1
		Londrina	1
IC001	2009	Mandaguari	1
MN001	2010	Cornélio Procópio	2
IE059	2010	Paranavaí	2