http://dx.doi.org/10.21577/0100-4042.20170014 *Quim. Nova,* Vol. 40, No. 4, 418-426, 2017

Artigo

AVALIAÇÃO DE ESPECTRÔMETRO NIR PORTÁTIL E PLS-DA PARA A DISCRIMINAÇÃO DE SEIS ESPÉCIES SIMILARES DE MADEIRAS AMAZÔNICAS

Liz F. Soares[a,b], Diego C. da Silva[a,b], Maria C. J. Bergo[a,b], Vera T. R. Coradin[a], Jez W. B. Braga[b,*] e Tereza C. M. Pastore[a]
[a]Laboratório de Produtos Florestais, Serviço Florestal Brasileiro, 70818-900 Brasília – DF, Brasil
[b]Instituto de Química, Universidade de Brasília, 70910-900 Brasília – DF, Brasil

Recebido em 12/10/2016; aceito em 20/12/2016; publicado na web em 10/02/2017

EVALUATION OF A NIR HANDHELD DEVICE AND PLS-DA FOR DISCRIMINATION OF SIX SIMILAR AMAZONIAN WOOD SPECIES. Supervising wood exploitation can be very challenging due to the existence of many similar species and the reduced number of wood identification experts to meet the demand. There is evidence that valuable endangered wood species are being smuggled disguised as other species. Near infrared spectroscopy (NIRS) and chemometrics has been successfully used to discriminate between Amazonian wood species using high resolution instruments. In this study, a handheld spectrometer was evaluated for the discrimination of six visually similar tropical wood species using PLS-DA. Woods of mahogany (*Swietenia macrophylla*) and cedar (*Cedrela odorata*), both high value tropical timber species included in Appendixes II and III of the CITES, respectively; crabwood (*Carapa guianensis*); cedrinho (*Erisma uncinatum*); curupixá (*Micropholis melinoniana*); and jatobá (*Hymenea coubaril*). The data for model development and validation take into account both laboratory and field measurements. Outlier exclusion was performed based on Hotelling T^2, residuals Q and errors in the estimated class values. The efficiency rates were higher than 90% for all species, showing that the handheld NIR combined with PLS-DA succeeded in discriminate between these species. These results stimulate the application of handheld NIR spectrometers in the supervision of wood exploitation, which can contribute to the species preservation.

Keywords: mahogany; cedar; crabwood; NIR; PLS-DA; amazon woods.

INTRODUÇÃO

A exploração e comercialização de madeiras ilegais contribuem para o crescimento contínuo das taxas de desmatamento das florestas Amazônica e demais do globo terrestre. Atualmente, há um esforço de várias instituições internacionais para combater a exportação de madeira ilegal, que envolve cifras da ordem de bilhões de dólares anuais, respeitando a legislação existente em cada país. Tal esforço tem como finalidades controlar, proibir ou desmotivar a exploração seletiva de espécies florestais produtoras de madeira ou de uma área específica explorada.[1]

Apesar de todo o empenho, existe carência em resolver uma questão básica e primordial, a de identificar rapidamente e de maneira confiável, a qual espécie florestal pertence a madeira que está sendo inspecionada. Para a identificação da madeira, geralmente desprovida de qualquer material botânico, da forma como ela é transportada e comercializada, fiscais e agentes ambientais treinados recorrem ao método convencional de anatomia de madeira, que compara os caracteres anatômicos e morfológicos da madeira examinada com a madeira de padrões depositados em xilotecas registradas.[2] As chaves de identificação, eletrônicas ou não, reúnem informações anatômica e física da madeira e facilitam a análise anatômica.[3] Contudo, ainda é necessária elevada experiência do analista para a aplicação do método com o nível de confiança necessário para realizar uma apreensão de carga ilegal. Adicionalmente, apesar dos ótimos resultados apresentados pelo método anatômico, em muitas regiões e postos de fiscalização não se dispõe de fiscais ou agentes treinados. A escassez de profissionais e o aumento contínuo da exploração e comercialização ilegais da madeira fazem com que seja urgente encontrar ferramentas eficientes que auxiliem a identificação de espécies florestais. Várias técnicas estão sendo estudadas e adaptadas para esse fim, tais como: espectroscopia de massas, determinação de isótopos estáveis, rádio-carbono, técnicas com DNA, espectroscopia no infravermelho próximo (NIRS, do inglês *Near Infrared Spectroscopy*), etc.[3-10]

A tecnologia NIRS, espectroscopia associada à analise multivariada dos espectros, permite a aquisição direta de medidas de reflectância que carregam informação dos diversos grupos funcionais presentes nas moléculas de alto peso (celulose, hemicelulose e lignina) e de menor peso molecular (extrativos) da madeira.[11] Além disso, trazem informações físicas e anatômicas, referentes à distribuição desses constituintes químicos na superfície, tornando-se uma "impressão digital" de cada espécie florestal produtora de madeira.

Por se tratar de um material complexo, o espectro de NIRS da madeira é constituído de um conjunto de bandas formadas pela sobreposição de várias transições vibracionais na região de sobretons e combinações de bandas, que requer o estabelecimento de um modelo matemático que relacione os espectros obtidos com uma ou mais propriedades de interesse, de maneira quantitativa ou qualitativa.[12] Surge, assim, a necessidade da aplicação de métodos quimiométricos como o de análise de componentes principais (PCA, do inglês *Principal Component Analysis*), regressão por mínimos quadrados parciais (PLSR, do inglês *Partial Least Squares Regression*), análise discriminante linear (LDA, do inglês *Linear Discriminant Analysis*), etc.[4-10] Esses modelos podem, então, ser utilizados para análises qualitativas ou quantitativas, dependendo dos objetivos ou do método quimiométrico empregado, de uma amostra em análise de rotina.

Em procedimentos qualitativos como a identificação de espécies produtoras de madeira, a tecnologia NIRS está intimamente ligada ao botânico, que é essencial para a construção do banco de dados de espectros necessários para o desenvolvimento dos modelos quimiométricos de classificação ou discriminação.[6]

A flora brasileira possui elevada diversidade de espécies produtoras de madeiras, sendo que muitas apresentam grande semelhança visual, mesmo em nível microscópico. A tecnologia NIRS destaca-se

*e-mail: jez@unb.br

por ser rápida, não destrutiva, reprodutível, precisa, requerer mínimo preparo da amostra, dispor de equipamento portáteis comerciais e exibir resultado da análise em tempo real.[6-8] Portanto, nos últimos anos esta técnica vem se consolidando como um método alternativo para discriminação de madeira. Trabalhos anteriores do nosso grupo de pesquisa evidenciam o potencial dessa tecnologia para a discriminação de madeiras amazônicas mogno, cedro, andiroba e curupixá, demonstrando sua aplicabilidade com amostras de diferentes países e avaliação dos fenóis totais e extrativos do mogno por NIRS.[6-9] É importante destacar que, devido à potencialidade demonstrada nos últimos anos pela NIRS aliada a métodos quimiométricos, essa tecnologia foi inserida como uma das técnicas recomendadas pelo guia de boas práticas para identificação de madeira para fins forenses, publicado no ano de 2016 pelo escritório das Nações Unidas sobre Drogas e Crime (UNODC, do inglês *United Nations Office on Drugs and Crime*) do Programa Global de Combate a crimes contra a vida selvagem e florestas (GPWLFC, do inglês *Global Programme for Combating Wildlife and Forest Crime*).[13]

Este trabalho tem como objetivo principal ampliar os estudos já realizados, apresentando resultados da discriminação de seis espécies brasileiras de madeiras anatomicamente similares: *Carapa guianensis* Aubl.(andiroba)*, Cedrela odorata* L. *(*cedro)*, Erisma uncinatum* Warm. (cedrinho)*, Micropholis melinoniana* Pierre (curupixá)*, Hymenea coubaril* L. (jatobá) e *Swetenia macrophylla* King. (mogno). Em relação aos estudos relatados anteriormente na literatura, além da adição de duas novas espécies (jatobá e cedrinho) e do uso do equipamento portátil, são apresentados modelos com um número maior e mais representativo de amostras e avanços nos critérios utilizados na identificação de amostras anômalas através do uso de limites com relação aos valores estimados de classe.[6-8]

As madeiras escolhidas são comercialmente conhecidas no mercado brasileiro e internacional. A andiroba, o mogno e o cedro são usados na construção civil, naval, móveis, instrumentos musicais e tonéis de cachaça. Possuem boa durabilidade e trabalhabilidade, podendo ser torneadas. O cedrinho, jatobá e cupurixá, também usados na construção civil, naval e em móveis, possuem baixa trabalhabilidade, são de difícil acabamento, mas são bastante duráveis. Todas são madeiras comercializadas e de difícil identificação ou separação visual.[1,2]

PARTE EXPERIMENTAL

O estudo foi realizado no Laboratório de Produtos Florestais (LPF) do Serviço Florestal Brasileiro, pertencente ao Ministério do Meio Ambiente (MMA), em colaboração com o Laboratório de Automação, Quimiometria e Química Ambiental (AQQUA) do Instituto de Química da Universidade de Brasília.

Obtenção e preparo das amostras

As espécies foram selecionadas com base no livro "Madeiras similares ao mogno (*Swietenia macrophylla* King); uma chave ilustrada para identificação anatômica em campo" editado pelo Serviço Florestal Brasileiro.[1] Das 15 espécies listadas, seis foram escolhidas para o estudo: *Carapa guianensis* Aubl. (andiroba)*, Cedrela odorata* L. (cedro)*, Erisma uncinatum* Warm. (cedrinho)*, Micropholis melinoniana* Pierre. (curupixá)*, Hymenea coubaril* L. (jatobá) e *Swietenia macrophylla* King (mogno). A maioria foi obtida na xiloteca Harry Van der Sloten da Área de Anatomia e Morfologia do LPF, em saídas de campo no município de Manuel Urbano (Acre) e nos países Guatemala, México e Peru.[14]

As amostras foram identificadas por anatomista de madeira e selecionadas para a construção do modelo de discriminação por serem madeiras similares anatomicamente. Foram analisados 922 indivíduos de árvores diferentes, sendo 103 de andiroba, 174 de cedro, 157 de cedrinho, 116 de curupixá, 61 de jatobá e 311 de mogno.

Posteriormente, as amostras foram secas à temperatura ambiente e a superfície foi polida com lixas nº 80. Essa granulação foi escolhida por ser uma lixa mais grossa que aumenta a superfície de contato da madeira e evita a formação de brilho (reflectância especular). Esse procedimento é importante para manter a uniformidade granulométrica e remover a camada externa oxidada. Os espectros NIR foram obtidos logo após o preparo.

Obtenção dos espectros NIRS

Os espectros de reflectância difusa das espécies estudadas foram obtidos com o auxílio do espectrômetro portátil MicroNir™ 1700 Spectrometer fabricado pela JDSU (Estados Unidos) com faixa espectral de 950 a 1.650 nm.

Para realização das medidas, os espectros foram obtidos a partir da superfície da madeira nas faces longitudinal, tangencial e transversal, sem que fosse feita distinção entre elas. Para a maioria das amostras, 3 espectros eram medidos em pontos distintos aleatórios. Devido a heterogeneidade natural presente nas amostras de madeira optou-se por não fazer médias dos espectros medidos em pontos diferentes de uma mesma amostra para ampliar a representatividade do conjunto de dados.

Para as espécies mogno, cedro e jatobá cerca de 20 a 30% dos espectros foram medidos em análises de campo. Cabe destacar que as saídas de campo foram planejadas com um foco maior nas espécies mogno e cedro, por serem espécies incluídas na CITES. Com relação à espécie jatobá, esta foi analisada por possuir uma madeira similar à da espécie mogno e que foi encontrada nas viagens a campo. Além disso, para essas amostras medidas em campo o número de replicatas medido não foi igual em todas as amostras, podendo variar de 5 a 10 replicatas por amostra.

Os seguintes parâmetros foram estabelecidos no software do equipamento: tempo de integração de 2.000 μs e 100 varreduras. A extremidade inicial do intervalo de comprimentos de onda dos espectros foi removida para minimizar ruídos e variações não relacionadas à diferença entre as espécies. Portanto, a região espectral selecionada para a construção dos modelos de discriminação correspondeu ao intervalo de 1.000 a 1.650 nm.

Análise Discriminante por Mínimos Quadrados Parciais

A análise de dados foi realizada empregando o modelo de Mínimos Quadrados Parciais para Análise Discriminante (PLS-DA, do inglês *Partial Least Squares for Discriminant Analysis*), sendo os cálculos efetivados no programa MATLAB versão 7.12.0 (R2011a) com pacote PLS toolbox 7.03. No desenvolvimento dos modelos foram avaliados os seguintes pré-processamentos: correção de espalhamento multiplicativo (MSC, do inglês *Multiplicative Scattering Correction*), Padronização Normal de Sinal (SNV, do inglês *Standard Normal Variate*), primeira e segunda derivada pelo algoritmo Savitzky-Golay e centragem dos dados na média).

Os espectros das amostras foram divididos em dois conjuntos, um para a calibração ou treinamento e outro para a validação, na proporção de dois terços e um terço, respectivamente. Para a divisão dos conjuntos os espectros de cada espécie foram colocados em ordem cronológica de aquisição e nesta sequência a cada três amostras, duas eram destinadas à fase de treinamento e uma para a validação. Desenvolveram-se 6 modelos PLS-DA binários, correspondendo à cada uma das 6 espécies estudadas.

Nos modelos PLS-DA, o valor de classe 1 foi atribuído às amostras de treinamento pertencentes à espécie que estava sendo

discriminada e o valor de classe 0 foi atribuído às amostras de treinamento das demais espécies.

O conjunto de treinamento foi composto por 614 amostras, incluindo amostras coletadas na xiloteca (medidas em condições de laboratório) e as amostras de mogno e cedro coletadas no Acre (medidas em campo). O restante das 308 amostras constituiu o conjunto de validação.

No modelo PLS-DA a matriz de dados (**X**) pode ser correlacionada com um vetor **y**, no qual cada classe é discriminada em relação às outras em modelos distintos, o qual é conhecido como PLS1-DA. Outra variação do modelo é quando os vetores que discriminam cada classe em relação às outras são reunidos em uma matriz **Y** e um único modelo de discriminação é construído, sendo essa variação referida como PLS2-DA na literatura.[15] Diversos trabalhos apresentam uma descrição detalhada da diferença entre esses modelos e suas propriedades.[16-18] Portanto, neste trabalho, apenas uma breve descrição do modelo será apresentada abaixo, dando enfoque maior para a otimização do modelo pela detecção de *outliers* e sua validação.

Desenvolvimento e otimização do modelo PLS-DA

Neste trabalho optou-se por utilizar apenas modelos PLS1-DA. Portanto, na fase de treinamento são utilizadas amostras cujas classes são conhecidas e, em cada modelo, o vetor **y** foi composto por valores de 0 e 1, sendo que o valor 1 foi atribuído às amostras que pertencem à classe que se pretende discriminar e o valor 0 atribuído às amostras pertencentes às outras classes. Na sequência a decomposição dos dados é realizada pelas equações abaixo:[16,17]

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \mathbf{E} \quad (1)$$

$$\mathbf{y} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{q}_a^{\mathrm{T}} + \mathbf{f} \quad (2)$$

em que $\mathbf{t}_a$ é o vetor de escores, $\mathbf{p}_a$ e q_a são os pesos referentes à primeira variável latente e **E** e **f** são as matrizes de erros de **X** e **y**, respectivamente. Para a otimização do modelo diversos pré-processamentos e métodos de seleção de variáveis podem ser utilizados, os quais consistem dos mesmos aplicados a problemas quantitativos com PLSR. O número de variáveis latentes *A*, assim como o melhor método de pré-processamento e seleção de variáveis da matriz **X** é usualmente determinado em PLS-DA através do modelo que apresenta o menor valor de erro de classificação empregando validação cruzada (CVCE, do inglês *Cross Validation Classification Error*). Contudo, é importante notar que o CVCE não penaliza amostras que apresentam elevados erros na estimativa do valor classe **y**, desde que as amostras sejam corretamente classificadas. No entanto, elevados erros nos valores de classe estimados podem ser uma indicação de que a amostra apresenta características diferentes das demais amostras do conjunto de treinamento ou da presença de uma amostra com um erro nos dados intrumentais, situações que caracterizam uma amostra anômala (do inglês *outlier*). Por outro lado, o valor da raiz quadrada do erro médio quadrático de validação cruzada (RMSECV, do inglês *Root Mean Square Error of Cros Validation*), frequentemente usado em PLSR, permite a otimização do modelo PLS-DA considerando a minimização dos erros de estimativa dos valores de classe, o qual, a princípio, tende a proporcionar uma maior separação dos valores estimados para a classe discriminada (y = 1) em relação às outras classes (y = 0). Portanto, considerando esse aspecto, neste trabalho foi empregado o RMSECV como critério para a otimização dos modelos PLS1-DA.

Outro aspecto relevante na otimização dos modelos PLS-DA é a identificação e exclusão de amostras anômalas. Com base no trabalho de Borin e Poppi e da Silva *et al.,* as amostras anômalas foram identificadas no conjunto de treinamento segundo os seguintes critérios:[19,20]

- Amostras que apresentaram valores dos parâmetros estatísticos de T^2 de Hotelling e resíduo espectral Q acima dos limites de 99,9% foram identificadas como um anômalas e removidas do conjunto de treinamento;
- Ao mesmo tempo, amostras que apresentaram resíduos Student que excederam o valor crítico do teste t descrito na ASTM E1655-05 com correção de viés proposta por da Silva *et al.* no nível de 99,9% de confiança também foram identificadas como um anômalas e excluídas do conjunto de treinamento.[20-22]

Os parâmetros T^2 de Hotelling e resíduos Q foram aplicados da mesma maneira para as amostras do conjunto de validação. Por outro lado, o teste t para resíduos de Student foi adaptado para estabelecer apenas o limite superior para as estimativas da classe 1 e o limite inferior para as estimativas da classe 0 de acordo com as equações 3 e 4:

$$y_{limite\ superior,classe\,1} = 1 \pm vies_{class\,1} + \left(t_{99,\mathrm{v}} \mathrm{RMSEC}_{vies} \sqrt{1-\bar{h}_c}\right) \quad (3)$$

$$y_{limite\ inferior,classe\,0} = 0 \pm vies_{class\,0} - \left(t_{99,\mathrm{v}} \mathrm{RMSEC}_{vies} \sqrt{1-\bar{h}_c}\right) \quad (4)$$

em que $\bar{h}_c$ é o valor médio da influência (do inglês *leverage*) observada nas amostras de treinamento, RMSEC_{vies} (do inglês *Root Mean Square Error of Calibration*) é a raiz quadrada do erro médio quadrático do conjunto de treinamento com correção de vies, $vies_{classe0}$, $vies_{classe1}$ são os vieses estimados para as classes 0 e 1, respectivamente, $t_{99,v}$ é o valor tabelado da distribuição de t-Student, com 99,9% de confiança e *n*-*A*-2 graus de liberdade e *n* é o número de amostras de treinamento. Portanto, para o conjunto de validação, foram consideradas amostras anômalas em relação à estimativa dos valores de classe aquelas que apresentaram valor de *y* maior do que $y_{limite\ superior,classe\ 1}$ ou inferior que $y_{limite\ inferior,classe\ 0}$. Em outras palavras, uma amostra de validação *i* que apresentar valor de classe (y_i) significativamente mais elevado do que os valores de **y** estimados para as amostras da classe 1 ou significativamente menor do que os valores estimados para classe 0 do conjunto de treinamento, tendo em conta o nível de confiança de 99,9%, serão anômalas.

Levando em conta esses critérios, a identificação e a exclusão de amostras anômalas foram realizadas em apenas uma etapa. Inicialmente, um primeiro modelo PLS1-DA foi construído para a discriminação de cada classe em relação às demais e os valores extremos foram excluídos do conjunto de treinamento. Em seguida, o modelo foi calculado com as amostras restantes e considerado otimizado.

Após otimização do modelo, um limite de discriminação foi calculado com base na dispersão dos valores estimados de **y** para as amostras de treinamento de forma a minimizar a ocorrência de erros positivos falsos e negativos de acordo com o teorema de Bayes.[16,18,23] Conforme definido anteriormente, é considerada a discriminação das amostras em duas classes, uma delas contendo as amostras da espécie que será discriminada, a qual será atribuído o valor de classe y=1 (classe A), e a outra contendo todas as amostras das demais espécies, as quais terão valor de classe y=0 (classe B). A partir dos valores estimados para o conjunto de treinamento são estimadas as probabilidades *a priori* (P(A) e P(B)) e as funções de densidade de probabilidade ($p(\hat{y}_i|A)$ e $p(\hat{y}_i|B)$) de cada classe. Considerando que as distribuições dos valores de classe estimados para as classes A e B se aproximam de uma distribuição normal, esses parâmetros podem ser definidos como:[23]

$$P(A) = \frac{I_A}{I_A + I_B} \quad (5)$$

$$P(B) = \frac{I_B}{I_A + I_B} \quad (6)$$

$$p(\hat{y}_i|A) = \frac{1}{s_A\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\hat{y}_i - \bar{\hat{y}}_A}{s_A}\right)^2} \quad (7)$$

$$p(\hat{y}_i|B) = \frac{1}{s_B\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\hat{y}_i - \bar{\hat{y}}_B}{s_B}\right)^2} \quad (8)$$

A partir desses parâmetros, a probabilidade de uma amostra *i* pertencer às classes A ou B podem ser determinadas, respectivamente, por:[18,23]

$$P(\hat{y}_i|A) = \frac{p(\hat{y}_i|A)P(A)}{p(\hat{y}_i|A)P(A) + p(\hat{y}_i|B)P(B)} \quad (9)$$

$$P(\hat{y}_i|B) = \frac{p(\hat{y}_i|B)P(B)}{p(\hat{y}_i|A)P(A) + p(\hat{y}_i|B)P(B)} \quad (10)$$

De acordo com a regra de Bayes uma amostra é atribuída à classe A se $P(\hat{y}_i|A) > P(\hat{y}_i|B)$, caso contrário a amostra é atribuída à classe B. Outra maneira de tomar essa decisão é através da determinação do limite de discriminação, o qual é obtido através do valor de y no qual $P(\hat{y}_i|A) = P(\hat{y}_i|B)$. Considerando que o denominador das equações 9 e 10 são iguais, a determinação do limite se simplifica a:[23]

$$p(\hat{y}_i|A)P(A) = p(\hat{y}_i|B)P(B) \quad (11)$$

O limite de discriminação é então obtido pela substituição das equações 5 a 8 na equação 11, substituição dos valores experimentais e determinação do valor de y. Detalhes dessas operações são apresentados no material suplementar. É importante destacar que se as probabilidades de ocorrência das classes A ou B forem iguais (P(A)=P(B)), as equações 9 e 10 passarão a depender apenas das funções de densidade de probabilidade de A e B. Na prática, em muitas situações o conjunto de treinamento não permite a obtenção de amostras que sejam boas aproximações das probabilidades de ocorrência das classes estudadas em amostras futuras. Nesses casos, é aconselhável a consideração de que P(A)=P(B).

Portanto, uma amostra genérica *i* é identificada como pertencente à classe 1 caso seu valor de classe estimado (y_i) for maior que o valor do limite de discriminação do modelo PLS1-DA correspondente. Caso contrário, essa amostra será identificada como pertencendo à classe 0, que contém as amostras de todas as demais classes modeladas.

Determinação de figuras de mérito dos modelos PLS-DA

A validação dos modelos de discriminação foi avaliada de acordo com o cálculo das figuras de mérito, conforme descrito por Botelho *et al.* e definidas a seguir.[18]

A taxa de falsos positivos (TFP) representa o percentual de amostras que apresentaram erros falso positivos e é calculada como a relação entre o número absoluto de falsos positivos (FP) e a soma do número absoluto de erros falso positivos (FP) e verdadeiros negativos (VN) multiplicada por 100, representada pela equação:

$$TFP = \frac{FP}{FP + VN} 100 \quad (12)$$

Por outro lado, a taxa de falsos negativos (TFN) representa o percentual de amostras que apresentou erros falso negativos, sendo calculada como a relação entre o número absoluto de falsos negativos (FN) e a soma do número absoluto de erros falso negativos (FN) e verdadeiros positivos (VP) multiplicada por 100, representada pela equação:

$$TFN = \frac{FN}{FN + VP} 100 \quad (13)$$

A especificidade (SPEC) representa o percentual de amostras pertencentes às outras classes (y=0) que foram identificadas como pertencentes a essas classes. Essa figura de mérito é calculada pela razão entre o número absoluto de verdadeiros negativos (VN) e a soma do número absoluto de verdadeiros negativos (VN) e dos erros falso positivos (FP) multiplicado por 100, representada pela equação:

$$SPEC = \frac{VN}{VN + FP} 100 \quad (14)$$

De forma complementar, a sensibilidade (SEN) representa o percentual de amostras pertencentes à classe discriminada que foram identificados como sendo dessa classe. Portanto, sendo calculada como a razão entre o número absoluto de verdadeiros positivos (VP) e a soma do número absoluto de verdadeiros positivos (VP) e dos erros falso negativos (FN) multiplicada por 100, representada pela equação:

$$SEN = \frac{VP}{VP + FN} 100 \quad (15)$$

Por fim, a taxa de eficiência (TEF) dos modelos de discriminação pode ser obtida pela diferença entre o valor de 100% e a soma das taxas de erros falso negativos (TFN) e falso positivos (TFP), representada pela equação:

$$TEF = 100 - (TFN + TFP) \quad (16)$$

O desenvolvimento dos modelos e posterior validação seguiram o procedimento descrito na Figura 1.

RESULTADOS E DISCUSSÃO

Os espectros médios das amostras de treinamento de cada uma das espécies florestais antes e após a aplicação do pré-processamento de 1ª derivada são apresentados na Figura 2. Visualmente, pode-se observar que os espectros são muito similares e que existe uma variação significativa de linha de base (Figura 2A). Tendo em vista essa grande semelhança, a simples identificação das espécies pela visualização dos espectros pode ser descartada, sendo necessária a análise dos dados por modelos quimiométricos. Além disso, tendo em vista que a madeira é composta por uma estrutura química complexa, é difícil realizar uma atribuição precisa às bandas que são observadas. Contudo, de acordo com o trabalho de Schwanninger *et al.*,[24] pode-se atribuir os principais sinais observados como se segue: (1) banda centrada em 1200 nm referente principalmente ao 2º sobretom do estiramento da ligação C–H das moléculas celulose e hemicelulose, mas contendo também absorção do 2º sobretom do estiramento assimétrico das ligações C–H e HC=CH das moléculas de lignina; (2) banda localizada entre 1350 a 1400 nm referente principalmente aos sinais do 1º sobretom do estiramento e deformação angular da ligação C–H das moléculas celulose e hemicelulose e (3) banda centrada em 1470 nm referente 1º sobretom do estiramento da ligação O–H das moléculas celulose, hemicelulose e água.

A Tabela 1 apresenta a composição química, em termos dos componentes majoritários da madeira (lignina, celulose, extrativos e teor de cinzas), e densidade básica das espécies estudadas. Esses dados permitem observar algumas das diferenças existentes entre

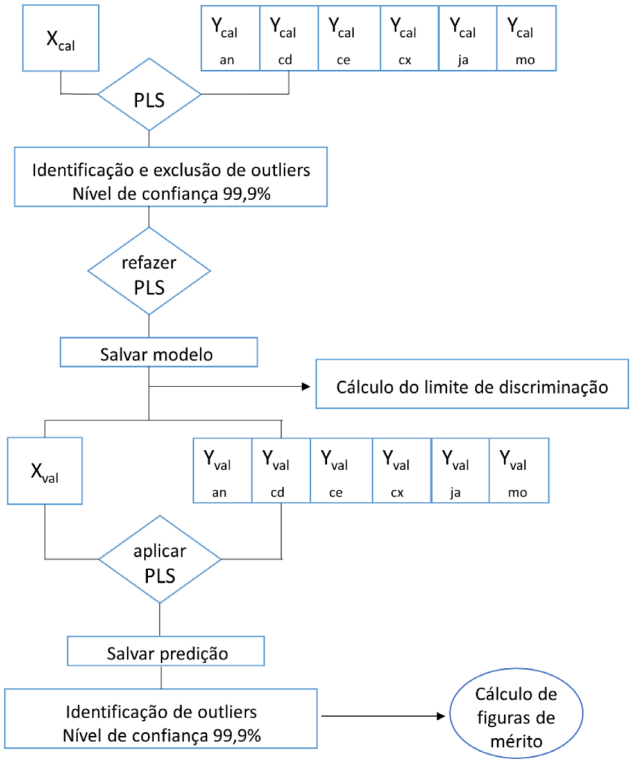

Figura 1. *Procedimento para realizar o desenvolvimento e validação dos modelos PLS-DA. (X_{cal}) matriz de espectros de treinamento, (y_{cal}) vetor de classes de treinamento para cada espécie (an = andiroba, cd = cedrinho, ce = cedro, cx = curupixá, já = jatobá e mo = mogno), (X_{val}) matriz de espectros de validação, (y_{val}) vetor de classes de validação*

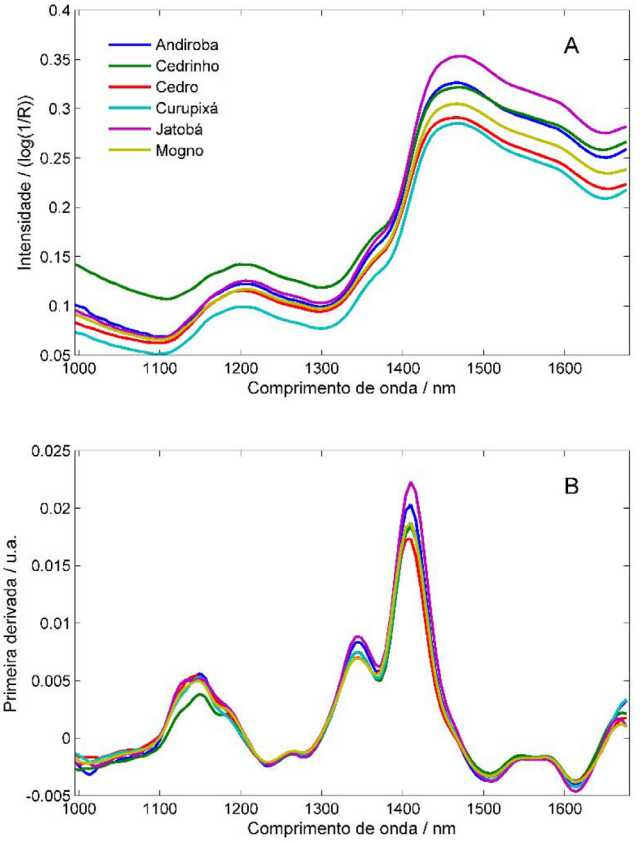

Figura 2. *Espectros médios do conjunto de treinamento sem pré-processamento (A) e pré-processados com primeira derivada (B) de cada uma das 6 espécies florestais*

as espécies e que justificam o sucesso da discriminação por NIR. Observa-se que a espécie mogno apresenta um teor de celulose maior que o observado nas outras espécies. Além disso, o teor de extrativos de mogno se diferencia das espécies andiroba, cedrinho e cedro, sendo semelhante com relação ao teor percentual ao jatobá. Contudo, deve-se destacar que apesar do teor percentual de extrativos das espécies mogno e jatobá ser semelhante, os compostos que formam os extrativos podem ser diferentes, de forma que, ao contrário dos componentes da Tabela 1, um teor próximo de extrativos não implica necessariamente em uma similaridade entre espécies. Observa-se ainda que as espécies andiroba, cedrinho e cedro apresentam variações significativas entre seus teores de extrativos. É importante ressaltar ainda que o espectro NIRS é uma resposta que varia não somente em relação à quantidade de componentes químicos presentes na madeira. A maneira com que esses compostos estão distribuídos e organizados também é um fator significativo. Parte desses aspectos se reflete na densidade dessas madeiras, as quais também apresentam variações significativas. Outro fato que chama a atenção na Tabela 1 é a ausência de dados para a espécie *Micropholis melinoniana* (curupixá). No melhor de nosso conhecimento, não existem dados na literatura sobre os parâmetros incluídos na Tabela 1 para essa espécie, que é nativa do Brasil.

A partir dos pré-processamentos avaliados para o desenvolvimento dos modelos (MSC, SNV, primeira e segunda derivada pelo algoritmo Savitzky-Golay e centragem dos dados na média), a combinação da primeira derivada por Savitzky-Golay (polinômio de 2ª ordem e janela de 5 pontos) e dados centrados na média foi o pré-processamento mais eficiente para minimizar os deslocamentos da linha de base e obter menores erros de classificação para todas as 6 espécies. A partir dos dados pré-processados, os modelos de discriminação foram desenvolvidos e otimizados conforme o procedimento descrito na Figura 1. Na otimização dos modelos

Tabela 1. Composição química em termos dos componentes majoritários da madeira e densidade básica observada nas espécies estudadas

	Andiroba	Cedrinho	Cedro	Curupixá	Jatobá	Mogno
Celulose (%)	46,5[25]	48,0[25]	48,1[25]	-	42,8[25]	63,3[26]
Lignina insolúvel (%)	31,1[25]	32,8[25]	32,7[25]	-	30,3[25]	-
Lignina solúvel (%)	1,8[25]	0,8[25]	0,7[25]	-	1,1[25]	-
Lignina total (%)	32,9[25]	33,6[25]	33,4[25]	-	31,4[25]	31,1[26]
Extrativos (%)	3,1[25]	1,7[25]	5,4[25]	-	8,5[25]	8,38[9]
Hemicelulose (%)	20,6[25]	18,4[25]	18,5[25]	-	25,8[25]	-
Teor de cinzas (%)	0,3[25]	0,8[25]	0,5[25]	-	0,3[25]	0,36[26]
Densidade (g/cm³)	0,59[3]	0,46[3]	0,38[3]	0,58[3]	0,76[3]	0,5 – 0,72[3]

tentou-se ainda a seleção de variáveis pela inspeção dos coeficientes de regressão, algoritmo de PLS por intervalos (IPLS, do inglês *Interval PLS*) e seleção de preditores ordenados (OPS do inglês *Ordered Predictors Selection*).[27,28] Contudo, nenhum dos métodos resultou em melhora nos resultados. Portanto, foi utilizada toda a região espectral.

Os principais parâmetros dos modelos de discriminação de cada espécie para a fase de calibração são apresentados na Tabela 2. Comparando-se o número de variáveis latentes utilizados na modelagem dos dados desse trabalho com os obtidos nos modelos de discriminação desenvolvidos anteriormente com quatro das espécies e realizava medidas em um espectrômetro NIR de bancada, pode-se constatar um aumento significativo de cerca de 7 variáveis latentes para 15.[6] Contudo, deve-se destacar que os dados analisados anteriormente foram obtidos em amostras na forma de serragem com tamanho de partícula controlado, umidade controlada e um conjunto de amostras significativamente menor. Por outro lado, os dados do presente trabalho foram obtidos em madeira sólida, sem controle de umidade, realização de parte das medidas em campo e de um número significativamente maior de amostras. Todos esses fatores fazem com que a modelagem desse conjunto de dados seja muito mais complexa, o que levou a uma menor taxa de eficiência e uso de maior número de variáveis latentes.

Observa-se, ainda na Tabela 2, que o número de amostras anômalas excluídas foi relativamente pequeno, sendo no máximo igual a 1,3% das amostras de calibração para a discriminação da espécie andiroba. Além disso, pode-se observar ainda que as taxas de eficiência foram sempre superiores a 90%, variando entre 90,4 a 99,7%, o que demonstra que os modelos apresentaram uma elevada taxa de acerto. De forma geral, os valores de TFN e TFP foram próximos para a maioria das espécies, revelando que não há uma tendência em erros em uma direção. A única exceção foi o modelo para a discriminação da espécie andiroba, para qual TFN foi o dobro de TFP.

Na Figura 3 é apresentado o gráfico dos valores obtidos para os parâmetros T^2 de Hotelling e resíduos Q. Para realizar a exclusão apenas de amostras que tenham uma alta probabilidade de serem anômalas foram considerados limites com 99,9% de confiança. Portanto, apenas as amostras localizadas no quadrante superior direito da Figura 3 foram excluídas por esses critérios. Pode ser observado que a maioria das amostras anômalas pertencem à espécie cedro, esse fato pode ser explicado devido a algumas amostras dessa espécie terem sido medidas em campo, o que pode ter acarretado em amostras em condições significativamente diferentes das amostras utilizadas na fase de treinamento do modelo, como por exemplo diferentes teores de umidade. Contudo, conforme pode ser observado na Tabela 3, que apresenta os resultados das figuras de mérito para as amostras de validação, a espécie cedro foi a que teve o maior número de amostras anômalas, sendo aproximadamente igual a 5,1% do total de amostras de validação, o que pode ser considerado aceitável levando em conta

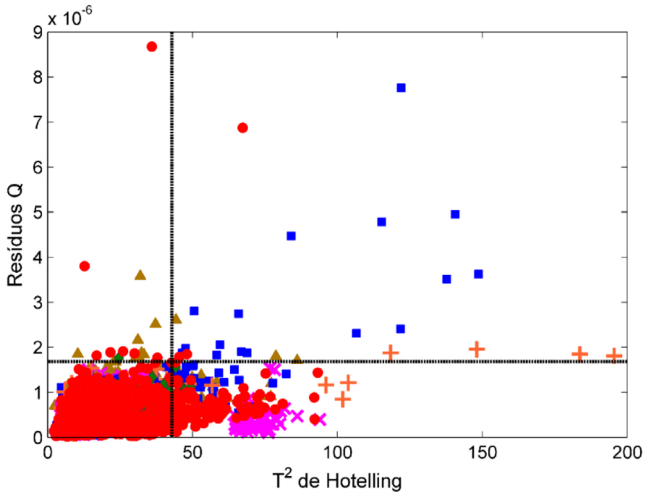

Figura 3. *Gráfico dos valores de T^2 de Hotelling e resíduos Q obtidos para o modelo PLS-DA para a discriminação da espécie mogno em relação às demais espécies. (△) andiroba, (×) cedrinho, (□) cedro, (+) curipixá, (◇) jatobá, (○) mogno, (símbolos vazios) calibração, (símbolos cheios) validação, (---) limites considerando 99,9% de confiança*

que a madeira é tipo de amostra que apresenta elevada variabilidade e heterogeneidade.

A Figura 4 apresenta a distribuição dos valores estimados obtidos tanto para o conjunto de treinamento quanto de validação. As amostras anômalas detectadas na etapa de validação são destacadas nessa figura como cor preta, na qual se observa a presença de amostras fora dos limites estabelecidos para as estimativas dos valores de classe e amostras dentro dos limites, que foram caracterizadas como anômalas pelos elevados valores de T^2 de Hotelling e resíduos Q. É interessante observar que nesse estudo muitas das amostras identificadas como anômalas por estarem fora dos limites definidos equações 3 e 4 não acarretariam em erros de discriminação. Contudo, como o PLS-DA, em sua essência, é um modelo de regressão, e em aplicações quantitativas elevados erros na estimativa da propriedade de interesse caracterizam amostras anômalas, o uso dos limites para os valores estimados de classe representa um parâmetro a mais de segurança que pode ajudar na prevenção de erros de classificação quando forem analisadas amostras em condições distintas das empregadas na fase de treinamento ou análise de amostras de classes/espécies não modeladas. As amostras identificadas como anômalas não foram consideradas para o cálculo dos valores de TFN, TFP e TEF.

Na Figura 4 ainda pode ser observado que a dispersão observada nos conjuntos de treinamento e validação não apresenta diferença significativa, indicando a ausência de sobre ajuste nos modelos PLS-DA.

Tabela 2. Descrição dos conjuntos de dados, parâmetros e figuras de mérito obtidas na fase de treinamento

Espécie	Andiroba	Cedrinho	Cedro	Curupixá	Jatobá	Mogno
Nº de espectros de treinamento			2211			
VL	15	15	15	14	14	15
Outliers treinamento	28	25	26	26	18	22
Limite de discriminação	0,271	0,374	0,412	0,374	0,234	0,463
TFP	3,0	0,6	0,2	0,3	0,1	3,4
TFN	6,6	0	0,2	0	0	2,8
TEF	90,4	99,4	99,6	99,7	99,9	93,7

TFP: Taxa de erros falso positivo; TFN: Taxa de erros falso negativo; TEF: Taxa eficiência; VL: variáveis latentes.

Tabela 3. Figuras de mérito obtidas pela análise das amostras do conjunto de validação

Espécie	Andiroba	Cedrinho	Cedro	Curupixá	Jatobá	Mogno
Nº de espectros de validação				1327		
Nº de *outliers* de validação	50	24	68	22	23	31
FP	40	8	2	13	8	31
FN	3	0	3	0	0	12
VP	93	155	337	111	26	543
VN	1141	1139	917	1181	1270	715
TFP	3,38	0,69	0,21	1,08	0,62	4,15
TFN	3,12	0,00	0,88	0,00	0,00	2,16
SPEC	96,62	99,31	99,79	98.92	99,38	95,85
SEN	96,88	100,00	99,12	100,00	100,00	97,84
TEF	93,50	99,30	98,90	98,91	99,37	93,68

FP (Falso Positivo), FN (Falso Negativo), VP (Verdadeiro Positivo), VN (Verdadeiro Negativo), TFP (Taxa de Falso Positivo), TFN (Taxa de Falso Negativo), SPEC (Especificidade), SEN (Sensibilidade) e TEF (Taxa de Eficiência).

Observa-se, ainda, que para algumas espécies foram encontrados erros sistemáticos negativos significativos em relação ao valor de classe estimado para a classe discriminada (y=1). Esses erros sistemáticos podem ser constatados visualmente pelo deslocamento do centro da distribuição da classe discriminada, sendo que os valores esperados de y devem ser iguais a 1, para valores menores. As médias dos valores estimados para cada classe discriminada nas amostras de calibração foram 0,6704, 0,8479, 0,9009, 0,8245, 0,6388 e 0,8530 para andiroba, cedrinho, cedro, curupixá, jatobá e mogno, respectivamente, o que indica que os maiores erros sistemáticos foram obtidos para as espécies andiroba e jatobá, respectivamente. A existência desse viés nos valores estimados de classe destaca a importância da correção para a identificação de amostras anômalas pelos limites estabelecidos pelas equações 3 e 4. Além disso, outra consequência direta desse viés é o deslocamento do limite de discriminação para valores menores, conforme pode ser observado na Tabela 2. Aparentemente, não há uma explicação para essas duas espécies terem apresentado um maior erro sistemático em relação às demais.

Com relação às figuras de mérito obtidas na etapa de validação, essas se mostraram compatíveis com as observadas nas amostras de treinamento, evidenciando que não há indício de sobreajuste. Baixos valores de TFP e TFN foram obtidos para todas as seis espécies, sendo que as melhores discriminações obtidas para as espécies cedrinho e jatobá. Essa melhor discriminação pode ser observada na Figura 4, na qual claramente uma menor dispersão e maior separação entre as distribuições foi observada para essas espécies. Observando os dados da Tabela 1, pode-se constatar que a espécie jatobá apresenta teor de celulose e densidade distinto das demais. Da mesma forma a espécie cedrinho apresenta alguns parâmetros da Tabela 1 distintos em relação às demais espécies, sendo estes o teor de extrativos e a densidade. Além disso, anatomicamente essas duas espécies, além da espécie curupixá, são as que apresentam maiores diferenças. Todos esses fatores são algumas das razões para a melhor discriminação dessas espécies. De forma geral, mesmo a espécie que apresentou menor TEF apresentou valores superiores a 90%, o que demonstra a eficiência dos modelos PLS-DA desenvolvidos com um espectrômetro NIRS portátil.

Novamente, realizando uma comparação entre os resultados obtidos nesse trabalho e os resultados obtidos com modelos com apenas quatro espécies em condições mais controladas e uso de um equipamento de bancada, pode-se constatar que anteriormente sempre se obteve uma perfeita discriminação, obtendo-se taxas de eficiência de 100%.[6] Conforme destacado anteriormente, os fatores que impactam na variabilidade dos espectros nos dados apresentados no presente trabalho é significativo (medidas em madeira sólida, amostras com teor de umidade variado, medidas em campo, variabilidade amostral significativamente maior, uso de equipamento portátil com menor resolução e sensibilidade). Contudo, mesmo com esses fatores a menor taxa de eficiência observada ainda foi de 90%, o que indica a viabilidade do método.

CONCLUSÕES

Os resultados apresentados demonstram que, mesmo com a utilização de um equipamento NIR portátil com uma faixa espectral restrita apenas à região de sobretom, foram obtidos modelos PLS-DA que permitiram a discriminação das seis espécies estudadas, ampliando dessa forma os estudos anteriores que abordavam apenas quatro espécies florestais produtoras de madeira que são nativas no Brasil. Para as seis espécies foram observadas taxas de eficiência acima de 90%, o que comprova a possibilidade do uso dessa técnica instrumental em escala portátil aliada a modelos PLS-DA para discriminação de madeiras em campo com elevada taxa de acerto.

Foi empregado um conjunto representativo de amostras, obtidas em diferentes localidades, o que contribui para o aumento da variabilidade das amostras de uma mesma espécie. Contudo, mesmo com essa grande variabilidade o método por NIRS e PLS-DA se mostrou eficiente.

As utilizações dos limites para as estimativas dos valores de classe permitiram a identificação e exclusão de amostras anômalas com valores estimados significativamente menores que 0 e maiores que 1, podendo contribuir para a identificação de amostras que não pertencem às populações das espécies estudadas.

MATERIAL SUPLEMENTAR

O conteúdo do material suplementar utilizado neste trabalho está disponível em http://quimicanova.sbq.org.br, na forma de arquivo PDF, com acesso livre.

AGRADECIMENTOS

Os autores agradecem ao programa ITTO-CITES, CNPq (processos 473936/2013-5 e 308748/2015-8), INCTBio, CAPES e FAPDF pelo auxílio financeiro.

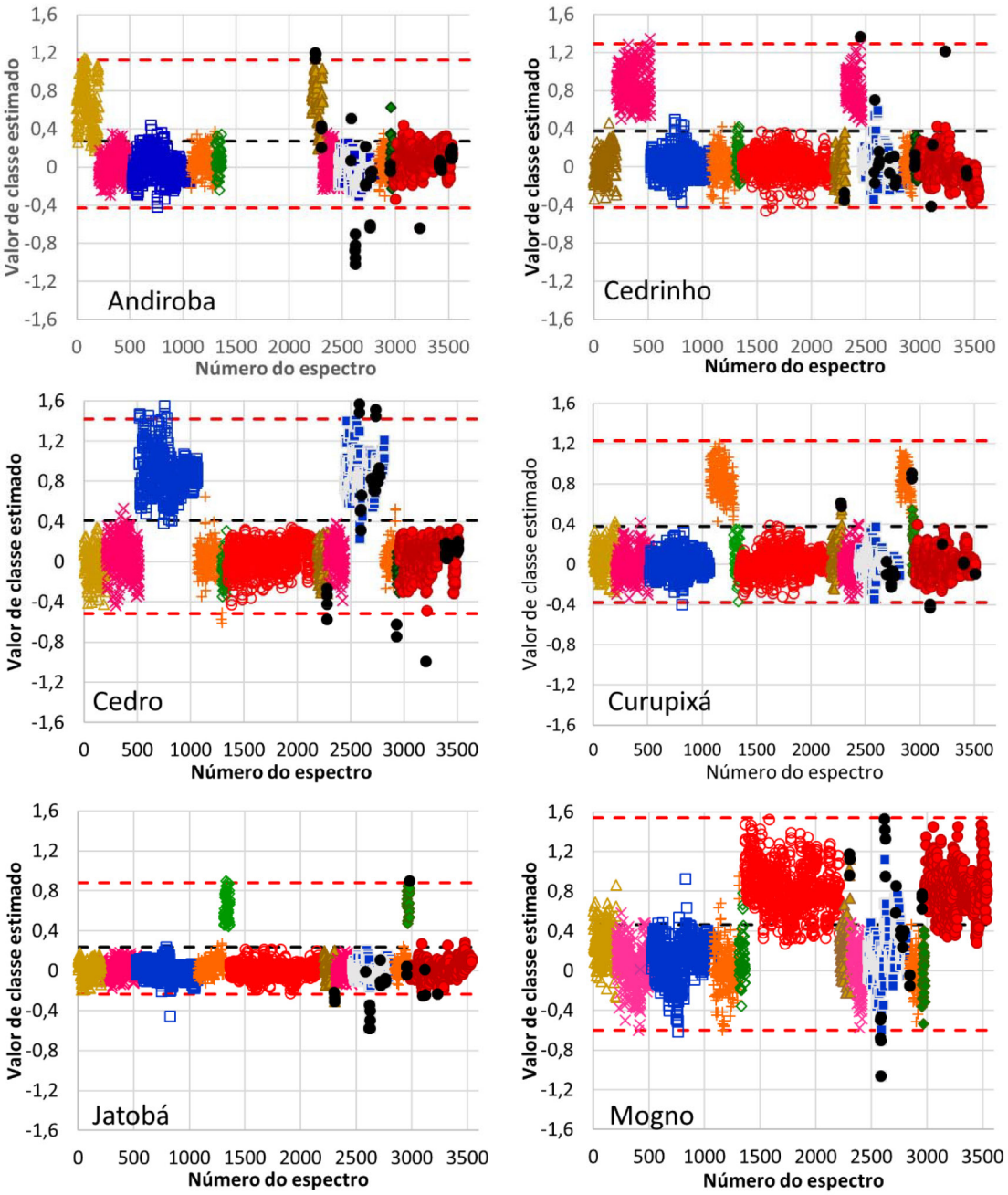

Figura 4. *Distribuição dos valores estimados do conjunto de treinamento e validação para os modelos PLS-DA para as seis espécies. (△) andiroba, (×) cedrinho, (□) cedro, (+) curipixá, (◇) jatobá, (○) mogno, (símbolos vazios) calibração, (símbolos cheios) validação e (símbolos em preto) amostras an*ômalas *do conjunto de validação*

REFERÊNCIAS

1. Dormantt, E. E.; Boner, M.; Braun, B.; Breulmann, G.; Degen, B.; Espinoza, E.; Gardner, S.; Guillery, P.; Hermanson, J. C.; Koch, G.; Lee, S. L.; Kanashiro, M.; Rimbawanto, A.; Thomas, D.; Wiedenhoeft, A. C.; Yin, Y.; Zahnen, J.; Lowe, A.; *Biological Conservation* **2015**, *191*, 790.
2. Sousa, M. H.; Megliano, M. M.; Camargos, J. A. A.; Sousa, M. R.; *Madeiras tropicais brasileiras*, 1th ed., v.1.; Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis, Laboratório de Produtos Florestais: Brasília, 1997.
3. Coradin, V. T. R.; Camargos, J. A. A; Marques, L. F.; Silva Jr., E. R.; Madeiras similares ao mogno (*Swietenia macrophylla* King.): chave ilustrada para identificação anatômica em campo, Serviço Florestal Brasileiro, Brasil, 2009.
4. Sandak, A.; Sandak, J.; Prądzń W.; Zborowska, M.; Negri, M.; *Folia For. Pol., Ser. B* **2009**, *40*, 31.
5. Shou, G.; Zhang, W.; Gu, Y.; Chao, D.; *J. Near Infrared Spectrosc.* **2014**, *22*, 423.
6. Pastore, T. C. M.; Braga, J. W. B.; Coradin, V. T. R.; Magalhães, W. L. E.; Okino, E. Y. A.; Camargo, J. A. A.; Muñiz, G. I. B.; Bressan, O. A.; Davrieux, F.; *Holzforschung* **2011**, *65*, 73.
7. Braga, J. W. B.; Pastore T. C. M.; Coradin, V. T. R.; Camargos, J. A. A.; da Silva, A. R.; *IAWA Journal* **2011**, *32*, 285.
8. Bergo, M. C. J.; Pastore, T. C. M.; Coradin, V. T. R.; Wiedenhoeft, A. C.; Braga, J. W. B.; *IAWA Journal* **2016**, *37*, 420.
9. da Silva, A. R.; Pastore, T. M. C.; Braga, J. W. B.; Davrieux, F.; Okino, E. Y. A.; Camargos, J. A. A.; Coradin, V. T. R.; do Prado, A. G. S.; *Holzforchung* **2013**, *67*, 1.

10. Kelley, S. S.; Rials, T. G.; Snell, R.; Groom, L. H.; Sluiter, A.; *Wood Sci. Technol.* **2004**, *38*, 257.
11. Tsuchikawa, S; *Appl. Spectrosc. Rev.* **2007**, *42*, 43.
12. Pasquini, C.; *J. Braz. Chem. Soc.* **2003**, *14*, 198.
13. United Nations Office on Drugs and Crime; Global Programme for Combating Wildlife and Forest Crime; *Best Practice Guide for Forensic Timber Identification*, UNODC, New York, 2016.
14. Stern, W. L.; *IAWA Bull.* **1988**, *9*, 209.
15. Brereton, R. G.; *Analyst* **2000**, *125*, 2125.
16. Barker, M; Rayens, W.; *J. Chemom.* **2003**, *17*, 166.
17. Brereton, R. G.; Lloyd, G. R.; *J. Chemom.* **2014**, *28*, 213.
18. Botelho, B. G.; Reis, N.; Oliveira, L. S.; Sena, M. M.; *Food Chem.* **2015**, *181*, 31.
19. Borin, A.; Poppi, R. J.; *J. Braz. Chem. Soc.* **2004**, *15*, 570.
20. da Silva, V. A. G.; Talhavini, M.; Zacca, J. J.; Maldaner, A. O.; Peixoto, I. C. F.; Braga, J. W. B.; *Microchem. J.* **2014**, *116*, 235.
21. da Silva, V. A. G.; Talhavini, M.; Zacca, J. J.; Trindade, B. R.; Braga, J. W. B.; *J. Braz. Chem. Soc.* **2014**, *25*, 1552.
22. Annual Book of ASTM Standards; *Standards Practices for Infrared Multivariate Quantitative Analysis, E1655-05*. ASTM International: West Conshohocken, 2012.
23. Ferreira, M. M. C.; *Quimiometria – Conceitos, Métodos e Aplicações*, 1ª ed., Editora Unicamp: Campinas, 2015.
24. Schwanninger, M.; Rodrigues, J. C.; Fackler K.; *J. Near Infrared Spectrosc.* **2011**, *19*, 287.
25. Santana, M. A. E.; Okino, E. Y. A.; *Holzforschung* **2007**, *61*,469.
26. Rutiaga-Quiñones, J. G.; Chemische und biologische Untersuchungen zum Verhalten dauerhafter Holzarten und ihrer Extrakte gegenüber holzabbauenden Pilzen, Buchverlag Gräfelfing:München, 2001.
27. Norgaard, L; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B.; *Appl. Spectrosc*. **2000**, *54*, 413.
28. Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C.; *J. Chemom*. **2009**, *23*, 32.