Talanta 117 (2013) 305–311

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Classification of Amazonian rosewood essential oil by Raman spectroscopy and PLS-DA with reliability estimation

Mariana R. Almeida, Carlos H.V. Fidelis, Lauro E.S. Barata, Ronei J. Poppi *

Institute of Chemistry, University of Campinas, POB 6154, 13084-971 Campinas, SP, Brazil

ARTICLE INFO

Article history:
Received 21 May 2013
Received in revised form
9 August 2013
Accepted 17 September 2013
Available online 25 September 2013

Keywords:
Rosewood oil
Raman spectroscopy
Chemometric
Uncertainty estimation
Reliability

ABSTRACT

The Amazon tree *Aniba rosaeodora* Ducke (rosewood) provides an essential oil valuable for the perfume industry, but after decades of predatory extraction it is at risk of extinction. The extraction of the essential oil from wood implies the cutting of the tree, and then the study of oil extracted from the leaves is important as a sustainable alternative. The goal of this study was to test the applicability of Raman spectroscopy and Partial Least Square Discriminant Analysis (PLS-DA) as means to classify the essential oil extracted from different parties (wood, leaves and branches) of the Brazilian tree *A. rosaeodora*. For the development of classification models, the Raman spectra were split into two sets: training and test. The value of the limit that separates the classes was calculated based on the distribution of samples of training. This value was calculated in a manner that the classes are divided with a lower probability of incorrect classification for future estimates. The best model presented sensitivity and specificity of 100%, predictive accuracy and efficiency of 100%. These results give an overall vision of the behavior of the model, but do not give information about individual samples; in this case, the confidence interval for each sample of classification was also calculated using the resampling bootstrap technique. The methodology developed have the potential to be an alternative for standard procedures used for oil analysis and it can be employed as screening method, since it is fast, non-destructive and robust.

1. Introduction

The Amazon tree *Aniba rosaeodora* Ducke, popularly known as rosewood, provides an essential oil valuable for the perfume industry. The rosewood oil is economically relevant by being a valuable source of monoterpene alcohol linalool (3,7-dimethyl-1,6-octadien-3-ol), as showed in Fig. 1. Since the beginning of its exploitation, around 1920, the extraction of the essential oil is performed with the cutting of trees and oil extraction from wood. After decades of predatory extraction and the intense illegal trade turned to exports, the species was included in the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) as a plant at risk of extinction [1].

In spite of the majority of studies on the rosewood essential oil have been based from the extraction of the wood, May and Barata [2] published a study in 2004 presenting alternatives for sustainable production of the rosewood essential oil. One of these alternatives is a production of essential oil using the leaves and branches, in a procedure without harming the tree. Other works of Barata and co-workers show the chemistry characterization of the essential oil extracted from leaves and branches [1,3–5]. In 2006, a study employing gas chromatography–mass spectrometry and enantioselective gas chromatography–olfactometry [3] showed that the rosewood leaf oil could replace the oil extracted from wood, as it presents similar chromatographic profile and olfactory quality to the oil extracted from wood. In another work, the characterization of the essential oil from the leaves of young plants has been made by comprehensive two-dimensional gas chromatography showing that young plants can produce essential oil with the quality required by the industry in a reasonable yield [1,4].

Gas chromatography is the main analytical technique to analyze the essential oil composition; however, the use of Raman spectroscopy has gained space in the analysis of essential oil, showing analytical advantages such as speed, cost and waste generation of analyzes. Daferera et al. [6] showed that the principal components of essential oil of some *Lamiaceae* can be recognized by FT-Raman. A review paper [7] shows the use of Raman spectroscopy to identify and quantify valuable plants substances. This review claimed that the main constituents can be identified through the spectral characteristics without the need for any physical separation. The main components of Eucalyptus oil were also analyzed by Raman spectroscopy [8], where it was possible to discriminate different essential oil of several Eucalyptus species. Raman spectroscopy was reported to be a unique tool for the chemical classification of unknown essential oil samples

* Corresponding author. Tel.: +55 19 35212134; fax: +55 19 35213126.
E-mail address: ronei@iqm.unicamp.br (R.J. Poppi).

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.09.025

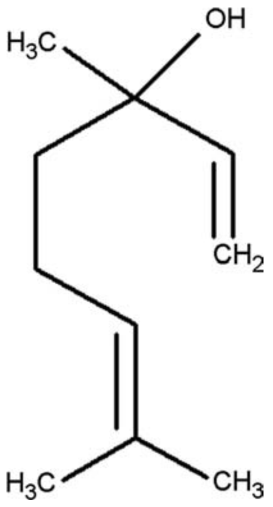

Fig. 1. Molecular structure of linalool.

with the purpose of fast quality control of essential oil from different species and aroma [9].

The majority of the studies previously mentioned employed Raman spectroscopy in combination with chemometrics tools. Various applications in the literature demonstrate the application of Raman spectroscopy combined with the supervised method PLS-DA (Partial Least Square for Discriminant Analysis) as possible methodology replacement of conventional qualitative analysis. The combined use of these techniques has shown a great potential in the biomedical field such as cancer diagnostics. Duraipandian et al. [10,11] applied Raman endoscopy together with PLS-DA for *in vivo* prospective diagnosis of gastric cancer and cervical precancer. In the area of food, it can also be found several applications, for instance, Guzman described the discrimination of olives before the oil processing stage employing Raman spectroscopy and supervised classification methods to detect if the olives have been collected directly from the tree or not [12]; Raman spectroscopy and multivariate analysis were also used to classify the adulterated milk powder samples as a rapid method for screening milk powder [13]. Samyn et al. [14] employed Fourier transform Raman, principal components analysis and partial least squares regression to classify different Brazilian vegetable oils, according to their degree of unsaturation based in the iodine value.

In spite of the application in various areas, most studies employing Raman spectroscopy and the supervised method PLS-DA only attribute object classifications; few papers evaluate the reliability estimation in these pattern recognition methods. Due to the importance of uncertainty estimation in analytical data, proposals have been published in the literature to estimate uncertainty in multivariate analysis, such as: linearization-based methods, resampling methods, U-deviation, among others [15].

In this work, we describe the use of Raman spectroscopy and supervised method PLS-DA as means to classify essential oil from wood, leaves and branches. The quality assessment of the results obtained was discussed by error of prediction. The evaluation of classification performance parameters was estimated by Confusion Tables, and parameters such as false positive and negative rates, sensitivity and specificity, accuracy, efficiency and Matthew's correlation coefficient. The degree of confidence of a classification was expressed by the resampling bootstrap technique. In this technique, it generates new data sets from the available one by an artificial perturbation [16]. From this new data set, the unknown distribution of a parameter could be estimated by mimicking the random mechanism through resampling of the data set.

2. Material and methods

2.1. Samples

The samples of essential oils originated from trees located in the States of Pará and Amazon, Brazil, where the characteristic climate is hot and humid. This rosewood material was steam distilled for 6 h in an industrial 1500-L iron reactor two-thirds filled with leaves. The oil was separated from water after reaching room temperature. The yield was, on average, 0.75%. The oil was transferred to glass flasks filled to the top and kept at a temperature of 4 °C.

The set of samples analyzed were composed of twenty samples of oil extracted from wood, sixty-five samples of oil from leaves, and twelve samples of oil from branches. Measurements were performed directly on the vial, without the need of sample preparation.

2.2. Raman measurements

The Raman spectra of essential oil were collected with a RamanStation 400F PerkinElmer dispersive spectrometer equipped with a cooled CCD detector using a Peltier system, a Echelle spectrograph and a 785 nm near-infrared diode laser of 250 mW (at source). A motorized positioner focused the laser beam to the sample, and a manual adjuster allowed sample adjustment for maximum optimal efficiency. The spectra were obtained in the Spectrum software version 6.3.5, with 3 s of exposure and 20 accumulations in the region of 3200–200 cm^{-1} with a spectral resolution of 4 cm^{-1}. Each spectrum was obtained in duplicate, an auto baseline correction and auto spectral subtraction of the glass spectrum was applied before the chemometric calculations.

2.3. Chemometric analysis

For all chemometric analyses, the Raman spectra were manipulated using MATLAB software Version 7.8 (R2009a) using routines developed in our laboratory and PLS Toolbox software, version 7.0.1, from Eigenvector Technologies.

Preprocessing is a crucial step in multivariate analysis. Thus, spectral preprocessing was carried out before any chemometric analysis. Several preprocessing methods were tried, such as the normalize method, derivative, smoothing, OSC (Orthogonal Signal Correction) filter and mean center. The objective of the preprocessing was to remove undesirable systematic variation in the data and to find classification models with the best performance. The choice of the better preprocessing was performed based on the error of the model. For more information about preprocessing methods can be encountered in Ref. [17].

Principal Component Analysis was used for exploratory analyses of the rosewood oil samples. PCA extracts the more important components of the data. These principal components describe the multivariate interactions between the variables and reveal tendencies of the samples. Similarities and differences between samples can be studied observing the scores of the principal components and the importance of variables for the model can be known by loadings plots. The number of principal components was choosing taking into account the explained variance. A component that explains a large amount of variance is important, while those explaining little variance consist of artefacts and noise.

Partial Least Square Discriminant Analysis (PLS-DA) is a supervised method, applied for classification purposes; the PLS-DA model is developed from the same PLS (Partial Least Squares) algorithms used for multivariate calibration [18]. In PLS, a **Y** matrix is necessary that is related to the interest property. In this case, a dummy matrix **Y** was created with 0 for leaf and branch oil and

1 for wood oil. For the development of the classification model, the data set was randomly split into two subsets for training and test. The matrices **X** and **Y** were preprocessed for training and test sets, and the number of latent variables for PLS-DA models was chosen by the criterion of lowest prediction error in leave-one-out cross-validation. The model quality was assessed by studying the root mean square error of calibration (RMSEC), of cross-validation (RMSECV) and of prediction (RMSEP). The classification model was validated with the test set.

The variable importance in projection (VIP) [19] was also analyzed. The VIP scores show the importance of the variables for the prediction ability of the model. When the VIP value is greater than one, the variable is considered as important for the model. The VIP score can be employed as a criterion for a variable selection.

2.4. *Reliability of the classification model*

The threshold value for the class separation is based on Bayes' Theorem [20]. The Bayesian threshold assumes that the predicted *y* values follow a distribution similar to what will be observed for future samples. Using these estimated distributions, a threshold is selected at the point where the two estimated distributions cross; this is the *y* value at which the number of false positives and false negatives should be minimized for future predictions [21]. The evaluation of classification performance parameters was performed by Confusion Table; parameters such as false positive and negative rates, sensitivity, specificity, accuracy, and efficiency were calculated and used to evaluate the performance of the model. It is important to note that some performance parameters have the same name, but differ from the definition used in analytical chemistry and metrology.

The false positive rate is the probability of a negative sample being classified as positive sample, for example, in the context of this work a sample of oil extracted from leaves and/or branches being classified as oil wood. In a similar way, the false negative rate is the probability of oil extracted from wood being classified as oil extracted from leaves and/or branches.

A similar way can be used to calculate true rates: sensitivity and specificity; the sensitivity is the ability of the model to correctly classify the wood oil samples. The specificity is the capacity of the model to correctly identify the samples of leaf and/or branch oil:

The accuracy is the proportion of correct classification, independent of the class. Efficiency and Matthew's correlation coefficient are other parameters, which can be used to provide a single measure of model performance. The efficiency combines all information carried by the sensitivity and specificity.

Matthew's correlation coefficient summarizes the quality of Confusion Table in a single numerical value which can be used even if the classes have different sizes. This expression returns a value between -1 and $+1$, where a value of $+1$ represents a perfect classification, 0 an erroneous classification and -1 an inverse classification.

However, the parameters mentioned give an overview of the behavior of the model, but do not estimate the confidence intervals for each sample. For this purpose, confidence interval estimations were obtained by using the bootstrap technique [22], according to Almeida et al. [23]. In this paper, we use residual bootstrap to calculate the uncertainties of prediction of the PLS-DA model.

The first step to estimate the uncertainties of the predictions is to calculate the residues (**F**) of the PLS-DA:

$$\mathbf{F} = \mathbf{Y} - \mathbf{Y}_{\mathrm{PLS}} \tag{1}$$

where **Y** is the reference value (0 or 1) and $\mathbf{Y}_{\mathrm{PLS}}$ is the value predicted by the model. The residues were corrected by the degrees of freedom, since the differences between the values predicted and observed are considerably smaller than those from the expected deviation values [16].

$$\mathbf{F}_{\mathrm{i}} = \frac{\mathbf{Y}_{\mathrm{i}} - \mathbf{Y}_{PLS\mathrm{i}}}{(1-(\mathrm{DF}/N))^{1/2}} \tag{2}$$

where DF is the number of degrees of freedom and *N* is the number of calibration samples.

The number of the degrees of freedom was calculated according to van der Voet [24], the pseudo-degrees of freedom (PDF), which take into account the difference between the MSEC (Mean Squared Error of Calibration) and MSECV (Mean Squared Error of Cross-Calibration). In this case, the higher this difference, the smaller the number of degrees of freedom, according to the following equations:

$$\mathrm{PDF} = N\left(1 - \sqrt{\frac{\mathrm{MSEC}}{\mathrm{MSECV}}}\right) \tag{3}$$

$$\mathrm{MSEC} = \frac{\sum (y - y_{\mathrm{PLS}})^2}{N} \tag{4}$$

$$\mathrm{MSECV} = \frac{\sum (y - y_{\mathrm{PLS-CV}})^2}{N} \tag{5}$$

where y_{PLS} is the value predicted by the PLS model and $y_{\mathrm{PLS\text{-}CV}}$ is the value predicted by cross-validation.

The residual of the PLS-DA model are bootstrap generated from random substitutions with replacement of the initial values. The bootstrap residues ($\mathbf{F}^*$) are added to the $\mathbf{Y}_{\mathrm{PLS}}$ values, generating a matrix $\mathbf{Y}^*$:

$$\mathbf{Y}^* = \mathbf{Y}_{\mathbf{PLS}} + \mathbf{F}^* \tag{6}$$

A new PLS model can be calculated from $\mathbf{Y}^*$, generating the regression coefficient bootstrap (β^*), which allows the calculation of the new values of $\hat{\mathbf{Y}}^*$, and then new residuals are obtained:

$$\hat{\mathbf{F}}^* = \mathbf{Y}_{PLS} - \hat{\mathbf{Y}}^* \tag{7}$$

The quantiles of the distribution $\hat{\mathbf{F}}^*$ are used to estimate the confidence interval for each sample. The percentile method was used, where the confidence intervals are asymmetric and specific, and the upper and lower limits are defined by

$$\hat{\mathbf{F}}^*_{B(\alpha/2)} \le y_{\mathrm{a}} \le \hat{\mathbf{F}}^*_{\mathrm{B}(1-(\alpha/2))} \tag{8}$$

where α is the degree of confidence and *B* is the number of bootstraps.

3. Results and discussion

3.1. *Raman spectra*

Monoterpenes ($C_{10}H_{16}$) are the main constituents of the rosewood oil; some sesquiterpenes ($C_{15}H_{24}$) also are encountered. Monoterpenes and sesquiterpenes are terpenoids, which are defined as substances composed of isoprene (2-methylbutadiene) units. The chemical constituents of essential oil from wood and leaves of rosewood are reported in literature [3–5]. Linalool is the major compound of the essential oil obtained from wood, leaves and branches of the Brazilian rosewood. Due to the high percentage of linalool (70–90%) contained in rosewood oil, the Raman spectrum presents the profile characteristic of this compound. Fig. 2a shows the Raman spectra of standard linalool, the oil extracted from wood, from leaves and from branches, where it is possible to observe a great similarity between samples. The

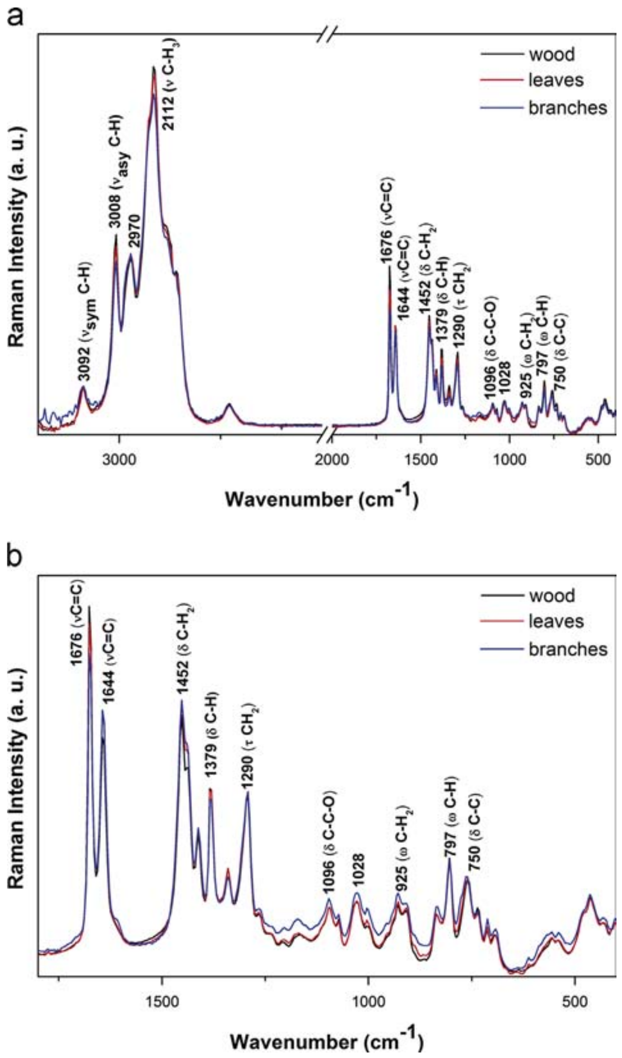

Fig. 2. (a) Raman spectra of oil extracted from wood, leaves and branches; (b) Raman spectra with enlargement of the region 1800–400 cm^{-1}.

tentative assignments of the main vibrational bands were based on the comparison with the literature [6–9]. Intense signals are observed in the region at 3000 cm^{-1} due to C–H stretching vibration; the presence of a weak band at 3092 cm^{-1} can be assigned to symmetrical =C–H stretching of the vinyl group; and the band observed at 3008 cm^{-1} corresponds to the asymmetric C–H stretching corresponding to C=C. The band at 2912 cm^{-1} was associated with the C–H vibrations of methylene groups. However, these bands cannot be used for the identification of rosewood oil; due to which most terpenes compounds have intense bands in this region.

The region between 1800 and 400 cm^{-1} is rich in structural information, thus, several works employed this region to distinguish essential oils [6–9]. The spectra show intense bands at 1676 cm^{-1} due to stretching vibration of disubstituted C=C bond, and the 1644 cm^{-1} band may be assigned to the stretching of the monosubstituted double-bond vibrations. These two bands are characteristics of acyclic monoterpenes, such as linalool, which presents the band at 1676 cm^{-1} more intense than the band at 1644 cm^{-1}. However, the presence and quantity present of other terpenes compounds $(C_5H_8)_n$ can cause changes in the intensities of these bands; this can lead to differentiation between oil types. The region below 1500 cm^{-1} shows bands which are due to single bond stretches and a wide variety of bending vibrations. The bands around 1379, 1293 and 797 cm^{-1} are characteristic of the linalool [9]. These bands are described as motions due to the angle deformations of the group with C–H and C–C bond. For the cyclic monoterpenes, the spectral features are observed in the region between 740–640 cm^{-1} due to ring deformation vibrations.

Comparing the Raman spectra of the essential oil extracted from wood, leaves and branches, it was not possible to visually observe differences in spectral profiles, as can be seen in Fig. 2b. Thus, the use of chemometric tools was necessary to build classification models for the essential oils extracted from different parts of the tree *A. Rosaeodora*.

3.2. *Multivariate analysis*

The Principal Components Analysis (PCA) is the first stage of chemometric processing and it is usually applied for exploratory data analysis. The PCA model was developed with 97 Raman spectra of essential oil from wood (20 samples), leaves (65 samples) and branches (12 samples). The better option for spectral preprocessing was the normalization method for vector of unit length [21]; after that, it was applied the second derivative, based on a 9-point window and a second-order polynomial for smoothing and finally the data was mean centered. The normalization allows an effective comparison across heterogeneous sets of samples, derivative filters remove unimportant baseline signal and identify overlapping bands and mean centering is an appropriate choice, once that variables with a great deal of variation are more important than those with small variation. These preprocessing caused gain of the robustness and better performance of classification analysis, with a best interpretability of the spectra.

Based on explained variance, four principal components that explain 93.7% of the total data variance were chosen. The projection of the samples onto the first principal component allows the observation of the distribution of the samples and the analysis of their grouping. However, with the projection of the principal components, it was not possible to observe the formation of clusters that can be used to distingue between wood oil and leaf or branch oil. Thereby, the inspection of the loadings to interpret the differences and similarities among the samples does not provide relevant information.

The great variability of samples, which contain variation from many sources and of several types, explains the need of four components to explain the variance of the data. The PCA does not provide satisfactory results in relation to the source of oil, and it does not show the presence of trends. PCA is an unsupervised method and it cannot be used to classify samples in one or more classes. However, the PCA can be used to identify outlier samples in the data set. In this case, three samples were identified as outliers and removed before the development of the classification model. These outlier samples had Q residual and T^2 values above the limits for 95% of confidence.

In particular, as the goal of our work was to develop a method for the classification of rosewood oil through the use of Raman spectra, a supervised pattern recognition approach should be employed. In this case, discriminant analysis using PLS is more appropriate when discrimination is the goal and dimension reduction is needed [25].

The spectra set were randomly divided into subsets of training (70% of the samples) and test (30% of the samples); it was used the training set to build the rule of classification and the test set to the validation model. The training set was composed of 13 samples of wood oil, 43 samples of the leaf oil and 8 samples of the branch oil. The dimensionality of the model was determined by leave-one-out cross-validation, based on the lowest value of the RMSECV. In this case, four latent variables were needed for the development of the classification model, representing 93% of the variance explained in

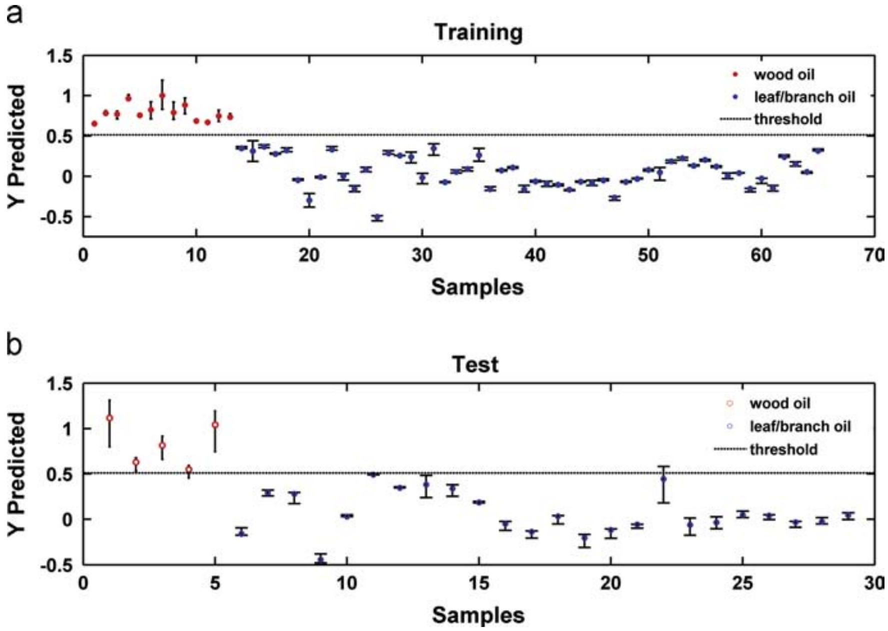

Fig. 3. Results of the training (a) and test (b) set of the PLS-DA model, showing wood oil (above dashed line) and leaf and branch oil (below dashed line) with prediction intervals for all samples analyzed by the bootstrap residual method.

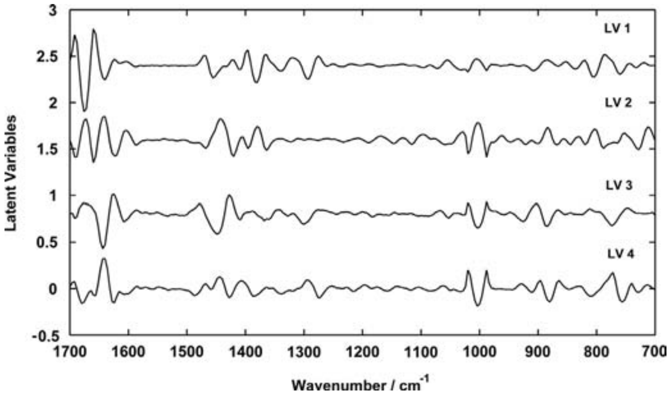

Fig. 4. Plot of the loadings of LV1, LV2, LV3 and LV4 *versus* wavenumber (variables) for the PLS-DA model.

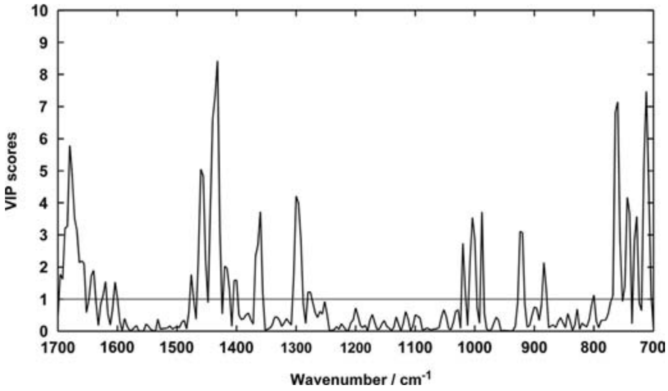

Fig. 5. Variable Importance in Projection (VIP) scores for PLS-DA model.

the **X** and 73% of the variance in the **Y** matrix, with RMSECV equal to 0.23 and mean error of 0.22 for the training set (RMSEC).

The results from the PLS-DA model for the training and test sets are shown in Fig. 3a and b, respectively. For the test set, the RMSEP was equal to 0.24. The parameters RMSEC and RMSEP show a good concordance, implying that the RMSEC value is a good estimate of the standard error of prediction observed in the test set.

The discriminant analysis was based on the profile of the Raman spectra of the rosewood essential oil. The loading plot, presented in Fig. 4, shows the Raman bands that contribute to the differentiation between classes. By analyzing the loadings of the latent variables, it is noted that the region around 1650 and 1450 cm^{-1} are important for the separation of samples. As previously discussed, other terpenes compounds $(C_5H_8)_n$ can cause changes in the intensities of these bands and this can lead to differentiation between oil types. Another method for judging the importance of the variables is the VIP scores, where interpretation is a useful and simple strategy tool for the evaluation of the importance of each variable. Fig. 5 shows the VIP scores obtained for the PLS-DA model when normalization, derivative and mean center were applied as pre-treatment. The values of VIP > 1 inform the spectral region that can be distinguished as important for optimal PLS-DA model performance. In this case, four regions can be selected for optimal performance: 1650, 1450, 1000 and 750 cm^{-1}.

The results above suggest that the difference between the sources of oil can be attributed to the variation of the quantities of terpenes compounds present in the samples. According to Zellner et al. [3], the oil extracted from leaves is characterized by a higher concentration of sesquiterpenes in relation to the oil from the wood, which contains a larger quantity of monoterpenes. The sesquiterpenes present Raman bands in the regions around 1650, 1450 and 750 cm^{-1} [7] and these regions showed the major contribution for a discrimination class in the loadings and VIP analysis.

Fig. 6 shows the graph of scores of the first three latent variables, where it is possible to observe the formation of two groups: the oil sustainably extracted from leaves and branches, and the oil extracted from wood, where it is necessary to cut the tree.

For the classification model, it is necessary to calculate a threshold value that separates the classes. Here, the cut-off calculated was estimated using Bayes' Theorem. The calculated threshold is shown as the horizontal line in Fig. 3, with a value of

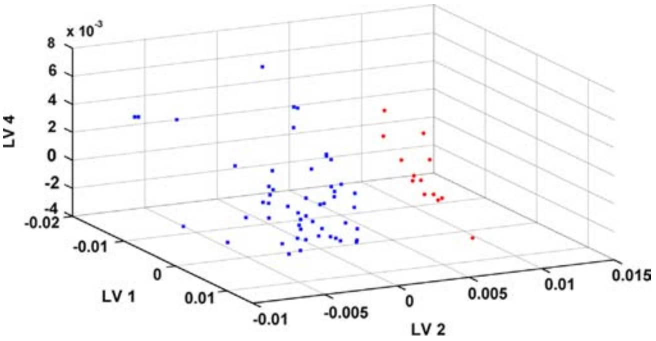

Fig. 6. Plot of the scores of LV1 × LV2 × LV4 of wood oil (●) and leaf and branch oils (■).

Table 1
Classification parameters obtained for PLS-DA.

	Training set	Test set
False positive rate (%)	0	0
False negative rate (%)	0	0
Sensitivity (%)	100	100
Specificity (%)	100	100
Accuracy (%)	100	100
Efficiency (%)	100	100
Matthew's correlation	1	1

0.51. Above this value, the oil is classified as wood oil and below as leaf and/or branch oil.

The performance of PLS-DA model was evaluated in terms of sensitivity, specificity, accuracy, efficiency, and Matthews's correlation coefficient, as discussed in Section 2.4. Table 1 presents the results obtained from the PLS-DA model for the training and test set samples. As shown in this table, the similarity in parameters obtained between the training and test sets indicate that overfitting did not occur, which assured the robustness and reliability of the PLS-DA model developed.

Some parameters can contribute towards false responses, for instance, the cut-off region of a class. The threshold value corresponds to the problem of committing type 1 and type 2 errors. Increasing threshold value reduces the probability of obtaining false positives results, but augments the false negatives results. Using the proposed methodology as screening test, a positive response could be confirmed by the reliable method, as for example, gas chromatographic–mass spectrometry, and the negative responses would be considered to be final, since this can add costs and time to the analysis. In this way, it is preferable that the probability of false negative results should be lower. In this work, it is important to avoid false negatives, as that would classify wood oil as oil extracted from leaves and branches.

We can assess the type of classification model using the statistics of false responses, such as false positive and false negative rates. For the training and test sets, there were no false answers. Another parameter that is possible to evaluate is the statistics of true answers, sensitivity and specificity in this case. As the statistics of true–false answers complement each other, the sensitivity and specificity of the test set were 100%.

Accuracy and efficiency were also evaluated. The accuracy shows the proportion of samples provided correctly, which on the test set was 100%. The efficiency combines all of the information carried by sensitivity and specificity; when a method is very sensitive to positive, it generates many false positives, and *vice versa*. The efficiency of the test set was 100%. Matthew's correlation coefficient also provides a single measure of classification performance. The value of Mathew's coefficient for the test set was 1, a value that can classify the methodology as having a perfect performance.

The results extracted from the Confusion Table also give an overall vision of the behavior of the model, but do not give information about individual samples; the error for each sample is not computed. Parameters such as RMSEC and RMSECV also evaluate the general model performance, but do not truly reflect prediction reliability. Several proposals are described in the literature to deal with the problem, as discussed in the work of Zhang and Garcia-Munoz [15]. The calculation of the reliability in the classification model is an important issue, since, besides assigning the object to a class, the knowledge of the uncertainty of this assignment can result in better ability in model predictability. For this proposal, the residual bootstrap technique was employed according to the description in Section 2.4.

Due to its stochastic nature, the bootstrap method yields (slightly) different results when applied repeatedly to the same problem (unless the same random numbers are reused). One of the ways to compensate for this variability is related to the adjustment of the number of trials (the number of evaluations of the measurement model), which should be large enough to ensure the reliability of the bootstrap results. The number of bootstrap was optimized to obtain the lowest standard deviations and reproducible results; with 10 repeated calculations, it was observed that after 1000 bootstrap samples the values of standard deviation varied very little. In this way, the calculation of bootstrap was performed with 1000 resampling. Each simulation was carried out using 1000 random trials for each calibration sample. All of the calculation samples were considered as a new observation and 95% bootstrap confidence intervals were computed. The degrees of freedom consumed by the PLS-DA model were 7, calculated by pseudo-degree of freedom (PDF), as in Eq. (3). The error bars shown in Fig. 3 are uncertain for the classification of each sample and allow better confidence in the test set. In the mean, the lower and upper limits ranged from ± 0.2. For the test set, some samples showed values near the threshold line and the uncertainty in the classification of two samples is noted, where the confidence intervals exceeded the threshold line. These results suggest that the spectra of these samples (class 0) have similarities with class 1 spectra and *vice versa*, leading to uncertainties in the classification. These results are not observed when evaluating only Confusion Table parameters and mean error of prediction, showing the importance of uncertainty calculation for each sample.

4. Conclusions

The proposed methodology, based on the characterization of rosewood oil extracted from different parts of the Amazon tree *A. rosaeodora*, employing Raman spectroscopy and PLS-DA, allows to distinguish between wood and leaf/branch oils.

The evaluation of the model performance parameters was performed by Confusion Tables, and parameters such as false positive and negative rates, sensitivity and specificity, accuracy, efficiency and Matthew's correlation coefficient were assessed for the training and test sets. This work takes into account the uncertainty in the classification of each sample by chemometric methods using residual bootstrap.

In this work, it was demonstrated that Raman spectroscopy can be used not only for chemotype classification of various essential oils and identification of their major oil constituents, but also for identification of essential oil of different tree parts. The results show that the Raman spectroscopy in conjunction with PLS-DA has the potential to be an alternative for standard procedures used for oil analysis, such as gas chromatography. Although stored under recommended conditions, samples are naturally susceptible to oxidation processes, once they were stored for about 3 years. For this reason, the proposed method showed robustness, since it was able to classify essential oil from different parts of *A. rosaeodora* Ducke even under such conditions.

The methodology can be used as a screening method to distinguish the oils extracted from different parts of the tree *A. rosaeodora*. The procedure is fast, non-destructive, robust and presents the possibility of *in situ* analysis using a portable instrument. The differentiation between the sources is important to ensure the sustainable origin of the oil (leaves and branches) and that it does not come from material that is currently not allowed for extraction (wood).

Acknowledgments

The authors thank CNPq, CAPES and INCTBio for the financial support.

References

[1] C.H.V. Fidelis, F. Augusto, P.T.B. Sampaio, P.M. Krainovic, L.E.S. Barata, J. Essent. Oil Res. 24 (2012) 245–251.
[2] P.H. May, L.E.S. Barata, Econ. Bot. 58 (2004) 257–265.
[3] B.A. Zellner, M. Lo Presti, L.E.S. Barata, P. Dugo, G. Dugo, L. Mondello, Anal. Chem. 78 (2006) 883–890.
[4] C.H.V. Fidelis, P.T.B. Sampaio, P.M. Krainovic, F. Augusto, L.E.S. Barata, Microchem. J. 109 (2013) 73–77.
[5] R.C.Z. Souza, M.M. Eiras, E.C. Cabral, L.E.S. Barata, M.N. Eberlin, R.R. Catharino, Anal. Lett. 44 (2011) 2417–2422.
[6] D.J. Daferera, P.A. Tarantilis, M.G. Polissiou, J. Agric. Food Chem. 50 (2002) 5503–5507.
[7] H. Schulz, M. Baranska, Vib. Spectrosc. 43 (2007) 13–25.
[8] M. Baranska, H. Schulz, S. Reitzenstein, U. Uhlemann, M.A. Strehle, H. Kruger, R. Quilitzsch, W. Foley, J. Popp, Biopolymers 78 (2005) 237–248.
[9] K.R. Strehle, P. Rosch, D. Berg, H. Schulz, J. Popp, J. Agric. Food Chem. 54 (2006) 7020–7026.
[10] S. Duraipandian, M.S. Bergholt, W. Zheng, K.Y. Ho, M. Teh, K.G. Yeoh, J.B.Y. So, A. Shabbir, Z. Huang, J. Biomed. Opt. (2012)081418-1–081418-8.
[11] S. Duraipandian, W. Zheng, J. Ng, J.J. Low, A. Ilancheran, Z. Huang, Anal. Chem. 84 (2012) 5913–5919.
[12] E. Guzmán, V. Baeten, J.A.F. Pierna, J.A. García-Mesa, Talanta 93 (2012) 94–98.
[13] M.R. Almeida, K.S. Oliveira, R. Stephani, L.F.C. Oliveira, J. Raman Spectrosc. 42 (2011) 1548–1552.
[14] P. Samyn, D.V. Nieuwkerke, G. Shoukens, L. Vonck, D. Stanssens, H.V. D. Abbeele, Appl. Spectrosc. 66 (2012) 552–565.
[15] L. Zhang, S. Garcia-Munoz, Chemon. Intell. Lab. Syst. 97 (2009) 152–158.
[16] N.M. Faber, Chemon. Intell. Lab. Syst. 64 (2002) 169–179.
[17] P. Lasch, Chemon. Intell. Lab. Syst. 117 (2012) 100–114.
[18] S. Wold, M. Sjöström, L. Eriksson, Chemon. Intell. Lab. Syst. 58 (2001) 109–130.
[19] I.G. Chong, C.H. Jun, Chemom. Intell. Lab. Syst. 78 (2005) 103–112.
[20] R.O. Duda, P.E. Hart, D.G. Store, Pattern Classification, 2nd ed., John Wiley & Sons, Inc., New York, 2001.
[21] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch, Chemometris Tutorial for PLS_Toobox and Solo Eigenvector Research, Inc., 3905 West Eaglerock Drive, Wenatchee, WA 98801, USA, 2006.
[22] A.M. Zoubir, B. Boashash, IEEE Signal Proc. Mag. 15 (1998) 55–76.
[23] M.R. Almeida, D.N. Correa, W.F.C. Rocha, F.J.O. Scafi, R.J. Poppi, Microchem. J. 109 (2013) 170–177.
[24] H. van der Voet, J. Chemom. 13 (1999) 195–208.
[25] M. Barker, W. Rayens, J. Chemom. 17 (2003) 166–173.