Fast Classification of Different Oils and Routes Used in Biodiesel Production Using Mid Infrared Spectroscopy and PLS2-DA

Sarmento J. Mazivila,*a,b Hery Mitsutake,*a Felipe B. de Santana,*a Lucas C. Gontijo,a,c Douglas Q. Santosd and Waldomiro Borges Netoa

*aInstitute of Chemistry, Federal University of Uberlândia, Campus Santa Mônica, 38408-100 Uberlândia-MG, Brazil
bJosina Machel Secondary School of Belane, Vilankulo-Inhambane, Mozambique
cGoiano Federal Institute of Education, Science and Technology, Rodovia Geraldo Silva Nascimento, km 2.5, 75790-000 Urutai-GO, Brazil
dTechnical School of Health, Federal University of Uberlândia, Campus Umuarama, 38400-902 Uberlândia-MG, Brazil

This work aimed at employing partial least square discriminant analysis (PLS2-DA), allied to mid-infrared (MIR) spectroscopy as an analytical method for simultaneous classification of biodiesels from different oils (soybean and used frying oil) and routes (methylic and ethylic). The evaluation of the model was verified through values of sensitivity and specificity for each parameter, in the interest class. PLS2-DA model showed 100% correct classification in the discrimination of types of biodiesels. Therefore, the proposed methodology is fast, because it allows simultaneous classification of different types of biodiesels. Consequently, it can be used in quality control of this type of biofuel.

Keywords: biodiesel, infrared, PLS2-DA classification

Introduction

Biodiesel is a fuel produced by a transesterification reaction in which oils from different oil seeds, animal fats and used frying oil react with alcohol, usually methanol or ethanol, in the presence of a catalyst.1-3

Among the feedstocks used in the production of biodiesel in Brazil, soy is the most commonly used (ca. 80.35%).4 However, it has the disadvantage of being used as food. Thus, just relying on cultivation as a main supplier of feedstocks for biodiesel production can limit the production of fuel for some regions according to climatic conditions and soil quality.5 Consequently, the use of other feedstocks (e.g., used frying oil) is necessary. The use of frying oil has promising potential for biodiesel production,6,7 because it is a disposal product of frying and avoids competition with vegetable oil employed in human consumption,8 once the use of used frying oils reduces the problem of contamination of waste water, that is, for the reuse of these fatty residues may contribute to decrease the burden on the government in remove these wastes in sewages.9 Furthermore, according to the European commission approved directive 2012/0288/EC,10 from 2020, biofuels produced from oils employed in human food should not be subsidized for the purpose of encouraging the use of feedstocks that do not take additional demand for land and do not enter the human food supply such as used frying oil.

The various feedstocks used to produce biodiesel, have different characteristics due to the chemical composition of fatty acids. Consequently, the final product has different physicochemical properties. Therefore, the quality control of biodiesel is very important,7 so it is necessary to develop methodologies that can identify the feedstock used in the biodiesel production.5,10

Wang et al.11 identified biodiesels with better properties due to their similarity with a reference sample using gas chromatographic fingerprint data and principal component analysis (PCA). However, chromatography is a sensitive
method but is slow when compared to vibrational spectroscopy techniques.\textsuperscript{13} Infrared spectroscopy methods have advantages, such as the use of relatively low cost equipment that allows field analysis; minimal or no sample treatment; expeditious analysis; causing no sample destruction and demanding no reagents.\textsuperscript{13} Also, these methods allow in situ analysis using portable equipment. In this sense, the use of infrared spectroscopy combined with multivariate analysis has been applied in several areas such as health and food. Yi \textit{et al.}\textsuperscript{14} employed near infrared (NIR) spectroscopy combined with principal component analysis (PCA) and hierarchical cluster analysis (HCA) to discriminate gastric cancer in relation to the type of tissue where 90 samples were classified as cancerous or normal tissues. The study showed the separation of different groups according to the type of tissue with an accuracy of 81.1\%. Borras \textit{et al.}\textsuperscript{15} developed a methodology to classify 160 samples of almonds in relation to their bitterness by NIR and partial least square discriminant analysis (PLS-DA), with an accuracy of 95.7\%. In this case, the PLS-DA model could not only classify the samples according to their similarity but was also able to recognize new samples not used in the development of the model.\textsuperscript{16}

Recent studies indicate that the use of mid-infrared (MIR) spectroscopy and NIR combined with supervised chemometric methods have potential as promising methodologies for classifying fuels. Veras \textit{et al.}\textsuperscript{17} used soft independent modeling of class analogy (SIMCA) to classify 108 samples of biodiesel using NIR spectroscopic data, which was able to classify, with an accuracy of 100\%, biodiesels in relation to the type of oil used: cotton, sunflower, soybean and rapeseed oil. Balabin \textit{et al.}\textsuperscript{18} employed NIR spectroscopic data and support vector machines discriminant analysis (SVM-DA) to classify motor oil according to the feedstock used: synthetic, semi-synthetic and mineral oil. The best result presented an error of 6\% for classification. Silva \textit{et al.}\textsuperscript{13} classified the common and additivated gasoline with a correct classification of 100\% of the test samples using MIR spectroscopic data combined with linear discriminant analysis and variable selection.

Notably, the simultaneous classification using MIR spectroscopy and PLS2-DA was not previously used to classify more than two types of biodiesels. The simultaneous classification using PLS2-DA is advantageous, for analysis of several properties of interest, in single calibration\textsuperscript{19} that can be used in industry in order to gain time. Thus, this work aims to classify the different types of biodiesels regarding their feedstock and the alcohol used in the transesterification process using MIR spectroscopy coupled with the multivariate technique PLS2-DA.

**Experimental**

**Biodiesel production and characterization**

The soybean refined oil used in the synthesis of biodiesel was acquired in the local market, and soybean used frying oil (UFO) were collected from local restaurants. To obtain each batch of biodiesels, 200.0 g of oil and a mixture containing potassium ethoxide or methoxide (60.0 g of ethanol or methanol (PA) and 2.0 g of KOH) were stirred at 400 rpm. After 1 h, the resultant solution was concentrated in a rotary evaporator to remove the excess of alcohol. Then the solution was allowed to stand for 24 h to separate production residues and coproducts. After the two-layer separation, the obtained esters were purified by washing with distilled water at 90 °C and drying using vacuum distillation. In a rotary evaporator with a bath temperature of 90 °C, the flash point was measured hourly until a constant value was reached.

For the production of biodiesel from UFO, esterification was performed before the transesterification process for the purpose of reducing the level of acidity and increasing the efficiency in the conversion of ethyl and methyl esters.

The methods used to characterize the biodiesel were as follows: the moisture content was analyzed using a Karl Fischer colorimetric titrator (model 831 KF) according to standard ASTM D-6304. Acidity was determined according to the official procedures recommended by ABNT NBR 14448. Free and total glycerin fractions were determined according to the methodology described by ABNT NBR 15771 and ABNT NBR 15344, respectively. The range of conversion in biodiesel were determined by ABNT NBR 15764.

**Sample preparation**

In this work, four models of classification were built: four using PLS2-DA. The types of biodiesels evaluated were: used frying oil ethyl esters (UFEE), soybean oil methyl esters (SME), used frying oil methyl esters (UFME) and soybean oil ethyl esters (SEE). All of the biodiesels used were synthesized in the Laboratory of Biofuels Institute of Chemistry of the Federal University of Uberlândia, Minas Gerais, Brazil. In constructing the model, 22 batches of different samples of each type of biodiesel (UFEE, SME, UFME and SEE) were used, making a total of 88 samples. The spectral data for each type of biodiesel were randomly split in 2/3 for the sample training (n = 15) and 1/3 for the sample test (n = 7).\textsuperscript{20}
Acquisition of spectral data

The MIR spectra were acquired using a PerkinElmer Spectrum Two spectrometer equipped with an attenuated total reflectance (ATR) sample holder and ZnSe crystal. The spectra were recorded in the range of 4000-600 cm\(^{-1}\) with 4 cm\(^{-1}\) resolution and were acquired using 16 scans for each of the quintuplicates. The average spectra of replicate (n = 5) for construction of the multivariate models were obtained.

Due to fluctuation of the spectral baseline these were submitted to correction by a baseline algorithm. For this, the spectral ranges 2500-1850 cm\(^{-1}\) and 4000-3150 cm\(^{-1}\) were selected. In order to minimize the unwanted systematic variation, in this study, before any chemometric analysis was performed, pre-processing of the data was carried out using a mean-centered approach for the variables in the X block and Y block.

Chemometrics analysis

To execute the multivariate procedures, MATLAB software version 7.5 (Mathworks Inc.) and PLS_Toolbox, version 7.5 (Eigenvector Research) were used.

The PLS2-DA models were developed based on the PLS algorithms. PLS2-DA is a version of the PLS-DA method, where the variables in the X block (spectral data) were related to classes contained in the Y matrix, in which each column vector represent the class of interest. The integer values of the class were arranged in each column of Y matrix, where 1 was used for the interest class and 0 for the class without interest.

The threshold value was predicted between 0 and 1 based on Bayes theorem, in order to minimize the errors in the prediction of the class through an appropriate number of latent variables (LV). The number of latent variables chosen for the PLS-DA models followed the criterion of the lowest prediction error in leave-one-out cross-validation and evaluation of the explained variance in the X and Y blocks. The outlier detection parameter was based on the identification of samples with high leverage values and Q residuals at 95% confidence.

Performance analysis

The criteria used to evaluate the quality of the PLS2-DA model were sensitivity (Sens) and specificity (Spec) in the test sample set. Sensitivity is number of samples predicted to be in the class divided by number actually in the class and specificity is number of samples predicted not to be in the class divided by actual number not in the class. These parameters were related to the concepts of statistical inference as false positive (FP) and false negative (FN) errors. Sens and Spec were calculated according to equations 1 and 2, respectively.

\[
\text{Sens} = \frac{TP}{TP + FN} \quad (1)
\]

\[
\text{Spec} = \frac{TN}{TN + FP} \quad (2)
\]

where TP and TN denote the numbers of true positives and true negatives, respectively. For example, in the classification of UFEE from the others type of biodiesels, UFEE was considered as “positive” and others types of biodiesels was considered as “negative”.

Results and Discussion

Table 1 presents the physicochemical properties of the UFEE, SME, UFME and SEE. The parameters are found to be within the established by National Petroleum Agency for Natural Gas and Biofuels of Brazil (ANP).

The MIR spectra of the samples of biodiesel are shown in Figure 1. According to Almeida et al., the 3000-2850 cm\(^{-1}\) is characteristic of the symmetric and asymmetric CH stretching vibrations of the methylene and terminal methyl groups. Thus, the prominent band at 2920 cm\(^{-1}\) is attributed to symmetric vibration of CH\(_3\) bond, near 2855 and 3010 cm\(^{-1}\) are attributed to methylene symmetric and asymmetric stretching vibrations, respectively; at 1740 cm\(^{-1}\) is observed an intense band assigned to the stretching of the C=O bond of an ester.

The region of 1500-900 cm\(^{-1}\) is known as the fingerprint region of complex spectra that include many coupled
vibrational bands, that is, particular bands in this region can hardly be attributed to a single chemical bond or group. In this perspective, peaks around 1460 cm\(^{-1}\) are due to bending vibration of the CH\(_2\) and CH\(_3\); near 1435 cm\(^{-1}\) is assigned to rocking vibration of CH bonds of cis-disubstituted olefins; near 1245 cm\(^{-1}\) may be attributed to antisymmetric axial stretching vibrations of CC(=O)–O bonds of the ester, while peaks near of 1195 cm\(^{-1}\) may be assigned to asymmetric axial stretching vibrations of O–C–C bonds.

In the region of low frequency from 900-700 cm\(^{-1}\), the significant vibrations were attributed to out of the plane deformation of the groups C=CH-\(\tau\) (olefinic CH) and \(\omega\) (olefinic CH) to the unsaturated derivatives. Therefore, as shown in Figure 1, it was observed that there is great similarity between the profiles of the different biodiesels due to the high degree of similarity between the spectral profiles of the biodiesels, and that the use of chemometric techniques is needed to discriminate between these biodiesels.

To verify the presence of outliers in the PLS2-DA model, the leverage versus graphic Q residual was used (Figure 2). We noted that none of the samples were considered an outlier.

During the analysis, it was found that the PLS1-DA models for each class showed 100\% correct classification, in other words, the sensitivity and specificity equal to 1. The PLS2-DA model developed also obtained values of sensitivity and specificity equal to 1, which shows 100\% correct classification. Nevertheless, when the dependent variables are strongly correlated, PLS2 provides models with predictive abilities similar to that of PLS1. Thus, the PLS2-DA model was used, once it is able to classify or predict multiple classes in the single calibration.

The 2461 variables from the 88 samples in the PLS2-DA model to classify each type of biodiesel (UFEE, SME, UFME and SEE) were reduced to three latent variables with 99.72\% of the variance explained. The first latent variable (LV1) is the main variable responsible for the separation

![Figure 1. The MIR spectra of the samples of biodiesel: (a) UFEE, (b) SME, (c) UFME and (d) SEE.](image-url)
routes (methyllic and ethyllic), explaining 93.71% of the original variance (Figure 3).

Figure 3 demonstrates a separation for the discriminated classes, showing that the MIR spectra present chemical information that makes it possible for PLS2-DA to discriminate between the four types of biodiesel. This result shows that the PLS2-DA model is able to differentiate samples of biodiesel obtained from soybean oil refined, as the frying oil used, from different routes (methyllic and ethyllic). However, if a sample of biodiesel of the same route was prepared from a mixture of soybean oil and used frying oil, the PLS2-DA model predicts in the class to which the sample presents greater similarity according to content of each of the types of oils in the mixture.

The loadings (Figure 4) show the wavenumbers responsible for this separation. Analyzing the graph of the loadings, the separation of the ethyllic route (UFEE and SEE) in relation to the methyllic route (SME and UFME) by LV1 was mainly determined by spectral vibration at 1740 and 1195 cm$^{-1}$, these vibration bonds can be assigned to C=O and O–C–C, respectively. The variables that have the greatest contribution to LV2 correspond to regions of low frequency from 900-700 cm$^{-1}$, which correspond to deformation out of the plane of the groups C=CH-τ (olefinic CH) and ω (olefinic CH) to the unsaturated derivatives. Finally, LV3 loadings indicate the contribution of the band at 2920 cm$^{-1}$ attributed to symmetric vibration of CH$_3$.

The class assignment of result is expressed in a value from 0 to 1. Thus, these values are normally distributed around 0, when the prevision is not the class of interest and near 1, when the prevision is the class of interest. For this reason is possible to obtain prevision value below 0 (negative value) and also values above 1. Therefore, it is necessary to calculate a threshold value to assume that a sample belongs or not to a determinate class of biodiesel. Its calculus is done by Bayes theorem, minimizing the number of false positives and false negatives for future prediction of model.

Table 2 shows the results of the classification parameters obtained for PLS2-DA model, where all types of biodiesel were separated between the estimated class value,
according the values of sensitivity and specificity, equal to 1 (maximum classification).

Table 2. Classification parameters obtained for PLS2-DA model

<table>
<thead>
<tr>
<th>Class</th>
<th>No. of LV</th>
<th>Threshold value</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>UFEE</td>
<td>3</td>
<td>0.57</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>SME</td>
<td>3</td>
<td>0.60</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>UFME</td>
<td>3</td>
<td>0.62</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>SEE</td>
<td>3</td>
<td>0.61</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The threshold value was used to discriminate between classes and was calculated based on the distribution of samples in the training prediction obtained using the PLS model. If the samples of the class of interest are above the threshold value, these are classified as belonging to the class of interest, and if below, they are classified as not belonging to the class.

The PLS2-DA showed excellent levels of sensitivity and specificity, which correctly classified 100% of the samples in the test set. According to Table 3, we can observe that in the PLS2-DA models, we have developed the samples in the test set could be classified as belonging to their respective classes.

Table 3. Result of prediction sample set of PLS2-DA model

<table>
<thead>
<tr>
<th>Class</th>
<th>N right/N total</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>UFEE</td>
<td>7/7</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SME</td>
<td>7/7</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>UFME</td>
<td>7/7</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SEE</td>
<td>7/7</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Conclusions

The use of mid-infrared spectroscopy in combination with multivariate technique using PLS2-DA classified with 100% efficiency samples of biodiesel in relation to the type of oil and the alcohol used in their production. Thus, this methodology is a viable alternative for the quality control of biodiesel, which offers fast analysis, non-destructive technique and potential in situ analysis using portable equipment to classify more than two types of biodiesels simultaneously.

Acknowledgements

The authors thank the Covenant CNPq/MCT-Mz, process No. 190802/2012-0, for financial support.

References


