In this work, PLS (Partial Least-Squares) models of calibration were compared to determine the amount of biodiesel and vegetable oil in biodiesel/vegetable oil/diesel blends using infrared spectra. The spectral data were collected in the 6550 - 4000 cm⁻¹ range with 4 cm⁻¹ resolution by Fourier Transform Infrared Spectroscopy using one Attenuated Total Reflectance accessory (FTIR/ATR). Initially, the spectral signals from the samples containing biodiesel and vegetable oil from between 0 and 8% were compacted using Discrete Wavelet Transform (DWT). A total of 80 binary and ternary blends containing biodiesel, vegetable oil and diesel were formulated, and 48 of these were applied to compose the calibration set and 32 the prediction set. The best models with the compacted or not compacted signals are studied, using the Interval Partial Least-Squares (iPLS) method, also identifying which spectral regions present better correlation. The results applying the compacted data presented similar errors in the determination of biodiesel and vegetable oil in the blends studied, indicating that the adulteration by vegetable oil in the mixtures B2 (diesel with addition of 2% of biodiesel) can be verified as long as the oil is also modeled. The combination of iPLS regression with the data obtained by FTIR/ATR has been promising to develop simpler, faster and non-destructive methodologies for the biodiesel determination and vegetable oil adulterations in blends commercialized since January 01 2008, when the addition of biodiesel to diesel was made mandatory in all national territory.

Palavras-chaves: biodiesel, diesel blends, multivariate regression, wavelet, adulteration, Infrared Spectroscopy
1. Introduction

Most part of all world energetic consumption comes from mineral sources and is not renewable, resulting in the future use up of these sources (FERRARI et al., 2005). And besides, environmental damage caused by the growing emission of toxic gases into atmosphere and the high price paid for oil are problems to be faced. The world society gets mobilized to search for new alternatives of energetic sources to reduce this problem (MUNIYAPPA et al., 1996). In this scenery one of the possibilities which has gotten world support is the use of clean and renewable fuels in automotive vehicles.

In Brazil great importance has been given to biodiesel, product derived from biomass, due to the great diversity and abundance of raw-material sources such as African palm oil, castor oil, pinewoods oil, canola oil, sunflower oil, palm oil, soybean oil and even animal grease (GERIS et al., 2007; DIAS, 2007; MARQUES et al., 2008). But great part of these oil plants, mainly the ones which are not used for human consumption, still depend upon research so that the ideal way of cultivating them in great scale is found (DIAS, 2007).

Another important factor is to assess the biodiesel quality. Among the most applied for this checking, we can see High Performance Liquid Chromatography (HPLC), Gas Chromatography (GC), Gel Permeation Chromatography (GPC) and Thin Layer Chromatography (TLC) (KNOTHE, 2006). However, some of these methodologies not only are destructive but also laborious, consume chemical reagents and destroy the analyzed samples. On the other hand, Infrared Spectroscopy applying Attenuated Infrared Reflection (ATR) is a methodology which is getting consolidated, mainly because it is a fast and non-destructive technique, not generating toxic chemical byproducts (COSTA FILHO e POPPI, 2002).

To evaluate the infrared signals obtained, Partial Least-Squares (PLS) Regression is commonly applied, being this the most common multivariate calibration technique used in the construction of models using spectra data (GELADI & KOWALSKI, 1986).

Interested in developing a clean, fast and non-destructive analytical methodology able to quantify the proportion of biodiesel in biodiesel/diesel blends (B2 - diesel with addition of 2% of biodiesel), as well as possible adulterations by raw vegetable oils in the blends used as fuel, this work proposes the use of FTIR/ATR signals. The results of the use of wavelets in compacting spectral data obtained by FTIR/ATR to build multivariate models by PLS are also presented.

2. Discrete Wavelet Transform

The signals obtained by FTIR/ATR usually present a considerable number of variables, around 2000 per spectrum, which cause a high computational consumption to build the diverse models of multivariate regression to be compared. This way, applying compacting algorithms to the generated data matrix leads to an expressive reduction of computational cost associated with the development of clean techniques through spectroscopic data.

One of the adequate techniques for the transformation of this data matrix is the use of Discrete Wavelet Transform (DWT) which treats the signal with all its characteristics independently of
the time or frequency. Acting directly upon the analytical signal frequency, this contributes to differentiate between signal and noise and the conservation of narrow peaks with the respective information (DANTAS FILHO, 2007).

According to Daubechies (1992), the mother wavelet function, from a mathematical point of view, generates the daughter wavelet functions which will decompose hierarchically according to the scheme presented in Figure 1. For each decomposition, two sub-signals of \( n/2 \) length are generated, where \( n \) represents the whole number. The first sub-signal corresponds to the scaling signals (scaling, \( A_1 \)), while the second to the wavelet signals, \( D_1 \).

To generate the daughter wavelet functions, besides choosing the kind of wavelets it is necessary to use a bank of filters composed of special coefficients (DAUBECHIES, 1992). Every DWT decomposes a discrete signal into two sub-signals with half its size (DAUBECHIES, 1992; MALLAT, 1998).

In this process low-pass digital filters are used, which produce the coefficients of signal approximation (\( A_1 \)), and high-pass filters which produce the coefficients of signal detailing (\( D_1 \)). The signal noise will be more concentrated on this second half, according to Dantas Filho (2007), and depending on the application this part may be rejected or not (JENSEN e COUR-HARBO, 2001). Processing the approximation coefficients (\( A_1 \)) again as input data with the same filter bank, another set of approximation and detailing coefficients is generated, with a lower resolution (DAUBECHIES, 1992; MALLAT, 1998; JENSEN e COUR-HARBO, 2001). But it is desirable that the signal which is intended to be compacted be in dimension \( 2^n \), with \( n \) belonging to natural numbers or being even, since this signal is always reduced to half in each level of compactness (DAUBECHIES, 1992).

A very common family of wavelets used in this article is DWT Daubechies with four coefficients (Daub4). The scaling and wavelet functions and their respective sub-signals are generated by these (DAUBECHIES, 1992):
To illustrate this process, in Figure 2a a signal function is presented, and in Figure 2b, the signal function in a first level transform (DAUBECHIES, 1992). With the adequate junction of scaling and wavelet sub-signals from the first transform it’s possible to recover the original signal, thus implicating in the preservation of this function information (DANTAS FILHO, 2007).

From the algorithmical point of view, DWT harmonizes the decomposition and smoothing techniques (CHAU et al., 2004). The selection of the adequate number of decomposition and the kind of wavelets will depend upon the signal and experience involved.

\[
\begin{align*}
    h_0 &= \frac{1+\sqrt{3}}{4\sqrt{2}} \\
    h_1 &= \frac{3+\sqrt{3}}{4\sqrt{2}} \\
    h_2 &= \frac{3-\sqrt{3}}{4\sqrt{2}} \\
    h_3 &= \frac{1-\sqrt{3}}{4\sqrt{2}} \\
    e
\end{align*}
\]

\[
\begin{align*}
    g_0 &= h_3 \\
    g_1 &= -h_2 \\
    g_2 &= h_1 \\
    g_3 &= -h_0
\end{align*}
\]

Figure 2 – Representation of a signal function $g(s)$, where (a) is the representation of this signal function and (b) the representation of this signal function with one transform level (OSORIO, 2008)

To obtain the scaling signals from Daub4 it is necessary to multiply the scaling coefficients by the signal $(s)$, according to equation 1 (DAUBECHIES, 1992; JENSEN e COUR-HARBO, 2001).

\[
\begin{align*}
    a_i &= h_0 s_{2^i} + h_3 s_{2^{i+1}} + h_2 s_{2^{i+2}} + h_1 s_{2^{i+3}} \quad \text{com } i = [0, \infty) \\
\end{align*}
\]

(Eq.1)

And to obtain the wavelet signals the same procedure is followed, being calculated according to equation 2.

\[
\begin{align*}
    c_i &= g_0 s_{2^i} + g_3 s_{2^{i+1}} + g_2 s_{2^{i+2}} + g_1 s_{2^{i+3}} \quad \text{com } i = [0, \infty) \\
\end{align*}
\]

(Eq.2)

As the obtained spectral signals were supposed to be compacted, only the scaling sub-signals were used because in this portion most part of the necessary spectral information is found to build the multivariate regression models.

3. Partial Least-Squares (PLS) Regression

In PLS the multivariate calibration is made using the whole spectrum information to build the regression model related to the interest property. For this reason it is called full-spectrum...
method (BORIN & POPPI, 2005). In this work PLS was used with MATLAB, which integrates mathematical computation, visualization and efficient language in a flexible environment for technical computation.

The PLS mathematical basis is the matrix $X$ decomposition, extended as the sum of various matrixes with unitary dimensions $M_i$ with $i = 1, 2,..., n$, added from the residual matrix (corresponding to errors), as represented in equation 3:

$$X = M_1 + M_2 + ... + M_n + E$$  \hspace{1cm} (Eq.3)

in which $n$ represents the number of latent variables used to truncate the equality and $E$ the residual matrix related to this number of latent variables chosen. The decomposition of matrix $X$ is made according to equations 4 and 5:

$$X = t_1p_1^t + t_2p_2^t + ... + t_np_n^t + E$$  \hspace{1cm} (Eq.4)

$$X = TP^t + E$$  \hspace{1cm} (Eq.5)

This way matrix $X$, which contains all spectral information, is now represented by a new matrix $X$, with smaller dimension, described by the matrixes $M_n$ without the loss of relevant information from the original variables.

And by analogy for matrix $Y$, resulting in equation 6:

$$Y = UQ^t + F$$  \hspace{1cm} (Eq.6)

where $T$ and $U$ are scores from $X$ and $Y$, and $P$ and $Q$ are loadings from $X$ and $Y$, respectively. The relation between $X$ and $Y$ results in the linear equation 7:

$$c = Xb + e$$  \hspace{1cm} (Eq.7)

where $X$ represents an answer matrix for the samples (spectra data), $c$ is a vector with the values of sample concentrations, $e$ indicates the model errors and spectra noises and $b$ is a vector representing the model parameters (KONZEN et al., 2003).

In this work the algorithm called PLS1 was applied, in which the matrix $Y$ represents a column matrix, that is a vector $y$, being each of the parameters (biodiesel and vegetable oil) modeled individually.

4. Interval Partial Least-Squares (iPLS) Regression

There are many ways of selecting the spectral region using objective criteria, such as evaluating the Root Mean Square Error of Cross Validation (RMSECV) and also using the algorithms which indicate the set of spectral variables which can give the best result based on RMSEP. Among the most applied there are heuristic methods such as genetic algorithm and the tabu search (KONZEN et al., 2003), or deterministic such as Interval Partial-Least Squares (iPLS).

The method iPLS is an extension from PLS, in which the set of data (spectra) is divided into a determined number of equidistant intervals, where for each interval a PLS model is built, besides the model built for the whole spectra, being the results presented in a graph to ease
understanding. The method is planned to give a general view of the data and can be useful to know which spectrum signals are more representative to build a good multivariate calibration model (NORGAAARD et al., 2000; LEARDI et al., 2004). However, attention must be given to the interval size because if it is too big the model might use information which is not representative for the problem, whereas if it is too small the opposite might happen, that is, it might not use all the necessary information to predict the interest property (NORGAAARD et al., 2000).

The following pseudocode for iPLS can be suggested:

\[
\text{Divide the spectra into } n \text{ equidistant intervals} \\
\text{While the number of intervals } \leq n \text{ do} \\
\quad \text{Build a PLS model for the } n \text{ interval} \\
\text{EndWhile} \\
\text{Select the model which interval presents the smallest RMSECV}
\]

5. Materials and Methods

5.1 Sampling

A total of 80 binary and ternary blends were formulated from diesel supplied by the LABCOM from Universidade Federal do Rio Grande do Sul – UFRGS, from palm biodiesel donated by Agropalma S/A, and from soybean biodiesel and its respective soybean oil, supplied by OleoPlan S.A.

For the calibration set, spectra obtained by FTIR/ATR from 48 blend samples were applied. The samples with extreme values of oil and/or biodiesel were included in this set, thus the respective RMSECV values were obtained.

For the prediction set, spectra obtained by FTIR/ATR from the 32 remaining blends were applied and RMSEP was obtained.

5.2. Acquisition of infrared spectra

The samples were submitted to homogenization before the spectrum acquisition, and they were obtained in duplicate in the range 4000-650 cm\(^{-1}\), with 4 cm\(^{-1}\) resolution and 32 scans, using one Nicolet Magna 550 spectrophotometer and one Attenuated Total Reflectance (ATR) accessory with a ZnSe crystal.

5.3 Compactness of signals with Discrete Wavelet Transform

The Daubechies Wavelet Transform with four coefficients (Daub4) was applied to treat the spectrum data from the binary and ternary blends containing biodiesel, vegetable oil and diesel, as stated before, with two levels of compactness being made: 1/4, 1/16, and only the scaling sub-signals were used to build the matrix to be modeled.

5.4. Multivariate regression models

To build the multivariate regression models the data were auto-scaled, which consists in centering the data on average and dividing them by the respective mean deviation, one for each wavelength. Models were built using the non-compacted spectrum data (matrix X) and the values % of biodiesel and vegetable oil (vectors y) used to formulate the blends, and these data were modeled using an environment MATLAB version 6.1 (The Math Works, Natick,
USA) applying the package iToolbox (NORGAARD et al., 2000) to develop the iPLS models, and the spectra were divided into 15 and 30 intervals in order to select the most representative spectral region to quantify the biodiesel and vegetable oil in the blends. The routines were run in an IBM microcomputer - Intel Pentium 4 CPU 3.00 GHz compatible with 1 GB RAM.

Later on new matrixes X were built with the signals deriving from 1/4 and 1/16 compactness, independently modeled, in analogy to what was described before.

To calculate the calibration model error RMSE (Root Mean Square Error) was applied, according to equation 8:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]  

(Eq.8)

where \(n\) is the number of spectra, \(y_i\) and \(\hat{y}_i\) are the % values used to formulate the blends and predicted by the PLS model, respectively, using the samples from the root mean square error of cross-validation set (RMSECV) and prediction set (RMSEP).

In Figure 3 a graph is presented supplied by the iPLS routine from the iToolbox package, to divide the spectrum into 15 intervals. The dotted line corresponds to the RMSECV value for the model containing the whole spectrum information, the bars correspond to the interval RMSECV and the number at the base of the bar corresponds to the number of latent variables (LVs) used to build the model.
The integration of productive chain with an approach to sustainable manufacturing.

Figure 3 - RMSECV values for a global model (dotted line) and for intervals (bars) generated by the iPLS routine from the iToolbox package. The numbers inside the bars correspond to the LVs for each interval

6. Results and Discussions

In Tables 1 and 2 the results for multivariate models are presented, in which the selection algorithm iPLS was applied to build these models with the use of compacted data subdivided into 15 and 30 intervals. It is verified that by varying the interval quantity to build the models the algorithm iPLS always uses a similar spectral region (Figures 4 and 6), which corresponds to a C-O deformation for present esters according to Barbosa (2007), for both the determination of biodiesel and vegetable oil, being this spectral region the one which identifies their qualitative and quantitative presence.

To determine the quantity of biodiesel in the blends, the root mean square error of cross-validation (RMSECV) had similar behaviors in relation to the total interval number and applied transforms. However, root mean square error of prediction (RMSEP) were lower for the models in which the spectra were subdivided into 15 intervals. In this determination, the most representative model (Figure 5) uses a 1/16 dimension from the original spectral matrix (108 variables), presenting a superior predictive capacity to the model without the compacted data (1736 variables).

In relation to the vegetable oil quantification, present on some samples on purpose, the RMSEP results for the models subdividing the spectra into 30 intervals showed more uniformity, which was not observed with the models built with the spectra divided into 15
intervals. For this determination the second transform level allowed the building of a model with smaller prediction error, as can be observed in Figure 7.
The integration of productive chain with an approach to sustainable manufacturing.

**Table 1 – Results obtained in the quantification of biodiesel in blends, applying iPLS**

<table>
<thead>
<tr>
<th>NT</th>
<th>NI</th>
<th>RMSECV</th>
<th>RMSEP</th>
<th>R²</th>
<th>LV</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15</td>
<td>0.1853</td>
<td>0.1352</td>
<td>0.9967</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>0.1883</td>
<td>0.1397</td>
<td>0.9966</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>0.1814</td>
<td>0.1279</td>
<td>0.9938</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>30</td>
<td>0.1866</td>
<td>0.1542</td>
<td>0.9967</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>0.1896</td>
<td>0.1413</td>
<td>0.9966</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>0.1812</td>
<td>0.1530</td>
<td>0.9969</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

**Table 2 – Results obtained in the quantification of vegetable oil in blends, applying iPLS**

<table>
<thead>
<tr>
<th>NT</th>
<th>NI</th>
<th>RMSECV</th>
<th>RMSEP</th>
<th>R²</th>
<th>LV</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15</td>
<td>0.1604</td>
<td>0.1524</td>
<td>0.9966</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>0.1500</td>
<td>0.0988</td>
<td>0.9980</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>0.1568</td>
<td>0.1427</td>
<td>0.9967</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>30</td>
<td>0.1428</td>
<td>0.1409</td>
<td>0.9973</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>0.1450</td>
<td>0.1321</td>
<td>0.9972</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>0.1796</td>
<td>0.1570</td>
<td>0.9957</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Figures 5 and 7 present the cross-validation and prediction samples well-adjusted in the curve, indicating that the models for biodiesel and oil can quantify their presence in the studied blends.

Another important aspect which must be pointed out is that even though two distinct kinds of biodiesel (palm biodiesel and soybean biodiesel) were used to formulate the binary and ternary blends, the models for the total quantification of biodiesel present in each sample presented determination coefficients bigger or equal to 0.9938, applying a few latent variables (from 2-4 latent variables).
Figure 4 – Selected spectral region (3) representing C-O deformation, characteristic of the biodiesel

Figure 5 – Reference values versus predicted values for the PLS model for the biodiesel determination using the interval 3 for the FTIR/ATR signals with 4 transforms
Figure 6 – Selected spectral region (3) representing C-O deformation, characteristic of the vegetable oil

Figure 7 – Reference values versus predicted values for the $i$PLS model for the vegetable oil determination using the interval 3 for the FTIR/ATR signals with 2 transforms

7. Conclusions
This paper shows the contribution from wavelets Daub4 to reduce the dimensionality of a spectral matrix with the information retention of the present chemical structures, allowing the construction of more robust $i$PLS calibration models for the quantification of biodiesel and
vegetable oil in biodiesel/vegetable oil/diesel blends applying FTIR/ATR signals.

Another important contribution from this work was to identify which infrared spectral region obtained by Attenuated Total Reflectance can model not only the presence of biodiesel in the blends but also the presence of a possible adulteration, which is the vegetable oil which originated this biodiesel.

The combination of iPLS regression with the data obtained by FTIR/ATR has been promising to develop simpler, faster and non-destructive methodologies for the biodiesel determination and vegetable oil adulterations in blends commercialized since January 1, 2008, when the addition of biodiesel to diesel was made mandatory in national territory.

Acknowledgments

We also gratefully acknowledge financial support from the Brazilian agencies FAPERGS, CNPq and CAPES.

References


DANTAS FILHO, H.A. Desenvolvimentos de técnicas quimiométricas de compressão de dados e redução de ruído aplicadas a óleo diesel e madeira de eucalipto empregando espectroscopia NIR. Tese, UNICAMP, Campinas, 2007.


