Talanta 119 (2014) 582–589

Contents lists available at ScienceDirect

Talanta

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

# Quantification of animal fat biodiesel in soybean biodiesel and B20 diesel blends using near infrared spectroscopy and synergy interval support vector regression

Paulo Roberto Filgueiras, Júlio Cesar L. Alves, Ronei Jesus Poppi *

*Institute of Chemistry, University of Campinas–UNICAMP, P.O. Box 6154, 13083-970 Campinas, SP, Brazil*

ARTICLE INFO

*Article history:*
Received 15 October 2013
Received in revised form
19 November 2013
Accepted 22 November 2013
Available online 28 November 2013

*Keywords:*
Biodiesel
Animal fat
Partial least squares
Support vector regression
Permutation test

ABSTRACT

In this work, multivariate calibration based on partial least squares (PLS) and support vector regression (SVR) using the whole spectrum and variable selection by synergy interval (siPLS and siSVR) were applied to NIR spectra for the determination of animal fat biodiesel content in soybean biodiesel and B20 diesel blends. For all models, prediction errors, bias test for systematic errors and permutation test for trends in the residuals were calculated. The siSVR produced significantly lower prediction errors compared to the full spectrum methods and siPLS, with a root mean squares error (RMSEP) of 0.18% (w/w) (concentration range: 0.00%–69.00%(w/w)) in the soybean biodiesel blend and 0.10%(w/w) in the B20 diesel (concentration range: 0.00%–13.80%(w/w)). Additionally, in the models for the determination of animal fat biodiesel in blends with soybean diesel, PLS and SVR showed evidence of systematic errors, and PLS/siPLS presented trends in residuals based on the permutation test. For the B20 diesel, PLS presented evidence of systematic errors, and siPLS presented trends in the residuals.

## 1. Introduction

Biodiesel is currently the main substitute of petroleum diesel oil due to its similar characteristics; it has been used in many countries as diesel engine fuel, usually in mixtures of up to 20%(v/v) in diesel oil, which is also called B20[1].

It is necessary to highlight that some low temperature properties (*i.e.*, pour point, cloud point and cold filter plugging point – CFPP) of biodiesel/diesel blends vary according to the biodiesel raw material; this variation can be a critical issue depending on the region of use of fuel and the climate temperature. The lower the climate temperature, the lower the pour point, cloud point and the cold filter plugging point should be to avoid faults related to engine fuel quality[2].

The vegetable oil or animal fat transesterification reaction produces a mixture of methyl esters. Depending on the biodiesel raw material used, it contains different amounts of long saturated carbon chain methyl esters, which tend to increase the pour point, cloud point and CFPP. As the animal fat has a relatively high content of saturated fatty acids, this biodiesel raw material leads to high contents of saturated carbon chain esters. In this manner, the raw material used to produce the biodiesel determines its low temperature quality properties, limiting the use of animal fat biodiesel to relatively cold climates.

This issue is especially important in the use of diesel fuel with a high content of biodiesel, such as B20, in order to avoid engine failures including difficulty starting the engine and clogging fuel filters. In Brazil, the biodiesel produced using soybean oil and animal fat as raw materials represents approximately 90% of the total production in 2012. The use of animal fat has been increasing due to several attractive aspects such as lower cost and abundant availability. Normally, the soybean biodiesel and animal fat biodiesel blends are commercially available and used to produce B20 diesel fuel, but depending on the mixture proportions and the operational temperatures, these blends can cause the cited engine operational failures.

The development of practical and reliable analytical methods for the determination of animal fat biodiesel content in soybean biodiesel or diesel B20 can facilitate the quality control of fuel. Also, it can avoid operational failures related to the quality of fuel in diesel engine vehicles.

Because of expansion of biodiesel production, several studies have been developed to determine quality parameters [3–10], either in biodiesel blends or biodiesel/petro diesel blends [11–17]. Near infrared spectroscopy (NIR) combined with partial least squares (PLS) regression has produced feasible results for analysis. However, other regression techniques, such as support vector regression (SVR), have been gaining attention due to their robustness, generalisation capabilities and ability to model nonlinear relationships [18–20].

---

* Corresponding author. Tel.: +55 019 35213126; fax: +55 019 35213023.
*E-mail addresses:* ronei@iqm.unicamp.br, rjpoppi@gmail.com (R.J. Poppi).

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.11.056

Considering the large number of variables involved in NIR spectra, variable selection based on spectral regions can improve the performance of the calibration model in relation to ones that employ the whole spectrum. Norgaard *et al.* [21] proposed a deterministic method for the selection of spectral regions in PLS models, in which the spectrum is divided into equidistant intervals. A combination of several intervals is then performed, leading to better predictions. This method was called the synergy interval or siPLS. The same interval selection idea can be used in the development of SVR models in an approach called siSVR, in which the combination of spectral regions can improve overall robustness with respect to the accuracy of the model where the whole spectrum is used.

Given the importance of the control of fuels containing animal fat biodiesel, this paper aims to apply multivariate calibration based on PLS and SVR using the full spectrum and variable selection by synergy interval in NIR spectra to quantify the content of animal fat biodiesel in blends with soybean biodiesel and B20 diesel. In addition, systematic errors were assessed by the bias test and trends in residuals of the developed models by the permutation test.

## 2. Data analysis

### 2.1. Calibration models

PLS regression [22–27] is the most common multivariate calibration method used to relate spectral information, such as NIR, to some properties of interest. Its theory has been widely described in the literature, and it is available in many chemometric packages. Hundreds of applications had been presented with excellent performance, but as the method is based on linear parameters, it has difficulty modelling non-linear trends or multi-class data. In this sense, SVR has proven to be an excellent alternative to PLS in such situations because it can address non-linear relationships and possesses excellent generalisation ability.

The SVR is a nonlinear regression method developed as an extension of the theory of support vector machines (SVM) used for binary classification problems [28–31]. From the data set $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ with $\mathbf{x} \in R^m$ and $y \in R$, the SVR aim is to find a function

$$f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b \tag{1}$$

where the input vectors **x** are mapped into a high-dimensional feature space Z through some nonlinear mapping, $\phi$:**x**i → zi and the parameters **w** and b are estimated by minimising the regularised risk function $R_f$

$$R_f = \frac{1}{2}||\mathbf{w}||^2 + C\left(\nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \tag{2}$$

$$\text{Subjectto}: \begin{cases} y_i - \langle \mathbf{w}, \phi(\mathbf{x}) \rangle - b \le \varepsilon + \xi_i^* \\ \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{3}$$

The parameter $\nu$ controls the number of support vectors, called $\nu$-SVR. The constant $C > 0$ considers the error function $(\xi_i, \xi_i^* \ge 0)$, which is enclosed by a tolerance $\varepsilon$, described by loss function $\varepsilon$-insensitive

$$L(\varepsilon) = \begin{cases} 0 & if, |L(\varepsilon) - f(\mathbf{x})| < \varepsilon \\ L(\varepsilon) - f(\mathbf{x}) & otherwise \end{cases} \tag{4}$$

Applying the Lagrange multipliers and solving the quadratic optimisation problem given in Eq. (2), the regression function is obtained

$$f(\mathbf{x}) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\mathbf{K}(\mathbf{x_i}, \mathbf{x}) + b \tag{5}$$

where $\alpha_i$ and $\alpha_i^*$ represent the Lagrange Multipliers. An important step in modelling is the mapping of the original data into a feature space of high dimensional data by a kernel function. Thus, the mapped data are analysed using conventional linear statistics in the feature space, which is equivalent to nonlinear analysis in the original space. Eq. (5) $\mathbf{K}(\mathbf{x}_i, \mathbf{x})$ represents the kernel mapping of the original data given by

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle. \tag{6}$$

The radial basis function (RBF) [32] is the kernel most used in SVR and is represented by

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2) \tag{7}$$

where the $\gamma$ parameter should be set by the analyst during modelling.

### 2.2. Accuracy assessment

The root mean squares error of prediction (RMSEP) expresses the accuracy [33] of the model. It reports the closeness of agreement between the reference value and the value found by the model

$$\text{RMSEP} = \sqrt{\sum_{i=1}^{n}\frac{(\text{y}_{\text{ref},i} - \text{y}_{\text{est},i})^2}{\text{n}}} \tag{8}$$

where $\text{y}_{\text{ref}}$ is the reference concentration value for each of the $n$ test samples and $\text{y}_{\text{est}}$ is the concentration value estimated by the model.

A comparison of the model accuracy can be evaluated by an *F*-test, according to the equation

$$Fcalc = \frac{RMSEP_1^2}{RMSEP_2^2} \tag{9}$$

where $\text{RMSEP}_1 > \text{RMSEP}_2$. The calculated $F_{\text{calc}}$ value is compared with the value of the Fisher–Snedecor distribution (F-statistic) with degrees of freedom equal to the number of prediction samples and an adopted significance level of 5% ($\alpha$=0.05). If the tabulated value of the F-statistic is less than $F_{\text{calc}}$, there is no statistical evidence of homogeneity of the values, and the method with $\text{RMSEP}_2$ presents better accuracy.

### 2.3. Systematic error assessment

The term bias is attributed to systematic errors that are calculated by the difference between the population mean and the true value and are all components of error that are not random. The average bias for the validation set can be calculated by Eq. (10) [34]

$$\text{bias} = \frac{\sum_{i=1}^{n}(y_{ref,i} - y_{est,i})}{n} \tag{10}$$

The standard deviation of the validation errors (SVD) is obtained as

$$\text{SVD} = \sqrt{\frac{\sum_{i=1}^{n}\left[\left[(y_{ref,i} - y_{est,i}) - bias\right]^2\right]}{n-1}} \tag{11}$$

The *t*-test for the validation samples is used for bias assessment

$$t = \frac{|\text{bias}|\sqrt{n}}{\text{SVD}} \tag{12}$$

The *t*-calculated is compared with the *t*-tabulated of the *t*-student distribution with $n-1$ degrees of freedom and a significance level of $\alpha$=0.05. If the *t*-calculated presents a value greater than its critical value, there is an indication that systematic errors are present in the model.

### 2.4. *Trends in residuals assessed by the permutation test*

In correctly adjusted calibration models, it is expected that the residuals (difference between the reference or true value and the value estimated by the model) remain roughly uniform in size as the measured value increases and is normally distributed about zero.

In this sense, a test was applied to verify the presence of the quadratic trend in the distribution of residuals as a function of the measured value. Other forms of trends other than quadratic can also be evaluated; however, in this work, only the quadratic trend was tested because this behaviour was observed in the multivariate models of this work.

The presence of trends in residuals generated by the multivariate model was assessed by a permutation test. The test is based on a repetitive reordering of the measured values, keeping the order of the residuals related to each estimated value unchanged [35].

The first step of the test is to adjust a second order polynomial function between the residuals ($e$) of the model and the measured values ($y_i$)

$$e(y_i) = b_0 + b_1 y_i + b_2 y_i^2 \tag{13}$$

Null and alternative hypotheses can be established: $H_0$: prediction residues do not follow a quadratic trend, but $H_1$: residues follow a quadratic trend.

From the adjusted Eq. (13), if the coefficient $b_2$ is significantly different from zero at a specified significance level, there will be evidence of a trend in the residuals. A permutation algorithm was developed to perform this test and consists of the following steps:

(i) coefficient $b_2$ is calculated by least squares and named $b_2$*;
(ii) $y$ is permuted randomly, maintaining the unchanged order of the residuals;
(iii) a new coefficient $b_2$ is calculated from the new permuted y and residuals;
(iv) steps (ii) and (iii) are repeated k times;
(v) $p$-value is determined by the ratio of the number of times that $b_2 \geq b_2$* and k. If the $p$-value is greater than the level of significance $\alpha = 0.05$, there is evidence to accept $H_0$ and that the residuals do not present a quadratic trend. Otherwise, there is a trend in the residuals. All calculations were performed in Matlab version R2009a.

## 3. Experimental

A total of 99 binary blends of animal fat biodiesel and soybean biodiesel were initially prepared, with concentrations of animal fat biodiesel ranging from 0% to 69%(w/w). In a second step, these 99 blends were mixed with S50 diesel fuel oil (petro diesel containing a maximum of 50 mg/kg of sulphur) in a fixed 20% of biodiesel blends and 80% of petro diesel, generating the B20 diesel fuel. Afterward, the animal fat biodiesel content in the diesel B20 ranged from 0.00% to 13.80%(w/w), and the soybean biodiesel content in diesel B20 ranged from 6.20% to 20.00%(w/w).

NIR spectra were obtained using a Perkin Elmer 100N spectrometer equipped with a liquid transflectance accessory in the range of 9000 to 4000 cm$^{-1}$, with a resolution of 4 cm$^{-1}$ and 32 scans per sample. Samples were measured in duplicate, and the average spectrum was used for building of calibration models.

The samples were randomly divided into two groups: 59 for calibration and 40 for prediction. Firstly, the NIR spectra were preprocessed using multiplicative signal correction (MSC) [36]. “k-fold” procedure was used to optimise the models. It consists of splitting the total dataset into k mutually exclusive subsets of the same size. A subset is used for testing (validation samples), and the remaining k − 1 is used for model development (calibration samples). This process is performed k times until all subsets have been tested. At the end of k iterations, the root mean square error of cross-validation (RMSECV) is calculated, generating a more reliable measure of the predictive ability of the model for future samples.

PLS models were built with mean-centered spectra and reference vector (biodiesel of animal fat or soy contents) auto-scaled. The optimal number of latent variables that minimise the RMSECV was achieved according to the cross-validation “7-fold” procedure. The PLS model was built using routines from the PLS Toolbox 6.7 for Matlab [37], and the siPLS model was developed based on the same data sets using the iToolbox [38]. Models were developed using the full spectrum divided into 5, 10 and 20 equally spaced intervals and were combined in two intervals.

The SVR model was developed using the LIBSVM routine for Matlab [39]. The spectral data matrix and the vector containing the biodiesel from animal fat (or soy) content was normalized in the range of 0–1. The Radial Basis Function was used as a kernel function, and a grid search was used to optimise the SVR parameters $C$, $\nu$ and $\gamma$, through the “7-fold” cross-validation. In the siSVR model, a grid search was also used to optimise the parameters $C$, $\nu$ and $\gamma$, based on a “7-fold” cross-validation, with routines developed in the laboratory. The same number of intervals and the combination of them tested in siPLS were applied in siSVR.

The linear fit of the adjusted line between the reference values against the predicted ones were assessed by the coefficient of determination$R^2$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2} \tag{14}$$

where $\overline{y}_i$ is the average value of prediction, $y_i$ is the reference value and $\hat{y}_i$ is the prediction value.

## 4. Results and discussion

### 4.1. *Spectral analysis*

Fig. 1 presents the NIR spectra of animal fat biodiesel with 31% (w/w) soybean biodiesel, pure soybean biodiesel and diesel B20 with 20%(w/w) of soybean biodiesel, split into 10 equidistant intervals used in the construction of siPLS and siSVR models. The numbers 1–10 refer to the interval number.

In the interval number 10, an absorption band approximately 4069 cm$^{-1}$ with a higher intensity for animal fat biodiesel is observed. This region is relative to the C3H deformation associated with the linear aliphatic, which is also associated with the lowest number of unsaturation in the animal fat biodiesel. Another difference can be observed in the spectral region of interval number 9 approximately 4662 cm$^{-1}$. This difference is related to C3H stretching, a CåO stretching combination and a C3H deformation combination, with a higher intensity for soybean biodiesel. In interval number 7, an absorption band at approximately 5865 cm$^{-1}$ is observed and is related to the second overtone of the C3H of the methyl terminal group, with a lower intensity for diesel B20.

### 4.2. *Quantification of animal fat biodiesel content in blends with soybean biodiesel*

Firstly, models were developed to quantify the animal fat biodiesel blends with soybean diesel because this knowledge is important to control the raw materials for biodiesel production. The results obtained using PLS, SVR, siPLS and siSVR for calibration

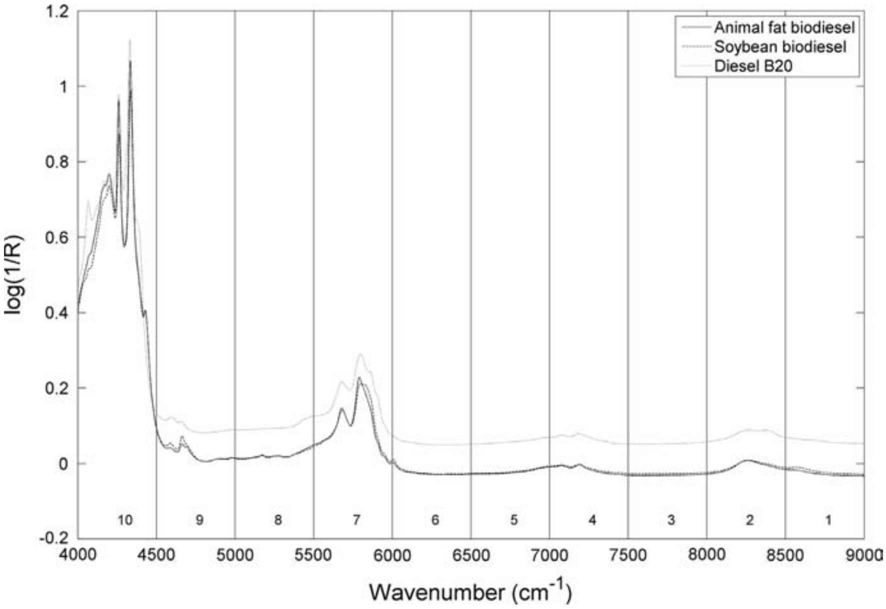

**Fig. 1.** NIR spectrum of animal fat biodiesel, soybean biodiesel and diesel B20 with 20% of soybean biodiesel, segmented into 10 intervals.

**Table 1**
Results of calibration models for the quantification of animal fat biodiesel blends with soybean biodiesel.

| Model | Intervals | Calibration | | | Prediction | | |
|---|---|---|---|---|---|---|---|
| | | RMSECV (%m/m) | bias (%m/m) | $R^2$cv | RMSEP (%m/m) | bias (%m/m) | $R^2$p |
| PLS (3)[a] | All | 0.84 | –0.0220 | 0.9983 | 0.82 | 0.0022 | 0.9983 |
| siPLS (7)[a] | 7 and 8 | 0.25 | –0.0132 | 0.9998 | 0.25 | 0.0065 | 0.9999 |
| SVR | All | 0.46 | –0.0308 | 0.9995 | 0.45 | –0.0938 | 0.9995 |
| siSVR | 7 and 9 | 0.20 | –0.0229 | 0.9999 | 0.18 | –0.0563 | 0.9999 |

[a] Number of latent variables in the model.

(in a cross-validation approach) and prediction are presented in Table 1. For all methods tested, it can be observed that the values of RMSECV and RMSEP are similar, indicating no overfitting in the models. The best results were obtained using the variable selection procedure, with a RMSEP of 0.25%(w/w) and 0.18%(w/w) for the siPLS and siSVR models, respectively. A reduction in error of more than 50% was achieved by using the variable selection, demonstrating the importance of adequately choosing the spectral interval. In addition, an *F*-test performed for accuracy comparison between siPLS and siSVR ($F_{cal}=1.93>$ F-statistic $(v_1=v_2=40,\ \alpha=0.05)=1.69$) indicated that the difference between the RMSEPs of these two models is statistically significant, with the siSVR model being more accurate.

To assess the systematic errors of the models, a test for bias significance was performed, and for all models, no systematic errors were found using the expression presented in Eq. (12).

The calibration and prediction residuals as a function of the measured value for each model are shown in Fig. 2, where the residuals for models with variable selection are smaller than the models using the full spectra. It also can be noted that the prediction residuals for the PLS and siPLS models have a quadratic trend, which underestimates of the results at the beginning and end of the curve, while, in the central part, the results are overestimated. For correctly adjusted models, it is expected that the residuals remain the size roughly uniform as the measured value increases and normally distributed about zero; thus, a permutation test was performed to verify trends in the residuals.

Fig. 3 shows the histograms of the permutation test results (10,000 permutations), where the distribution of the values of the permuted coefficients relative to the quadratic adjustment ($b_2$ coefficient of Eq. (13)) is presented; the vertical line is the coefficient $b_2$ calculated by least squares ($b_2$*) before the permutation. A *p*-value is calculated by the ratio of the area to the right (because the positive quadratic coefficient) of the $b_2$* and the total area. If the *p*-value is smaller than the significance level adopted (0.05 in this work), the quadratic coefficient is significant. In this case, it was confirmed that the PLS and siPLS models present evidence of quadratic trends in their residuals, with a *p*-value of 0.0389 and 0.0001, respectively, while the SVR and siSVR models presented no evidence of this type of trend in the residuals.

The siSVR model selected intervals 7 and 9. Interval 7 has two absorption bands: 5797 $\text{cm}^{-1}$ from the first overtone of a C–H stretch of a hydrocarbon and 5675 $\text{cm}^{-1}$ from the first overtone of C–H symmetrical stretching of an aliphatic hydrocarbon [40]. Interval 9 contains an absorption band at approximately 4662 $\text{cm}^{-1}$, which is related to C3H stretching, a CåO stretching combination and a C3H deformation combination. The siPLS model selected interval 7 as the template siSVR; interval 8 apparently does not present absorption bands in the spectrum (see Fig. 1).

### 4.3. *Quantification of animal fat biodiesel content in diesel B20*

Table 2 presents the results of the quantification of animal fat biodiesel content in diesel B20. The strong performance of the models, except for PLS, can be seen by the high values of the determination coefficient (above 0.99) for the adjusted line between the reference and predicted values. It is possible to note

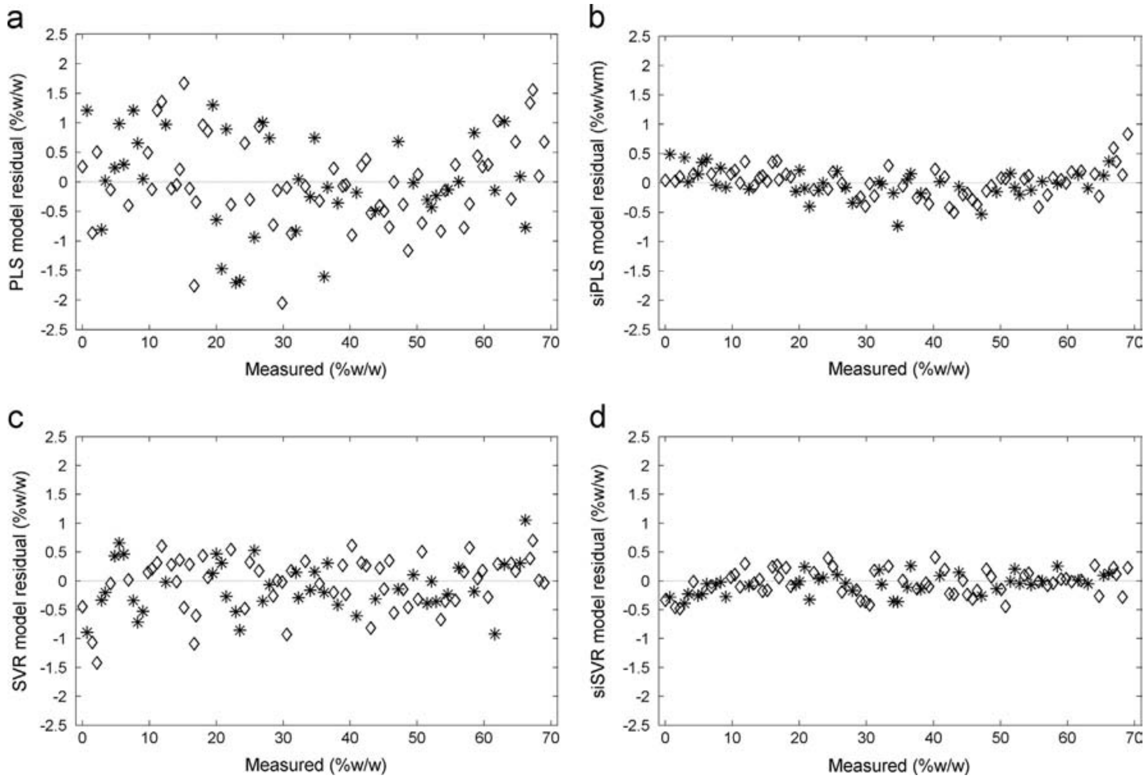

**Fig. 2.** Plot of the residuals of the multivariate calibration models against the measured values for the determination of the animal fat biodiesel content in blends with soybean biodiesel. (a) PLS, (b) siPLS, (c) SVR, (d) siSVR. Residuals of the calibration model (◊) and residuals of the prediction model (*).

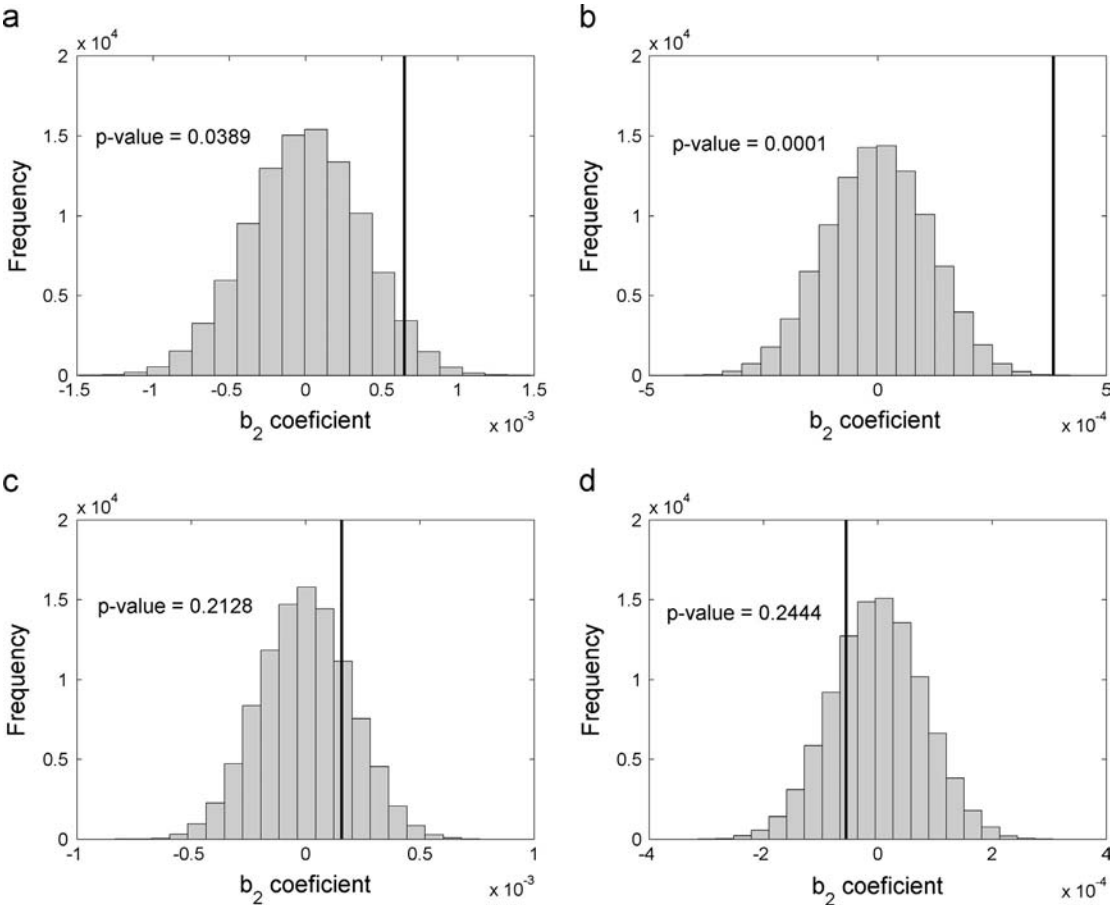

**Fig. 3.** Permutation test results for the evaluation of trends in the residuals of calibration models for the quantification of animal fat biodiesel in blends with soybean biodiesel. (a) PLS, (b) siPLS, (c) SVR and (d) siSVR.

**Table 2**
Results of calibration models for the quantification of animal fat biodiesel in diesel B20.

| Model | Intervals | Calibration | | | Prediction | | |
|---|---|---|---|---|---|---|---|
| | | RMSECV (%m/m) | bias (%m/m) | $R^2$cv | RMSEP (%m/m) | bias (%m/m) | $R^2$p |
| PLS (4)[a] | All | 0.53 | 0.0057 | 0.9829 | 0.55 | −0.0999 | 0.9817 |
| siPLS (5)[a] | 7, 8 | 0.16 | 0.0015 | 0.9984 | 0.15 | −0.0390 | 0.9986 |
| SVR | All | 0.36 | 0.0179 | 0.9921 | 0.29 | 0.1010 | 0.9953 |
| siSVR | 7, 9 | 0.16 | −0.0162 | 0.9984 | 0.10 | −0.0213 | 0.9994 |

[a] Number of latent variables in the model.

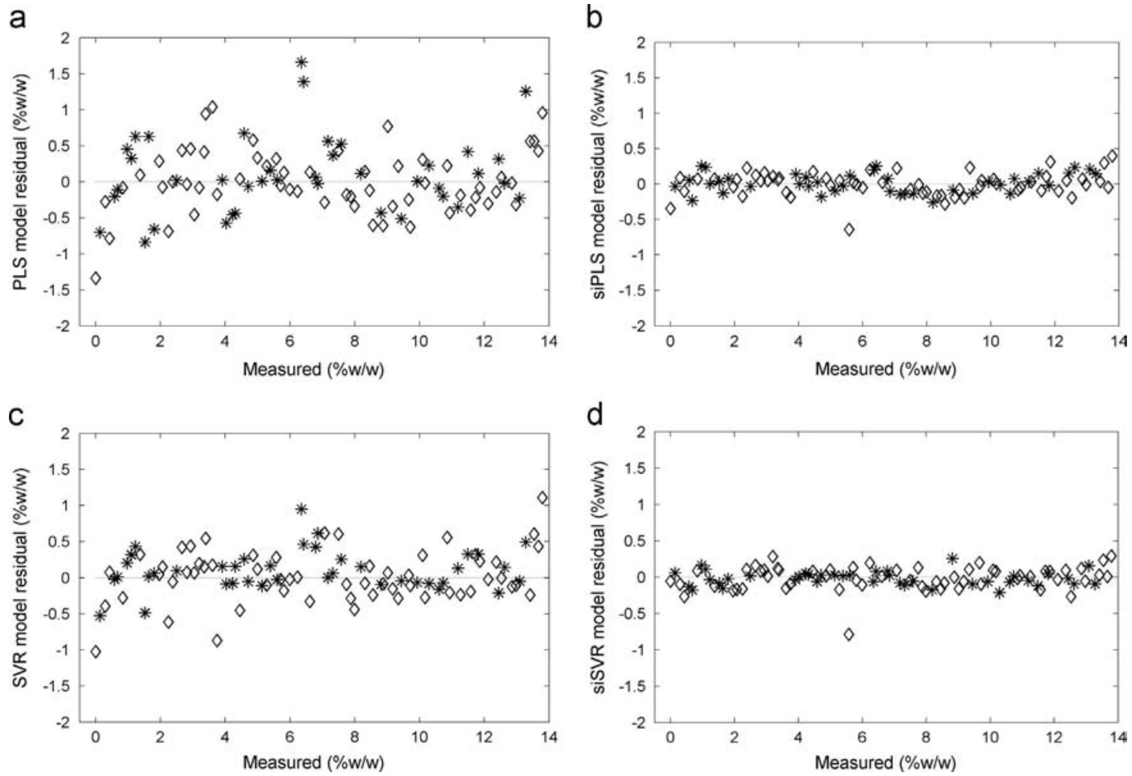

**Fig. 4.** Plot of the residuals of the multivariate calibration models against the measured values for the determination of the animal fat biodiesel content in B20. (a) PLS, (b) siPLS, (c) SVR and (d) siSVR. Residuals of the calibration model (◊) and residuals of the prediction model (⁎).

that even changing the matrix with the addition of petro diesel, the siPLS model selected intervals 7 and 8, and the siSVR model selected intervals 7 and 9 (see Fig. 1). However, the bias test showed evidence of systematic errors at a significance level of 0.05 in the results for PLS and SVR without variable selection. For PLS, $t_{cal}=2.71 > t\text{-statistic}\ (\nu=39,\ \alpha=0.05)=2.02$, and for SVR, $t_{cal}=2.31 > t\text{-statistic}(\nu=39,\alpha=0.05)=2.02$.

The prediction results presented the same trends observed in a previous calibration, where petro diesel was absent. The calibration and prediction residuals as a function of the measured value for each model are shown in Fig. 4, where the residuals for models with variable selection are smaller than the models using the full spectra. The lowest prediction errors obtained by the siPLS and siSVR models with RMSEPs of 0.15%(w/w) and 0.10%(w/w), respectively. An *F*-test applied to the RMSEPs of these models with variable selection showed that the errors are not similar at a 0.05 level of significance ($F_{cal}=2.25 > \text{F-statistic}(\nu_1=\nu_2=40,\ \alpha=0.05)=1.69$). This finding is an indication that the siSVR model produces more accurate results.

The permutation test was conducted to evaluate trends in residuals of the models. As the above test demonstrated, 10,000 permutations were performed to build the histogram of the distribution of the $b_2$ coefficient values. In this case, only the siPLS model showed evidence of quadratic trends in the residuals, with a *p*-value of 0.0156, as seen in Fig. 5, which is lower than the significance level of 0.05.

## 5. Conclusions

The non-linear multivariate calibration combined with selection of variables is an adequate methodology for prediction of animal fat biodiesel content in blends with soybean biodiesel or diesel B20. The SVR with synergy interval (siSVR) produced significantly lower prediction errors in comparison with the full spectrum method or linear approach such as PLS (based on the full spectrum or interval selection with synergism of intervals). The siSVR method selected spectral intervals related to the C–H bond of the aliphatic hydrocarbon and the CåO carbonyl group, indicating the main variations between samples and correlating with the type of biodiesel in B20.

The bias test for systematic error analysis and the permutation test for trends in residuals were applied in all models, even in those with a high coefficient of determination in the relationship of reference and predicted values. The PLS method, which is currently the most used algorithm for this type of modelling, requires a more careful analysis of the results because there was

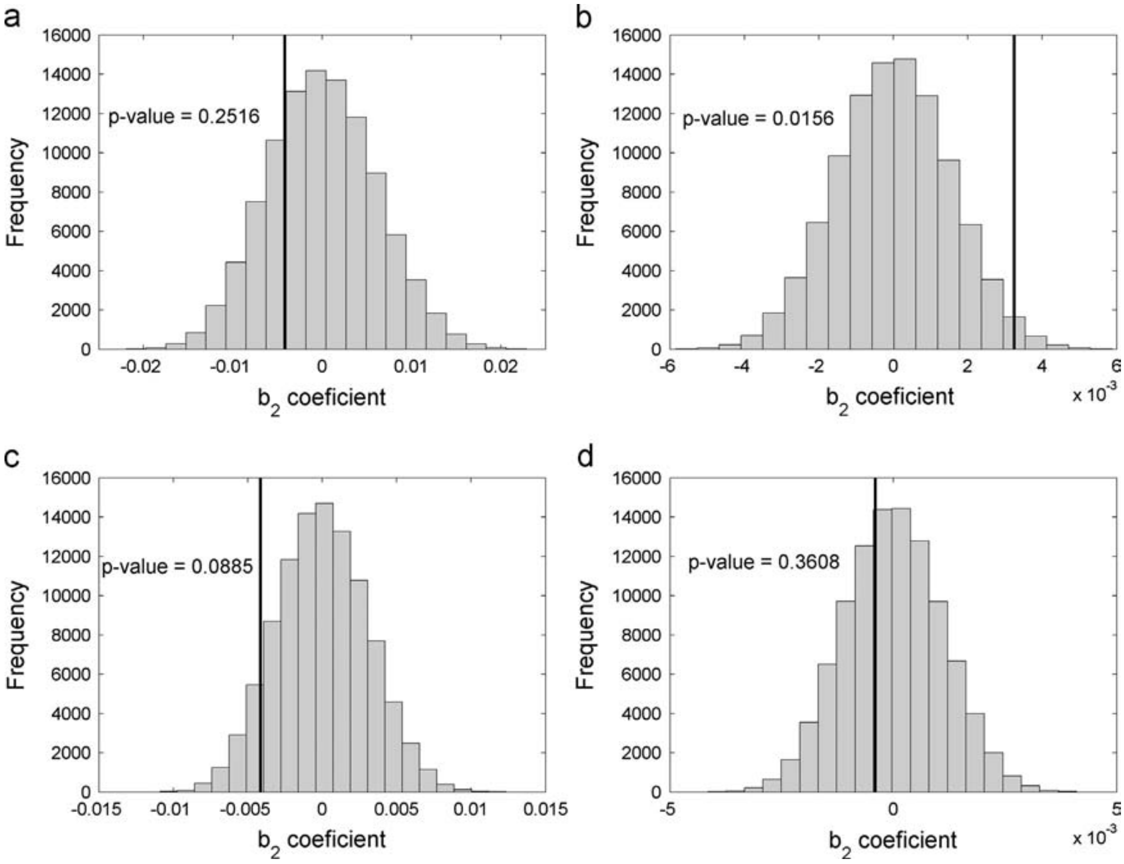

**Fig. 5.** Permutation test results for the evaluation of trends in the residuals of calibration models for the quantification of animal fat biodiesel in diesel B20. (a) PLS, (b) siPLS, (c) SVR and (d) siSVR.

evidence of systematic errors and trends in the residuals in all of the tested calibrations. These results corroborate the importance of the evaluation of quality parameters in multivariate calibration models other than errors of prediction related in RMSEP because these does not provide information about bias or lack of adjustment of the models. Using the bias test for determining systematic errors and the permutation test for detecting trends in the residuals, additional information can be incorporated into the comparison of models, assisting in choosing the best calibration methodology.

## Acknowledgements

The authors would like to thank Petróleo Brasileiro S.A.–PETROBRAS for providing the oil blends and CNPq for their financial support.

## References

[1] ANP 2012. National agency of petroleum, Natural Gas and Biofuels, ⟨http://nxt.anp.gov.br/NXT/gateway.dll/leg/resolucoes_anp/2011/janeiro/ranp%202%20-%202011.xml?f=templates$fn=document-frame.htm$3.0$q=$x=$nc=9362⟩ (August, 2013).
[2] ANP 2012. National Agency of Petroleum, Natural Gas and Biofuels, ⟨http://nxt.anp.gov.br/nxt/gateway.dll/leg/resolucoes_anp/2012/maio/ranp%2014%20-%202012.xml⟩ (August, 2013).
[3] P. Felizardo, P. Baptista, J.C. Menezes, M.J.N. Correia, Anal. Chim. Acta 595 (2007) 107–113.
[4] B.Q. Araújo, R.C.R. Nunes, C.V.R. Moura, E.M. Moura, A.M.G.L. Citó, J.R. Santos Júnior, Energy Fuels 24 (2010) 4476–4480.
[5] L.S.G. Teixeira, M.B. Couto, G.S. Souza, M. Andrade Filho, J.C.R. Assis, P.R. B. Guimarães, L.A.M. Pontes, S.Q. Almeida, J.S.R. Teixeira, Biomass Bioenergy 34 (2010) 438–441.
[6] L.F.B. Lira, F.V.C. Vasconcelos, C.F. Pereira, A.P.S. Paim, L. Stragevitch, M.F. Pimentel, Fuel 89 (2010) 405–409.
[7] L.F.B. Lira, M.S. Albuquerque, J.G. Andrade Pacheco, T.M. Fonseca, E.H.S. Cavalcanti, L. Stragevitch, M.F. Pimentel, Microchem. J. 96 (2010) 126–131.
[8] M.F. Ferrão, M.S. Viera, R.E.P. Pazos, D. Fachini, A.E. Gerbase, L. Marder, Fuel 90 (2011) 701–706.
[9] N. Canha, P. Felizardo, J.C. Menezes, M.J.N. Correia, Fuel 97 (2012) 352–357.
[10] J.C.L. Alves, C.B. Henriques, R.J. Poppi, Fuel 97 (2012) 710–717.
[11] D.D.S. Fernandes, A.A. Gomes, G.B. Costa, G.W.B. Silva, G. Véras, Talanta 87 (2011) 30–34.
[12] F.V.C. Vasconcelos, P.F.B. Souza Jr., M.F. Pimentel, M.J.C. Pontes, C.F. Pereira, Anal. Chim. Acta 716 (2012) 101–107.
[13] M.J.C. Pontes, C.F. Pereira, M.F. Pimentel, F.V.C. Vasconcelos, A.G.B. Silva, Talanta 85 (2011) 2159–2165.
[14] W.F.C. Rocha, R. Nogueira, B.G. Vaz, J. Chemom. 26 (2012) 456–461.
[15] V. Gaydou, J. Kister, N. Dupuy, Chemom. Intell. Lab. Syst. 106 (2011) 190–197.
[16] V.O. Santos Jr., F.C.C. Oliveira, D.G. Lima, A.C. Petry, E. Garcia, P.A.Z. Suarez, J.C. Rubim, Anal. Chim. Acta 547 (2005) 188–196.
[17] I.P. Soares, T.F. Rezende, R.C. Silva, E.V.R. Castro, I.C.P. Fortes, Energy Fuels 22 (2008) 2079–2083.
[18] S.S. Fong, V.S. Kiss, R.G. Brereton, Talanta 83 (2011) 1269–1278.
[19] H. Li, Y. Liang, Q. Xu, Chemom. Intell. Lab. Syst. 95 (2009) 188–198.
[20] D. Zhu, B. Ji, C. Meng, B. Shi, Z. Tu, Z. Qing, Anal. Chim. Acta 598 (2007) 227–234.
[21] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.
[22] A. Höskuldsson, J. Chemom. 2 (1998) 211–228.
[23] Q.S. Xu, Y.Z. Liang, H.L. Shein, J. Chemom. 15 (2001) 135–148.
[24] R.G. Brereton, Analyst 125 (2000) 2125–2154.
[25] S. Jong, Chemom. Intell. Lab. Syst. 18 (1993) 251–263.
[26] S. Wold, M. Sjöströma, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[27] M. Andersson, J. Chemom. 23 (2009) 518–529.
[28] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Statist. Comput. 14 (2004) 199–222.
[29] V.N. Vapnik, IEEE Tran. Neural Netw. 5 (1999) 988–999.
[30] J.H. Kalivas, J. Chemom. 13 (1999) 111–132.
[31] C. Cortes, V. Vapnik, Mach. Learn. 20 (3) (1995) 273–297.
[32] B. Schölkopf, K.K. Sung, C.J.C. Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik, IEEE Trans. Neural Netw. 11 (1997) 2758–2765.

[33] P. Valderrama, J.W.B. Braga, R.J. Poppi, J. Agric. Food Chem. 55 (2007) 8331–8338.
[34] Annual book of ASTM Standards, standards practices for infrared, multivariate, quantitative analysis, E1655-05, vol. 03.06, ASTM International, West Conshohocken, Pennsylvania, USA, 2005.
[35] H. van der Voet, Chemom. Intell. Lab. Syst. 25 (1994) 313–323.
[36] P. Geladi, D. MacDougall, H. Martens, Appl. Spectrosc. 39 (1985) 491–500.
[37] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch, PLS toolbox Version 4.0 for use With Matlab, Eigenvector research Inc, Wenatchee, 2006.
[38] itoolbox, ⟨http://www.models.life.ku.dk/ipls⟩ 2000.
[39] C.C. Chang, C.J. Lin,;1; LIBSVM: a library for support vector machines, software available at ⟨http://www.csie.ntu.edu.tw/∼cjlin/libsvm⟩ 2001.
[40] J. Workman Jr., L. Weyer, Practical Guide to Interpretive Near Infrared Spectroscopy, CRC Press, Boca Raton, 2008.