Talanta 104 (2013) 155–161

Contents lists available at SciVerse ScienceDirect

Talanta

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

Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM)

Julio Cesar L. Alves*, Ronei J. Poppi

Institute of Chemistry, State University of Campinas – UNICAMP, PO Box 6154, 13083-970 Campinas, SP, Brazil

ARTICLE INFO

Article history:
Received 20 June 2012
Received in revised form
27 September 2012
Accepted 14 November 2012
Available online 23 November 2012

Keywords:
Biodiesel
Diesel fuel
Support vector machines
Near infrared (NIR) spectroscopy

ABSTRACT

This work verifies the potential of support vector machine (SVM) algorithm applied to near infrared (NIR) spectroscopy data to develop multivariate calibration models for determination of biodiesel content in diesel fuel blends that are more effective and appropriate for analytical determinations of this type of fuel nowadays, providing the usual extended analytical range with required accuracy. Considering the difficulty to develop suitable models for this type of determination in an extended analytical range and that, in practice, biodiesel/diesel fuel blends are nowadays most often used between 0 and 30% (v/v) of biodiesel content, a calibration model is suggested for the range 0–35% (v/v) of biodiesel in diesel blends. The possibility of using a calibration model for the range 0–100% (v/v) of biodiesel in diesel fuel blends was also investigated and the difficulty in obtaining adequate results for this full analytical range is discussed. The SVM models are compared with those obtained with PLS models. The best result was obtained by the SVM model using the spectral region 4400–4600 cm^{-1} providing the RMSEP value of 0.11% in 0–35% biodiesel content calibration model. This model provides the determination of biodiesel content in agreement with the accuracy required by ABNT NBR and ASTM reference methods and without interference due to the presence of vegetable oil in the mixture. The best SVM model fit performance for the relationship studied is also verified by providing similar prediction results with the use of 4400–6200 cm^{-1} spectral range while the PLS results are much worse over this spectral region.

1. Introduction

The use of renewable energy resources for production of fuels such as ethanol and biodiesel, the so-called alternative fuels or biofuels obtained from biomass, has been increasing in Brazil and in many countries in recent years due to attractive environmental [1–4], economic and social [5–7] issues.

Economic issues such as reducing importations of diesel oil, social issues such as encouraging family farms or small producers and the growing attention to the environmental impact caused by the use of fossil energy resources and the effort to decrease atmospheric pollutant gases led the Brazilian government, in 2005, to introduce biodiesel as an energy resource with mandatory use of 2% (v/v) in diesel fuel in 2008, increasing the mandatory use to 5% (v/v) in 2010.

The environmentally friendly characteristic of biodiesel is mainly related to decrease of pollutant gases put into the atmosphere of large cities and metropolitan areas. Due to the raw material origin of soybean biodiesel, the carbon dioxide issued during its production and consumption is recycled in the growing process. In this manner, replacing petroleum diesel by soybean biodiesel decreases the total amount of this greenhouse gas put into atmosphere by means of combustion emissions. Moreover the use of biodiesel can decrease carbon oxide, sulfur oxide and particulate matter emissions into the atmosphere [1–4].

In 2010 Brazil produced 2.4 billion liters of biodiesel, becoming the second largest producer in the world. In 2011 Brazil produced 2.6 billion liters of biodiesel of which 81% were produced using soybean oil. Studies show that in 2020 the consumption of diesel oil in Brazil will be around 70 billion liters per year [8] and now there is a government study to make mandatory the use of 10% of biodiesel content in diesel fuel blend by that year.

Biodiesel is currently the main substitute for petroleum diesel fuel due to its similarities in physico-chemical properties, allowing its use in diesel engines in blends up to 20% (v/v) in diesel oil (B20) without any engine modification. Biodiesel blends up to 30% (v/v) in diesel oil (B30) can also be used but may require some adjustments in the injection system, fuel filters and elastomer materials, depending on the engine manufacturer. Vehicles with adequately adapted engines can use pure biodiesel (B100) as fuel [4].

Many countries now use biodiesel/diesel fuel blends. In the United States the most common use of this fuel is with 20% (v/v)

* Corresponding author. Tel.: +55 11 9 9612 7438.
E-mail address: julio@iqm.unicamp.br (J.C.L. Alves).

0039-9140/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.11.033

or less biodiesel content [9] and the American Society for Testing and Materials (ASTM) standard D7467 [10] provides the reference specifications for blends containing 6–20% of biodiesel. In Europe the most common use of this blend is between 5 and 7% (v/v) of biodiesel although there are experiments that use 30% (v/v) of biodiesel in the diesel fuel used in some buses and in light vehicles [11]. In Brazil the experiences in the cities of São Paulo and Curitiba stand out. São Paulo has been using blends with up to 30% (v/v) of biodiesel in its bus fleet. A municipal ordinance from 2009 determines that the entire urban transport bus fleet will gradually be replaced and by 2018 will use renewable fuels instead of fossil fuels. In Curitiba buses that use pure biodiesel are part of the urban transport fleet.

The biodiesel is produced by a transesterification reaction of a triglyceride, an ester from vegetable oils or animal fats, with a short chain alcohol such as methanol or ethanol in the presence of a catalyst, yielding a mixture of fatty acid alkyl esters and glycerol [12,13].

The use of biodiesel does not imply changes in the distribution and storage structure related to petroleum diesel fuel, although operational care similar to the use of petroleum diesel fuel in storage and in operation and maintenance of engines must be carefully controlled due to some distinct biodiesel properties in relation to petroleum diesel fuel, such as greater hygroscopicity and solvency, lower oxidative stability and physico-chemical characteristics at low temperatures such as higher pour point and cold filter plugging point [4,14–16].

1.1. Biodiesel content determination in diesel fuel blends

Traditional methods for biodiesel content determination in diesel fuel blends use mid-infrared spectroscopy, through measurements of transmittance or attenuated total reflectance (ATR) and partial least squares (PLS) calibration models, as described by ASTM D 7371 [17] and the Associação Brasileira de Normas Técnicas (ABNT) NBR 15568 [18] reference methods. Due to the difficulty in obtaining the required accuracy the use of two or three models and narrow analytical ranges are recommended. The ASTM method suggests the development of models for analytical ranges of 0–10% (v/v), 10–30% (v/v) and 30–100% (v/v) biodiesel content in diesel fuel. The reproducibility is specified according to the biodiesel content in the sample and varies from 0.76 to 1.66% (v/v) for samples with 1% and 20% (v/v) of biodiesel, respectively. The ABNT NBR method suggests the development of models for analytical ranges of 0–8% (v/v) and 8–30% (v/v) biodiesel content in diesel fuel and these models must have root mean square errors of prediction (RMSEP) which cannot be greater than 0.1% (v/v) and 1% (v/v), respectively.

However, analytical methods based on near infrared (NIR) spectroscopy combined with chemometric methods have also been developed for analysis of petroleum products, such as lubricant oil [19,20], gasoline [21,22], diesel [23–25] and biodiesel/diesel fuel blends [26–34], providing efficient determinations. Concerning the analysis of biodiesel/diesel fuel blends some papers report quality parameter determinations [31,32], the identification of the vegetable oil in the biodiesel/diesel fuel blends [29,33,34] and the quantification of biodiesel in mixtures with petroleum diesel using linear chemometric methods [27–30] such as PLS [35] or nonlinear chemometric methods [26] such as artificial neural networks (ANN) [36]. These studies report biodiesel determination in narrow analytical ranges of 0–5% [27] and 0–10% [29], providing RMSEP values close to 0.1%. Other studies present biodiesel determination in comprehensive analytical ranges but with the exclusion of lower and upper extreme values of the full analytical range and report obtained RMSEP values of 0.6% for determination in the range of 5–50% [28] and up to 0.05% for determination in the range of 2–90% [30]. Only one study has reported results for biodiesel determination over the full analytical range. With the use of variable selection and PLS an RMSEP value of 0.06% was obtained [26] for determining the analytical range of 0–100%. These results demonstrate the importance of considering the analytical range, data pre-treatment and the chemometric method used when comparing the obtained accuracy by these methods. For convenience it is adequate to know some statistical evidence of a good model fit such as the absence of prediction bias.

For biodiesel/diesel fuel blend determinations good results are obtained with the use of linear or nonlinear methods, depending on the spectral range used, data pre-treatment and, mainly, the analytical range. Furthermore the performance of different methods may be associated with several factors involving the nonlinearity of the relationship in this type of determination, such as instrumental factors (nonlinearity of the detection system), or sample related factors, such as changes in hydrogen bonding patterns as the concentrations of the various species undergo relative concentration changes [37,38], for example by changing the raw material of biodiesel and/or the type of petroleum in the diesel production. Thus, the use of a chemometric method able to properly model linear and nonlinear relationships and with a high generalization performance can provide more efficient and effective models.

Support vector machines (SVM) [39] involve learning algorithms based on statistical learning theory [40] and have been introduced in chemometrics recently with success in applications using near infrared (NIR) spectroscopy data for regression problems with superior performance related to reference algorithms such as PLS [23,24,41,42]. One of the major features of SVM models is that they can operate in a kernel-induced feature space allowing nonlinear modeling and good generalization performance can be obtained even with relatively small data sets. These characteristics can provide a better performance for SVM in relation to linear regression algorithms like PLS.

This work is a study to evaluate if the performance of the SVM algorithm applied to near infrared (NIR) spectroscopy data for the development of calibration models for the determination of biodiesel content in diesel fuel blends is simpler and more effective for analytical purposes, avoiding the construction of two or three calibration models for determination over an extended analytical range and using minimal data pre-treatment.

Nowadays the use of biodiesel/diesel fuel blends occurs most often between 0% and 30% (v/v). Thus a calibration model for biodiesel content determination in the range 0–35% (v/v) in diesel fuel is suggested. A study of the use of only one calibration model for biodiesel content determination in the range 0–100% (v/v) in diesel fuel is also presented. The SVM model results are compared with the results obtained with PLS models.

1.2. Support vector machines

Support vector machines were initially developed to treat classification problems and then extended to treat regression problems. Support vector regression (SVR) [43–46] estimation seeks to estimate the function:

$$\mathrm{f}(\boldsymbol{x}) = (\boldsymbol{w}\,\boldsymbol{x}) + \mathrm{b} \tag{1}$$

based on data $(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_\mathrm{n},y_\mathrm{n})$, by minimizing the regularized risk functional:

$$\frac{1}{2}||\boldsymbol{w}||^2 + C\,R_{\mathrm{emp}} \tag{2}$$

where C is a constant determining the trade-off between minimizing the training error, or empirical risk R_{emp}, and the model complexity term $||\boldsymbol{w}||^2$.

The main insight of statistical learning theory is that, in order to obtain a small risk, both training error and model complexity need to be controlled. The minimization of Eq. (2) is equivalent to the following constrained optimization problem.

minimize:

$$\tau\left(\boldsymbol{w},\ \xi^{(*)},\varepsilon\right) = \frac{1}{2}||\boldsymbol{w}||^2 + C\left(\nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \tag{3}$$

subject to the following constraints:

$$(\boldsymbol{w}\,\boldsymbol{x}_i + b) - y_i \le \varepsilon + \xi_i, \tag{4}$$

$$y_i - (\boldsymbol{w}\,\boldsymbol{x}_i + b) \le \varepsilon + \xi_i^*, \tag{5}$$

$$\xi_i^{(*)} \ge 0, \quad \varepsilon \ge 0. \tag{6}$$

Each point $\boldsymbol{x}_i$ is associated to an error of magnitude ε. Errors above ε are captured by the slack variables $\xi^{(*)}$ and they are penalized in the objective function via the regularization parameter *C*, chosen *a priori*. Here the so-called ε-insensitive loss function [43] is used, which tends to limit the calibration errors.

In ν-SVR [45,46] the size of ε is not defined *a priori* but is a variable to be optimized via the adequate choice of parameter ν which has to be chosen in accordance with the noise that is in the *y*-values.

In order to solve nonlinear functions the following generalization can be done: input vectors $\boldsymbol{x}_i$, are mapped into a high-dimensional feature space Z through some nonlinear mapping, ϕ:$\boldsymbol{x}_i \rightarrow \boldsymbol{z}_i$, chosen *a priori*. The optimization problem is solved in the feature space Z. The calculation of the inner product in a high-dimensional space is performed by a suitable linear or nonlinear function k, leading to a regression function. The linear or nonlinear function k is called a kernel [23,47]. A suitable kernel function makes it possible to map the input space to a high-dimensional feature space where some nonlinear relationships can be performed as a linear problem.

In this work the adequate choice of the ν-SVR parameters, **ν** and *C* were optimized and the specific kernel parameters were kept constant.

2. Experimental

Near infrared spectra were obtained in the range of 4400–6200 cm^{-1} for the 81 samples in the analytical range 0–100% (v/v) of soybean biodiesel in ultra low sulfur diesel (ULSD) oil, both supplied by Petrobras Distribuidora S.A. from its Barueri, SP, Brazil facilities. This local distributor has in its facilities a very large diesel storage tank which continuously receives diesel oil contributions from different Petrobras refineries in São Paulo state, making a blend of different refinery productions which process different petroleum feedstocks. The same distributor has a biodiesel storage tank which receives contributions of different biodiesel manufacturers making a blend of different production batches. The main soybean biodiesel manufacturer is Camera S.A. from Ijuí, RS, Brazil but there are also contributions from BSBios S.A. from Passo Fundo, RS, Brazil. The soybean biodiesel has 97% (w/w) of fatty acid methyl ester (FAME). Thus each tank provides representative samples of the fuels used.

The transflectance spectra in the near infrared region were obtained by a Perkin–Elmer Spectrum 100 MID/NIR spectrometer with a halogen source and a deuterated triglycine sulfate (DTGS) detector. A Petri dish combined with an aluminum reflector with 0.5 mm path length was used as the transflectance cell. Each spectrum was obtained as an average of 32 scans with 4 cm^{-1} resolution.

For development of calibration models for 0–35% (v/v) of biodiesel in diesel fuel 41 calibration samples and 25 validation samples were used. For development of calibration models for 0–100% (v/v) of biodiesel in diesel fuel 50 calibration samples and 31 validation samples were used. The validation set was used only after model development for checking its prediction accuracy.

The samples were prepared by mixing a total volume of 20 ml placed in dark glass bottles of 100 ml. The calibration and validation sample sets are determined by means of experimental design as follows: in the calibration set the biodiesel content in each sample increases by 0.5% up to 5% and then increases by 1% up to 35%. For samples with biodiesel content above 40% the biodiesel content increases by 10% up to 100%. There are no replicates. For the validation sample set were prepared samples with biodiesel contents that are not in the calibration set. In this work all sample analyses were run on the same day and a week after sample preparation. Many Petri dishes and aluminum reflectors with identical characteristics were used and for cleaning purposes just water and detergent were used.

Different data preprocessings were carried out to verify which provides the best model using ν-SVR and PLS. The tested preprocessings were baseline correction and mean centering; standard normal variate (SNV); SNV and mean centering; first derivative and first derivative and mean centering. A blocked cross-validation of the calibration set was used for model development. The LIBSVM package [48] was employed in this study to develop ν-SVR models and the genetic algorithm (GA) [49] was applied for parametric optimizing. All routines were for Matlab version 7.7 [50] and computational work was run on a HP Pavilion dv6000 [51] equipment with Windows Vista System [52].

For the ν-SVR model development, different kernel functions [23,47], such as radial basis function (RBF), polynomial, sigmoid and linear, were tested and the data set was previously scaled between 0 and 1. The LIBSVM default value of γ parameter for the RBF kernel ($\gamma = 1/k$, where k means the number of variables in the calibration data set) was used and the polynomial degree in the polynomial kernel function was three. The parameters C and ν were selected by GA with parametric optimizing ranges from 0 to 10^4 and 10^{-4} to 1, respectively. For parametric optimization with GA, the following parameters were used: number of 20 individuals and maximum of 15 generations, since it was observed that with these settings the value of the cross-validation error was stabilized not being improved by increasing the number of generations. The fitness function assigned a fitness for each individual according to their rank in the population and the objective function to be optimized by GA was the minimum error obtained by cross-validation with three subsets of the training set. Each individual or chromosome has a string 30 bits long based on a binary code using 15 bits for each parameter or gene. As the minimizing error of cross-validation in the training set does not guarantee obtaining the best parameters, a manual grid search was further performed from the values previously selected by the GA.

3. Results and discussion

The spectral range of 4400–6200 cm^{-1} allows the calibration of biodiesel content in diesel fuel due to the occurrence of combination bands, first and second overtones of vibrational modes of C–H bond in methyl and methylene groups and C=C bond of unsaturated compounds [37]. There is also a combination band of C–H bond stretching and C=O bond stretching near 4650 cm^{-1} [37]. Moreover the difference in NIR spectra in the regions of 4425 cm^{-1} and 6005 cm^{-1}, probably related to stretching of terminal methyl groups [37], where methyl esters have peaks while triglycerides exhibit only shoulders, provides the biodiesel quantification in a selective manner, as demonstrated in an earlier study [53], without the interference of the possible presence of vegetable oil in the mixture. Fig. 1 illustrates the difference between the spectrum of petroleum diesel, soybean biodiesel and soybean oil using spectra with SNV preprocessing. The development of calibration models using the regions (i) 4400–6200 cm^{-1} and (ii) 4400–4600 cm^{-1} was tested. The spectra of the 81 samples used, with SNV preprocessing are shown in Fig. 2.

3.1. *Calibration models for 0–35% (v/v) of biodiesel in diesel fuel*

The best result obtained with PLS used spectral region (ii) and SNV as data preprocessing. Three latent variables were used that explain 99.9% of data variance. The results are shown in Table 1.

The best result obtained with ν-SVR used spectral region (ii), the linear kernel function and SNV preprocessed data. The results are also shown in Table 1. Using nonlinear kernel functions the results obtained were very poor, being worse than the PLS results. For the best ν-SVR model the selected parameters were $C=20$ and $\nu=0.007$ and in this model 12 support vectors were used. It was considered that the lower the number of support vectors used, the lower the possibility of model overfit. The ideal is that this number be at most two thirds of the calibration samples. In this model the appropriate number of support vectors used demonstrates the good fit of the model. Fig. 3 illustrates the prepared concentration values versus predicted concentration values obtained with the ν-SVR model.

It is possible to see the better fit of the ν-SVR model compared to the PLS model through the absolute residual distribution for the calibration and validation sample sets of models using PLS and ν-SVR, shown in Fig. 4(a) and (b), respectively. It appears that the ν-SVR model provides a better fit throughout the analytical range, with a better residual distribution and smaller residues, mainly for the validation sample set.

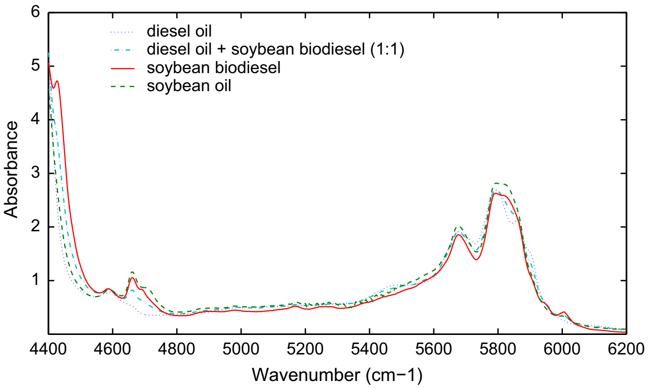

Fig. 1. Characteristic NIR spectra of diesel fuel, soybean biodiesel, biodiesel/diesel (1:1) blend and soybean oil.

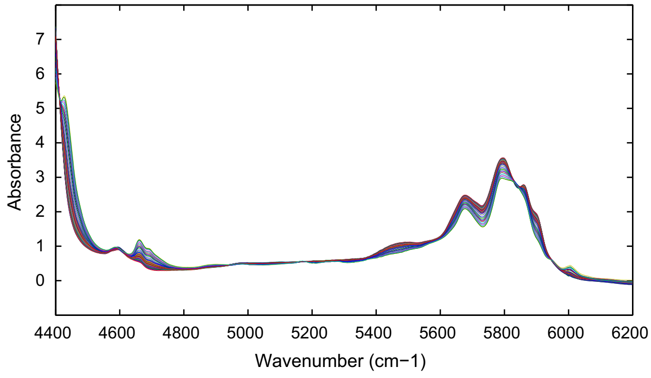

Fig. 2. NIR spectra of the 81 samples used.

The ν-SVR model obtained provides an RMSEP value which is approximately 13% better compared with the value obtained with the PLS model. The obtained RMSEP values for PLS and ν-SVR models match the value required by the ABNT NBR reference method for biodiesel determination from 0% (v/v) in diesel fuel and the RMSEP values are smaller than the minimum reproducibility required by the ASTM method for biodiesel determination.

Related to previously cited studies concerning the quantification of biodiesel in diesel fuel and papers that include this analytical range [27–30] a model was obtained that provides an RMSEP value quite similar to or better than the cited results, but with minimal data preprocessing and with the advantage of providing determinations for a larger analytical range and/or including the lower extreme values of the analytical range, which is suitable for today's practical needs. This model is also selective for biodiesel determination due to use of the spectral region without the presence of a band related to the stretching of the carbonyl group.

On the other hand with the use of spectral region (i) the best model obtained with PLS used first derivative preprocessed data and provides poorer results with higher root mean square error of calibration (RMSEC) and RMSEP. In this model two latent variables were used that explain 99.7% of data variance. Fig. 5 shows the absolute residual distribution for calibration and validation sample sets for PLS model using spectral region (i) and the results are shown in Table 1. The use of ν-SVR and spectral region (i) provides very good results, similar to those obtained by the use of spectral region (ii). In this model linear kernel function and SNV preprocessed data were used. The selected parameters $C=2$ and $\nu=0.005$ and 25 support vectors were used. The results are

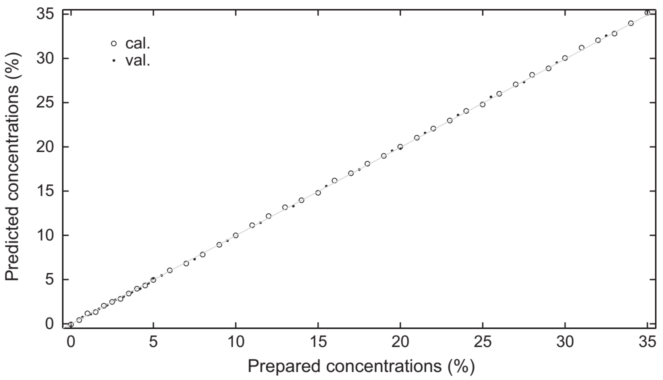

Fig. 3. Prepared concentration values versus predicted concentration values using the best ν-SVR model for 0–35% (v/v) of biodiesel in diesel fuel blends with the calibration (○) and validation (●) samples.

Table 1
ν-SVR and PLS model results for 0–35% (v/v) of biodiesel in diesel fuel blends.

Model	Spectral region (cm^{-1})	RMSEC (%) R^2	RMSEP (%)
PLS	(i) 4400–6200	0.58 0.997	0.76
	(ii) 4400–4600	0.12 0.999	0.13
ν-SVR	(i) 4400–6200	0.05 0.999	0.12
	(ii) 4400–4600	0.12 0.999	0.11

shown in Table 1 and the absolute residual distribution for the calibration and validation sets are shown in Fig. 5.

In order to get further insight into the accuracy of the developed methods, linear regression analyses of prepared concentration values versus PLS and ν-SVR predicted concentration values for the validation set were applied. The estimated intercept (b) and slope (a) were compared with their ideal values of 0 and 1, respectively, using the elliptical joint confidence region (EJCR) test, in this case by using an ordinary least squares fitting of the prepared concentration values versus predicted concentration values for each model.

The boundary of the ellipse is determined by the magnitude of experimental errors and by the degrees of confidence chosen, and is described by the following equation:

$$n(\mathrm{b}-\beta)^2+2\left(\sum y_{\mathrm{i}}\right)(\mathrm{b}-\beta)(\mathrm{a}-\alpha)+\left(\sum y_{\mathrm{i}}^2\right)(\mathrm{a}-\alpha)^2=2s^2F_{2,\mathrm{d}} \qquad (7)$$

where n is the number of data points, y_{i} are the prepared concentration values, s^2 the regression variance and $F_{2,\mathrm{d}}$ is the critical F value with 2 and $\mathrm{d}=n-2$ degrees of freedom at a given confidence level. In this work 95% confidence level was used.

The center of ellipse is (b,a) and any point (β,α) that lies inside the EJCR is compatible with the data at the chosen confidence level. In order to check constant (translational) or proportional (rotational) bias, the values $\beta=0$ and $\alpha=1$ are compared with the estimates b and a using EJCR. If the point (0,1) lies inside the EJCR, then biases are not present. This can be done from easy calculations [54–59].

Fig. 6 shows the EJCR for PLS and ν-SVR results for the best models for 0–35% of biodiesel content using region (ii) and Fig. 7 shows the EJCR for PLS and ν-SVR results for the best models for 0–35% of biodiesel content using region (i). There are no significant differences between prepared concentration values and predicted concentration values by PLS and ν-SVR models in the validation set and there is no evidence of bias with the 95% confidence level. For region (i) the ellipse for the ν-SVR results presents a smaller size, showing that the ν-SVR results are in better agreement than the PLS results.

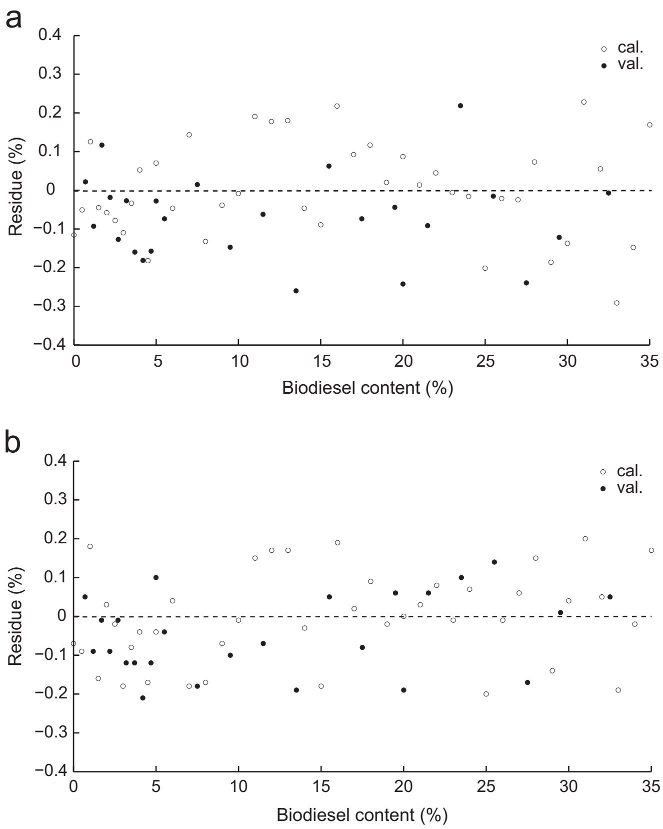

Fig. 4. Absolute residual distribution of predictions for PLS (a) and ν-SVR (b) models using the spectral region (ii) for the calibration (○) and validation (●) samples.

The better performance of linear kernel function in ν-SVR models shows that the nonlinear mapping of data input space in a higher dimensional feature space does not provide the best results in this case. However the optimization of two ν-SVR parameters and the use of ε-insensitive loss function, limiting

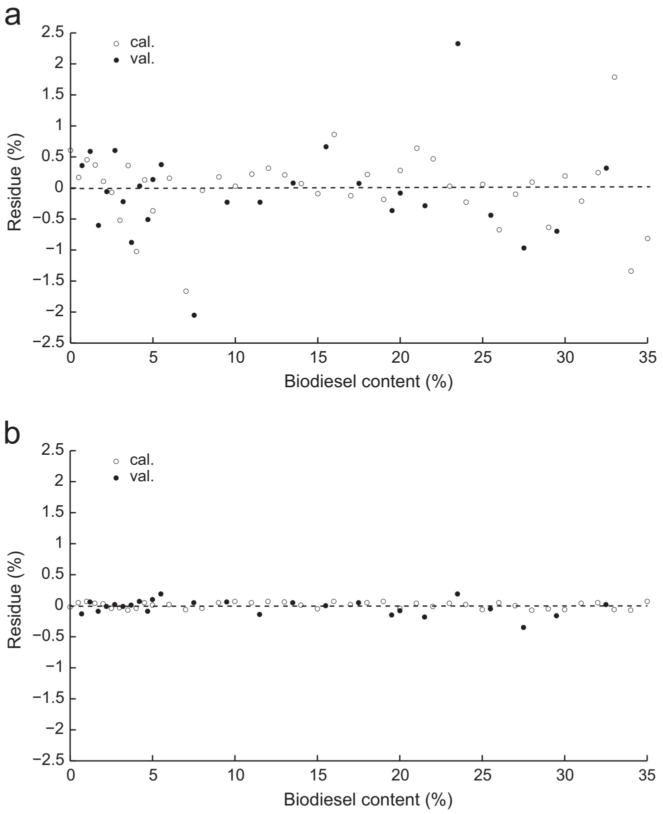

Fig. 5. Absolute residual distribution of predictions for PLS (a) and ν-SVR (b) models using the spectral region (i) for the calibration (○) and validation (●) samples.

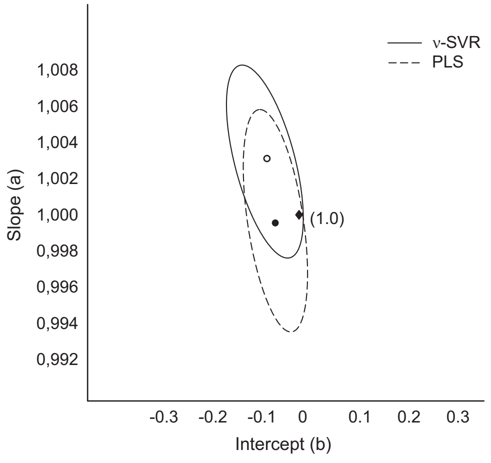

Fig. 6. Elliptical joint confidence regions for the intercept and slope corresponding to regressions of prepared concentration values versus PLS and ν-SVR model predicted concentration values for 0–35% (v/v) biodiesel in diesel fuel blends using spectral region (ii). The estimated (b,a) for PLS (●) and ν-SVR (○) models.

calibration errors, provided the better model fit in relation to PLS. When using a large number of variables (♯ variables ≫ ♯ samples) it can be expected that nonlinear mapping cannot improve the model performance and that the linear kernel provides similar results related to RBF or polynomial kernel functions [60] but it is interesting to see that even with the use of a reduced number of variables (spectral region (ii)=200 variables) instead of a larger number of variables (spectral region (i)=1800 variables) the RBF, polynomial or sigmoid kernel functions do not provide better performance in relation to the linear kernel function, suggesting that nonlinear mapping is not necessary. The good results provided by the linear kernel function are not surprising, since studies presented and discussed in a recent paper [42] have obtained better results using ν-SVR and the linear kernel function compared to PLS results, although the RBF and polynomial kernel provided better results than the linear kernel due to the relationship particularities of the studied problem that suggested some degree of nonlinearity.

3.2. Calibration models for 0–100% (v/v) of biodiesel in diesel fuel

The best result obtained with PLS used spectral region (ii) and SNV as data preprocessing. Three latent variables were used that explain 99.9% of data variance. The results are shown in Table 2. The best result with ν-SVR used spectral region (ii), the linear kernel function and SNV preprocessed data. The results are also shown in Table 2. The selected parameters were $C=2$ and $\nu=0.1463$ and in this model 14 support vectors were used.

The ν-SVR model provides an RMSEP value which is approximately 10% better compared with the RMSEP value obtained with the PLS model. The RMSEP values obtained for both the ν-SVR and PLS models are smaller than required by the ABNT NBR reference method for biodiesel determination in the analytical range of 8–30% (v/v) and are also smaller than the minimum reproducibility required by the ASTM reference method. However there is a nonconstant variance of the residual values of the validation set throughout the analytical range and a tendency to negative errors in the predicted values of the validation sample set for lower biodiesel contents with both models. The use of spectral region (i) cannot solve the heteroscedasticity problem of the validation sample set residuals and did not provide better models.

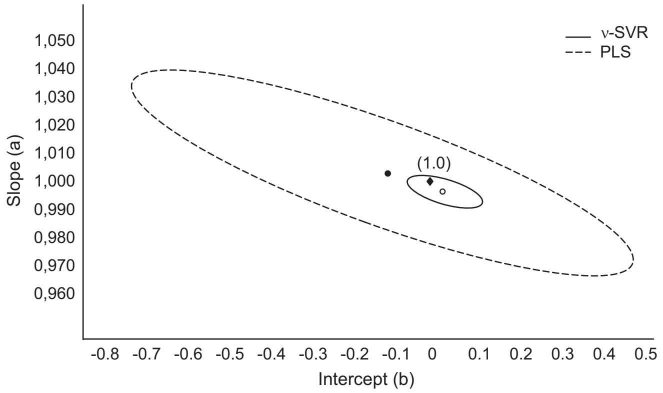

Fig. 7. Elliptical joint confidence regions for the intercept and slope corresponding to regressions of prepared concentration values versus PLS and ν-SVR model predicted concentration values for 0–35% (v/v) biodiesel in diesel fuel blends using spectral region (i). The estimated (b,a) for PLS (●) and ν-SVR (○) models.

Table 2
ν-SVR and PLS model results for 0–100% (v/v) of biodiesel in diesel fuel blends.

Model	Spectral region (cm^{-1})	RMSEC (%) R^2	RMSEP (%)
PLS	(i) 4400–6200	0.61 0.999	0.83
	(ii) 4400–4600	0.18 0.999	0.32
ν-SVR	(i) 4400–6200	0.05 0.999	0.30
	(ii) 4400–4600	0.17 0.999	0.28

Fig. 8 shows the EJCR for PLS and ν-SVR models for 0–100% content of biodiesel using spectral region (ii). The results shows that there are statistical differences between the prepared concentration values and predicted concentration values with both models in the validation sample set with 95% confidence level. The large distance between the theoretical point (0,1) and the boundary of the joint confidence region indicates that results from both the PLS and ν-SVR models have important bias.

4. Conclusion

The use of near infrared spectral region (ii) that includes the absorption band at 4425 cm^{-1} related to the vibrational mode of the terminal methyl group in fatty acid methyl esters provides similar prediction results with ν-SVR and PLS; however, the ν-SVR model, due to a better fit throughout the analytical range provides an improvement of 13% in RMSEP value for the 0–35% biodiesel/diesel fuel blend calibration model. The near infrared spectral region (i) provides very similar results related to those obtained with the spectral region (ii) in terms of RMSEP using the ν-SVR but, on the other hand, PLS does not provide a good model fit using spectral region (i), showing that ν-SVR gives better performance, adequately fitting the relationship. The absence of bias in the validation sample set prediction results was demonstrated by the EJCR test.

The RMSEP value obtained is suitable for this type of determination, as compared to the ABNT NBR reference method, and is smaller than the reproducibility of the ASTM reference method. Furthermore the method developed is suitable for determining biodiesel/diesel fuel blends in the extended analytical ranges of biodiesel content used nowadays, requires minimal data preprocessing and is selective for biodiesel, not being affected by a possible mixture with vegetable oil.

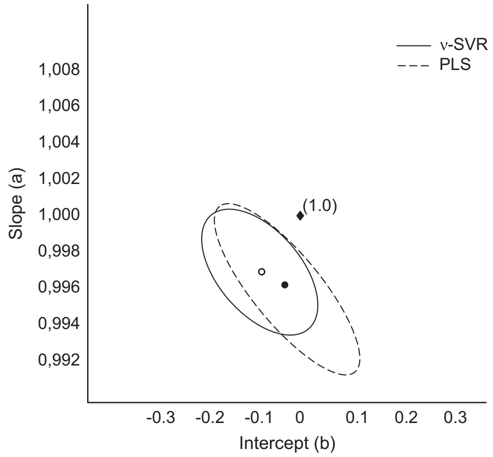

Fig. 8. Elliptical joint confidence regions for the intercept and slope corresponding to regressions of prepared concentration values versus PLS and ν-SVR model predicted concentration values for 0–100% (v/v) biodiesel in diesel fuel blends using the spectral region (ii). The estimated (b,a) for PLS (●) and ν-SVR (○) models.

The development of a model for the determination of biodiesel content in the full analytical range of 0–100% in diesel fuel blends proved to be quite difficult using either linear or nonlinear models. Although it is possible to obtain good RMSEP values a bias was found by the EJCR test that suggests further studies are needed to improve the performance of this type of model.

Acknowledgments

The authors would like to thank Petrobras Distribuidora S.A. for providing the diesel and biodiesel samples for development of this work and CNPq, CAPES and FAPESP for financial support.

References

[1] A Comprehensive Analysis Of Biodiesel Impacts On Exhaust Emissions, Draft technical report, US Environmental Protection Agency – EPA, 2002.
[2] X. Shi, Y. Yu, H. He, S. Shuai, J. Wang, R. Li, Fuel 84 (2005) 1543.
[3] X. Pang, X. Shi, Y. Mu, H. He, S. Shuai, H. Chen., R. Li, Atmos. Environ. 40 (2006) 7057.
[4] M.A. Fazal, A.S.M.A. Haseeb, H.H. Masjuki, Renew. Sustain. Energy Rev. 15 (2011) 1314.
[5] D.F. Amaral, Desmistificando O Programa Nacional de Produção E Uso Do Biodiesel. A Visão da Indústria Brasileira De Óleos Vegetais, ABIOVE, São Paulo, available at 〈http://www.abiove.com.br/palestras/ abiove_relatorio_biodiesel_ago09_br.pdf〉, 2009 (accessed March 10, 2012).
[6] A.F. Porte, R.C.S. Schneider, J.A. Kaercher, R.A. Klamt, W.L. Schmatz, W.L.T. Silva, W.A. Severo Filho, Fuel 89 (2010) 3718.
[7] O que é o Programa Nacional de Produção e Uso do Biodiesel (PNPB)?, Ministério do Desenvolvimento Agrário, Brasilia, available at: 〈http://www.mda.gov.br/portal/saf/programas/biodiesel〉, (accessed March 10, 2012).
[8] Plano decenal de expansão de energia 2020, Ministério de Minas e Energia. Empresa de Pesquisa Energética, Brasilia, 2011, available at: 〈http://www.epe.gov.br〉, (accessed March 15, 2012).
[9] T.L. Alleman, L. Fouts, R.L. McCormick, Fuel Process. Technol. 92 (2011) 1297.
[10] ASTM Standard D7467, 2009a, Standard specification for diesel fuel oil, biodiesel blend (B6 to B20), ASTM International, West Conshohocken, PA, 2009, DOI: 10.1520/D7467-09a, 〈www.astm.org〉.
[11] A. Macor, F. Avella, D. Faedo, Appl. Energy 88 (2011) 4989.
[12] J.M. Marchetti, Process Saf. Environ. Prot. 90 (2012) 157.
[13] S.A. Basha, K.R. Gopal, S. Jebaraj, Renew. Sustain. Energy Rev. 13 (2009) 1628.
[14] G. Knothe, Fuel Process. Technol. 86 (2005) 1059.
[15] H. Tang, S.O. Salley, K.Y. Simon Ng, Fuel 87 (2008) 3006.
[16] C. Boshui, S. Yuqiu, F. Jianhua, W. Jiu, W. Jiang, Biomass and Bioenergy 34 (2010) 1309.
[17] ASTM Standard D7371, 2007, Standard test method for determination of biodiesel (fatty acid methyl esters) content in diesel fuel oil using mid infrared spectroscopy (FTIR-ATR-PLS method), ASTM International, West Conshohocken, PA, 2007, DOI: 10.1520/D7371-07, 〈www.astm.org〉.
[18] Norma Brasileira ABNT NBR. 15568, Biodiesel – Determinação Do Teor De Biodiesel Em Óleo Diesel Por Espectroscopia Na Região Do Infravermelho Médio, Associação Brasileira De Normas Técnicas, Rio De Janeiro, 2008, www.abnt.org.br.
[19] F.S.G. Lima, M.A.S. Araujo, L.E.P. Borges, J. Near Infrared Spectrosc. 12 (2004) 159.
[20] R.M. Balabin, R.Z. Safieva, Fuel 87 (2008) 2745.
[21] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Chemom. Intell. Lab. Syst. 88 (2007) 183.
[22] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Anal. Chim. Acta 671 (2010) 27.
[23] H. Li, Y. Liang, Q. Xu, Chemom. Intell. Lab. Syst. 95 (2009) 188.
[24] J.C.L. Alves, C.B. Henriques, R.J. Poppi, Fuel 97 (2012) 710.
[25] F.B. Gonzaga, C. Pasquini, Anal. Chim. Acta 670 (2010) 92.
[26] J.S. Oliveira, R. Montalvão, L. Daher, P.A.Z. Suarez, J.C. Rubim, Talanta 69 (2006) 1278.
[27] M.F. Pimentel, G.M.G.S. Ribeiro, R.S. da Cruz, L. Stragevitch, J.G.A. Pacheco Filho, L.S.G. Teixeira, Microchem. J. 82 (2006) 201.
[28] D.D.S. Fernandes, A.A. Gomes, G.B. Costa, G.W.B. Silva, G. Véras, Talanta 87 (2011) 30.
[29] 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.
[30] W.F.C. Rocha, R. Nogueira, B.G. Vaz, J. Chemom. 26 (2012) 456.
[31] W. Zhang, W. Yuan, X. Zhang, M. Coronado, Appl. Energy 98 (2012) 122.
[32] L.F.B. Lira, F.V.C. Vasconcelos, C.F. Pereira, A.P.S. Paim, L. Stragevitch, M.F. Pimentel, Fuel 89 (2010) 405.
[33] V. Gaydou, J.K.N. Dupuy, Chemom. Intell. Lab. Syst. 106 (2011) 190.
[34] 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.
[35] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1.
[36] D. Svozil, Chemom. Intell. Lab. Syst. 39 (1997) 43.
[37] J. Workman Jr., L. Weyer, Practical Guide to Interpretive Near Infrared Spectroscopy, CRC Press, Boca Raton, 2008.
[38] C. Pasquini, J. Braz. Chem. Soc. 14 (2003) 198.
[39] C. Cortes, V. Vapnik, Mach. Learn. 20 (1995) 273.
[40] B. Scholkopf, A.J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.
[41] U. Thissen, M. Pepers, B. Ustun, W.J. Melssen, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 73 (2004) 169.
[42] J.C.L. Alves, R.J. Poppi, J. Near Infrared Spectrosc. 20 (2012) 419.
[43] A.J. Smola, B. Scholkopf, Stat. Comput. 14 (2004) 199.
[44] N. Chen, W. Lu, J. Yang, G. Li, Support Vector Machines in Chemistry, Word Scientific Publishing, Singapore, 2004.
[45] B. Scholkopf, A.J. Smola, R.C. Williamson, P.L. Bartlett, Neural Comput. 12 (2000) 1207.
[46] A. Chalimourda, B. Scholkopf, A.J. Smola, Neural Netw. 17 (2004) 127.
[47] A.J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans, Introduction to large margin classifiers, in: A.J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, 2000.
[48] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, 2001, 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉, (accessed April 20, 2009).
[49] R. Wehrens, L.M.C. Buydens, Trends Anal. Chem. 17 (1998) 193.
[50] Anon. The Mathworks Inc., Natick, MA, USA, 〈http://www.mathworks.com〉.
[51] Anon. Hewlett-Packard Company, Palo Alto, CA, USA, 〈http://www.hp.com〉.
[52] Anon. Microsoft Corporation, Mountain View, C.A., USA, 〈http://www.microsoft.com〉.
[53] G. Knothe, J. Am. Oil Chem. Soc. 78 (2001) 1025.
[54] J. Mandel, F.J. Linning, Anal. Chem. 29 (1957) 743.
[55] A.G. Gonzalez, M.A. Herrador, A.G. Asuero, Talanta 48 (1999) 729.
[56] A. Martinez., J. Riu, O. Busto, J. Guasch, F.X. Rius, Anal. Chim. Acta 406 (2000) 257.
[57] P. Valderrama, J.W.B. Braga, R.J. Poppi, J. Agric. Food Chem. 55 (2007) 8331.
[58] M. Grunhut., M.E. Centurion, W.D. Fragoso, L.F. Almeida, M.C.U. Araujo, B.S.F. Band, Talanta 75 (2008) 950.
[59] H.C. Goicoechea, A.C. Olivieri, Analyst 126 (2001) 1105.
[60] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification, National Taiwan University, 2009 (accessed April 20, 2009).