Fuel 261 (2020) 116344

Contents lists available at ScienceDirect

Fuel

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

Full Length Article

Multivariate regression models obtained from near-infrared spectroscopy data for prediction of the physical properties of biodiesel and its blends

Camilla L. Cunha[a], Alexandre R. Torres[b], Aderval S. Luna[a,*]

[a] Rio de Janeiro State University, Chemical Engineering Graduate Program, Rua São Francisco Xavier, 524, Maracanã, Rio de Janeiro, RJ 20550-013, Brazil
[b] Rio de Janeiro State University, Faculty of Technology, Rodovia Presidente Dutra Km 298, Pólo Industrial, Resende, Rio de Janeiro 27537-000, Brazil

ARTICLE INFO

Keywords:
Biodiesel
Near-infrared spectroscopy
A kinematic viscosity at 40 °C
Cold filter plugging point
Partial least squares
Support vector machine

ABSTRACT

Multivariate calibration based on Partial Least Squares (PLS), Random Forest (RF) and Support Vector Machine (SVM) methods combined with variable selections tools were used to model the relation between the near-infrared spectroscopy data of biodiesel fuel to its physical-chemical properties. The cold filter plugging point (CFPP) and a kinematic viscosity at 40 °C of the biodiesel samples and its blends were evaluated using spectroscopic data obtained with a near-infrared reflectance accessory (NIRA/NIR-FT-IR). Therefore, one hundred forty-nine blends were prepared using biodiesel from different sources, such as canola, corn, sunflower, and soybean. Furthermore, biodiesel samples purchased from the Brazil South Region were added to the study. One hundred samples were used for the calibration set, whereas the remaining samples were used as an external validation set. The results showed that the SVM model with baseline correction + mean centering preprocessing gave the best prediction for the CFPP, with a root-mean-square of error (RMSEP) equal to 0.9 °C. Among the models presented, the best result for predicting the kinematic viscosity at 40 °C was obtained by the PLS regression method using an interval selected by UVE with baseline correction + derivative preprocessing, with the RMSEP equal to 0.0133 $mm^2.s^{-1}$. The results in this work showed that the proposed methodologies were adequated in predicting the biodiesel fuel properties. The figure of merit Sum of Wilcoxon Test Probability (SWTP) presented in this study was necessary for the conclusion of the best model.

1. Introduction

The American Society for Testing and Materials defined biodiesel as a mono-alkyl ester derivative of long-chain fatty acids (ASTM D6751). Biodiesel is synthesized, generally from vegetable oils or animal fats, by the transesterification reaction. During this chemical process, vegetable oil or animal fat is reacted with short-chain alcohols (methanol or ethanol) to form methyl or ethyl esters (biodiesel) and glycerin (a valuable byproduct used in soaps and other products) [1]. It is worth mentioning that methanol is mainly employed to produce methyl esters in Brazil. The alcohol is used in excess to facilitate the extraction from glycerin to achieve higher yields in the transesterification reaction.

Biodiesel is an example of the use of biomass for energy production. This biofuel has advantages over diesel oil, as it is not toxic and is obtained from renewable sources with little sulfur contents, providing a better quality of emissions during the combustion process [2]. Even though biodiesel provides approximately 10% less energy than petroleum diesel, engine performance is practically the same, concerning power and torque, since biodiesel provides 10% (wt) oxygen [3]. Thus, it is essential to note that biodiesel can be used in any diesel cycle engine, with little or no need for adaptation. According to reports in the literature [4], an increase in the emissions of nitrogen oxides and carbonyl compounds is observed when using this biofuel. Another biodiesel disadvantage is that the cold flow properties may cause problems in starting the engines, limiting the use of biodiesel in cold climates.

For biodiesel production, the quality control of the final product is critical because the presence of certain impurities significantly affect the engine performance. For this reason, biodiesel must have some properties which must be monitored for the production, distribution, and use to be authorized. According to the Petroleum, Natural Gas and Biofuel Brazilian National Agency (ANP) standards [5], some physical-chemical properties that are used to ensure biodiesel quality are cold filter plugging point (CFPP) and a kinematic viscosity at 40 °C.

The prediction of cold filter plugging point is a valuable property to be studied, mainly in cold regions where reduced temperature causes the biodiesel to partially solidify or lose its the fluidity (reducing the fuel flow). It is leading to interruption of the flow of fuel and clogging of the filtration system, creating difficulties in engine starting [6,7].

* Corresponding author.
E-mail address: asluna@uerj.br (A.S. Luna).

https://doi.org/10.1016/j.fuel.2019.116344
Received 6 June 2019; Received in revised form 6 September 2019; Accepted 2 October 2019
Available online 17 October 2019
0016-2361/

The viscosity of biodiesel increases with the length of the carbon chain and the saturation degree [8] and influences the burning process in the combustion chamber of the engine. High viscosity promotes heterogeneity in the combustion of biodiesel, due to the reduction of the atomization efficiency in the combustion chamber, causing the deposition of waste inside the engine [9–11]. Thus, biodiesel viscosity being higher than that diesel viscosity implies a decrease in engine performance and an increase in exhaust emissions, especially NOx [3,12]. The kinematic viscosity of biodiesel is measured at 40 °C, according to ASTM D6751 (analytical method D445), which establishes the permitted viscosity range of 1.9 to 6.0 mm^2/s.

Some researchers worked on predicting biodiesel properties using spectroscopic methods (infrared, visible UV, Raman or other spectral data) combined with multivariate calibration tools (principal component regression, multiple linear regression, partial least squares, and others). For instance, Balabin & Safieva predicted the fractional composition, iodine value, and cold filter plugging point of biodiesel using chemometric tools, such as multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS) regression, artificial neural network (ANN) and correlated the properties with the spectra in the near-infrared region [13]. The authors obtained a prediction model with 11 latent variables and RMSEP equal to 1.64 °C for cold filter plugging point using PLS regression. The MLR and PCR models for prediction the CFPP exhibited RMSEP equal to 4.96 °C and 1.63 °C, respectively.

In another example, Inan, Al-Hajji & Koseoglu (2016) applied Fourier Transform mid-infrared (FT-MIR) spectroscopy in association PLS chemometrics to determine the properties of nine groups of middle distillates (diesel) boiling in the range 180–370 °C. Among the thirty-two properties studied by the authors were viscosity at 40 °C, distillation (0 to 100% w) and others [14]. The authors obtained a prediction model with six latent variables and RMSEP equal to 0.00215 cSt, for kinematic viscosity.

Cunha et al. (2017) developed models to determine density, refractive index and cold filter plugging point in biodiesel samples using the combination of mid-FT-IR spectroscopy coupled with partial least squares regression (PLS) and support vector machine regression (SVM). The authors reached good results for CFPP using the SVM regression method, in which the root-mean-square error of prediction (RMSEP) was equal to 0.6 °C [15].

Another author, like Santos Jr. *et al.* (2005), employed a dataset prepared from diesel samples. The authors studied the prediction of sulfur content, cetane index, density, viscosity, and distillation temperatures using FTNIR, FTIR-ATR, and FT-Raman spectral measurements. The models were built using PLS and artificial neural network (ANN) regression. For viscosity, the authors built prediction models with ANN/FTNIR and ANN/FTIR-ATR, with RMSEP equal to 0.16 and 0.23, respectively [16].

Balabin & Smirnov (2011) evaluated the use of FT-NIR for the prediction of properties of biodiesel fuel, including viscosity, density, methanol content, and water concentration. The authors performed MLR and PLS regression methods with 16 different variable selection tools, among which are regression by interval partial least squares (iPLS), successive projections algorithm (SPA), uninformative variable elimination (UVE) and genetic algorithm (GA). The results have shown that FT-NIR spectrometry, combined with multivariate calibration methods and variable selection tools, can be used to predict biodiesel properties. The models with iPLS, UVE-PLS, and GA-PLS exhibited RMSEP equal to 0.195, 0.152, and 0.153 $mm^2.s^{-1}$, it was possible to predict the viscosity, respectively [17].

Furthermore, Baptista *et al.* (2008) evaluated the use of near-infrared spectroscopy (NIRS) to predict the following biodiesel properties: iodine number, cold filter plugging point, kinematic viscosity at 40 °C and the density at 15 °C. The principal components analysis (PCA) was used to perform an exploratory data analysis and regression by PLS was used to predict the biodiesel properties of the mixtures of palm, soybean, rapeseed, and oils used for frying. The authors prepared seventy-one samples of biodiesel produced from mixtures of oils. The spectral region between 9000 and 4500 cm^{-1} was used to build a calibration model for the biodiesel CFPP, allowing a validation model with RMSEP equal to 1.1 °C. For the viscosity, the RMSEP for validation model obtained was 0.09 $mm^2.s^{-1}$ [18].

Typically, the tests are time-consuming. Therefore, it is necessary to develop a methodology in which more than one property can be predicted by a single and cheaper method of measurement. Infrared spectroscopy can provide these properties when combined with multivariate calibration techniques. Also, it is a versatile and nondestructive technique.

Based on these considerations, this research aimed to provide models for predicting physical-chemical properties (kinematic viscosity and cold filter plugging point) from samples of biodiesel blends based on data from NIR spectroscopy coupled with multivariate analysis techniques such as PCA, PLS (and its variations), Random Forest and SVM. Besides, this work proposes the application of a new approach to the Wilcoxon test for comparison between the obtained models.

2. Theoretical background

2.1. Preprocessing

The spectra may exhibit systematic variations which are not explained by the concentration of the analyte in question, such as random noise and baseline deviation, as reported [19,20]. Thus, the pretreatment of the spectrum can improve the interpretation, reduce the complexity of the model, and make it many times more robust and reliable against undesirable variations in new samples.

The baseline correction is a preprocessing step often used in the treatment of spectra that exhibit deviations from the baseline. This method subtracts from each variable of the spectrum a unique value (1st variable) or the average of a range of variables. The absence of this fix may cause damage to the model built since the presence of noise will hinder the spectral analysis. The spectra are usually adjusted using a baseline correction to eliminate negative values. That way, for each sample the value of the lowest point on the spectrum is subtracted from all variables, compelling the bottom end of the set to be zero, and the others take positive sign [21,22]. Following baseline correction, a signal with a distinct peak is obtained, and when the process terminates, the number of primary variables is maintained.

Mean centering consists of calculating the average intensity for each wavelength and subtracting each intensity from the mean value. In addition to that, each variable will have zero means. In other words, the coordinates are moved to the center of the data, allowing differences in the relative intensities of the variables to be more accessible to observe. The objective of the average removal is eliminating the intensity value of each variable, revealing data fluctuations around the mean value [23].

The first derivative is the simplest form of derivative preprocessing, where each variable in a sample is subtracted from its nearest neighbor variable. This preprocessing cause significant changes in the processed information and can show a positive effect when it highlights an analytical signal of interest, or negative, by showing instrumental noise. The derivation also has the effect of offset or baseline alignment.

Standard normal variate (SNV) is a preprocessing step applied to every spectrum individually and used quite often in near-infrared to eliminate the scatter. The average and standard deviation of all the data points for that spectrum are calculated. The mean value is subtracted from the absorbance for every data point, and the result is divided by the standard deviation. The SNV improves prediction accuracy, but it does not simplify the model or reduce systematic interference [23].

2.2. Principal component analysis (PCA)

The exploratory analysis methods used for the initial assessment of the data determine the data behavior and what kind of information can be extracted from them [24]. Principal Component Analysis (PCA) is the best known statistical method and was chosen to carry out this analysis. This method can be applied to the variables of the matrix **X** that have a high degree of collinearity. As a consequence, redundant information and the small variability of noise can be removed [25].

The reduction in the dimensionality of the data is shown by the principal components (PC), which portrays the original set using a combination of variables describing the data trend. The main feature of this new set is the orthogonality, but it is easily reconstructed from the linear combination of the original variables (spectra) [26–28]. The principle of PCA is the approach of transforming the initial matrix **X** into a product of two smaller matrices: loadings (**P**) and scores (**T**), and an error matrix (**E**), as shown in Eq. (1).

$$\boldsymbol{X} = \boldsymbol{T}\bullet\boldsymbol{P}^{\boldsymbol{t}} + \boldsymbol{E} \tag{1}$$

Knowing that **X** is the matrix (m × n), **T** is the scores matrix (m × PC), $\mathbf{P}^{t}$ is the loadings matrix transposed (PC × n), and **E** is the errors matrix (m × n).

PCA is an important tool in statistics studies, but it is known that the classical PCA is sensitive to gross errors. According to Ma & Aybat (2018), the robust PCA (ROBPCA) can be used to remove the effect of sparse gross errors [29]. Thus, for a data matrix **X** (m × n), ROBPCA decomposes this matrix into two matrices, as shown in Eq. (2).

$$\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S} \tag{2}$$

where **L** is a low-rank matrix, and **S** is a sparse matrix. The robust PCA consider that **X** is a superposition of **L** and **S**. Like this, the gross errors will be captured by the sparse matrix **S**, allowing the low-rank matrix **L** can still approximate **X** well. In summary, ROBPCA provides a low-dimensional approximation which is robust to outliers.

2.3. Multivariate regression models

The tools in this work used to construct regression models can be divided into two types: linear and nonlinear methods. Also, it was used variable selection tools in an attempt to improve forecasting models.

2.3.1. Linear methods

The PLS is a regression method which provides models that relate the blocks of variables **X** and **Y**. Thus, the information of the spectral measures (**X**) and the concentrations or properties (**Y**) are simultaneously used in the calibration phase. This method is one of the most widely used and was first proposed by Herman Wold [30]. Regression through the PLS method can solve collinearity problems with satisfactory predictive ability. The information of the variables are compressed, making it more robust, and as a result, the models are more accessible to interpret, and the spectral noise can be kept out of the model, in the form of residuals [25]. In a simplified manner, the PLS model consists of regression between the scores of the matrices **X** and **Y**. It uses the **X** matrix (m × n) on the X-axis and the **Y** vector (n × 1) on the y-axis in building the model. This model is considered primarily as an external relation between the matrices **X** and **Y** individually and subsequently as an internal relationship that relates the two matrices (**X** and **Y**). In practical terms, the PLS assumes that there are errors in both matrices, which are of equal importance. The external relation for **X** can be expressed as the sum of the new matrices, originating from the decomposition of **X**, as shown in Equation (3).

$$\boldsymbol{X} = \boldsymbol{T}\bullet\boldsymbol{P}^{\boldsymbol{t}} + \boldsymbol{E}_{\boldsymbol{x}} \tag{3}$$

The Y external relation follows the same path, as presented in Eq. (4):

$$\boldsymbol{Y} = \boldsymbol{U}\bullet\boldsymbol{Q}^{\boldsymbol{t}} + \boldsymbol{E}_{\boldsymbol{y}} \tag{4}$$

where **X** and **Y** are the decomposed matrices, **T** and **U** are the scores matrices (m × LV), $\mathbf{P}^{T}$ and $\mathbf{Q}^{T}$ have transposed weight matrices (LV × n), and $\mathbf{E}_{X}$ and $\mathbf{E}_{Y}$ are the errors matrices of the matrices **X** and **Y**, respectively [21].

In this study, the Wold's R criterion was applied to define the appropriate number of latent variables to be included in the PLS model construction. According to the studies of Stone and Wold, the Wold's R criterion is based on cross-validation, where the data (X and Y) are divided into k blocks, and a latent variable model is constructed from (k-1) data blocks [31,32]. In this way, the excluded block is used for testing, and a single PRESS is calculated. This process is repeated by excluding each block at a time, and then the total PRESS is calculated for a latent variable by adding the individual PRESS values. For each of the latent variables, the total PRESS is calculated, and then a series of PRESS values are obtained [31,32].

Also, according to Krzanowski, the Wold's R criterion can be applied in the comparison [33]. The difference is that the adjusted Wold's R criterion uses 0.95 and 0.90 as thresholds, rather than adopting the unit as a threshold as in the Wold's R criterion, given the variability of sampling. The Wold's R criterion suggests that an additional latent variable will not be included in the PLS model unless it offers significantly better predictions [33].

2.3.2. Variable selection

The use of variable selection techniques allows the construction of a robust and easily interpreted model since the choice of specific regions can minimize the errors of the multivariate model.

Variable selection methods can be based on chemical information or algorithms. In this work, methods based on chemical information take into account the identification of functional groups characteristic of biodiesel in near-infrared spectral data. In addition to the full spectrum (10000–4000 cm^{-1}), the spectral regions with chemical information about biodiesel will be considered for the construction of the models.

On the other hand, the methods based on algorithms used in this paper were ensemble PLS (enPLS), sparse PLS (sPLS), interval partial least squares (iPLS), variable importance in projection (VIP), genetic algorithm (GA) and uninformative variable elimination (UVE).

The application of ensemble methods can be used in multivariate regression and classification studies, mainly because those can improve the accuracy of unstable predictors. According to some authors, the enPLS method seeks to develop a robust algorithm to build models based on ensemble learning and statistical distribution concepts. The statistical distribution provides some information about variables behavior. Monte Carlo or bootstrap sampling methods are the most used methods due to its asymptotic properties [34–37].

Sparse PLS is a technique where PLS models are obtained from the full set of predictor variables adding penalties in the objective function to decrease the number of variables kept in the PLS model. Some authors impose the sparsity using a penalty in the procedure of dimension reduction [38–40].

The iPLS method is an extension of PLS, in which the algorithm defragments the spectral data set in a given number of intervals of the same width, constructing a PLS model for each interval, and exposing the results in a graph to simplify the comparison with the model constructed for the entire spectral profile. Thus, the method can offer a complete view of the data, which makes it very useful in the interpretation of which signals of the spectrum are more relevant and should be included in the calibration model [41,42].

The variable importance in projection (VIP) is a method that accumulates the importance of each variable in the model, which is projected by the vector of weights (w) of each latent variable [43,44]. The elimination of the variable occurs when it is below a cut-off threshold.

The genetic algorithm (GA) applies a sequence iteratively until it

reaches the subsets that provide the lowest RMSECV (root mean square error of cross-validation) value. According to Wise et al. (2003), for a given matrix **X** (independent variables) and a vector or matrix Y (dependent variables), the algorithm determines a subset of variables in **X**, using a multivariate regression method and a cross-validation strategy, the RMSECV value for the subset of variables is determined [45].

For uninformative variable elimination (UVE), artificial noise variables are added to the prediction set before the application of the regression model [46]. After that, all original variables with less "importance" than the noise variables are eliminated before the procedure is repeated until a stop criterion is reached.

2.3.3. Nonlinear methods

Random forest (RF) algorithm was developed by Breiman [47], and since then it has been widely applied in several scientific fields, such as environmental science [48], food [49], engineering [50] and medicine [51–54]. The random forest can be defined as a supervised learning algorithm which combines several decision trees such that to build a forest, improving the accuracy, obtaining a stable prediction and straightforward interpretation. In short, the method adds randomness to the model and search the most critical feature while splitting a node, and it usually results in a better model [47,55,56].

The support vector machine (SVM) was developed by Vapnik, and it is a method based on statistical learning theory and the concept of structural risk minimization. Initially, it was applied to solve pattern recognition problem, and since then has been used to the case of nonlinear regression, time series prediction, classification, and regression problems [57–60]. This method has the advantage of being universal approximator of any multivariate function to any required degree of accuracy. In short, to understand the theory of SVM, it may consider a regression problem containing dataset training and validation data. The SVM regression works similarly to the SVM classification when constructing the algorithm from the calibration data set, which can provide results for new information inputs, as far as the algorithm learn to group the data according to similarities found among them, in order to minimize error.

For SVM models, the full spectrum and the intervals selected by variable selection techniques were used, based on chemical information and algorithms (sPLS, iPLS, VIP, GA, and UVE).

2.4. Figures of merit

Some figures of merit were employed in this work, which is critical for the evaluation and comparison of models such as RMSEC, RMSECV, and RMSEP. The figures of merit presented have been used to evaluate which model showed more significant predictive capability.

The root mean square error of calibration (RMSEC) is a parameter that provides an overall view of the ability of the model since it is a measure of the average difference between the predicted value and the actual value. The RMSEC values were estimated according to Eq. (5):

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{I_c} (y_i - \hat{y}_i)^2}{I_c}} \quad (5)$$

Given that $\hat{y}_i$is the estimated value of sample i, it is not included in the built model; $y_i$is *i-th* the y values and i_c is the number of calibration samples.

Another parameter was considered in this work; the root mean square error of cross-validation (RMSECV) given by Eq. (6). This parameter was obtained from estimated values by 10-fold cross-validation and the known values, in conformity with literature reports [61,62].

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{s} (y_i - \hat{y}_i)^2}{s}} \quad (6)$$

Given that $\hat{y}_i$ is the estimated value of sample i, that it is not included in the model building, s is the number of segments or partitions of the dataset, and y_i is *i-th* the y values. In this paper, s is the number of segments used in the interleaved cross-validation (segments equal to 10).

According to the detailed report by Mevik & Cederkvist, for the external validation stage, a set of independent variables is provided to the model with h components to evaluate their predictive power [63]. The model prediction was assessed in this work using root mean square error of prediction (RMSEP) as presented in Eq. (7).

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I_P} (\hat{y}_i - y_{i,measured})^2}{I_P}} \quad (7)$$

Knowing that $\hat{y}_i$ is the estimated value by the predicted model for sample i, I_p is the number of sample in the prediction data set, and $y_{i,measured}$ is obtained by the measured value. Soon after the model has been built and validated, it can be applied for predictions of new samples.

2.5. The elliptical joint confidence region

The elliptical joint confidence region (EJCR) is applied to detect the occurrence of systematic errors present in the calibration and validation models. The elliptical joint confidence region (with 95%) was used to compare the current values and the predicted values concerning intercept and slope obtained from the regression. A two-dimensional plot of the elliptic region is constructed and checks to see if it contains the point (1,0). If the point (1,0) is not contained within the ellipse, it indicates that the method is not accurate and presents systematic errors [64,65].

2.6. Wilcoxon test

The Wilcoxon test was used to compare and choose the best model for each property. It is a nonparametric test that compares two paired groups (samples or models), calculates the difference between each set of pairs and analyzes these differences. This test assumes that there is information in the signs and the magnitudes of the differences in the values of each pair [66].

In this work, the Wilcoxon test was applied in a certain way like a new approach. The Wilcoxon test results on a probability value when applied to compare only two models. Residual values of each prediction model were used as input data for Wilcoxon test probability calculation (WTP). In case that i model was better than j model, the $\mathbf{WTP}_{i,j}$ will be a small figure. Then, it is possible to construct a $N \times N$ matrix applying the Wilcoxon test to a set of N models, where $i = 1,\ldots, N$ and $j = 1, \ldots, N$. Considering some line of this matrix, it contains the results of comparing i model with all models on the set. Summation, all the probabilities in the i-th matrix line, can be obtained a value that summarizes i model behavior. This summation will be expressed in Eq. (8).

$$SWTP_i = \sum_{j=1}^{N} WTP_{i,j} \quad (8)$$

The minimum value of this vector will reveal the best model on the set of studied models.

3. Material and methods

3.1. Biodiesel production

Biodiesel samples were synthesized from different refined oils such as soybean, canola, corn, and sunflower. The biodiesel was produced by the transesterification reaction of the refined oils mentioned above with methanol, using a volumetric ratio of 1:3 alcohol (MeOH)/vegetable oil. Lately, it is soon after placed in a constant-temperature bath and heated to 45°C. The catalyst (1.0% m/m of potassium hydroxide (KOH),

by mass of vegetable oil) was dissolved in the methanol under stirring, and then the solution was added to the reactor. The duration of the reaction was 30 min after the addition of the solution, with agitation set at a constant speed throughout the experiment (280 rpm). Later, following the reaction, the mixture was transferred to a separatory funnel, where the glycerol phase was removed by gravity. Transesterification reaction parameters such as % catalyst concentration, the volumetric ratio of alcohol (MeOH)/vegetable oil, agitation, temperature and reaction time can be compared with the parameters used in some literature articles [12,67–71].

After this, the biodiesel was neutralized by an HCl solution (0.5% v/v) followed by phase separation, and then it was washed with a saturated KCl solution to avoid the formation of an emulsion. Later, the biodiesel was washed with distilled water. After that, the fatty acid methyl esters were dried with the addition of anhydrous sodium sulfate (Na_2SO_4) to remove any trace of water present in the biodiesel. Last, the biodiesel was filtered using a separatory funnel with two layers of cotton. Additionally, some pure standard biodiesel samples were provided by the supplier, located in the Brazilian South Region.

3.2. Preparation of the blends

A total of 149 samples were prepared by mixing biodiesel from different sources to compose binary, ternary and quaternary blends, including samples of each pure biodiesel. The compositions of the binary blends proposed in this work were made from blending 10, 30, 50, 70 and 90% (v/v) of soybean biodiesel with biodiesel from canola, corn, sunflower, and Brazil South Region biodiesel. For the ternary and quaternary blends, the compositions are presented in Table 1. Fractions in biodiesel blends were determined to represent the different biodiesel compositions produced in different regions of Brazil, considering some of the different oilseeds available.

Near-infrared spectra were obtained for all samples and measurements of the kinematic viscosity at 40℃, and cold filter plugging point was carried out.

3.3. Kinematic viscosity measurements

The kinematic viscosity at 40 °C ($mm^2.s^{-1}$) test was performed in a petroleum products kinematic viscosity tester (Ray-265D, Raylabel Instrument Co.) according to ASTM D445. The instrument consists of a thermostatic bath and the capillary viscometers (Cannon-Fenske, 0.63 mm). The temperature of the thermostatic bath was maintained at 40 °C throughout the procedure.

Samples were previously filtered (0.22 µm pore size) and homogenized. The flow time in the capillary was measured using a stopwatch, considering the descent time (in seconds) of the sample between the two markings (lines) observed in the capillary. N-hexadecane was chosen as the standard substance for the calculation of viscosity of the samples. Since the viscosity at 40 °C for n-hexadecane is known in the literature (2.9457 $mm^2.s^{-1}$). Knowing that flow time is directly proportional to the viscosity and that each capillary tube has its conversion factor between time and viscosity, for all the capillaries used were carried out measures of time (s) with the chosen standard. The determination was repeated ten times, and the average was used as the reference value. For this test, the reproducibility and repeatability values calculated are 0.0279 $mm^2.s^{-1}$ and 0.0155 $mm^2.s^{-1}$, respectively. Moreover, the range of variation observed for this property of the dataset was 4.1449 to 4.7951 $mm^2.s^{-1}$ with measuring error equal to 0.0167 $mm^2.s^{-1}$

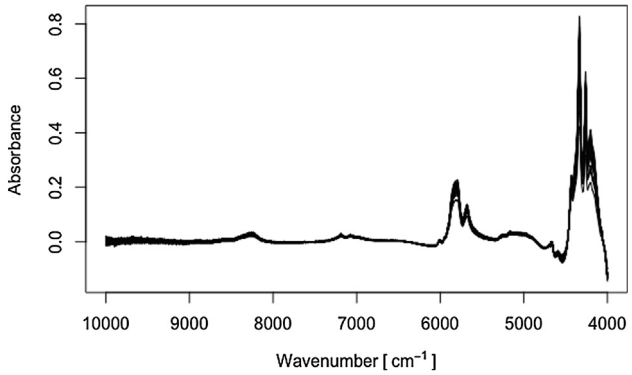

Fig. 1. Near-infrared spectra of pure biodiesel and blends with baseline correction.

3.4. Cold filter plugging point measurements

The CFPP is the temperature in °C at which 20 ml of the sample fails to pass through a standard metal filter in 60 s or less [72].

The CFPP test was performed in a Tanaka AFP-102 according to ASTM D-6371. The method is based on cooling at a rate of 1 °C/min and a sample volume of 45 ml, which is sucked by a pipette to a standardized metallic mesh filter under vacuum controlled. This procedure is repeated as often as possible until some crystals, which separate from the solution, is sufficient to stop or reduce the flow of the sample through the filter. Alternatively, this procedure is repeated if the time required to fill the pipette exceeds 60 s or to allow the sample to return entirely to the test vessel, before being cooled over 1 °C. Thus, the temperature at which the last filtration was started is called the cold filter plugging point. The determination was repeated four times, and the average was used as the reference value. Additionally, the

Table 1
The compositions for ternary and quaternary blends.

Sample	Oil	%v/v	Oil	%v/v	Oil	%v/v	Oil	%v/v
1	Soybean	50	Corn	40	Canola	10	–	–
2	Soybean	40	Corn	30	Canola	30	–	–
3	Soybean	50	Corn	10	Canola	40	–	–
4	Soybean	50	Corn	40	Sunflower	10	–	–
5	Soybean	40	Corn	30	Sunflower	30	–	–
6	Soybean	50	Corn	10	Sunflower	40	–	–
7	Soybean	50	Canola	40	Sunflower	10	–	–
8	Soybean	40	Canola	30	Sunflower	30	–	–
9	Soybean	50	Canola	10	Sunflower	40	–	–
10	Soybean	25	Corn	25	Canola	25	Sunflower	25
11	Soybean	40	Corn	40	Canola	10	Sunflower	10
12	Soybean	40	Corn	10	Canola	40	Sunflower	10
13	Soybean	40	Corn	10	Canola	10	Sunflower	40
14	Soybean	10	Corn	40	Canola	40	Sunflower	10
15	Soybean	10	Corn	40	Canola	10	Sunflower	40
16	Soybean	10	Corn	10	Canola	40	Sunflower	40

Table 2
Identification of the near-infrared spectral bands of the biodiesel samples [82].

Band	Wavenumber (cm^{-1})	Probable group	Class of substances
1	9000–8000	3νC−H	Alkanes
2	7500–6150	2νO−H	Carboxylic acids, alcohols
3	5750–5700	2νC−H	Alkanes
4	5350–4900	3νC=O	Carbonylated derivatives
5	4370–4260	The 1st region of C−H combination	Alkanes

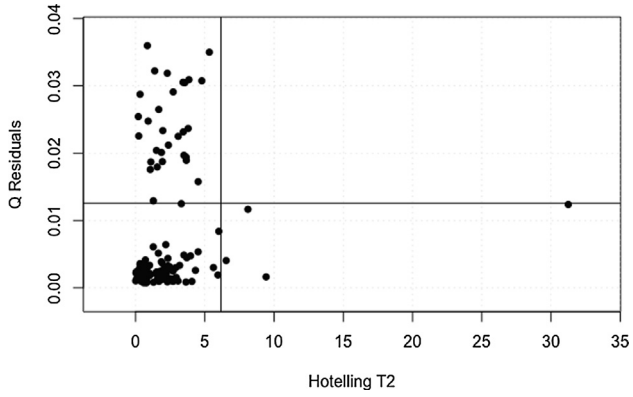

Fig. 2. Q residual *versus* Hotelling T^2 plot of PCA model using baseline correction + mean centering preprocessing.

Table 3
Comparison of the results achieved by the PCA using mean centering preprocessing combined with baseline correction.

PCs	Parameters		
	Eigenvalues	Captured Variance (%)	Accumulated Variance (%)
1	8.83 E−02	87.99	87.48
2	5.01 E−03	4.99	92.83
3	1.53 E−03	1.52	94.34
4	9.95 E−04	0.99	95.33
5	3.46 E−04	0.34	95.67
6	2.87 E−04	0.29	95.96
7	2.75 E−04	0.27	96.24
8	2.34 E−04	0.23	96.47
9	2.09 E−04	0.21	96.68
10	2.04 E−04	0.20	96.88

Legend: PCs = Number of Principal Components.

repeatability of the reference method was 1.8 °C, and the reproducibility was 2.0 °C. The range of variation observed for the CFPP in the data set under consideration was −7.0 to 6.0 °C with measuring error equal to 1.6 °C.

3.5. Near-infrared analysis

The NIR-FT-IR spectra were achieved using an Infrared Spectrometer with Fourier Transform (Perkin-Elmer, model Frontier, USA) in the range 10000–4000 cm^{-1} with a 4 cm^{-1} resolution and data interval of 1 cm^{-1}. The FTIR was equipped with a transflectance accessory for near-infrared (NIRA) with a zinc selenide crystal (ZnSe) from Perkin-Elmer. Each spectrum is an ensemble average of 20 scans. The sample holder used for measurements consists of a small Petri plate and a limiter of sample film thickness (0.224 mm). The cleaning of the sample holder was made by washing with water and detergent, followed by rinsing with distilled water and ethanol and finally drying in the laboratory oven.

3.6. Software applied for chemometric analysis

The PCA, PLS, Random Forest, SVM and variables selection calculations were performed with R Core Team [73], and the packages mdatools [74], rrcov [75], pls [76], plsVarSel [44], randomForest [77], enpls [78], sgPLS [79], e1071 [80], prospectr [81] and ellipse [82] were used in this work.

3.7. Calibration and validation sets

The Kennard-Stone (KS) algorithm begins by selecting the two samples with the greatest Euclidean distance between them in the sample set. For each remaining sample, the minimum distance between the samples already selected is calculated. Soon after, the sample with the highest minimum distance is maintained, and the procedure is repeated until some samples are selected [83].

KS algorithm was applied to choose the set of calibration and validation samples for the NIR spectral database. From the original dataset, the algorithm divided the total sample in calibration set, which is responsible for building the regression model, and validation set that aims to evaluate the predictive ability of the model [83]. The samples were divided into the two datasets according to this algorithm, 100 samples for the calibration set, and the remaining samples were used for the external validation set. At first, the calibration and validation NIR spectra data set have matrices of size, 100 × 6001 and 49 × 6001 (samples × variables), respectively, considering that the initial dataset had 149 samples with 6001 variables.

3.8. Near-infrared spectra preprocessing

For building the predictive models of the properties studied, cross-validation contiguous blocks (with ten splits) were chosen, and some preprocessing tested in the matrix of near-infrared spectra (X-block) such as baseline correction combined with mean centering, first derivative and standard normal variate (SNV).

3.9. Exploratory analysis

It is well established that the loadings are related to the signals (variables), and the scores are linked to the samples in the PCA model. In this context, PCA was used to identify possible outliers of biodiesel blends from different sources, such as canola, sunflower, corn, soybean, and from the south region. All variables studied were on the same scale, and the PCs were obtained from the covariance matrix. The PCA model was constructed using baseline combined with mean centering preprocessing in the near-infrared spectra matrix (X-Block). Eigenvalues were used as a criterion in the choice of the number of PCs for the PCA model.

Besides classic PCA, robust PCA was used were used in the identification and removal of samples with anomalous behavior, since this is more robust to outliers.

The methods employed for anomalous sample detection and elimination in this dataset were based on studies of leverages and residual analysis, following the recommendations of ASTM E 1655 [84].

Table 4
Summary of results obtained by the PLS, sPLS, iPLS, GA-PLS, VIP-PLS and UVE-PLS and enPLS models for both properties.

Tool	Intervals	PP	Kinematic Viscosity at 40 °C ($mm^2.s^{-1}$)				Cold Filter Plugging Point (°C)			
			LVs	RMSEP	RSEP	SWTP	LVs	RMSEP	RSEP	SWTP
PLS	full	mc	4	0.0157	0.067	13.07	5	1.0	2.23	29.13
	full	deriv	4	0.0166	0.129	31.82	6	1.4	11.62	46.68
	full	snv	4	0.0157	0.057	29.56	4	0.9	2.51	21.14
	1	mc	4	0.0185	0.359	50.59	6	2.4	12.56	66.11
	1	deriv	4	0.0248	0.394	67.91	6	3.3	13.84	71.15
	1	snv	4	0.0207	0.434	48.61	6	2.6	12.07	68.93
	2	mc	6	0.0180	0.047	50.28	6	2.2	8.65	65.65
	2	deriv	6	0.0246	0.431	67.28	4	2.9	4.61	70.36
	2	snv	5	0.0183	0.016	52.70	6	2.5	6.61	67.31
	3	mc	6	0.0216	0.142	62.56	3	1.5	12.21	49.26
	3	deriv	3	0.0216	0.083	60.31	3	1.7	10.11	56.49
	3	snv	4	0.0217	0.143	63.11	2	1.6	9.33	56.29
	4	mc	6	0.0162	0.224	37.20	6	1.5	0.27	43.31
	4	deriv	4	0.0193	0.302	57.50	4	2.1	2.80	62.95
	4	snv	4	0.0166	0.250	36.66	4	1.5	2.84	43.74
	5	mc	4	0.0144	0.191	22.39	5	1.6	2.87	55.41
	5	deriv	4	0.0144	0.036	32.16	6	2.0	7.67	64.34
	5	snv	4	0.0148	0.110	33.26	4	1.9	5.47	61.72
sPLS	spls	mc	4	0.0170	0.193	38.66	4	1.1	1.30	34.60
	spls	deriv	5	0.0148	0.009	14.81	4	1.1	2.06	26.62
	spls	snv	5	0.0157	0.011	26.26	5	1.0	1.92	34.53
iPLS	ipls	mc	2	0.0288	0.405	70.05	6	1.0	9.17	38.89
	ipls	deriv	6	0.0223	0.127	63.11	4	1.0	1.38	15.54
	ipls	snv	3	0.0144	0.004	22.08	5	1.2	5.35	30.34
GA-PLS	ga	mc	4	0.0162	0.062	21.47	5	0.9	2.59	12.78
	ga	deriv	5	0.0172	0.096	24.84	5	1.3	4.36	45.47
	ga	snv	5	0.0156	0.037	19.87	4	0.9	0.53	20.14
VIP-PLS	vip	mc	5	0.0160	0.118	27.30	5	0.9	5.91	13.61
	vip	deriv	5	0.0161	0.064	22.52	6	1.4	11.09	46.70
	vip	snv	5	0.0165	0.059	23.59	4	0.9	5.22	12.15
UVE-PLS	uve	mc	6	0.0178	0.209	25.59	6	1.6	7.59	54.36
	uve	deriv	4	0.0133	0.023	1.89	5	1.9	0.50	59.63
	uve	snv	5	0.0149	0.003	18.96	3	0.9	3.70	26.34
enPLS	enpls	mc	8	0.0142	0.003	2.68	8	1.2	1.34	15.26
	enpls	deriv	7	0.0157	0.085	12.38	7	1.5	7.75	37.97
	enpls	snv	10	0.0161	0.091	23.10	10	1.1	1.69	25.18

Legend: LVs = Latent Variables; PP = Preprocessing; RMSEP = Root Mean Square Error of Prediction; RSEP = Relative Standard Error of Prediction; SWTP = Sum of Wilcoxon Test Probability; mc = mean centering; deriv = first derivative; snv = standard normal variate.

3.10. Linear regression models

For PLS models, with or without variable selection, the data were preprocessed (baseline correction combined with mean centering, first derivative or standard normal variate), and the interleaved cross-validation (10 segments) method was used for external validation. The Wold's R criterion was applied in the choice the appropriate number of latent variables (LV) to be considered in the PLS model construction.

PLS models were built considering the selected intervals based on the spectral regions since they present information about functional groups characteristic of biodiesel.

In this paper, PLS models were also built using intervals selected by algorithms: ensemble PLS (enPLS), sparse PLS (sPLS), interval partial least squares (iPLS), variable importance in projection (VIP), genetic algorithm (GA) and uninformative variable elimination (UVE).

Sparse PLS and ensemble PLS are tools that have parameter tuning function, which allows finding the combination of parameters to be used in the construction of the models with lower error values.

The root mean square error of prediction (RMSEP) were used to evaluate the accuracy of the models. Besides that, the fit of the PLS models was evaluated by correlating the predicted *versus* measured values from the validation set.

Initially, PLS combined with variable selection tools was used to obtain prediction models. In an attempt to improve the prediction of CFPP and viscosity, which are nonlinear behavior properties, nonlinear models applying SVM and random forest, were constructed.

3.11. Non-linear regression models

Likewise, the same preprocessing combining used for PLS were applied to random forest and SVM models. For both tools, their tuning functions were used to identify the most suitable parameters for the construction of the models with lower error values.

As well as for PLS models, SVM models were also built considering the selected intervals based on chemical information and the intervals selected by the variable selection tools (sPLS, iPLS, GA, VIP, UVE).

4. Results and discussion

4.1. Interpretation of the NIR spectra

In this work, near-infrared spectra of pure biodiesel from five distinct sources and their blends (binary, ternary and quaternary) was obtained in the spectral region of 10000–4000 cm^{-1}, as shown in Fig. 1.

The spectral regions in Fig. 1 were identified as presented in Table 2, indicating the presence of axial deformation bands of C—H and O—H bonds, and carbonyl derivatives with axial deformation of the C=

Table 5
Summary of results obtained by nonlinear models, RF and SVM, considering full spectra, five regions and the intervals obtained by iPLS, sPLS, GA, VIP and UVE models for both properties.

Tool	Intervals	PP	Kinematic Viscosity at 40 °C				Cold Filter Plugging Point (°C)			
			nSVs	RMSEP	RSEP	SWTP	nSVs	RMSEP	RSEP	SWTP
RF	full	mc	–	0.0155	0.074	18.28	–	1.0	4.46	14.55
	full	deriv	–	0.0153	0.122	20.67	–	1.2	5.33	24.24
	full	snv	–	0.0150	0.110	13.44	–	1.0	3.07	24.68
SVM	full	mc	100	0.0152	0.010	24.47	100	0.9	3.54	0.98
	full	deriv	100	0.0153	0.069	14.21	100	1.2	6.54	24.80
	full	snv	100	0.0152	0.049	17.85	100	0.8	2.48	2.57
	1	mc	45	0.0186	0.298	51.97	84	2.0	14.98	54.00
	1	deriv	69	0.0261	0.537	70.69	81	3.0	18.40	58.09
	1	snv	51	0.0181	0.225	48.38	90	2.1	16.19	54.78
	2	mc	100	0.0183	0.056	52.06	91	2.0	10.83	59.70
	2	deriv	64	0.0254	0.492	68.83	100	2.7	5.47	68.15
	2	snv	87	0.0182	0.074	49.85	76	1.7	7.18	50.09
	3	mc	46	0.0185	0.080	47.99	38	1.2	3.64	45.67
	3	deriv	84	0.0223	0.046	61.33	89	1.2	2.44	29.68
	3	snv	38	0.0191	0.155	53.77	49	1.0	2.27	27.77
	4	mc	48	0.0200	0.027	56.73	100	1.2	2.33	17.30
	4	deriv	100	0.0199	0.179	62.66	100	2.0	5.80	56.77
	4	snv	100	0.0188	0.171	56.71	100	1.2	7.80	38.12
	5	mc	87	0.0147	0.099	11.38	100	1.3	0.54	44.31
	5	deriv	100	0.0163	0.054	47.86	100	1.3	4.03	35.21
	5	snv	67	0.0166	0.011	42.28	52	0.8	5.04	8.39
	ipls	mc	40	0.0256	0.532	69.18	78	1.0	1.516	22.24
	ipls	deriv	99	0.0200	0.085	62.08	99	1.1	6.147	19.94
	ipls	snv	100	0.0156	0.185	23.74	100	1.1	1.964	22.29
	ga	mc	99	0.0156	0.003	30.32	100	0.9	3.129	5.05
	ga	deriv	100	0.0162	0.102	31.93	100	1.2	5.844	34.66
	ga	snv	100	0.0153	0.056	9.55	100	0.9	3.977	2.37
	spls	mc	100	0.0170	0.065	40.63	100	0.9	3.608	10.47
	spls	deriv	100	0.0153	0.045	19.41	100	1.0	3.272	23.04
	spls	snv	100	0.0168	0.042	37.32	100	0.9	2.020	9.95
	vip	mc	72	0.0157	0.006	21.54	100	1.1	0.868	28.37
	vip	deriv	100	0.0152	0.021	7.30	100	0.9	4.917	13.41
	vip	snv	87	0.0148	0.008	7.52	100	0.8	2.414	12.56
	uve	mc	84	0.0170	0.081	45.76	100	1.1	0.594	35.43
	uve	deriv	99	0.0143	0.079	2.80	88	1.7	7.276	55.95
	uve	snv	100	0.0149	0.090	15.04	100	0.8	1.870	6.33

Legend: nSVs = number of Support Vectors; PP = Preprocessing; RMSEP = Root Mean Square Error of Prediction; RSEP = Relative Standard Error of Prediction; SWTP = Sum of Wilcoxon Test Probability; mc = mean centering; deriv = first derivative; snv = standard normal variate.

O bond.

4.2. Principal components analysis (PCA)

For the NIR spectra database (149 samples × 6001 variables), the study was performed using baseline correction combined with mean centering preprocessing. In the model, the validation step was using the cross-validation procedure with the selection of interleaved segments (with ten segments), which was adequate for the database size.

After construction of the model using the baseline correction (base) + mean centering (mc), three samples with atypical behavior were identified, samples of soybean biodiesel (Am058), ternary blend (Am113) and a quaternary blend (Am154). Those samples were removed because it showed a combination of high residual values and high leverage in the model. It was the criterion adopted for the removal of samples with atypical behavior, which was confirmed with the Robust PCA.

Later, a new model was generated after removal of these atypical samples. In this second model, a sample presenting atypical behavior was also identified as a sample of corn biodiesel (Am042). This sample was removed under the same criterion adopted in the removal of samples in the previous model.

The Robust method for Principal Components Analysis (ROBPCA) was applied to a conference tool to ensure that removed samples by PCA were correctly identified as outliers. In R, the rrcov package [75] was used to apply the ROBPCA.

It should be noted that when generating a new model without the atypical sample of the dataset, the new model can present atypical samples not initially detected in the previous model, this may continuously occur with each new model built. This behavior is explained by the standard ASTM E1655-05, which ranks this occurrence as the "snowballing effect." In this case, the norm recommends detecting and removing atypical samples only to the second model [84].

When removing these four samples, a new model PCA was generated. This model was built with two PCs and proved to be suitable for the description of the dataset. As presented in Q residuals vs. Hotelling T^2 plot (Fig. 2), the PCA model with the baseline correction combined mean centering preprocessing showed the absence of outliers.

Table 3 was constructed to compare the eigenvalues and captured variance (%) values obtained from the PCA model to facilitate the choice of the most appropriate principal components number for model. From Table 3, it can be seen that two PCs are enough in the construction of the PCA model.

After PCA, the one hundred forty-five remaining samples were divided into two sets, the calibration (one hundred samples) and validation (forty-five samples), using the selection Kennard-Stone (KS) algorithm.

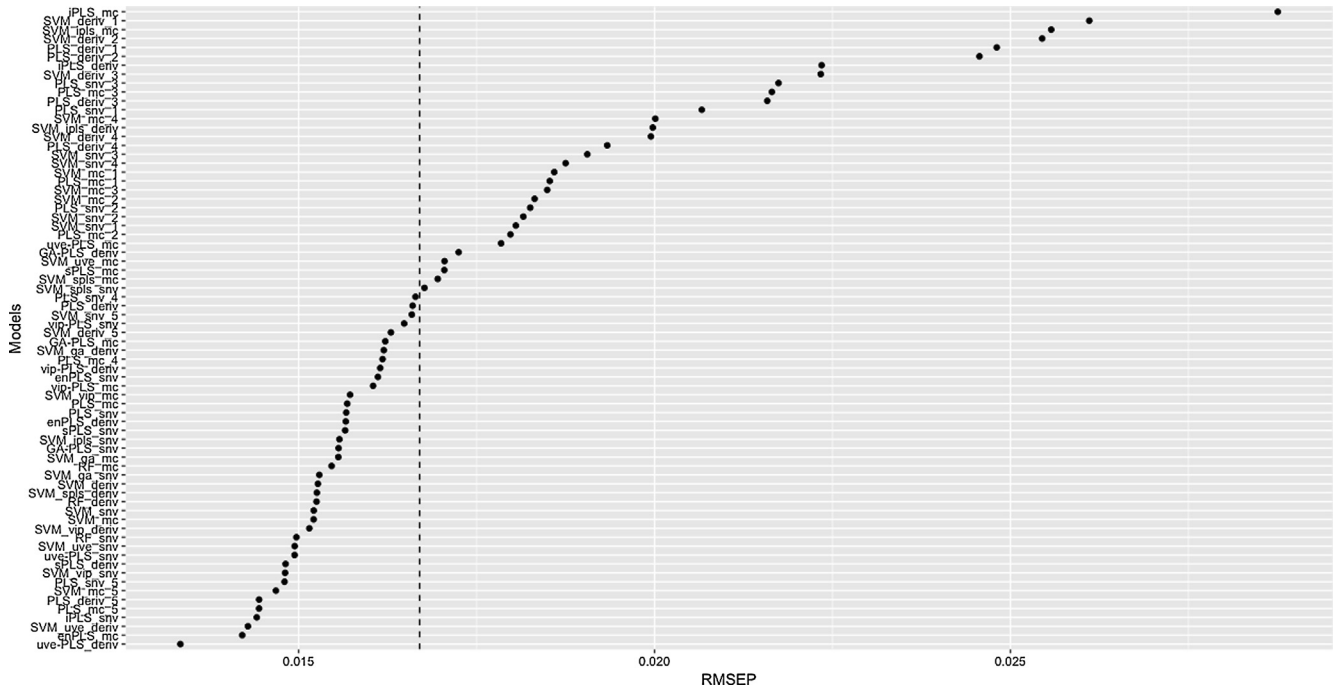

Fig. 3. Models *versus* RMSEP for kinematic viscosity at 40 °C ($mm^2.s^{-1}$).

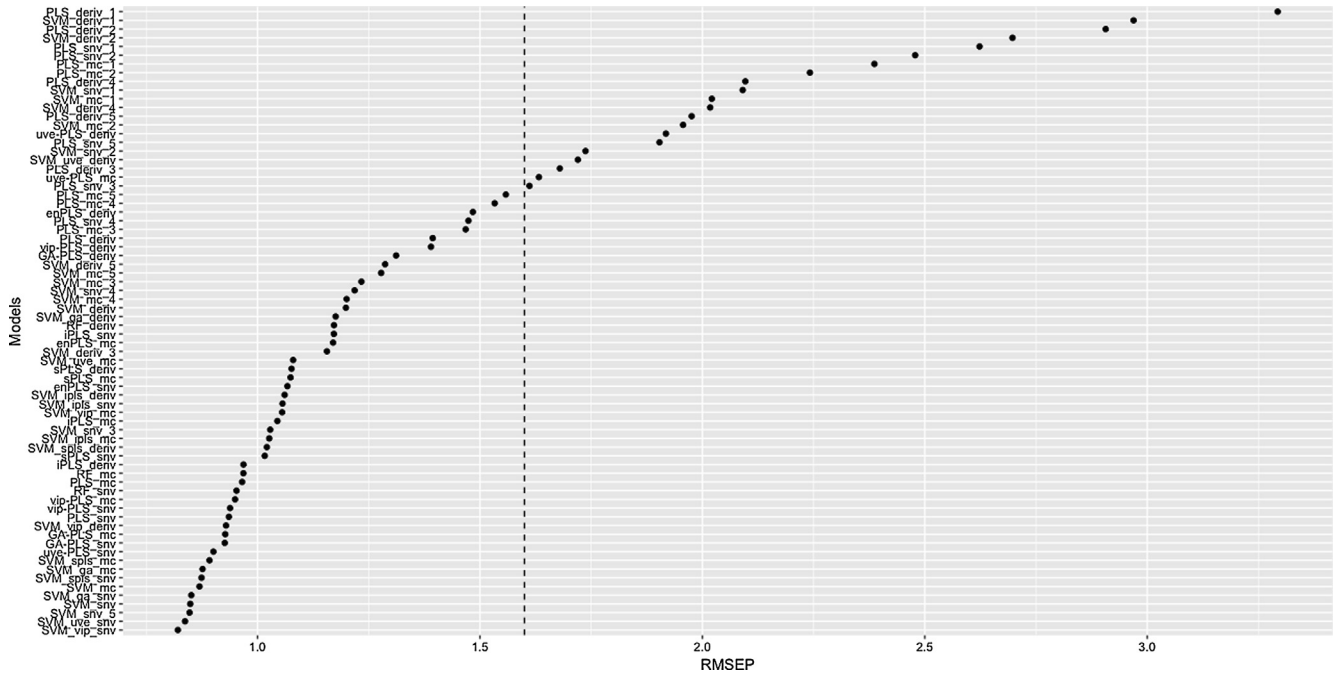

Fig. 4. Models *versus* RMSEP for CFPP (°C).

4.3. Regression models

First of all, it is worth pointing out that viscosity data was used after applying the natural logarithm.

PLS and SVM models were built considering full spectra and the intervals selected by variables selection based on algorithms and chemical information. The selected intervals based the five spectral regions presented in Table 2: (1st) 9000–8000 cm^{-1}, (2nd) 7500–6150 cm^{-1}, (3rd) 5750–5700 cm^{-1}, (4th) 5350–4900 cm^{-1} and (5th) 4370–4260 cm^{-1}, since they presented information about functional groups characteristic of biodiesel. On the other hand, the intervals selected based on algorithms were: ensemble PLS (enPLS), sparse PLS (sPLS), interval partial least squares (iPLS), variable importance in projection (VIP), genetic algorithm (GA) and uninformative variable elimination (UVE).

The Wold's R criterion was applied in choosing the appropriate number of latent variables to be included in each forecast model. In Tables 4 and 5 are presented the number of latent variables for all models.

A priori, the models that presented RMSEP less than the experimental error of studied properties can be considered adequate for prediction of these properties. In Figs. 3 and 4, the vertical line indicates the experimental error of viscosity and PEFF, respectively. Thus, the models that are to the left of the vertical line show RMSEP less than

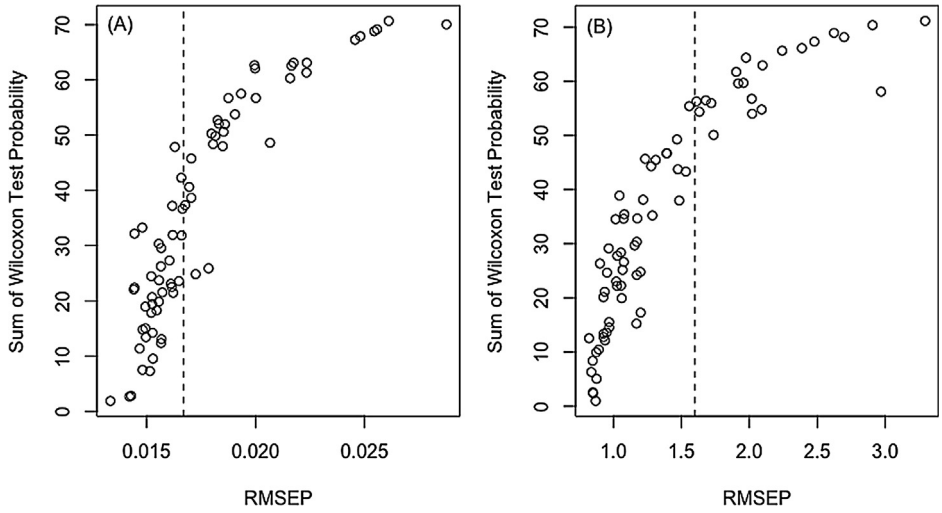

Fig. 5. Sum of Wilcoxon Test Probability *versus* RMSEP of all models for the properties: (A) natural logarithm of kinematic viscosity at 40 °C ($mm^2.s^{-1}$); (B) CFPP (°C).

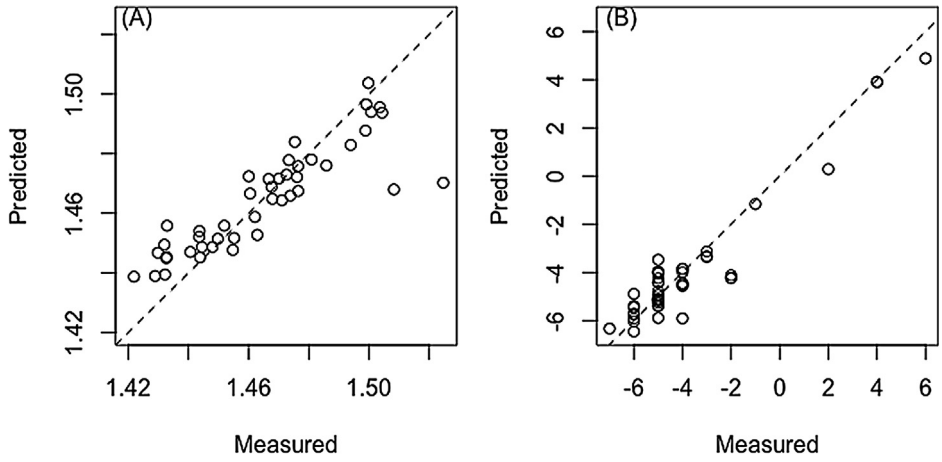

Fig. 6. Predicted *versus* measured plot models for the properties: (A) natural logarithm of kinematic viscosity at 40 °C ($mm^2.s^{-1}$) – PLS with first derivative (deriv) using intervals selected by UVE; (B) CFPP (°C) – SVM with baseline + SNV using interval 5.

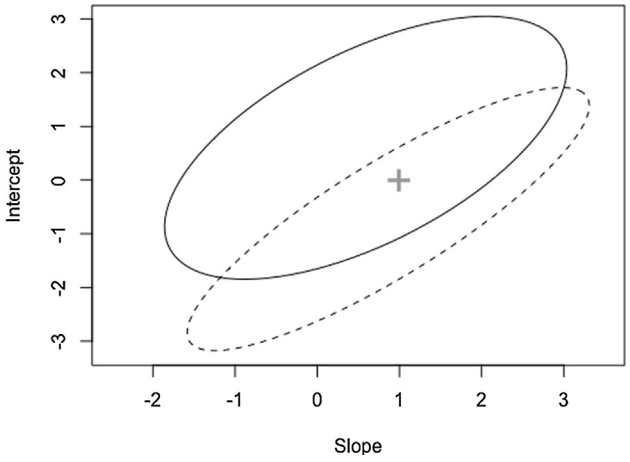

Fig. 7. The elliptical joint confidence region (EJCR) of the best model for each property, applied in prediction models: (–) natural logarithm of kinematic viscosity at 40 °C ($mm^2.s^{-1}$) - PLS with first derivative (deriv) using intervals selected by UVE; (–) CFPP (°C) - SVM with baseline + SNV using interval 5).

the experimental error, 0.0167 $mm^2.s^{-1}$ for viscosity and 1.6 °C for PEFF.

Among the models with RMSEP value smaller than the experimental error, it is necessary to compare the models and to identify the model with the smaller residual value among the models studied.

The paired Wilcoxon test was chosen to compare the residual values between the models. In a new approach, a matrix (72 × 72) with the probability values resulting Wilcoxon test was built. A vector containing the Sum of Wilcoxon Test Probability (SWTP) all lines were obtained and, the minimum value of this vector revealed the best model for each property.

The SWTP values for each model are in Tables 4 and 5, and the minimum value indicates the best model among those studied models. In the case of kinematic viscosity, the PLS model using the intervals obtained by the UVE algorithm and first derivative preprocessing (uve_PLS_deriv) showed the lowest SWTP value (1.887) among the evaluated models. For the CFPP, the lowest SWTP value (0.98) was obtained in the SVM model with the centered spectra in the mean (SVM_mc). However, the SVM_mc calibration model for CFPP was built with 100 support vectors from a dataset containing 100 samples. The number of support vectors (nSV) of a model must be less than the number of samples that built it. The equality between nSV and number of samples means that in the construction of the model was used a different support vector to explain each sample, which indicates a

memorization behavior and not a learning behavior. Then, following the sequence of lowest SWTP values, it is observed that the SVM model for interval 5 with SNV preprocessing (SVM_snv_5) is the most appropriate since it was obtained with 52 support vectors and presented low RMSEP (0.8 °C) and SWTP (8.39). Fig. 5 shows the behavior of the models when comparing SWTP values with RMSEP values for (A) kinematic viscosity and (B) CFPP, respectively.

Due to the difficulty to predict kinematic viscosity, since it is a property of nonlinear behavior, several authors have made use of more robust tools, such as a variables selection [17] or non-linear techniques [16] for the prediction models of this property. Thus, this study also needed to apply a variables selection and nonlinear techniques such as Support Vector Machine and Random Forest.

For kinematic viscosity prediction models, when comparing the models by SWTP obtained, the uve_PLS_deriv model was selected as being the best model, which showed RMSEP equal 0.0133 $mm^2.s^{-1}$ (Table 4). Predicted *versus* measured values plot for kinematic viscosity is showing in Fig. 6(A).

This result can be compared with other studies in the literature, such as Baptista *et al.* [16]. In their work, the authors studied a methodology for determining the kinematic viscosity of biodiesel blends using near-infrared spectroscopy (7700–5400 cm^{-1}) and PLS model with two latent variables and RMSEP equal to 0.10 $mm^2.s^{-1}$.

For the CFPP prediction model, the best result was obtained using SVM for interval 5 with baseline correction + SNV preprocessing and 52 support vectors, which showed RMSEP equal to 0.8 °C, as shown in Table 5. Predicted *versus* measured values plot for CFPP is showing in Fig. 6(B).

This work can be compared with the study developed by other authors. For example, Baptista et al. [18] studied a methodology for determining the CFPP for biodiesel using NIR spectra (9000–4500 cm^{-1}) and constructed a PLS model with three latent variables, resulting in an RMSEP equal to 1.1 °C. Cunha et al. [15] developed models to predict cold filter plugging point in biodiesel samples using the combination of mid-FT-IR spectroscopy coupled with support vector machine regression (SVM), in which the RMSEP was equal to 0.6 °C.

The ellipse of the joint confidence region (EJCR) was employed in the best prediction model for each property, to evaluate the existence of systematic errors and the method accuracy. Moreover, in Fig. 7, it is possible to observe that selected forecast models for each property did not present systematic error.

5. Conclusions

Summarizing, the results of this work are presented as follows:

1. Results indicated that NIR spectroscopy is a valid tool for predicting these parameters in biodiesel samples, with errors similar to the maximum errors allowed by the experimental error.
2. Suitable models for predicting biodiesel fuel properties based on near-infrared (NIR) spectroscopy data using the partial least squares (PLS), random forests (RF) and support vector machines (SVM) techniques were built.
3. For kinematic viscosity, the best model was uve_PLS_deriv with RMSEP equal 0.0133 $mm^2.s^{-1}$.
4. For CFPP, The SVM_snv_5 was considered the best prediction model, with 52 support vectors and lower SWTP and RMSEP (0.8 °C) values.
5. Random forest model prediction for the CFPP and viscosity were considered important methods since they presented low RMSEP values with intermediary SWTP values and without systematic error.
6. The methodologies proposed in this study proved to be useful for monitoring the quality of biofuels, mainly because it is a quick, non-destructive, and inexpensive analysis.

Acknowledgements

The authors would like to thank CENPES (Centro de Pesquisas e Desenvolvimento Leopoldo A. Miguez de Mello) and CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior) for their financial support. Luna, A. S. thanks UERJ (Programa Prociência), FAPERJ (Fundação de Amparo à Pesquisa no Rio de Janeiro), and CNPq (Conselho Nacional de Pesquisa) for their support.

References

[1] Oliveira IK, Rocha WF de C, Poppi RJ. Application of near infrared spectroscopy and multivariate control charts for monitoring biodiesel blends. Anal Chim Acta 2009;642:217–21. https://doi.org/10.1016/j.aca.2008.11.003.
[2] Ma F, Hanna MA. Biodiesel production: a review. Bioresour Technol 1999;70:1–15. https://doi.org/10.1016/s0960-8524(99)00025-5.
[3] Lotero E, Liu Y, Lopez DE, Suwannakarn K, Bruce DA, Goodwin JG. Synthesis of biodiesel via acid catalysis. Ind Eng Chem Res 2005;44:5353–63. https://doi.org/10.1021/ie049157g.
[4] Alptekin E, Canakci M. Determination of the density and the viscosities of biodiesel-diesel fuel blends. Renew Energy 2008;33:2623–30. https://doi.org/10.1016/j.renene.2008.02.020.
[5] Agência Nacional do Petróleo Gás Natural e Biocombustíveis (ANP). ANP Resolution No. 14 of May, 2012 2012. www.anp.gov.br (accessed May 21, 2015).
[6] Torres-Jimenez E, Svoljšak-Jerman M, Gregorc A, Lisec I, Dorado MP, Kegl B. Physical and chemical properties of ethanol-biodiesel blends for diesel engines. Energy Fuels 2010;24:2002–9. https://doi.org/10.1021/ef901158a.
[7] Imahara H, Minami E, Saka S. Thermodynamic study on cloud point of biodiesel with its fatty acid composition. Fuel 2006;85:1666–70. https://doi.org/10.1016/j.fuel.2006.03.003.
[8] Knothe G. Dependence of biodiesel fuel properties on the structure of fatty acid alkyl esters. Fuel Process Technol 2005;86:1059–70. https://doi.org/10.1016/j.fuproc.2004.11.002.
[9] Knothe G. Viscosity of biodiesel. In: Knothe G, Gerpen J Van, Krahl J, editors. Biodiesel Handb. 2nd ed., 2005, p. 89–90.
[10] Knothe G. The History of Vegetables Oil-based Diesel Fuels. In: Knothe G, Gerpen J Van, Krahl J, editors. Biodiesel Handb. 2nd ed., 2005, p. 12–24.
[11] Freitas SVD, Pratas MJ, Ceriani R, Coutinho AP. Evaluation of predictive models for the viscosity of biodiesel. Energy Fuels 2011:352–8. https://doi.org/10.1021/ef101299d.
[12] Abed KA, El Morsi AK, Sayed MM, El Shaib AA, Gad MS. Effect of waste cooking-oil biodiesel on performance and exhaust emissions of a diesel engine. Egypt J Pet 2018;27:985–9. https://doi.org/10.1016/j.ejpe.2018.02.008.
[13] Balabin RM, Safieva RZ. Near-infrared (NIR) spectroscopy for biodiesel analysis: fractional composition, iodine value, and cold filter plugging point from one vibrational spectrum. Energy Fuels 2011;25:2373–82. https://doi.org/10.1021/ef200356h.
[14] Inan TY, Al-Hajji A, Koseoglu OR. Chemometrics-based analytical method using FTIR spectroscopic data to predict diesel and diesel/diesel blend properties. Energy Fuels 2016;30:5525–36. https://doi.org/10.1021/acs.energyfuels.6b00731.
[15] Cunha CL, Luna AS, Oliveira RCG, Xavier GM, Paredes MLL, Torres AR. Predicting the properties of biodiesel and its blends using mid-FT-IR spectroscopy and first-order multivariate calibration. Fuel 2017;204:185–94. https://doi.org/10.1016/j.fuel.2017.05.057.
[16] Santos VO, Oliveira FCC, Lima DG, Petry AC, Garcia E, Suarez PAZ, et al. A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis. Anal Chim Acta 2005;547:188–96. https://doi.org/10.1016/j.aca.2005.05.042.
[17] Balabin RM, Smirnov SV. Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. Anal Chim Acta 2011;692:63–72. https://doi.org/10.1016/j.aca.2011.03.006.
[18] Baptista P, Felizardo P, Menezes JC, Neiva Correia MJ. Multivariate near infrared spectroscopy models for predicting the iodine value, CFPP, kinematic viscosity at 40 °C and density at 15 °C of biodiesel. Talanta 2008;77:144–51. https://doi.org/10.1016/j.talanta.2008.06.001.
[19] Menezes JC, Ferreira AP, Rodrigues LO, Brás LP, Alves TP. Chemometrics role within the PAT context: examples from primary pharmaceutical manufacturing. Compr Chemom 2010;4:313–55. https://doi.org/10.1016/B978-044452701-1.00012-0.
[20] Zhao C, Gao F, Wang F. Phase-based joint modeling and spectroscopy analysis for batch processes are monitoring. Ind Eng Chem Res 2010;49:669–81.
[21] Brereton RG. Signal Process 2003;8. https://doi.org/10.1002/0470863242.ch3.
[22] Gaydou V, Kister J, Dupuy N. Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil. Chemom Intell Lab Syst 2011;106:190–7. https://doi.org/10.1016/j.chemolab.2010.05.002.
[23] Naes T, Isaksson T, Fearn T, Davies T. A user-friendly guide to multivariate calibration and classification. Chichester-UK: NIR Publications; 2002.
[24] Malinowskimon ER. Factor analysis in chemistry. vol. 5. 1991. doi:10.1002/cem.1180050607.
[25] Martens H, Naes T. Multivariate calibration. Spectrochim Acta Part A Mol Biomol Spectrosc 1989;44:287–321.
[26] Ferrer-Riquelme AJ. Statistical control of measures and processes. Compr Chemom

2010;1:97–126. https://doi.org/10.1016/B978-044452701-1.00096-X.
[27] Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat 2010;2:433–59. https://doi.org/10.1002/wics.101.
[28] Cordella CBY. PCA: the basic building block of chemometrics. Anal Chem 2012:1–46. https://doi.org/10.5772/3086.
[29] Ma S, Aybat NS. Efficient optimization algorithms for robust principal component analysis and its variants. . Proc IEEE 2018:1–16. https://doi.org/10.1109/JPROC.2018.2846606.
[30] Wold H. Soft modelling: the basic design and some extensions. Syst. under Indirect Obs. Causality-structure-prediction, 1982, p. 1–54.
[31] Wold S. Cross-validation estimation of the number of compo- nents in factor and principal component analysis. Technometrics 1978;24:397–405.
[32] Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc 1974;B36:111–47.
[33] Krzanowski WJ. Croos-validation in principal component analysis. Biometrics 1987;43:575–84.
[34] Cao DS, Deng ZK, Zhu MF, Yao ZJ, Dong J, Zhao RG. Ensemble partial least squares regression for descriptor selection, outlier detection, applicability domain assessment, and ensemble modeling in QSAR/QSPR modeling. J Chemom 2017;31:1–17. https://doi.org/10.1002/cem.2922.
[35] Mevik B, Segtnan VH, Næs T. Ensemble methods and partial least squares regression. J Chemom 2005;18:498–507. https://doi.org/10.1002/cem.895.
[36] Bi YM, Chu GH, Wu JZ, Yuan KL, Wu J, Liao F, et al. Ensemble partial least squares algorithm based on variable clustering for quantitative infrared spectrometric analysis. Chin J Anal Chem 2015;43:1086–91. https://doi.org/10.1016/S1872-2040(15)60842-8.
[37] Hu Y, Peng S, Peng J, Wei J. An improved ensemble partial least squares for analysis of near-infrared spectra. Talanta 2012;94:301–7. https://doi.org/10.1016/j.talanta.2012.03.047.
[38] Xu X, Cheng K, Deng L, Dong J. Chemometrics and Intelligent Laboratory Systems A sparse partial least squares algorithm based on sure independence screening method. Chemom Intell Lab Syst 2017;170:38–50. https://doi.org/10.1016/j.chemolab.2017.09.011.
[39] Lee D, Lee W, Lee Y, Pawitan Y. Sparse partial least-squares regression and its applications to high-throughput data analysis. Chemom Intell Lab Syst 2011;109:1–8. https://doi.org/10.1016/j.chemolab.2011.07.002.
[40] Liquet B, De Micheaux PL, Hejblum BP, Thiébaut R. Gene expression Group and sparse group partial least square approaches applied in genomics context. Bioinforma Adv Access 2017;32:35–42. https://doi.org/10.1093/bioinformatics/btv535.
[41] Xiaobo Z, Jiewen Z, Povey MJW, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. Anal Chim Acta 2010;667:14–32. https://doi.org/10.1016/j.aca.2010.03.048.
[42] Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl Spectrosc 2000;54:413–9. https://doi.org/10.1366/0003702001949500.
[43] Godinho MS, Blanco MR, Gambarra FF, Lião LM, Sena MM, Tauler R, et al. Talanta Evaluation of transformer insulating oil quality using NIR, fl uorescence, and NMR spectroscopic data fusion. Talanta 2014;129:143–9. https://doi.org/10.1016/j.talanta.2014.05.021.
[44] Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in Partial Least Squares Regression. Chemom Intell Lab Syst 2012;118:62–9.
[45] Wise BM, Gallagher NB, Bro R, Shaver JM. PLS_Toolbox 3.01. Inc Manson, Eig Res 2003.
[46] Centner V, Massart D-L, de Noord OE, de Jong S, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. Anal Chem 1996;68:3851–8. https://doi.org/10.1021/ac960321m.
[47] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/a:1010933404324.
[48] Zhang H, Wu P, Yin A, Yang X, Zhang M, Gao C. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: a comparison of multiple linear regressions and the random forest model. Sci Total Environ 2017;592:704–13. https://doi.org/10.1016/j.scitotenv.2017.02.146.
[49] Liu W, Liu C, Yu J, Zhang Y, Li J, Chen Y, et al. Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics. Food Chem 2018;251:86–92. https://doi.org/10.1016/j.foodchem.2018.01.081.
[50] Shahbazi B, Chehreh Chelgani S, Matin SS. Prediction of froth flotation responses based on various conditioning parameters by Random Forest method. Colloids Surfaces A Physicochem Eng Asp 2017;529:936–41. https://doi.org/10.1016/j.colsurfa.2017.07.013.
[51] Ni D, Ji X, Wu M, Wang W, Deng X, Hu Z, et al. Automatic cystocele severity grading in transperineal ultrasound by random forest regression. Pattern Recognit 2017;63:551–60. https://doi.org/10.1016/j.patcog.2016.09.033.
[52] Deconinck E, Cauwenbergh T, Bothy JL, Custers D, Courselle P, De Beer JO. Detection of sibutramine in adulterated dietary supplements using attenuated total reflectance-infrared spectroscopy. J Pharm Biomed Anal 2014;100:279–83. https://doi.org/10.1016/j.jpba.2014.08.009.
[53] Huang JH, Yan J, Wu QH, Ferro MD, Yi LZ, Lu HM, et al. Selective of informative metabolites using random forests based on model population analysis. Talanta 2013;117:549–55. https://doi.org/10.1016/j.talanta.2013.07.070.
[54] Li B, Wei Y, Duan H, Xi L, Wu X. Discrimination of the geographical origin of Codonopsis pilosula using near infrared diffuse reflection spectroscopy coupled with random forests and k-nearest neighbor methods. Vib Spectrosc 2012;62:17–22. https://doi.org/10.1016/j.vibspec.2012.05.001.
[55] Wang Y, Huang HY, Zuo ZT, Wang YZ. Comprehensive quality assessment of Dendrubium officinale using ATR-FTIR spectroscopy combined with random forest and support vector machine regression. Spectrochim Acta - Part A Mol Biomol Spectrosc 2018;205:637–48. https://doi.org/10.1016/j.saa.2018.07.086.
[56] Amjad A, Ullah R, Khan S, Bilal M, Khan A. Raman spectroscopy based analysis of milk using random forest classification. Vib Spectrosc 2018;99:124–9. https://doi.org/10.1016/j.vibspec.2018.09.003.
[57] Bishop CM. Pattern Recognition and Machine Learning. vol. 4. 2006. doi:10.1117/1.2819119.
[58] Müller K, Smola A, Rätsch G, Schölkopf B, J. Predicting time series with support vector machines. Artif Neural Networks—ICANN'97 1997;1327:999–1004. doi:10.1007/BFb0020283.
[59] Vapnik VN. The Nature of Statistical Learning Theory. vol. 8. 1995. doi:10.1109/TNN.1997.641482.
[60] Steinwart I, Christmann A. Support vector machines New York: Springer-Verlag; 2008. https://doi.org/10.1007/978-0-387-77242-4.
[61] Wise BM, Gallagher NB, Bro R, Shaver JM, Windig W, Koch RS. Chemometrics Tutorial for PLS_Toolbox and Solo, 2006.
[62] Ballabio D. Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal Methods 2013;5:3790–8. https://doi.org/10.1039/C3AY40582F.
[63] Mevik BH, Cederkvist HR. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J Chemom 2004;18:422–9. https://doi.org/10.1002/cem.887.
[64] De Souza LM, Mitsutake H, Gontijo LC, Borges Neto W. Quantification of residual automotive lubricant oil as an adulterant in Brazilian S-10 diesel using MIR spectroscopy and PLS. Fuel 2014;130:257–62. https://doi.org/10.1016/j.fuel.2014.03.051.
[65] Valderrama P, Braga JWB, Poppi RJ. Variable selection, outlier detection, and figures of merit estimation in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy. J Agric Food Chem 2007;55:8331–8. https://doi.org/10.1021/jf071538s.
[66] Gehan EA. A generalized two-sample wilcoxon test for doubly censored data. Biometrika 1965;52:650–3.
[67] Clark WM, Medeiros NJ, Boyd DJ, Snell JR. Biodiesel transesterification kinetics monitored by pH measurement. Bioresour Technol 2013. https://doi.org/10.1016/j.biortech.2013.03.089.
[68] Hossain ABMS, Mazen MA. Effects of catalyst types and concentrations on biodiesel production from waste soybean oil biomass as renewable energy and environmental recycling process. Aust J Crop Sci 2010;4:550–5.
[69] Gülüm M, Bilgin A. Density, flash point and heating value variations of corn oil biodiesel – diesel fuel blends. Fuel Process Technol 2015;134:456–64. https://doi.org/10.1016/j.fuproc.2015.02.026.
[70] Gülüm M, Bilgin A. Two-term power models for estimating kinematic viscosities of different biodiesel-diesel fuel blends. Fuel Process Technol 2016;149:121–30. https://doi.org/10.1016/j.fuproc.2016.04.013.
[71] Gülüm M, Bilgin A. Measurements and empirical correlations in predicting biodiesel-diesel blends' viscosity and density. Fuel 2017;199:567–77. https://doi.org/10.1016/j.fuel.2017.03.001.
[72] Edith O. Factors affecting the cold flow behaviour of biodiesel and methods for improvement – a review. Pertanika J Sci Technol 2012;20:1–14.
[73] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2016.
[74] Kucheryavskiy S. mdatools: Multivariate Data Analysis for Chemometrics. R package version 0.7.0 2015.
[75] Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis. J Stat Softw 2009;32:1–47.
[76] Mevik BH, Wehrens R, Liland KH. pls: Partial Least Squares and Principal Component Regression. R package version 2.5-0 2015.
[77] Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2:18–22.
[78] Xiao N, Cao D-S, Li M-Z, Xu Q-S. enpls: Ensemble Partial Least Squares Regression. R Packag Version 60 2018.
[79] Liquet B, Micheaux PL de, Broc C. sgPLS: Sparse Group Partial Least Square Methods. R Packag Version 17 2017.
[80] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R Packag Version 17-0 2018.
[81] Stevens A, Ramirez-Lopez L. An introduction to the prospectr package. R Packag Vignette 2013.
[82] Murdoch D, Chow ED. ellipse: Functions for drawing ellipses and ellipse-like confidence regions. R package version 0.3-8 2013.
[83] Kennard RW, Stone LA. Computer Aided Design of Experiments. Technometrics 1969;11:137–48. https://doi.org/10.2307/1266770.
[84] ASTM E1655-05. Standard practices for infrared multivariate quantitative analysis ASTM Int; 2012. p. 29. https://doi.org/10.1520/E1655-05R12.2.