Talanta 125 (2014) 233–241

Contents lists available at ScienceDirect

Talanta

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

Application of correlation constrained multivariate curve resolution alternating least-squares methods for determination of compounds of interest in biodiesel blends using NIR and UV–visible spectroscopic data

Rodrigo Rocha de Oliveira [a,b], Kássio Michell Gomes de Lima [a], Romà Tauler [c], Anna de Juan [b,*]

[a] Grupo de Pesquisa em Quimiometria Aplicada, Instituto de Química, Universidade Federal do Rio Grande do Norte, Av. Senador Salgado Filho 3000, Natal, CEP 59078-970, Brazil
[b] Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Diagonal 645, Barcelona 08028, Spain
[c] Institute of Environmental Assessment and Water Studies, IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, Spain

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form
25 February 2014
Accepted 28 February 2014
Available online 7 March 2014

Keywords:
Biodiesel analysis
Near infrared spectroscopy
Visible spectroscopy
Multivariate curve resolution alternating least-squares
Correlation constraint
Multivariate calibration
Sample matrix effect

ABSTRACT

This study describes two applications of a variant of the multivariate curve resolution alternating least squares (MCR-ALS) method with a correlation constraint. The first application describes the use of MCR-ALS for the determination of biodiesel concentrations in biodiesel blends using near infrared (NIR) spectroscopic data. In the second application, the proposed method allowed the determination of the synthetic antioxidant *N,N′*-Di-*sec*-butyl-*p*-phenylenediamine (PDA) present in biodiesel mixtures from different vegetable sources using UV–visible spectroscopy. Well established multivariate regression algorithm, partial least squares (PLS), were calculated for comparison of the quantification performance in the models developed in both applications. The correlation constraint has been adapted to handle the presence of batch-to-batch matrix effects due to ageing effects, which might occur when different groups of samples were used to build a calibration model in the first application. Different data set configurations and diverse modes of application of the correlation constraint are explored and guidelines are given to cope with different type of analytical problems, such as the correction of matrix effects among biodiesel samples, where MCR-ALS outperformed PLS reducing the relative error of prediction RE (%) from 9.82% to 4.85% in the first application, or the determination of minor compound with overlapped weak spectroscopic signals, where MCR-ALS gave higher (RE (%)=3.16%) for prediction of PDA compared to PLS (RE (%)=1.99%), but with the advantage of recovering the related pure spectral profile of analytes and interferences. The obtained results show the potential of the MCR-ALS method with correlation constraint to be adapted to diverse data set configurations and analytical problems related to the determination of biodiesel mixtures and added compounds therein.

1. Introduction

Increasing consumption of fossil fuels and environmental concerns about the consequent increased greenhouse gas emissions have propelled the development of biofuels as an alternative energy source to meet future needs and reduce the environmental impact [1,2]. Biodiesel consists of a mixture of alkyl esters of long chain fatty acids; it is a biofuel produced via the transesterification of fatty acids (triglycerides) from vegetable oils and animal fats with short chain alcohols via homogeneous, heterogeneous or enzymatic catalysis. Because of its natural properties, biodiesel can partially or completely replace diesel made from fossil fuels. Nowadays, many countries worldwide have commercially exploited and used up to 20% (v/v) of biodiesel in mixtures with diesel oil in their vehicles [3]. The Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) established via Resolution no. 42, 2009 the specifications for diesel oil types A (without biodiesel) and B (blends of diesel with biodiesel) and the mandatory usage of blends of diesel with 5% (v/v) of biodiesel occurs in Brazil since the beginning of 2010 and the specifications established by the ANP must be met. Therefore, analytical methods for biodiesel determination should be well established, rapid and accessible to meet the growing demand for this product [2].

Spectroscopic techniques have been applied for the determination of several parameters in biodiesel. All the spectroscopic range

* Corresponding author. Tel.: +34 934 03 9778.
E-mail address: anna.dejuan@ub.edu (A. de Juan).

http://dx.doi.org/10.1016/j.talanta.2014.02.073
0039-9140

from ultraviolet to mid infrared absorption spectroscopy has been used in many works for determination of biodiesel parameters from different feedstocks [4–11], as well as molecular fluorescence spectroscopy [12,13]. Infrared spectroscopy has been the subject of many works recently reviewed by Zhang [7]. Near infrared (NIR) spectroscopy and chemometric is one of the most used combinations for biodiesel analysis, due to the easy, rapid, and non-destructive sample analysis [5]. Several chemometric methods have been applied to spectroscopic biodiesel analysis. Multivariate calibration methods, such as Multivariate Linear Regression (MLR), Principal Component Regression, partial least squares (PLS) regression, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) have been often used to extract information from NIR spectra for determination of quality parameters in biodiesel and biodiesel blends [5,6,8–10]. Chemometric methods for classification, such as soft independent modeling of class analogy (SIMCA), hierarchical cluster analysis (HCA), and successive projections algorithm (SPA) with linear discriminant analysis and PLS discriminant analysis (PLS-DA) have been used to classify biodiesel according to the feedstock [11,13]. Variable selection methods, such as Genetic Algorithm, interval-PLS, SPA and others, have been used to reduce the number of variables and improve the abilities of calibration and classification models [5,6,13].

One of the main biodiesel problems is the low stability to oxidation because of its high content of unsaturated esters. The oxidation is mainly due to air contact, metallic ions contamination, light exposure or long-term storage. Therefore, synthetic antioxidants must be added to biodiesel fuels to maintain their quality parameters [14,15]; if not, the oxidation may lead to increase of viscosity, corrosion of fuel system components and formation of gums and sediments that may clog the engine fuel filter. Aromatic amines and phenolic compounds are two families of antioxidant compounds that react and stabilize the free radicals formed during the biodiesel oxidation. Many works have been devoted to study the effect of adding synthetic antioxidants to biodiesel [14,16,17]. Therefore, determination of the antioxidant concentration is an important task in order to evaluate the stability of biodiesel to oxidation.

Different analytical methodologies were proposed for biodiesel antioxidant analysis. Tormin et al. developed methods based on the amperometric determination of mixtures of tert-butylhydroquinone (TBHQ) and butylated hydroxyanisole (BHA) by batch-injection analysis [18] in synthetic samples of biodiesel. The aromatic amine *N*,*N*′-Di-*sec*-butyl-*p*-phenylenediamine (PDA) has been proven to be an efficient antioxidant and a versatile artificial marker for biodiesel and has been analyzed by easy ambient sonic-spray ionization mass spectrometry (EASI-MS) [19]. Peaks in the mid infrared region were also used for calibration and determination of PDA antioxidant in sunflower biodiesel mixtures [20].

Most spectroscopic measurements in biodiesel samples have been analyzed by multivariate exploratory analysis and calibration methods, but very few use multivariate resolution techniques, such as multivariate curve resolution with alternating least squares (MCR-ALS). MCR-ALS has been applied in very few works for biodiesel analysis. The authors found two works where MCR-ALS has been applied for resolving spectrophotometric sequential injection analysis (SIA-DAD) data in the determination of sulfate and acidity of biodiesel samples [21,22] and for quantitative analysis of biodiesel and diesel blends using two dimensional chromatography [23]. MCR-ALS decomposes a data table (matrix) of multivariate mixed measurements into a bilinear model of meaningful pure component contributions. In spectroscopy, this is analogous to recover the underlying Beer–Lambert model, i.e., extracting the pure spectra of the sample constituents and the related concentration profiles from the information contained in the raw measured spectra [24,25].

MCR-ALS has been proven to be efficient to resolve and provide relative determination of compounds in different types of complex processes and mixtures [24], such as liquid chromatography with diode array detection [26,27] and spectral data from industrial processes [28,29]. A variant of MCR-ALS with a correlation constraint was proposed by Antunes et al. to obtain quantitative information of analytes in real concentration units in the presence of interferences [30]. This methodology has been applied successfully to quantify metal ions [30], industrial mixtures in the production process of vinyl acetate monomer [31], ascorbic acid in powder juices and tetracycline in serum samples [32], steroid drugs in pharmaceutical samples, moisture and protein in forage samples [33] and in pharmaceutical industrial process control [28]. This constraint was further extended for first order data and correction of matrix effect in the determination of paracetamol in tablets contained in blister packages using Raman spectroscopy [34].

The present work reports the use of MCR-ALS with the recently proposed modified correlation constraint for determination of biodiesel content in diesel blends using NIR spectroscopy. The new methodology was applied to overcome the sample matrix effect problem observed due to differences of ageing between two batches of biodiesel blends used to build the calibration model. This work also shows the potential of the MCR variant with correlation constraint to adapt to the diverse data set configurations and analytical problems related to the determination of minor compound, synthetic antioxidant PDA, in biodiesel mixtures from different vegetable sources with overlapped spectral signal using UV–visible spectroscopy. PLS [35,36] models were calculated for comparison of the quantification performance in the models developed.

2. Experimental

2.1. Raw materials and sample preparation

Two sets of samples were used in this work. The first set of samples contained mixtures of neat diesel and soybean biodiesel provided by the Laboratory of Fuels and Lubricants (LCL) of the Federal University of Rio Grande do Norte (UFRN), RN, Brazil. Biodiesel was prepared by the basic catalyzed transesterification reaction of commercial soybean vegetable oil with methanol. With the prepared biodiesel and the neat diesel fuel, 38 blends were prepared in two batches of 30 and eight samples. The first batch of blends was prepared and submitted to natural aging for about three months before measurement. The second batch was freshly prepared and measured at the same time as the first batch. Percentage of biodiesel in samples was determined following the European method EN 14078 and ranged from 0% to 20.5% (v/v).

The second set of samples was formed by 62 samples containing mixtures of biodiesels from four different feedstocks (peanut, sesame, *Jatropha curcas* and soybean oil seeds) and two commercial synthetic antioxidants (butylated hydroxytoluene – BHT and *N*, *N*′-Di-*sec*-butyl-*p*-phenylenediamine – PDA). All raw products were provided by the Laboratory of Engines and Fuels (Lamoc) in the National Institute of Metrology, Quality and Technology (Inmetro), RJ, Brazil. Oil seed extraction and biodiesel synthesis were carried out by Lamoc following the method used in [20]. A cubic D-optimal mixture design was developed with Design-Expert® (Stat-Ease Inc., Minneapolis, MN, USA) software to set the composition of the samples. All samples were prepared according to the required composition for a total sample mass of 4 g using an analytical balance with a precision of 0.1 mg, Marte®, model AY220. The concentration of antioxidants covered the range commercially used for biodiesel fuels. To achieve low concentration levels for antioxidants, diluted stock solutions of

Table 1
Statistics for the 62 samples mixtures.

	Kind of biodiesel (%wt)				Antioxidant (ppm)	
	PN	SE	JC	SB	BHT	PDA
Min.	0.18	0.19	0.18	0.18	2	1
Max.	99.38	99.30	99.39	53	2632	1006
Mean	24.42	26.12	23.56	22.33	892	302
Std.	18.07	19.71	19.23	15.01	712	232

PN, peanut; SE, sesame; JC, *Jatropha curcas*; SB, soybean; BHT, butylated hydroxytoluene; PDA, *N*,*N*′-Di-*sec*-butyl-*p*-phenylenediamine.

each antioxidant were prepared using each biodiesel as solvent. The range of concentrations for each compound is described in Table 1.

2.2. Instrumentation and experimental measurements

Near infrared spectra of biodiesel blends were recorded using a FT-NIR spectrophotometer model MB 160 (Bomem) with a resolution of 8 cm^{-1} and an average of 50 scans. Spectra were collected in cells with two optical pathlengths: 10 mm (for the spectral range between 1105 and 1677 nm) and 1.0 mm (for the spectral range between 2111 and 3216 nm) to compensate the different signal intensity in the two spectral ranges acquired.

UV–visible spectra of biodiesel and antioxidant mixtures were acquired with a UV–visible spectrophotometer model Evolution 60 S (Thermo Scientific) in the spectral range 370–670 nm, with a wavelength increment of 2 nm among consecutive measurements. A 10 mm pathlength quartz cuvette was used.

Pure compound NIR and UV–visible spectrum were also recorded to be used afterwards in the chemometric analysis.

3. Data treatment

3.1. Data sets

The first set of samples gave a matrix formed by the NIR spectra collected. Two samples were removed as spectral outliers from the first batch, thus the final size of the matrix was (36 × 1224), with the rows containing the samples spectra and the columns designing the wavelength variables.

The first 28 spectra were from the first aged batch and the last eight from the second fresh batch. The first 801 columns were associated with the spectral range (1105–1677 nm), referred to spectra collected with the 10 mm pathlength cell and the last 423 columns covered the range (2111–3216 nm), used in the spectra recorded with the 1.0 mm pathlength cell. Spectral preprocessing consisted of offset correction to remove negative values in the spectra followed by Multiplicative Signal Correction (MSC) [37] to correct baseline fluctuations in the NIR spectra. Both training and test samples were baseline corrected.

The second set of samples was formed by the UV–visible spectra collected from the mixtures of biodiesels and antioxidants and was sized (62 × 151), accounting for (samples × variables). The best preprocessing method found for this data was the first order Savitzky–Golay derivative [38] with a second order polynomial fit and 11 points window.

Calibration samples were selected in the NIR data set by the Kennard–Stone algorithm [39]. It is worth noting that the calibration set covered samples from both, aged and fresh samples, the same for the test set. In the UV–visible data set, selection was done covering the full antioxidant concentration range. About two thirds of the total number of samples was selected for the calibration set, and the rest were used to test the calibration model.

3.2. Chemometric methods

3.2.1. Multivariate curve resolution alternating least squares (MCR-ALS)

MCR assumes a bilinear model that is the multiwavelength extension of the Lambert–Beer's law [27,29,40,41] and is described in matrix form by the expression,

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \quad (1)$$

where $\mathbf{D}(ij)$ is a data matrix containing the NIR or UV–visible spectra of the i samples for the j wavelengths recorded, $\mathbf{C}(in)$ and $\mathbf{S}^{\mathrm{T}}(nj)$ are the matrices with the concentration and spectral profiles of the n pure components in the samples, respectively. **E** has the same size of **D** and contains the variance not explained by the bilinear model, related to the experimental error. The same bilinear model of MCR holds for multiset analysis, which consists of the simultaneous analysis of multiple data matrices coming from different techniques and/or from different experiments or batches [26,28,29,40].

The first step before MCR-ALS optimization is estimating the number of components by singular value decomposition (SVD) or other methods [42]. To start the ALS optimization, a method based on SIMPLISMA [43] was applied to obtain an $\mathbf{S}^{\mathbf{T}}$ matrix of initial estimates. Matrices **C** and $\mathbf{S}^{\mathbf{T}}$ are alternatingly optimized by the ALS procedure solving Eq. (1). If the algorithm starts with an $\mathbf{S}^{\mathbf{T}}$-type initial estimate, the unconstrained least squares solution for the **C** matrix is obtained by the expression,

$$\mathbf{C} = \mathbf{D}(\mathbf{S}^{\mathrm{T}})^{+} \quad (2)$$

If **C** is used, instead, the unconstrained least squares solution for the $\mathbf{S}^{\mathbf{T}}$ matrix is obtained by the expression,

$$\mathbf{S}^{\mathrm{T}} = \mathbf{C}^{+}\mathbf{D} \quad (3)$$

where $(\mathbf{S}^{\mathrm{T}})^{+}$ and $\mathbf{C}^{+}$ are the pseudoinverses of the full rank matrices $\mathbf{S}^{\mathbf{T}}$ and **C**, respectively. In each iteration some constraints must be applied in order to reduce the number of possible solutions for **C** and $\mathbf{S}^{\mathrm{T}}$ and to give physicochemical meaning to the results. Natural constraints, such as non-negativity and selectivity were applied in this work [27,41]. Non-negativity constraint forces the concentration and/or spectral profile to be equal or larger than zero [44]. Selectivity or local rank constraints are used when some species are not present in certain sample spectra improving the definition of profiles during the iterations [27]. Another less common constraint is the correlation constraint that builds internal univariate calibration models between reference concentration values in calibration samples and the analogous values in MCR concentration profiles. This constraint allows prediction of concentration values in unknown samples and provides concentration profiles in real concentration units. A modified correlation constraint was recently proposed to correct matrix effects between sample sets due to the presence of blister packages in the determination of paracetamol content and excipients using Raman spectroscopy [34]. Detailed explanation of this constraint and the recently modified method is found in Section 3.2.1.1.

The ALS optimization procedure finishes when a certain convergence criterion is achieved [41]. Usually, the convergence is reached when the relative difference between the root mean square of the residuals matrix **E** between consecutive iterations is lower than a threshold value, commonly set to 0.1%. The quality of the MCR-ALS fit to the experimental data matrix is calculated by

the percentage of lack of fit as stated in Eq. (4),

$$lof(\%) = 100\sqrt{\frac{\sum_{ij}(d_{ij}-\hat{d}_{ij})^2}{\sum_{ij}d_{ij}^2}} \quad (4)$$

where d_{ij} are the elements of the original data matrix and $\hat{d}_{ij}$ those reproduced by a MCR-ALS model.

3.2.1.1. Correlation constraint. The correlation constraint builds internal univariate calibration models between the concentration values calculated by the MCR models and reference values from calibration samples. These models are afterwards used to predict concentration in validation and test samples. As a consequence, the concentration profiles are expressed in real concentration units. In each iteration, the relative concentration values $\mathbf{c}_{\mathbf{cal}}^{\mathbf{ALS}}$ of calibration samples, obtained from the suitable MCR **C** concentration profile, are regressed against the respective reference concentration values $\mathbf{c}_{\mathbf{cal}}^{\mathbf{ref}}$ of the analyte in these samples. The slope b and offset b_0 are obtained by fitting a linear least squares regression model between $\mathbf{c}_{\mathbf{cal}}^{\mathbf{ref}}$ and $\mathbf{c}_{\mathbf{cal}}^{\mathbf{ALS}}$ values.

$$\mathbf{c}_{\mathbf{cal}}^{\mathbf{ALS}} = b\mathbf{c}_{\mathbf{cal}}^{\mathbf{ref}} + b_0 \quad (5)$$

$$\hat{\mathbf{c}}_{\mathbf{test}} = \frac{\mathbf{c}_{\mathbf{test}}^{\mathbf{ALS}} - b_0}{b} \quad (6)$$

Once the parameters b and b_0 in Eq. (5) are obtained, a vector $\hat{c}_{test}$ with the predicted concentrations is obtained by Eq. (6) using the relative concentration values in the **C** profile for the test samples, $\mathbf{c}_{\mathbf{test}}^{\mathbf{ALS}}$. The unconstrained $\mathbf{c}^{\mathbf{ALS}}$ vector is updated by the vector of reference values for calibration samples $\mathbf{c}_{\mathbf{cal}}^{\mathbf{ref}}$ and by the predicted values for test and/or unknown samples $\hat{\mathbf{c}}_{\mathbf{test}}$. The same procedure is repeated in the next ALS iterations until the ALS optimization converges. As any other constraint, the correlation constraint can be applied to one or more analyte concentration profiles. Therefore, a different calibration model Eq. (5) is obtained for each component.

The correlation constraint can be applied in a flexible way when the data set contains sample subsets with different behavior [34] The modes used in this work are described below:

a. Correlation constraint with a single regression model for all sample subsets. This is the conventional way to apply the correlation constraint, when all the calibration samples in the data set are used to build a single calibration model per analyte.
b. *Correlation constraint with local models per individual sample subset or group of sample subsets and matrix effect correction.* Separate models for different sample subsets can be calculated to overcome matrix effect problems among samples of different subsets [34]. Matrix effect means that there are different linear relationships between concentration values and signal responses of the analytes for each sample subset affected by a different matrix effect (please note that $\mathbf{D_i}$ and $\mathbf{C_i}$ here refer to sample subsets). In such a case, the pure spectrum of the analyte in would be different in scale for each sample subset. To overcome this modeling problem, one of the sample subsets should be taken as a reference and a rescaling procedure must be applied to the concentration values of the other sample subsets affected by a different matrix effect before updating its corresponding concentration profile during ALS iterations. Real unscaled concentration values predicted by each local regression model for calibration and test samples are stored in a separate output and can be recovered at the end of the MCR-ALS optimization.

Sample matrix effects can be caused by different reasons, such as temperature changes, time ageing, different batches and variation in the analysis or instrumental conditions. This type of effect was observed in the present work for the NIR analysis of the two biodiesel blend batches. Fig. 1 shows, in detail, the application of the correlation constraint and matrix effect correction in the data investigated in this work. Data matrix **D** is formed by two sample subsets, $\mathbf{D}_1$ and $\mathbf{D}_2$, which contain both calibration (cal) and test samples (test).

Once the concentration profile for a certain analyte is selected for the correlation constraint, two local regression models are built as described above: one for the concentration profile coming from $\mathbf{D}_1$ (relating $\mathbf{c}_{\mathbf{cal}_1}^{\mathbf{ALS}}$ to $\mathbf{c}_{\mathbf{cal}_1}^{\mathbf{ref}}$, and with b_1 and $b_{0,1}$ as slope and intercept, respectively) and one for the concentration profile coming from $\mathbf{D}_2$ (relating $\mathbf{c}_{\mathbf{cal}_2}^{\mathbf{ALS}}$ to $\mathbf{c}_{\mathbf{cal}_2}^{\mathbf{ref}}$, and with b_2 and $b_{0,2}$ as slope and intercept, respectively). These models are used to predict real concentration values in test samples, $\hat{\mathbf{c}}_{\mathbf{test}_1}$ and $\hat{\mathbf{c}}_{\mathbf{test}_2}$.

Since sample matrix effects exist, direct update in **C** matrix by the actual concentrations predicted by the two models is not possible. Therefore, $\hat{\mathbf{c}}_{\mathbf{test}_1}$ and $\hat{\mathbf{c}}_{\mathbf{test}_2}$ are stored in a separate output and they are used to obtain the constrained concentration profile to be introduced in the ALS optimization by means of a rescaling step. In this step, the model coming from subset $\mathbf{D}_1$ (with b_1 and $b_{0,1}$ parameters) is adopted as reference, and $\mathbf{c}_{\mathbf{cal}_2}^{\mathbf{ref}}$ and $\hat{\mathbf{c}}_{\mathbf{test}_2}$ are rescaled according the following equation:

$$\mathbf{C}_2^{\mathbf{new}} = \frac{b_2\begin{bmatrix}\mathbf{c}_{\mathbf{cal}_2}^{\mathbf{ref}}\\ \hat{\mathbf{c}}_{\mathbf{test}_2}\end{bmatrix} + b_{0,2} - b_{0,1}}{b_1} \quad (7)$$

Rescaled values for calibration and test samples, $\mathbf{C}_2^{\mathbf{new}}$, are then updated in the corresponding concentration profile in **C** matrix, and ALS optimization continues.

3.3. Figures of merit (FOM)

In order to evaluate the prediction ability of the multivariate calibration models, a set of validation samples was used. The number of validation samples was about one third of the total number of samples. From the predicted $\hat{c}$ values for these samples, some figures of merit [45] were calculated according to the following expressions.

Root mean square error of prediction (RMSEP),

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(c_i-\hat{c}_i)^2}{n}} \quad (8)$$

Standard error of prediction (SEP),

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(c_i-\hat{c}_i-\mathrm{bias})^2}{n-1}} \quad (9)$$

Bias,

$$\mathrm{bias} = \frac{\sum_{i=1}^{n}(c_i-\hat{c}_i)}{n} \quad (10)$$

Relative percentage error in concentration predictions (RE %),

$$\mathrm{RE}\ (\%) = 100\sqrt{\frac{\sum_{i=1}^{n}(c_i-\hat{c}_i)^2}{\sum_{i=1}^{n}c_i^2}} \quad (11)$$

where c_i and $\hat{c}_i$ are the actual and predicted analyte concentration in sample i, respectively, and n is the total number of samples used in the validation set.

A linear regression fit was performed between actual and predicted analyte concentration. Slope, offset and square correlation coefficient were also calculated. To check the similarity between experimental and MCR-ALS recovered pure component spectral

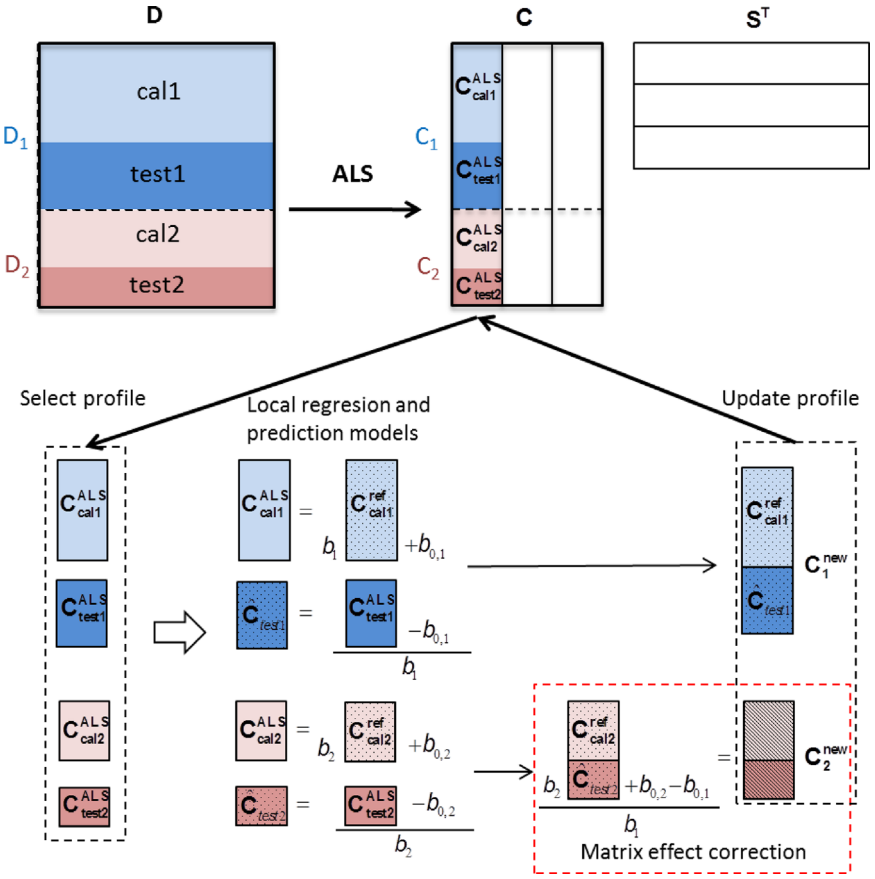

Fig. 1. Description of the correlation constraint with matrix effect correction in a data set formed by sample subsets with different behaviors.

profile, a correlation coefficient was calculated. This parameter gave a measure of how similar the shape of the individual recovered spectral profile is to the real experimental pure component spectrum.

3.4. Chemometrics software

Data pre-processing and PLS analysis were carried out using PLS Toolbox software package (Eigenvector Research, Manson, WA, USA) for MATLAB (The MathWorks, Natick, MA, USA). Calculation of figures of merit and MCR-ALS models were performed with laboratory-written MATLAB routines and functions. A graphical user interface for MCR-ALS was proposed by Jaumot et al. [41] and can be freely downloaded from the MCR web page [46]. The MCR-ALS variant with the correlation constraint is available from request to the authors.

4. Results and discussion

4.1. Analysis of diesel and biodiesel blends by NIR spectroscopy

Fig. 2a and b shows the NIR spectra for the 36 diesel and biodiesel blends before and after preprocessing, respectively. The first spectral range was multiplied by a constant scaling factor of 2.5 in order to balance the intensity differences between the two ranges. The biodiesel determination was possible because the presence of second overtone, combination mode and part of the second overtone of the vibrational modes of C–H bond stretching located in the 1150–1250 nm, 1350–1450 and 1630–1677 nm range, respectively. The combination band of the C=O stretching and aliphatic C–H bond stretching is also observed at about 2150 nm [47,48]. NIR spectra of biodiesel have some specific band regions comparing with the feedstock vegetal oil, that can be used to detect a probable adulteration [49,50].

As mentioned in Section 3.1 these spectra form a data set with two batches of 28 aged biodiesel blend samples and eight fresh biodiesel samples. Considering the potential possibility that a calibration model could be constructed with sets of calibration samples with different ages of preparation, the natural aging of the first batch was studied to test whether there were differences on the measurement compared to a freshly prepared batch. The time chosen was arbitrary, but can be studied in further experiments. It is worth to mention that the samples were stored in suitable conditions recommended by fuel agencies and in three months the samples did not degrade.

Two major components were detected in this data set by SVD analysis, assigned to diesel and biodiesel, respectively. Initial estimates of NIR pure spectra of these two components were obtained as described in section 3.2.1. Before starting MCR-ALS, a single pure spectrum of neat biodiesel was added to the data set, producing a final data structure formed by the sample subset of the first batch, the sample subset of the second batch and the pure biodiesel spectrum.

Classical MCR-ALS was applied to this data set using non-negativity constraints in the spectral and concentration profiles. Selectivity constraint was applied to the pure biodiesel spectrum. Fig. 3a shows the scatter plot of MCR-ALS relative concentration values vs. real reference concentrations. MCR-ALS recovered biodiesel and diesel spectral profiles are shown in the Fig. 3b. Correlation coefficients higher than 0.999 were achieved for both components when the similarity of the recovered and experimental pure spectra was compared. Fig. 3a shows different linear trends between $\mathbf{c}^{ALS}$ concentration values and $\mathbf{c}^{ref}$ values in the two batches of biodiesel

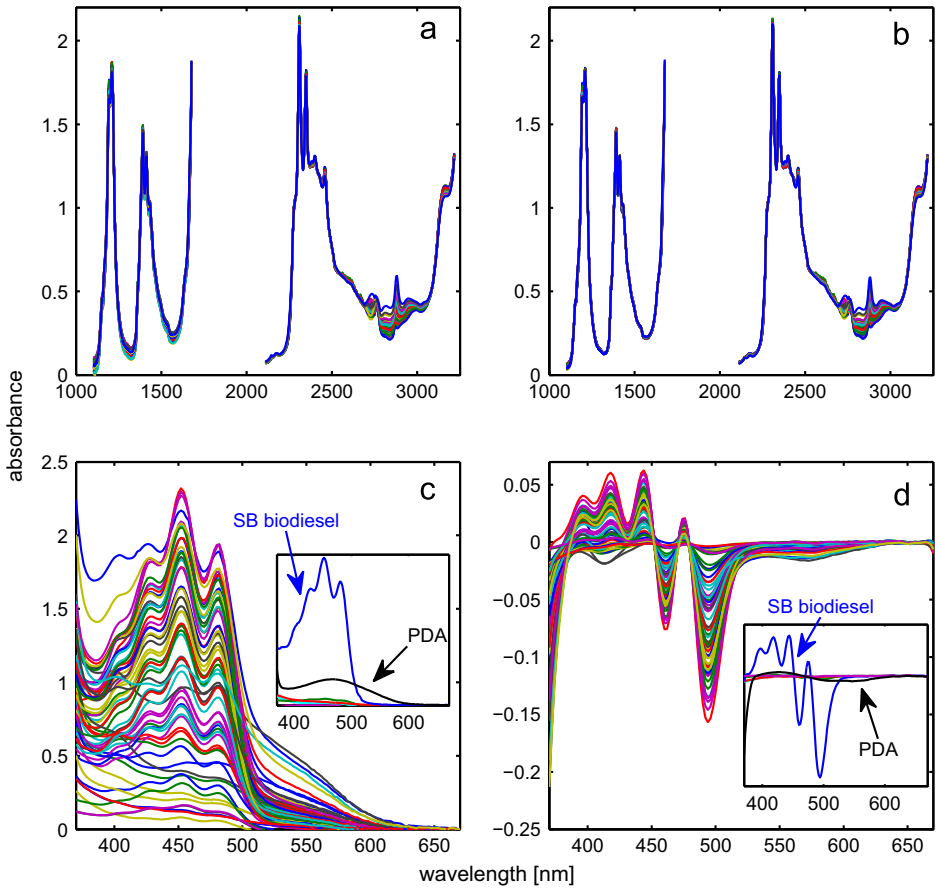

Fig. 2. NIR spectra of biodiesel blends before (a) and after MSC preprocessing (b). UV–visible spectra of the 62 antioxidant and biodiesel mixtures before (c) and after first order Savitzky–Golay derivative preprocessing (d). Inset figures in (c, d) show the pure experimental spectra of the compounds used to prepare the synthetic mixtures in the second set of samples before and after derivation, respectively. The spectra of the two compounds with high absorptivity, SB biodiesel and PDA, are assigned with the arrows.

samples. The reason for this could be the long time of storage of the first batch compared to the second one, and possible natural ageing of sample mixtures. This indicates that a single calibration model involving both batches would lead to erroneous results.

An external calibration curve was constructed, similar to a univariate calibration, with relative concentration values obtained by application of classical MCR-ALS (without correlation constraint) and real concentrations of biodiesel in calibration samples. Concentrations of test samples were then predicted using this calibration curve and concentration values for these samples estimated by MCR-ALS. Figures of merit for this prediction model are displayed in Table 2 (model 1). As suspected, this model did not provide very good results, as it can be seen from a relative low correlation coefficient (r^2=0.930) and a relative high error (RE (%)=13.58%) for biodiesel prediction in the test samples.

MCR-ALS with a single regression model and correlation constraint (applied to first and second subsets) gave similar results (Table 2, model 2) because of the presence of the sample matrix effect. Fig. 4a shows the quality of the predictions obtained using this model.

Results obtained by MCR-ALS with the new correlation constraint strategy involving two local regression models to correct sample matrix effects are shown in Table 2 for model 3. Sample matrix effects were suppressed and improvement of the correlation coefficient (r^2=0.992) and a decrease of the relative error (RE (%)=4.85%) was observed compared with previous results where a single regression model was applied (model 2). Fig. 5 shows the regression plots of MCR-ALS predicted biodiesel concentrations vs. reference concentrations when MCR-ALS with correlation constraint involving local regression models and matrix effect correction is applied.

The same preprocessed data sets for calibration and test used in MCR-ALS analysis were also used for comparison with PLS regression models. When PLS was applied, the matrix of preprocessed data was divided in two input matrices, one with the calibration sample spectra and the other with the test samples spectra as required by the algorithm. The leave-one-out cross validation method was used for determination of the number of PLS latent variables by evaluating the evolution of the root mean square error of cross-validation (RMSECV). The optimum number of latent variables was that with the lowest RMSECV or when no reduction is observed. The cross-validation model indicated two latent variables, but better results were achieved using three, as explained below.

Model 5 in Table 2 shows the results obtained when PLS regression with two latent variables was employed in the NIR data for prediction of biodiesel concentration. Fig. 4c shows the regression plot for the predictions of biodiesel content. The same matrix effect problem was observed in the representation of predicted vs. reference values. A slight reduction in the error parameters was observed in comparison with model 1, but the model was clearly worse than the MCR-ALS model correcting sample matrix effects (model 3). A possible strategy to alleviate sample matrix effects in PLS modeling is including more latent variables in the PLS models. Model 6 shows the results when three latent variables were used. A reduction in the error parameters was observed and there were no differences in the linear trends of the predicted vs. actual concentration values between the two batches, as observed in the Fig. 4d.

MCR-ALS with correlation constraint and a single regression model was also applied with three components (model 4) for

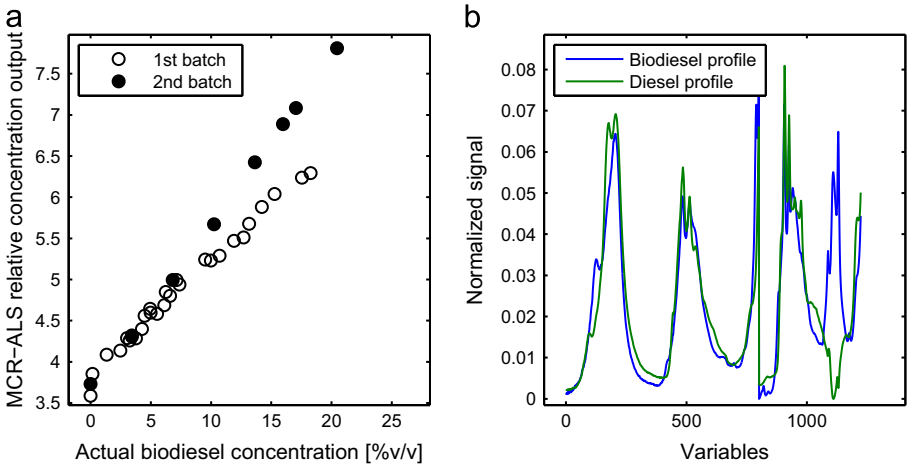

Fig. 3. (a) Plot of relative concentration profile of biodiesel obtained by classical MCR-ALS optimization vs. actual biodiesel concentration. (b) MCR-ALS recovered spectral profiles.

Table 2
Figures of merit of MCR-ALS and PLS regression models for prediction of biodiesel concentration.

Model description	#Comp.	RMSEC[a]	RMSEP[a]	SEP[a]	Bias[a]	RE (%)	r^2
1. MCR-ALS (classical model without correlation constraint and with external calibration)	2	1.43	1.40	1.40	0.392	13.58	0.930
2. MCR-ALS (single calibration model involving the first and second subsets)	2	1.43	1.41	1.41	0.411	13.63	0.930
3. MCR-ALS (different local calibration models with matrix effect correction)	2	0.442	0.502	0.507	0.126	4.85	0.992
4. MCR-ALS (single calibration model involving the first and second subsets)	3	0.866	1.04	1.06	0.210	10.06	0.961
5. PLS 2LV	2	1.28	1.36	1.37	0.363	13.14	0.938
6. PLS 3LV	3	0.797	1.02	1.04	0.196	9.82	0.965

#Comp. is the number of components or latent variables used in the MCR-ALS or PLS model, respectively.
r^2 is the correlation coefficient between predicted and actual concentration values of test samples.
[a] The values are given in (%v/v) units.

comparison with PLS 3LV model. Results shown in Table 2 for this model exhibit similar behavior when compared to PLS 3LV model. Despite the fact that the number of chemical compounds was two, three components were necessary to reduce matrix effects when a single regression model was used, as shown in Fig. 4b. None of the single models (either PLS or MCR-ALS) with three components outperformed the MCR-ALS model with two components when correlation constraint considering matrix effect correction was applied. The strategy including sample matrix effect correction requires that a subset of samples could be described by the same local calibration model, i.e., the matrix effect should be identical in all samples of the subset, such as in a batch. When this is not the case, the problem of sample matrix effect can be alleviated increasing the size of single MCR-ALS or PLS models [33].

4.2. Analysis of biodiesels and antioxidant mixtures by UV–visible spectroscopy

UV–visible absorption data were employed to show the application of the correlation constraint for the determination of one kind of biodiesel and one antioxidant at low concentration level in a mixture with biodiesels from different feedstocks. Fig. 2c and d shows the original and the preprocessed spectra of the 62 sample mixtures, respectively. For the biodiesel mixture, the main spectral contribution was from SB biodiesel due to the high absorptivity of chromophores that absorb in broad bands from 400–500 nm. Antioxidant concentrations in the investigated samples were rather low and only PDA (*N*,*N*′-Di-*sec*-butyl-*p*-phenylenediamine) contributed appreciably to the overall signal, but very overlapped by SB biodiesel bands, as shown in the inset figure in Fig. 2c. The rest of compounds had lower signals and were completely overlapped by SB biodiesel compound signal, as observed in the inset figures in Fig. 2c and d. First order Savitzky–Golay derivative enhances overlapped band differences between SB biodiesel and PDA in the original spectra, as observed in Fig. 2d.

SVD analysis estimated a number of components between four and six. Further analysis by MCR-ALS showed that six components provided better prediction results. MCR-ALS spectral initial estimates for the six components were estimated by a method based on SIMPLISMA using non-preprocessed data, avoiding negative values present in derivative spectra, not suitable for SIMPLISMA. Due to negative spectra values in preprocessed data, non-negativity constraints were only applied to concentration profiles. Pure experimental SB biodiesel and PDA antioxidant spectra were added to the data set to introduce the selectivity constraint.

This strategy allows a better recovery of the spectral profiles by MCR-ALS models, mainly for PDA, because of its low spectral signal intensity in comparison to SB biodiesel spectrum.

MCR-ALS was firstly applied in the traditional way to the data and external univariate calibration models were built with final output concentration profiles. Results for this traditional approach are depicted in Table 3 for both SB biodiesel and PDA antioxidant. A high relative error was found for PDA prediction (RE (%)= 40.08%), because of its low concentration range (1–1006 ppm) in the analyzed mixtures and of its low spectral contribution to the global measured signal.

Table 3 shows the MCR-ALS results including the correlation constraint applied to the sample subset formed by the mixtures of biodiesels and antioxidants for the determination of SB biodiesel and PDA antioxidant. The figures of merit associated with these determinations proved that there was an improvement with

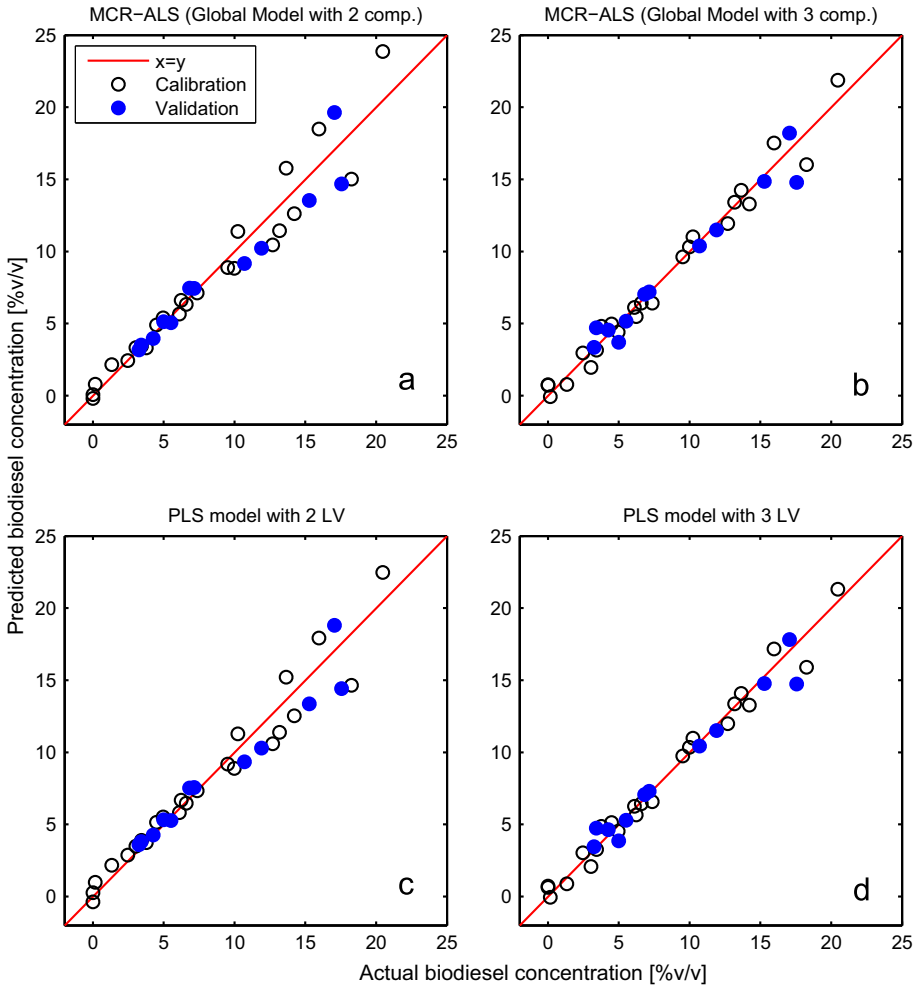

Fig. 4. Plot of predicted vs. actual biodiesel concentration. Results from MCR-ALS models with (a) two components (comp.) (model 2) and (b) three components (model 4); results from PLS models with (c) two latent variables (LV) (model 5) and (d) three latent variables (model 6). See model description in Table 2.

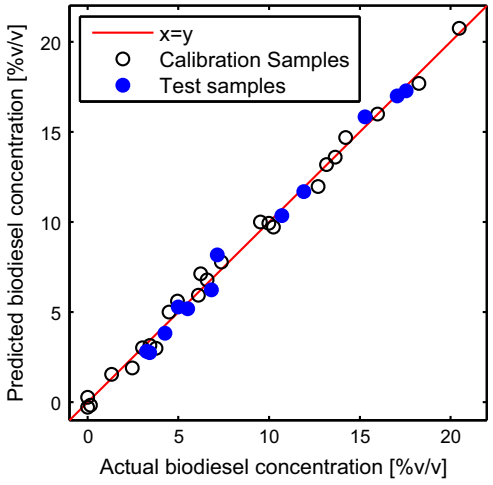

Fig. 5. Plot of predicted vs. actual biodiesel concentrations obtained using MCR-ALS with local regression models and matrix effect correction (see model 3 in Table 2).

respect to the results obtained by MCR-ALS analysis without using the correlation constraint. Slight reduction of error parameters for SB biodiesel calibration and a great improvement for the calibration model of the minor compound PDA were obtained. In this case, for prediction of PDA in the test set, a reduction of RE (%) from 40.08% to 3.16% was observed as well as an increasing of the r^2 from 0.929 to 0.997. The recovered spectral profiles of the analytes were in agreement with pure experimental spectra with correlation coefficients above 0.999.

To assess the performance of the correlation constraint for quantitative analysis, analogous PLS regression models were also constructed using the same preprocessed calibration and test samples. The number of latent variables chosen by cross-validation was equal to 5. Results in Table 3 show that MCR-ALS with correlation constraint and PLS have comparable performances. However, MCR-ALS has the advantage of providing additional meaningful information associated with recovered spectral profiles, which can be used for component identification and confirmation.

5. Conclusions

MCR-ALS with the correlation constraint has been demonstrated to be a useful and accurate tool for determination of compounds of interest in biodiesel samples when sample matrix effect exists and when minor compounds with overlapped signals have to be determined.

Thus, MCR-ALS was employed for determination of biodiesel in biodiesel blends using NIR spectroscopy. The proposed modification of the correlation constraint using local regression models to correct batch-to-batch sample matrix effects due to ageing of biodiesel blends highly improved the predictive ability of the algorithm when compared to the application of this constraint

Table 3
Figures of merit of MCR-ALS models for prediction of SB biodiesel and PDA concentration.

Compound	Model description	#Comp.	RMSEC[a]	RMSEP[a]	SEP[a]	Bias[a]	RE (%)	r^2
SB biodiesel								
	MCR-ALS (classical with external calibration)	6	0.986	0.728	0.637	0.38	2.68	0.998
	MCR-ALS (correlation constraint)	6	0.585	0.515	0.508	0.140	1.9	0.999
	PLS	5	0.237	0.2666	0.272	0.028	0.983	0.999
PDA								
	MCR-ALS (classical with external calibration)	6	82.99	176.4	158.5	-85.2	40.08	0.929
	MCR-ALS (correlation constraint)	6	9.16	13.90	14.26	0.432	3.16	0.997
	PLS	5	6.57	8.75	8.97	0.282	1.99	0.999

[a] Values given in (%wt) and (ppm) units for SB biodiesel and PDA, respectively.

with a single regression model. This implementation of MCR with correlation constraint and matrix effect correction also outperformed standard multivariate calibration models built with PLS.

When the sample matrix effect is variable or cannot be easily associated with sample subsets, increasing the number of components in the model, both in MCR-ALS with correlation constraint and a single regression model and PLS, improved the results obtained. In this situation, MCR-ALS provided slightly worse results than PLS, but had the advantage of providing qualitative (spectral) information about the compounds analyzed.

The correlation constraint was also applied to a rather complex case where a minor antioxidant compound was determined in the presence of a biodiesel mixture with highly overlapped UV–visible spectra. It was shown that the introduction of selective spectral information for the minor compound and the use of the correlation constraint were crucial to improve both the quality of the UV–visible recovered pure spectrum and the quantitative prediction of this compound in the analyzed mixtures. The relative error of prediction obtained by MCR-ALS with correlation constraint was higher than that provided by PLS, but MCR-ALS could provide spectral profiles of the analytes and interferences.

Acknowledgments

This work was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil, 238577/2012-0 grant). R.R. de Oliveira acknowledges the Brazilian interchange scholarship program (Ciência sem Fronteiras – CAPES/CNPq). R. Tauler acknowledges the CAPES (70/2012) for the visiting researcher grant financial support. LCL-UFRN and Lamoc-Inmetro for providing samples and sample analysis. A. de Juan acknowledges support of Spanish project CTQ 2012-38616.

References

[1] G. Knothe, R.O. Dunn, M.O. Bagby, ACS Symp. Ser. 666 (1997) 172.
[2] M.R. Monteiro, A.R.P. Ambrozin, L.M. Lião, A.G. Ferreira, Talanta 77 (2008) 593.
[3] J. Janaun, N. Ellis, Renew. Sustain. Energy Rev. 14 (2010) 1312.
[4] F.H.N. Souza, L.R. de Almeida, F.S.C.L. Batista, M.A. de S. Rios, Energy Sci. Technol. 2 (2011) 56.
[5] F.V.C. de Vasconcelos, P.F.B. de Souza, M.F. Pimentel, M.J.C. Pontes, C.F. Pereira, Anal. Chim. Acta 716 (2012) 101.
[6] D.D.S. Fernandes, A.A. Gomes, G.B. da Costa, G.W.B. da Silva, G. Véras, Talanta 87 (2011) 30.
[7] W. Zhang, Renew. Sustain. Energy Rev. 16 (2012) 6048.
[8] J.C.L. Alves, R.J. Poppi, Talanta 104 (2013) 155.
[9] R.M. Balabin, E.I. Lomakina, R.Z. Safieva, Fuel 90 (2011) 2007.
[10] V. Gaydou, J. Kister, N. Dupuy, Chemom. Intell. Lab. Syst. 106 (2011) 190.
[11] G. Veras, A.D.A. Gomes, A.C. da Silva, A.L.B. de Brito, P.B.A. de Almeida, E.P. de Medeiros, Talanta 83 (2010)565 83 (2010).
[12] A.R.L. Caires, V.S. Lima, S.L. Oliveira, Renew. Energy 46 (2012) 137.
[13] M. Insausti, A.A. Gomes, F.V. Cruz, M.F. Pistonesi, M.C.U. Araujo, R.K.H. Galvão, C.F. Pereira, B.S.F. Band, Talanta 97 (2012) 579.
[14] R.O. Dunn, Fuel Process. Technol. 86 (2005) 1071.
[15] I.P. Aquino, R.P.B. Hernandez, D.L. Chicoma, H.P.F. Pinto, I.V. Aoki, Fuel 102 (2012) 795.
[16] W.W. Focke, I. Van Der Westhuizen, A.B.L. Grobler, K.T. Nshoane, J.K. Reddy, A.S. Luyt, Fuel 94 (2012) 227.
[17] M. Lapuerta, J. Rodríguez-Fernández, Á. Ramos, B. Álvarez, Fuel 93 (2012) 391.
[18] T.F. Tormin, R.R. Cunha, E.M. Richter, R.A.A. Muñoz, Talanta 99 (2012) 527.
[19] R.M. Alberici, R.C. Simas, P.V. Abdelnur, M.N. Eberlin, V. de Souza, G.F. de Sá, Energy Fuels 24 (2010) 6522.
[20] L.N. Batista, V.F. Silva, M.G. Fonseca, E.C.G. Pissurno, R.J. Daroda, V.S. Cunha, C.N. Kunigami, L.C. de Santa Maria, Microchem. J. 106 (2013) 17.
[21] V. del Río, M.S. Larrechi, M.P. Callao, Anal. Chim. Acta 676 (2010) 28.
[22] V. del Río, M.S. Larrechi, M.P. Callao, Talanta 81 (2010) 1572.
[23] N.G.S. Mogollon, F.A.D.L. Ribeiro, M.M. Lopez, L.W. Hantao, R.J. Poppi, F. Augusto, Anal. Chim. Acta 796 (2013) 130.
[24] A. de Juan, R. Tauler, Anal. Chim. Acta 500 (2003) 195.
[25] A. de Juan, R. Tauler, Crit. Rev. Anal. Chem. 36 (2006) 163.
[26] A. de Juan, R. Tauler, J. Chromatogr. A 1158 (2007) 184.
[27] R. Tauler, A. Smilde, B.R. Kowalski, J. Chemom. 9 (1995) 31.
[28] J. Jaumot, B. Igne, C.A. Anderson, J.K. Drennen, A. de Juan, Talanta 117 (2013) 492.
[29] R. Tauler, B.R. Kowalski, S. Fleming, Anal. Chem. 65 (1993) 2040.
[30] M.C. Antunes, J.E.J. Simão, A.C. Duarte, R. Tauler, Analyst 127 (2002) 809.
[31] S.E. Richards, E. Becker, R. Tauler, A.D. Walmsley, Chemom. Intell. Lab. Syst. 94 (2008) 9.
[32] H.C. Goicoechea, A.C. Olivieri, R. Tauler, Analyst 135 (2010) 636.
[33] T. Azzouz, R. Tauler, Talanta 74 (2008) 1201.
[34] L.B. Lyndgaard, F. van den Berg, A. de Juan, Chemom. Intell. Lab. Syst. 125 (2013) 58.
[35] S. Wold, A. Ruhe, H. Wold, I.I.I. Dunn, W. SIAM, J. Sci. Stat. Comput. 5 (1984) 735.
[36] H. Martens, T. Næs, Multivariate Calibration, 1st ed., Wiley, Chichester, UK, 1992.
[37] K.E. Kramer, R.E. Morris, S.L. Rose-Pehrsson, Chemom. Intell. Lab. Syst. 92 (2008) 33.
[38] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627.
[39] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137.
[40] R. Tauler, Chemom. Intell. Lab. Syst. 30 (1995) 133.
[41] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, Chemom. Intell. Lab. Syst. 76 (2005) 101.
[42] G.H. Golub, C.F. van Loan, Matrix Computation, JHU Press, Baltimore, US, 1996.
[43] W. Windig, J. Guilment, Anal. Chem. 63 (1991) 1425.
[44] R. Bro, S. Jong, J. Chemom. 11 (1997) 393.
[45] A.C. Olivieri, N.M. Faber, J. Ferré, R. Boqué, J.H. Kalivas, H. Mark, Pure Appl. Chem. 78 (2006) 633.
[46] ⟨http://www.mcrals.info⟩ (accessed 24.02.14).
[47] J.S. Oliveira, R. Montalvão, L. Daher, P.A.Z. Suarez, J.C. Rubim, Talanta 69 (2006) 1278.
[48] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14.
[49] G. Knothe, J. Am. Oil Chem. Soc. 76 (1999) 795.
[50] F.C.C. Oliveira, C.R.R. Brandão, H.F. Ramalho, L.A.F. da Costa, P.A.Z. Suarez, J.C. Rubim, Anal. Chim. Acta 587 (2007) 194.