Available online at www.sciencedirect.com

ScienceDirect

Analytica Chimica Acta 595 (2007) 107–113

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Multivariate near infrared spectroscopy models for predicting methanol and water content in biodiesel

Pedro Felizardo[a], Patrícia Baptista[a], José C. Menezes[b], M. Joana Neiva Correia[a,*]

[a] Centre of Chemical Processes, IST, Technical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
[b] Centre for Biological and Chemical Engineering, IST, Technical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

Received 18 October 2006; received in revised form 31 January 2007; accepted 21 February 2007

Abstract

The transesterification of vegetable oils, animal fats or waste oils with an alcohol (such as methanol) in the presence of a homogeneous catalyst (sodium hydroxide or methoxyde) is commonly used to produce biodiesel. The quality control of the final product is an important issue and near infrared (NIR) spectroscopy recently appears as an appealing alternative to the conventional analytical methods. The use of NIR spectroscopy for this purpose first involves the development of calibration models to relate the near infrared spectrum of biodiesel with the analytical data. The type of pre-processing technique applied to the data prior to the development of calibration may greatly influence the performance of the model. This work analyses the effect of some commonly used pre-processing techniques applied prior to partial least squares (PLS) and principal components regressions (PCR) in the quality of the calibration models developed to relate the near infrared spectrum of biodiesel and its content of methanol and water. The results confirm the importance of testing various pre-processing techniques. For the water content, the smaller validation and prediction errors were obtained by a combination of a second order Savitsky–Golay derivative followed by mean centring prior to PLS and PCR, whereas for methanol calibration the best results were obtained with a first order Savitsky–Golay derivative plus mean centring followed by the orthogonal signal correction.

Keywords: Biodiesel; Near infrared; Calibration models; Data pre-processing

1. Introduction

There are several advantages of using biodiesel as a fuel in diesel motors, such as the reduction of the greenhouse gases emissions, the increasing of the eco-efficiency and, if waste frying oils (WFO) are used as the raw-material for biodiesel production, the treatment of industrial and household wastes [1,2].

The production of biodiesel may be achieved by a homogeneous (sodium hydroxide or methoxyde) catalysed transesterification reaction between a lipid (vegetable oils and fats) and a short chain alcohol, such as methanol, to produce an ester and a by-product, glycerol. This reaction occurs stepwise, with mono and diglycerides as intermediate products [3]. At the end of the reaction period, the glycerol rich-phase is separated from the ester layer by decantation or centrifugation. After separation, the biodiesel phase is contaminated with mono, di and triglycerides, methanol, catalyst, free glycerol and soaps and has to be purified to comply with the European Standard EN 14214 [4]. The washing of the esters phase with water followed by vacuum drying is the most commonly used process for biodiesel purification [5].

Since it is possible to produce biodiesel from several different feedstocks and technologies, the quality control of the final product is of great concern and the European Standard EN 14214 [4] establishes 25 parameters that have to be analysed to certify biodiesel quality. Among these, the contents in water and methanol are two important parameters [3]. In fact, the use of biodiesel contaminated with water can cause corrosion in the engine or the reaction with the glycerides to produce soaps and glycerine. The EN imposes, therefore, a maximum content of 0.05% (m/m) of water in fuels. Methanol is responsible for metal corrosion, particularly of aluminium, as well as for the decreasing of the fuel flash point. As such, a maximum content of 0.2% (m/m) of methanol in biodiesel is specified [4]. Biodiesel analyses are very expensive and time consuming and

---

[*] Corresponding author. Tel.: +351 21 8417344; fax: +351 21 8417246.
E-mail address: qjnc@ist.utl.pt (M.J.N. Correia).

0003-2670/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2007.02.050

NIR spectroscopy appears as a cheaper and faster alternative to perform the quality control of biodiesel [6–9].

The use of NIR spectroscopy in combination with multivariate data analysis for the analysis of biofuels and other complex matrices has been reported in recent papers [3–9]. NIR spectroscopy is a well-established analytical technique based on the absorption of electromagnetic energy in the region from 700 to 2500 nm. This technique enables the analysis of multicomponent samples in a fast and non-destructive way, without requiring complex pre-treatments. The use of partial least squares (PLS) or principal components regression (PCR) allows the development of calibration models between spectral and analytical data [10–13]. This work analyses the effect of applying several commonly used pre-processing techniques, prior to the application of PLS and PCR, in the quality of the calibration models developed to relate the near infrared spectrum of biodiesel and its content of methanol and water.

2. Experimental

Industrial-scale and laboratory-scale samples of biodiesel produced from soybean, mixtures of soybean and palm, and from waste frying oils were prepared according to the procedure presented elsewhere [3,9]. Industrial samples of biodiesel produced from soybean, palm and waste frying oils were supplied by two Portuguese industrial companies.

The reference method for water determination was the Karl Fisher titration [14] that was performed in a Metrohm 682 titroprocessor, while methanol content was analysed by Head Space Gas Chromatography [15] using a HP 5890 equipped with a PoraPlot Q packed column (3 m long).

The near-infrared diffuse transflectance spectra of the biodiesel samples were acquired using an ABB BOMEM MB160 spectrometer equipped with an InGaAs detector and a transflectance probe from SOLVIAS. Spectra were recorded in duplicate for each sample at room temperature (22–24 °C), with the aid of the Galactic Grams software package, in the wave number range of 12,000–4000 cm^{-1}, with a spectral resolution of 16 cm^{-1}. The average of the two measurements was used for model development.

2.1. Data analyses and calibration development

All calculations were carried out using Matlab Version 6.5 (MathWorks, Natick, MA) and the PLS Toolbox Version 3.0 (Eigenvector Research Inc., USA) for Matlab. Partial least squares (PLS) and principal components regressions (PCR) were used to develop the calibration models with the spectral and the analytical data. Both methods search for linear combinations, named as factors or components, of the original X-values and use only these linear combinations in the regression equation of the model that relates the spectra X with a given property of interest, y, in this case the content of water or methanol in biodiesel [11]. However, the approach followed in each case to derive the model is different. In fact, PCR selects the components according to their ability to account for the variability in X, without using information about y. On the other hand, instead of using components, PLS uses factors determined by employing both X and y in estimation; for PLS regression each component is obtained by maximizing the covariance between y and all possible linear combinations of the columns of X. This leads to components, which are more directly related to the variability in y than the principal ones of the PCR approach [11].

Prior to PCR or PLS regressions, various widely used pre-processing techniques described in literature [12], were applied to the data. This work presents the calibration results obtained from untreated data (identified in the tables below as none) and from pre-treated values, using the following methods: Mean Centering (MC); Mean Scattering Correction (MSC) followed by MC (MSC + MC); Standard Normal Variate scaling (SNV) plus MC (SNV + MC); first and second order Savitsky–Golay derivative followed by mean centering (SV1 + MC and SV2 + MC, respectively); MC followed by the Orthogonal Signal Correction method (MC + OSC) and, finally, the SV1 + MC and SV2 + MC followed by OSC (SV1 + MC + OSC, SV2 + MC + OSC, respectively).

The Orthogonal Signal Correction is a method developed to reduce the data variance in the spectra (X) due to light scatter effects and to more general types of interferences that have no correlation with the measured property y (water or methanol content). The idea is that all the information in the spectrum related to y should be considered rather than removed [11].

One of the most important steps in developing a reliable calibration model between the NIR spectrum and the analytical data is the selection of the optimum number of latent variables (LV) or principal components (PC) to be used. There are several methods described in the literature to select this number, such as the akaike information criterion [16], bootstrap [17,18], cross validation [11,17,19], ICOMP criterion for PCR [20] and conditional model dimensionality test for PLS [21]. Among these, the method used in this work, LOOCV or leave one out cross-validation, is one of the most commonly used criteria. Therefore, all the PCR and PLS regressions were developed using the LOOCV method to determine the optimum number of latent variables or principal components. Additionally, the conditional model dimensionality test was also used for the PLS regressions to check any possible overfitting of the data. For both water and methanol models, this test confirmed the choice of the latent variables number obtained by applying the cross-validation method, whereas the calculated risk of over-fitting was lower than 1%.

The detection of outliers was performed based on the leverage values, Q-residuals, and Studentized y-residuals. Thus, a sample was considered to be an outlier if its leverage value was twice as large as the average leverage value (given by $2(1 + \mathrm{LV})/N$ where LV is the number of latent variables and N the number of samples), or if its Q- residual falls above the 95% confidence limits for the considered model, or yet if y-residual of the sample was larger than twice the residual standard deviation [12].

The calibration models were developed taking into account the number of LV or PC, and by calculating the root mean square errors of the cross-validation, RMSECV, and of the external validation, RMSEP and the determination coefficients, Q_{Y}^2, between the predicted and the measured values [10,12]. The later coefficient, calculated using Eq. (1), quantifies the amount of variance