Talanta 104 (2013) 155–161

Contents lists available at SciVerse ScienceDirect

Talanta

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

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

Julio Cesar L. Alves *, Ronei J. Poppi

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

ARTICLE INFO

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

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

ABSTRACT

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



## 1. Introduction

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

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

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

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

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

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

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

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

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

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

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

### 1.1. Biodiesel content determination in diesel fuel blends

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

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

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

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

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

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

### 1.2. Support vector machines

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

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

based on data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, by minimizing the regularized risk functional:

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

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