Fuel 116 (2014) 123–130

Contents lists available at ScienceDirect

Fuel

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

Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration

Paulo R. Filgueiras [a], Cristina M.S. Sad [b], Alexandre R. Loureiro [b], Maria F.P. Santos [b], Eustáquio V.R. Castro [b], Júlio C.M. Dias [c], Ronei J. Poppi [a,*]

[a] Institute of Chemistry, State University of Campinas, POB: 6154, 13084-971 Campinas, Brazil
[b] Department of Chemistry, Federal University of Espírito Santo, Laboratory of Research and Development of Methodologies for Analysis of Oils, Av. Fernando Ferrari, 514 Goiabeiras, Vitória 29075-910, Espírito Santo, Brazil
[c] CENPES/PETROBRAS, Av. Jequitiba 950, Rio de Janeiro 21941-598, Brazil

HIGHLIGHTS

- API gravity, kinematic viscosity and water content were determined in petroleum oil.
- ATR-FTIR technique associated with multivariate calibration was applied for determinations.
- SVR and PLS were used for multivariate calibration.
- The SVR model was more accurate than PLS for API gravity determination.
- For kinematic viscosity and water content the two methods were equivalent.

ARTICLE INFO

Article history:
Received 24 October 2012
Received in revised form 26 July 2013
Accepted 29 July 2013
Available online 14 August 2013

Keywords:
Crude oil
ATR-FTIR
Partial least squares regression
Support vector regression

ABSTRACT

In this work, API gravity, kinematic viscosity and water content were determined in petroleum oil using Fourier transform infrared spectroscopy with attenuated total reflectance (FT-IR/ATR). Support vector regression (SVR) was used as the non-linear multivariate calibration procedure and partial least squares regression (PLS) as the linear procedure. In SVR models, the multiplication of the spectra matrix by support vectors resulted in information about the importance of the original variables. The most important variables in PLS models were attained by regression coefficients. For API gravity and kinematic viscosity these variables correspond to vibrations around 2900 cm^{-1}, 1450 cm^{-1} and below to 720 cm^{-1} and for water content, between 3200 and 3650 cm^{-1}, around 1650 cm^{-1} and below to 900 cm^{-1}. The SVR model produced a root mean square error of prediction (RMSEP) of 0.25 for API gravity, 22 $mm^2\ s^{-1}$ for kinematic viscosity and 0.26% v/v for water content. For PLS models, the RMSEP values for API gravity was 0.38 $mm^2\ s^{-1}$, for kinematic viscosity was 27 $mm^2\ s^{-1}$ and for water content was 0.34%. Using the *F*-test at 95% of confidence it was concluded that the SVR model produced better results than PLS for API gravity determination. For kinematic viscosity and water content the two methods were equivalent. However, a non-linear behavior in the PLS kinematic viscosity model was observed.

1. Introduction

Petroleum is a complex mixture of organic compounds with heterogeneous chemical composition [1]. Due to this complexity its quality in primary processing is evaluated by physicochemical properties, such as API (American Petroleum Institute) gravity, kinematic viscosity and water content. Knowledge of these parameters is essential to indicate possible changes that might occur in oil composition, and they can aid the development of transportation and refining strategies [1–4]. Also, API gravity and kinematic viscosity strongly affect the economic viability of producing fields, since, in addition to oil value, they aid in the design of the equipment used in exploration and field productivity. Even after the decision to exploit an oil field has been taken, API gravity and kinematic viscosity continue to influence the decision process, since these properties control the choice of the reservoir interval that must be completed and in which wells.

Water coming from producing wells presents suspended solids, salts, dissolved gases and microorganisms [2,5,6]. Water and sediments are undesirable contaminants that might cause problems in transportation and refining, such as corrosion of equipment, accidents during the distillation process or adverse effects on final product quality. Its measurement allows evaluating selling price,

* Corresponding author. Tel.: +55 019 35213126; fax: +55 019 35213023.
E-mail address: ronei@iqm.unicamp.br (R.J. Poppi).

0016-2361/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fuel.2013.07.122

production rates, custody transfer, pipeline oil quality control and royalties [1].

Fast determination of the physicochemical parameters of oil is necessary in order to expedite a decision on increasing production in Brazil. In recent years, infrared spectroscopy has emerged as a tool in quantitative analysis of petroleum, diesel, biodiesel or mixtures of diesel–biodiesel [7–15], whose main advantages are the need for small sample quantities and quick procedures with minimal pretreatment of sample. In these systems, the conversion of the given instrumental response of interest requires the use of multivariate calibration techniques.

The standard methodology usually used in multivariate calibration for spectral data treatment is partial least squares regression (PLS) [16,17]. This methodology has been used in several applications with infrared analysis of oil samples [12,18,19]. Although good results are often obtained, there are situations where PLS cannot be implemented in routine analysis. The main drawbacks are the presence of non-linearities or complex data samples. In these cases, many strategies have been implemented to overcome these difficulties such as: processing strategies, use of local modeling, and use of multivariate non-linear modeling based on neural networks [20] or support vector regression [21].

In this work Fourier transform infrared spectroscopy with attenuated total reflectance (FT-IR/ATR) in association with multivariate calibration based on support vector regression (SVR) and partial least squares regression (PLS) was used for determination of API gravity, kinematic viscosity and water content in medium and heavy petroleum oil.

1.1. Support vector regression (SVR)

The support vector is a machine learning method developed by Cortes and Vapnik [22], originally for solving binary classification problems. However, the technique was extended to handle multiclass problems [23,24] and regression [21,25–29]. Support vector regression (SVR) is machine learning based on statistical learning theory and seeks to maximize the ability to generalize using the structural risk minimization principle.

For ε-SVR the aim is to find a function $f(\mathbf{x})$ that has at most ε-sensitive deviation from the desired targets y_i for all the training data, and at the same time is as smooth as possible. We can describe a linear function $f(\mathbf{x})$ by the form:

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b \tag{1}$$

where the input vectors $\mathbf{x_i}$ are mapped into a high-dimensional feature space Z by the transfer kernel function ϕ. This function serves as a technique for increasing dimensions and transforming a linearly inseparable dataset, its original space, into linearly separable entities within high dimension feature space Z by the nonlinear mapped function: ϕ: $\mathbf{x_i} \rightarrow \mathbf{z_i}$. The kernel function is an important step to transform a non-linear dataset into a linear one in a high dimension feature space.

The optimal linear function is the one that minimizes the restriction function. We can write this problem as a convex optimization problem:

$$\text{minimize}: \frac{1}{2}\|\mathbf{w}\| + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{2}$$

$$\text{subject to}: \begin{cases} y_i - \mathbf{w} \cdot \phi(\mathbf{x_i}) - b \leqslant \varepsilon + \xi_i \\ \mathbf{w} \cdot \phi(\mathbf{x_i}) + b - y_i \leqslant \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geqslant 0 \end{cases} \tag{3}$$

where ε-sensitive deviated represent the amount up to which deviations are tolerated. Constant $C > 0$ represent a cost parameter, the higher its value the greater the penalty on the error of the samples outside the ε-tube. ξ_i and ξ_i^* are the slack variables introduced to account for samples that do not lie in the ε-sensitive zone. The formulation of the error function is equivalent to dealing with a so-called ε-insensitive loss function defined by:

$$L(\varepsilon) = \begin{cases} 0 & \text{if}, |L(\varepsilon) - f(\mathbf{x})| \leqslant \varepsilon \\ L(\varepsilon) - f(\mathbf{x}) & \text{otherwise} \end{cases} \tag{4}$$

That is, only the data points outside the ε-tube cause loss. With the application of the Lagrange multiplier method, the solution of this problem leads to the following regression model:

$$f(\mathbf{x}) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\mathbf{K}(\mathbf{x_i}, \mathbf{x}) + b \tag{5}$$

where α_i and α_i^* represent the Lagrange multipliers satisfying the subject to $0 \leqslant \alpha, \alpha_i^* \leqslant C$, these values are determined by solving a quadratic programming (QP) problem. The regularization parameter C should be optimized by the analyst. Only non-zero Lagrange multipliers α_i contribute to the final regression model. These data points (samples) are called support vectors, where $\mathbf{K}(\mathbf{x_i}, \mathbf{x})$ represents a kernel function. The most commonly used kernel function is the Radial Basis Function (RBF) [30]. This function is defined in Eq. (6):

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2) \tag{6}$$

For the RBF kernel, γ is a tuning parameter controlling the width of the kernel function, that can be optimized by the analyst.

However, the disadvantage of using the kernel function is that the correlation between the SVR model obtained and the original input space is lost. Üstun [29] developed a methodology for obtaining information from the original variables after SVR modeling, by using product of the spectral matrix by the support vectors on the SVR model (Eq. (7)).

$$\mathbf{p}vector_{(nx1)} = \mathbf{x}^T_{(nxm)} \cdot \alpha_{(mx1)} \tag{7}$$

The p-vectors relate information of the original variables with the support vectors generated in the SVR modeling. Therefore it is interpreted similarly to the regression coefficients in the PLS model [29].

1.2. Partial least squares

Partial least squares regression (PLS) is currently the most widely used method for multivariate calibration and is used in many applied sciences. Its theory has been widely described in the literature [17,31] and it is available in many statistical software packages.

To construct the calibration model, spectra matrix $\mathbf{X}$, as well as, the matrix of interest variables $\mathbf{Y}$ are both decomposed into a sum of latent variables h:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} = \mathbf{t}_h\mathbf{p}_h^T + \mathbf{E} \tag{8}$$

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F} = \mathbf{u}_h\mathbf{q}_h^T + \mathbf{F} \tag{9}$$

where $\mathbf{T}$ and $\mathbf{U}$ are analogous to scores matrices, and $\mathbf{P}$ and $\mathbf{Q}$ are matrices analogous to loadings of the principal component analysis. The linear relationship between the two blocks can be performed correlating scores for each component using a linear model.

The regression vector $\mathbf{b}$ is determined by the following relationship:

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q} \tag{10}$$

where $\mathbf{W}$ is the matrix of weights of the PLS model. The regression vector $\mathbf{b}$ considers the contribution of each variable to the PLS