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 $\mathrm{cm^{-1}}$, 1450 $\mathrm{cm^{-1}}$ and below to 720 $\mathrm{cm^{-1}}$ and for water content, between 3200 and 3650 $\mathrm{cm^{-1}}$, around 1650 $\mathrm{cm^{-1}}$ and below to 900 $\mathrm{cm^{-1}}$. The SVR model produced a root mean square error of prediction (RMSEP) of 0.25 for API gravity, 22 $\mathrm{mm^2\,s^{-1}}$ for kinematic viscosity and 0.26% v/v for water content. For PLS models, the RMSEP values for API gravity was 0.38 $\mathrm{mm^2\,s^{-1}}$, for kinematic viscosity was 27 $\mathrm{mm^2\,s^{-1}}$ 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*(**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*(**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: $\mathbf{x_i} \to \mathbf{z_i}$. 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:

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

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 **X**, as well as, the matrix of interest variables **Y** are both decomposed into a sum of latent variables *h*:

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

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

where **T** and **U** are analogous to scores matrices, and **P** and **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 **b** is determined by the following relationship:

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

where **W** is the matrix of weights of the PLS model. The regression vector **b** considers the contribution of each variable to the PLS model, i.e., the higher the value of **b** the more important is the variable for model calibration.

2. Experimental

In this study, 68 petroleum blend samples from three off-shore and one on-shore oil field located in the sedimentary basin of the Brazilian coast were used. These samples were analyzed in the Laboratory of Research and Development of Methodologies for Analysis of Heavy Oil (Labpetro) – Department of Chemistry (DQUI) of the Federal University of Espirito Santo (UFES), following their respective standard analyses techniques:

API gravity – API gravity (141.5 – specific gravity – 131.5) of the samples was determined according to ISO 12185-96 standard [32]. Density was determined by injecting a sample into the digital automatic densimeter analyzer Anton Paar model DMA 5000. It was measured at 50 °C then estimated at 20 °C for calculating API gravity.

Kinematic viscosity – The kinematic viscosity was determined according to ASTM D 7042-04 standard [33]. It was analyzed by injecting a sample into the digital automatic viscosimeter analyzer Anton Paar Stabinger SVM 3000. It was measured at 50 °C and 60 °C then estimated at 40 °C by regression, as described in the technical bulletin Petrobras (2004). In the sector for exploration and production of crude oil, kinematic viscosity is analyzed at 40 °C, but for very viscous oils their direct measurement at this temperature generates large errors. Thus for these oils it is measured at two higher temperatures and the value extrapolated to 40 °C.

The water content – The water content was determined by the Karl Fischer (KF) reagent method, in accordance with ASTM D 4377 standard procedures [34]. The solvent used during the analysis was a mixture of dry methanol and chloroform (20% v/v). For standardization of the KF reagent, distilled water was solubilized into the solvents. A Metrohm KF titrator (model 836 Titrando) equipped with a double platinum electrode was employed during the water content determination tests. The ASTM D 4377-00 standard covers results in the range of 0.02–2% v/v water in oil. Samples with results above this limit can be analyzed by the technique, but are not covered by this standard.

IR spectra were acquired in a BOMEM SPLA model 2000-102 mid-infrared spectrometer with a ZnSe crystal attenuated total reflectance accessory (ATR-FTIR). The spectra were measured in the region between 4000 and 646.10 cm^{-1} with 16 scans and 4 cm^{-1} resolution. A reference spectrum was recorded for room air and subtracted from the sample. The relative humidity and ambient room temperature were around 36% and 24 °C, respectively.

2.1. Model development

PLS and SVR models for API gravity, kinematic viscosity and water content were developed from the ATR-FTIR spectra. For building of calibration models, the 68 samples were split into calibration (48 samples) and prediction (20 samples) sets by Kennard-Stone algorithm [35]. The data were processed in MATLAB 7.8.0 (MathWorks Inc., Natick, MA). Multiplicative signal correction (MSC) was used for baseline correction prior to application of the models. The PLS models were prepared on the platform PLS Toolbox from Eigenvector [36] and the SVR models with the package LIBSVM [37]. In PLS modeling the data were mean centered and the "leave one out" cross-validation procedure in the calibration samples was used to determine the number of latent variables. The modeling SVR was accomplished using the kernel function RBF (radial basis function) through the routine ε-SVR. The parameters C and ε were optimized by a grid search, where one value was fixed while the other was changed. The models were compared according to the results of the statistical parameters: coefficient of multiple determination (R^2), root mean square error of cross-validation (RMSECV) and root mean square error of prediction (RMSEP), calculated by Eq (11):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{11}$$

where n is the number of samples and y_i and $\hat{y}_i$ are the reference values determined by standard methods and those predicted by PLS or SVR model, respectively. The percentage error of prediction was determined by:

$$RMSEP\% = 100\frac{RMSEP}{\bar{y}_{prediction}} \tag{12}$$

where $\bar{y}_{prediction}$ is the average of prediction set samples.

The models were compared by F-test statistics. In this case, the F-test is applied to by the ratio of two RMSEPs, verifying different sources of variability, or difference in accuracy. The improvement in accuracy can be evaluated by hypothesis testing to verify if variances are homogeneous, according to Eq. (13):

$$F_{calculed} = \frac{RMSEP_1^2}{RMSEP_2^2} \tag{13}$$

where $RMSEP_1 > RMSEP_2$. If $F_{calculed}$ is greater than $F^{critical}_{\nu1,\nu2,\alpha}$ the hypothesis of homogeneity of variances is rejected, otherwise the null hypothesis is maintained.

3. Results and discussion

The 68 petroleum samples studied have API gravities ranging between 16° and 23°, corresponding to medium and heavy oil, according to the ANP (Brazilian Agency of Petroleum, Natural Gas and Biofuels), kinematic viscosities in the range 57–429 $mm^2\ s^{-1}$ and water content ranging from 0.1 to 6.1% v/v. The results obtained by standard ASTM methods were used as reference for the development of calibration models. The spectra of the calibration samples are shown in Fig. 1.

The API gravity is the form usually employed in petroleum exploration and production sector to represent the density. Fig. 2A and B show the plot of the API gravity predicted by PLS and SVR against the reference values, respectively. Both methods showed good calibration results: R^2cv of 0.9292 and R^2p of 0.9461 for the PLS model and R^2cv of 0.9817 and R^2p of 0.9751 for the SVR model (Table 1), indicating good agreement between results predicted by the models and measured by ISO 12185-96. It can also be observed that the data have random distribution of points around the straight linear relationship.

Using the leave-one-out procedure, the PLS model showed lower cross-validation error (RMSECV) using 6 latent variables (0.42, Table 1). In SVR, the optimal values of C and ε parameters for minimal RMSECV (0.22, Table 1) were 134.3 and 0.0452, respectively. It is possible to note that SVR prediction error was lower than PLS with RMSEP value of 0.25 for SVR and RMSEP value of 0.38 for PLS (Table 1). The F-test for these errors was performed, resulting in: $F_{calculed} = 2.31 > F^{critical}_{\nu1=20,\nu2=20,\alpha=0.05} = 2.12$ and the null hypothesis was rejected at 95% confidence level indicating that the SVR model was more accurate than PLS for API gravity determination.

Based on the concept of multivariate net analyte signal (NAS) [38], figures of merit as described in [39,40] for PLS model were calculated. The values were: root mean squares error of calibration (RMSEC) of 0.30, sensibility of 0.12, limit of detection (LOD) of 0.06,

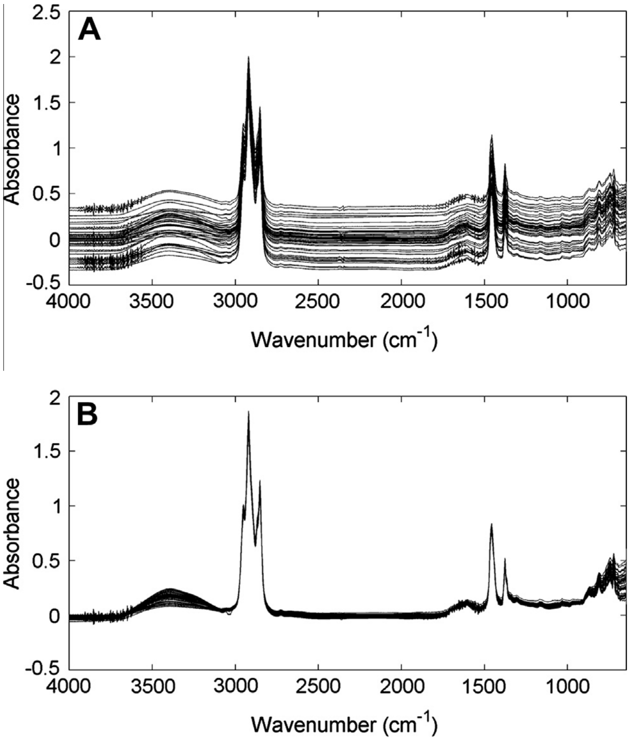

Fig. 1. Infrared spectra of petroleum. (A) Original spectra and (B) preprocessed spectra by multiplicative signal correction.

limit of quantification (LOQ) of 0.19 and selectivity range of 0.12–0.17.

Abbas et al. [41] determined the API gravity from crude oil by FTIR-ATR spectroscopy and obtained RMSEP of 1.66 and RMSEP% of 4.61%. The authors used oils from seven geographic locations that contains a wide variation of physical chemical characteristics of oils. In this work, it was used petroleum of one geographic location from sedimentary basin of the Brazilian coast. Due to use of different samples in two works, a better parameter to compare the accuracy of models is RMSEP%. The results obtained in this work for PLS and SVR models were 2.0% and 1.3%, respectively .

The relationship of the kinematic viscosity with temperature is exponential, thereby to minimize this source of variation, it was measured at two temperatures close to the reference for all oils and models were developed based on the calculated value at 40 °C.

Using the leave-one-out procedure, the PLS model showed lower error cross- validation (RMSECV) using 6 latent variables (20 mm^2 s^{-1}, Table 1). In SVR, the optimal values of C and ε parameters for minimal RMSECV (8 mm^2 s^{-1}, Table 1) were 18.0 and 0.0001, respectively. For PLS model, the figures of merit obtained were: RMSEC of 14 mm^2 s^{-1}, sensibility of 0.13 mm^{-2} s, LOD of 0.05 mm^2 s^{-1}, LOQ of 0.18 mm^2 s^{-1} and selectivity range of 0.01–0.05.

Table 1
SVR and PLS parameter results.

Model		API gravity	Kinematic viscosity (mm^2 s^{-1})	Water content (% v/v)
PLS	R^{2}cv	0.9292	0.8183	0.9189
	R^{2}p	0.9461	0.7811	0.9670
	RMSECV	0.42	20	0.38
	RMSEP	0.38	27	0.34
SVR	R^{2}cv	0.9817	0.9661	0.9768
	R^{2}p	0.9751	0.8584	0.9767
	RMSECV	0.22	8	0.20
	RMSEP	0.25	22	0.26

The SVR prediction error was lower than for PLS, with RMSEP of 22 mm^2 s^{-1} for SVR and 27 mm^2 s^{-1} for PLS (Table 1), but this difference is not statistically significant: $F_{calculed} = 1.51 < F^{critical}_{v1=20,v2=20,\alpha=0.05} = 2.12$, and there is no evidence to reject the hypothesis of equal variances at the 95% confidence level. However, it can be observed in Fig. 3A (PLS model) that the results do not seem to distribute randomly about the line of linear relationship, as shown in Fig. 3B for SVR modeling. The residual plot (difference between reference and predicted values) for the PLS model is shown in Fig. 4A. The residuals suggest a quadratic behavior, in which the central region on the graph focuses mostly positive residues while the end zones have negative residues. This behavior of residuals plot is not observed for SVR (Fig. 4B), where the points are randomly distributed around zero. This is an indication that non-linear calibration models must be applied in such data analyses and PLS is not the best choice.

For water content determination, using the leave-one-out procedure, the PLS model presented lower errors of cross-validation (RMSECV) with 4 latent variables (0.38% v/v, Table 1). In SVR, the optimal values of C and ε parameters for minimal RMSECV (0.20% v/v, Table 1) were with 138.3 and 0.0141, respectively. For PLS model, the figures of merit obtained were: RMSEC of 0.33 (% v/v), sensibility of 0.53 (% v/v)$^{-1}$, LOD of 0.02 (% v/v), LOQ of 0.06 (% v/v) and selectivity range of 0.01–0.19.

The SVR prediction error was lower than for PLS, with RMSEP of 0.26% v/v for SVR and 0.34% v/v for PLS (Table 1), but this difference is not statistically significant: $F_{calculed} = 1.71 < F^{critical}_{v1=20,v2=20,\alpha=0.05} = 2.12$, and there is no evidence to reject the hypothesis of equal

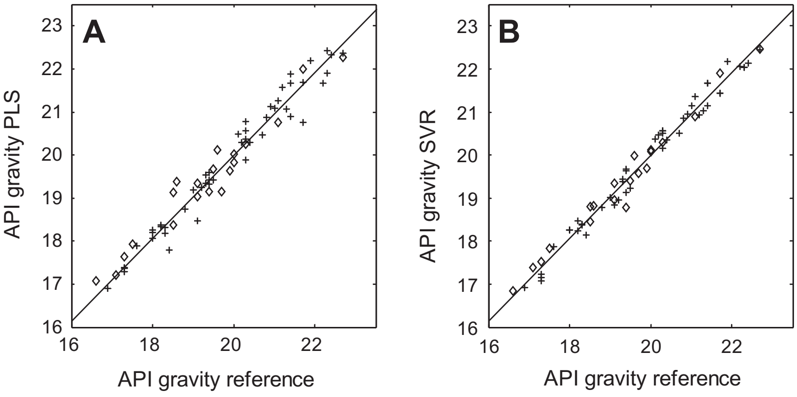

Fig. 2. API gravity values obtained using ISO 12185-96 versus predicted API gravity values by (A) PLS and (B) SVR. (+) Calibration samples and (◇) prediction samples.

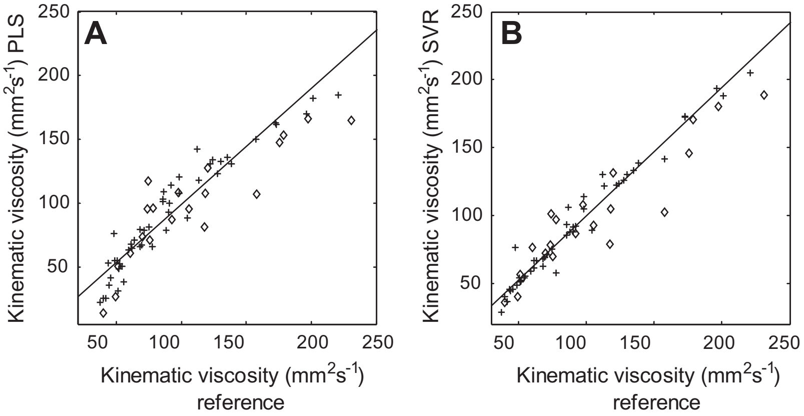

Fig. 3. Kinematic viscosity values obtained using ASTM D 7042 versus kinematic viscosity values predicted by (A) PLS and (B) SVR. (+) Calibration set and (◇) prediction set.

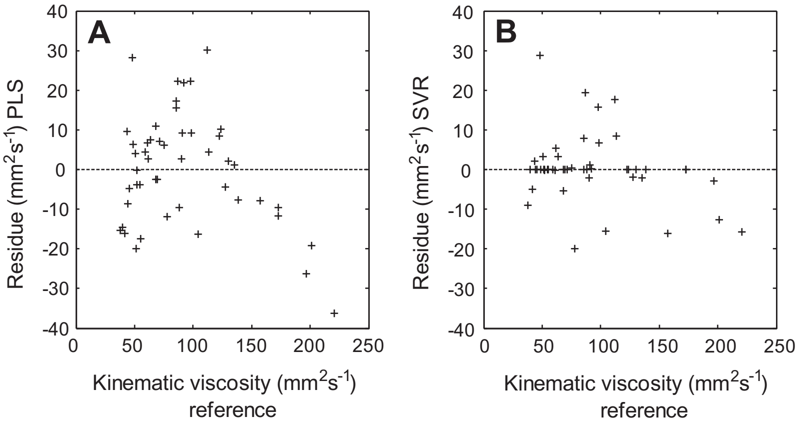

Fig. 4. Residuals plot for kinematic viscosity. (A) PLS model and (B) SVR model.

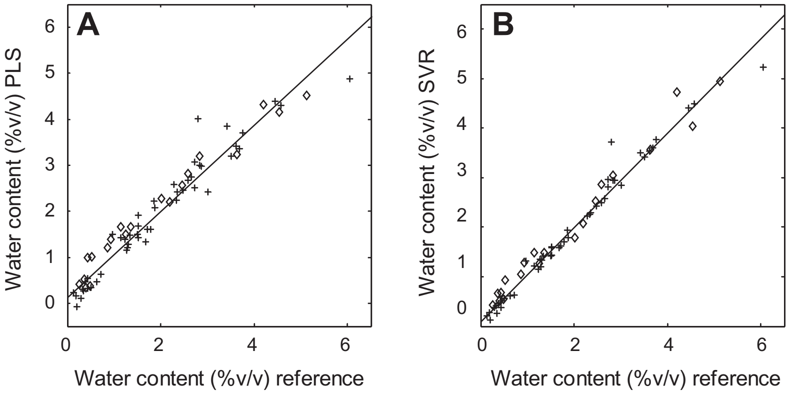

Fig. 5. Water content values obtained using ASTM D 4377-00 versus predicted water content values by (A) PLS model (B) SVR model. (+) Calibration set and (◇) prediction set.

variances at 95% confidence level. However it can be observed in Fig. 5A and B that the SVR model has a better linear relationship between results obtained by the standard method and results of the modeling of both calibration and prediction (R^{2}p of 0.9768 and R^{2}cv of 0.9767 for the SVR against R^{2}p of 0.9670 and R^{2}cv of 0.9189 for PLS, Table 1).

3.1. Variable analysis

The PLS model is well established in the area of multivariate calibration showing good predictive ability and easy model interpretability, as well as good indicators of which variables most contribute to the model development. Different the PLS, SVM is a technique popularly known a "black box" for the lack of interpretation of the model, mainly because it does not indicate the variables with the greatest contribution to its development. This occurs due to the change of the original space of variables to a feature space of high dimension by applying the kernel function, which is a preliminary step in the SVM modeling. However, by applying the method developed by Üstun et al. [27], it is possible to interpret the SVR models by examining the p-vector generated in the calculations.

Figs. 6 and 7 refer to the analysis of the variables in API gravity and kinematic viscosity, respectively. The average spectrum of 48

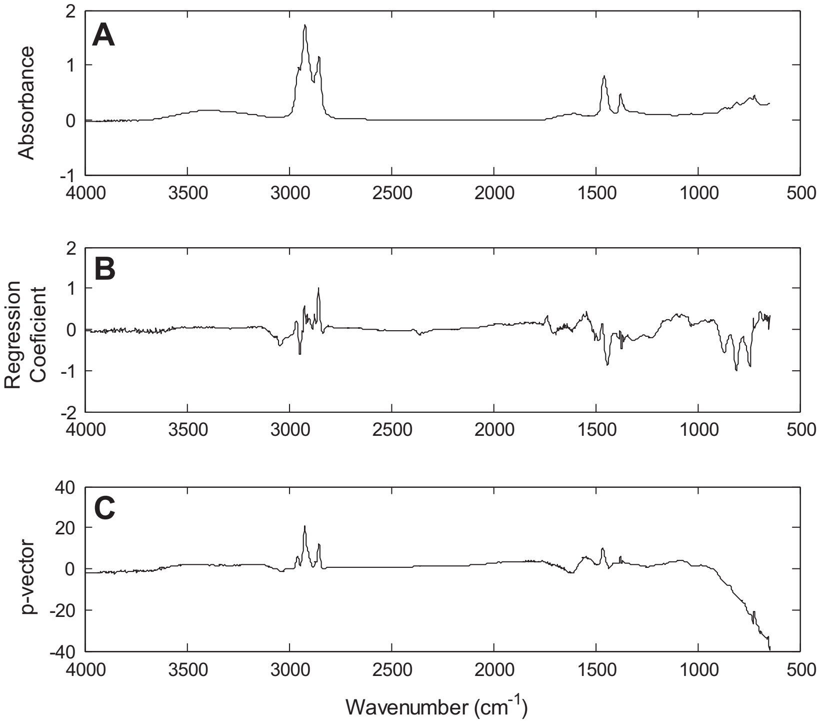

Fig. 6. (A) Average spectrum for the calibration set. (B) Regression coefficients for the PLS model for API gravity. (C) P-vector plot of the SVR model for API gravity.

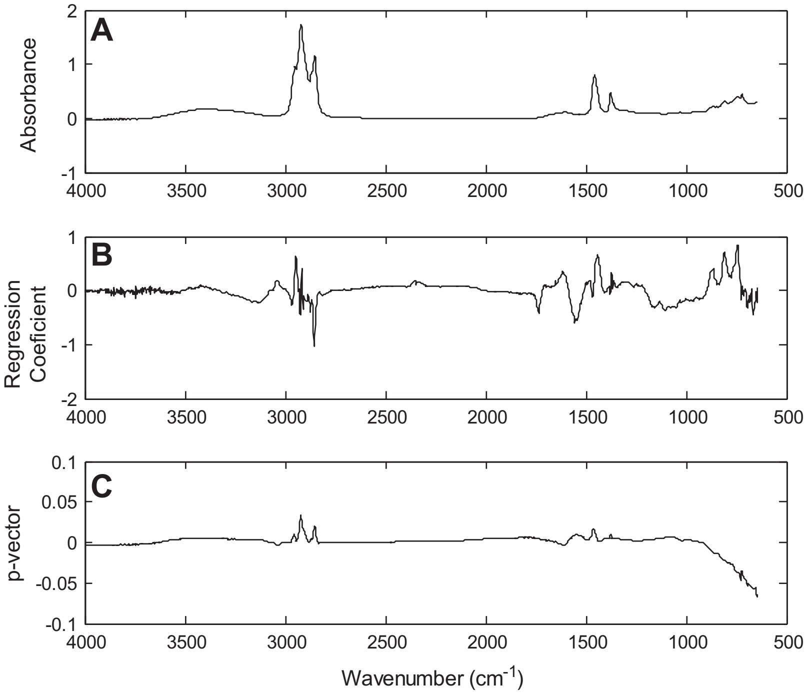

Fig. 7. (A) Average spectrum for the calibration set. (B) Regression coefficients for the PLS model for kinematic viscosity (mm^2 s^{-1}). (C) P-vector plot of the SVR model for kinematic viscosity (mm^2 s^{-1}).

calibration samples are shown in Figs. 6A and 7A, the plot of regression coefficients of the PLS model are shown in Figs. 6B and 7B, where higher values of the coefficients indicates that the variable has greater importance for the model. The plot of p-vectors of the SVR model is shown in Figs. 6C and 7C. The p-vector also indicates the importance of the variable in the model. It can be observed that the same variables are important for PLS and SVR models. The most important of these parameters are from the spectral regions of the C–H stretching band of primary or secondary carbon (region around 2900 cm^{-1}), angular deformation of –(CH$_2$)$_N$– (around 1450 cm^{-1}) and bands appearing in the region around 720 cm^{-1} related to angular deformation of the chain when $N > 3$. These regions are correlated to the density and kinematic viscosity of oil, because they are indicative of a greater amount

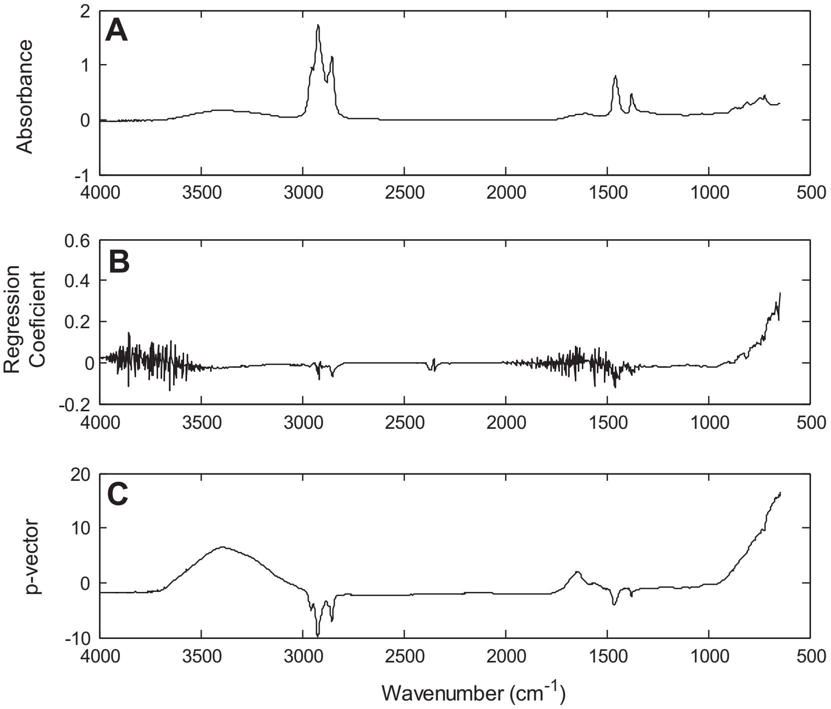

Fig. 8. (A) Average spectrum for the calibration set. (B) Regression coefficients for the PLS model for water content. (C) P-vector plot of the SVR model for water content.

of linear chains and therefore a higher resistance to translational movement of the molecules of oil, thereby increasing the density and kinematic viscosity of oil.

Fig. 8A shows the average spectrum of the 48 calibration samples, Fig. 8B shows the regression coefficients of the PLS model and Fig. 8C shows SVR p-vector for modeling the water content in oil. Fig. 8B and C present as significant variables, with positive values, the region below to 900 cm^{-1} relative to the fingerprint region, which can be related to the sloped O–H outside the angular plane of deformation. Also, the region around 1650 cm^{-1} associated with the broad band 3200 to 3650 cm^{-1} is strong evidence of the presence of water in oil. For the first region, PLS regression coefficients and SVR p-vector indicate that there is a significant and direct relationship with water content in the sample. The second region, from 3200 to 3650 cm^{-1}, more characteristic of the O–H stretching, is very well defined and with great importance in the SVR p-vector, whereas for PLS it is not defined. These results can show that SVR uses the fundamental spectrum region for water prediction and it can correlate to the best results obtained for this parameter.

4. Conclusion

The ATR-FTIR technique associated with multivariate calibration methodologies was efficient for determining the API gravity, kinematic viscosity and water content in medium and heavy oils, featuring models with low prediction errors. The simplicity in the sample preparation and the ability to determine simultaneously the three physicochemical properties of oil in the samples with only a single spectrum are major advantages in the use of the proposed methodology. From the *F*-test at 95% of confidence it was concluded that the SVR model was more accurate than PLS for API gravity determination. For kinematic viscosity and water content the two methods were equivalent. However, a non-linear tendency in the kinematic viscosity model was observed. The matrix multiplication of spectra by support vectors of SVR model made possible the observation of the spectral regions with a greater contribution in the modeling, leading to results as interpretable as PLS.

Acknowledgments

The authors would like to thank LABPETRO-UFES for provide the crude oils samples for development of this work and CNPq for financial support.

References

[1] Speight JG. Handbook of petroleum product analysis. Hoboken: Wiley-Interscience; 2002.

[2] Simanzhenkov V, Idem R. Crude oil chemistry. New York: Marcel Dekker, Inc.; 2003.

[3] Lyons WC, Plisga GJ. Standard handbook of petroleum & natural gas engineering. 2nd ed. Amsterdam: Elsevier; 2005.

[4] Riazi MR. Characterization and properties of petroleum fractions. Philadelphia: American Society for Testing and Materials (ASTM); 2005.

[5] Dantas TNC, Neto AAD, Moura EF. Microemulsion systems applied to breakdown petroleum emulsions. J Petrol Sci Eng 2001;32:145–9.

[6] Fortuny M, Silva EB, Filho AC, Melo RLFV, Nele M, Coutinho RCC, et al. Measuring salinity in crude oils: evaluation of methods and an improved procedure. Fuel 2008;87:1241–8.

[7] Santos Jr 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.

[8] Ferrão MF, Viera MS, Pazos REP, Fachini D, Gerbase AE, Marder L. Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. Fuel 2011;90:701–6.

[9] Parisotto G, Ferrão MF, Müller ALH, Müller EI, Santos MFP, Guimarães RCL, et al. Total acid number determination in residues of crude oil distillation using ATR-FTIR and variable selection by chemometric methods. Energy Fuels 2010;24:5474–8.

[10] Teixeira LSG, Oliveira FS, Santos HC, Cordeiro PWL, Almeida SQ. Multivariate calibration in fourier transform infrared spectrometry as a tool to detect adulterations in Brazilian gasoline. Fuel 2008;87:346–52.

[11] Lira LFB, Vasconcelos FVC, Pereira CF, Paim APS, Stragevitch L, Pimentel MF. Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration. Fuel 2010;89:405–9.

[12] Soares IP, Rezende TF, Pereira RCC, dos Santos CG, Fortes ICP. Determination of biodiesel adulteration with raw vegetable oil from ATR-FTIR data using chemometric tools. J Braz Chem Soc 2011;7:1229–35.

[13] Balabin RM, Lomakina EI, Safieva RZ. Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. Fuel 2011;90:2007–15.
[14] Fodor GE, Mason RA, Hutzler SA. Estimation of middle distillate fuel properties by FT-IR. Appl Spectrosc 1999;53:1292–8.
[15] Chung H, Ku M-S. Comparison of near-infrared, infrared, and Raman spectroscopy for the analysis of heavy petroleum products. Appl Spectrosc 2000;54:239–45.
[16] Höskuldsson A. PLS regression methods. J Chemom 1998;2:211–28.
[17] ASTM E1655. Standard practices for infrared multivariate quantitative analysis. West Conshohocken, PA: ASTM International; 2005.
[18] Ruiz MD, Bustamante IT, Dago A, Hernández N, Núñez AC, Porro D. A multivariate calibration approach for determination of petroleum hydrocarbons in water by means of IR spectroscopy. J Chemom 2010;24:444–7.
[19] Canha N, Felizardo P, Menezes JC, Correia MJN. Multivariate near infrared spectroscopy models for predicting the oxidative stability of biodiesel: effect of antioxidants addition. Fuel 2012;97:352–7.
[20] Blanco M, Coello J, Iturriaga H, Maspoch S, Page's J. NIR calibration in non-linear systems: different PLS approaches and artificial neural networks. Chemom Intel Lab Syst 2000;50:75–82.
[21] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comp 2004;14:199–222.
[22] Cortes C, Vapnik VN. Support-vector network. Mach Learn 1995;20:273–97.
[23] Wu YC, Lee YS, Yang JC. Robust and efficient multiclass SVM models for phrase pattern recognition. Pat Recog 2008;41:2874–89.
[24] Chen PC, Lee KY, Lee TJ, Huang SY. Multiclass support vector classification via coding and regression. Neurocomputing 2010;73:1501–12.
[25] Fong SS, Kiss VS, Brereton RG. Self-organizing maps and support vector regression as aids to coupled chromatography: illustrated by predicting spoilage in apples using volatile organic compounds. Talanta 2011;83:1269–78.
[26] Barman I, Kong CR, Dingari NC, Dasari RR, Feld MS. Development of robust calibration models using support vector machines for spectroscopic monitoring of blood glucose. Anal Chem 2010;82:7000–7.
[27] Li H, Liang Y, Xu Q. Support vector machines and its applications in chemistry. Chemom Intel Lab Syst 2009;95:188–98.
[28] Devos O, Ruckebusch C, Durand A, Duponchel L, Huvenne JP. Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation. Chemom Intel Lab Syst 2009;96:27–33.
[29] Üstün B, Melssen WJ, Buydens LMC. Visualisation and interpretation of support vector regression models. Anal Chim Acta 2007;595:299–309.
[30] Schölkopf B, Sung KK, Burges CJC, Girosi F, Niyogi P, Poggio T, et al. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans Sign Process 1997;45:2758–65.
[31] Andersson M. A comparison of nine PLS1 algorithms. J Chemom 2009;23:518–29.
[32] ISO 12185. Crude petroleum and petroleum products – determination of density – oscillating U-tube method. Geneva: International Organization for Standardization; 1996.
[33] Astm, D 7042. Standard test method for kinematics viscosity in crude oil. West Conshohocken, PA: ASTM International; 2004.
[34] Astm, D 4377. Standard test method for karl fischer in crude oil. West Conshohocken, PA: ASTM International; 2006.
[35] Kennard RW, Stone LA. Comput Aided Des Exp Technom 1969;11:137–48.
[36] Wise BM, Gallagher NB, Bro R, Shaver JM, Windig W, Koch RS. PLS Toolbox Version 4.0 for Use with Matlab. Wenatchee: Eigenvector Research Inc.; 2006.
[37] Chang CC, Lin CJ. LIBSVM: A Library for support vector machines; 2001. Software available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
[38] Lorber A. Error propagation and figures of merit for quantification by solving matrix equations. Anal Chem 1986;58:1167–86.
[39] Olivieri AC, Faber NKM, Ferré J, Boqué R, Kalivas JH, Mark H. Uncertainty estimation and figures of merit for multivariate calibration. IUPAC 2006;78:633–61.
[40] Valderrama P, Braga JWB, Poppi RJ. Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy. J Braz Chem Soc 2007;18:259–66.
[41] Abbas O, Rebufa C, Dupuy N, Permanyer A, Kister J. PLS regression on spectroscopic data for the prediction of crude oil quality: API gravity and aliphatic/aromatic ratio. Fuel 2012;98:5–14.