energy&fuels Article
pubs.acs.org/EF

Predictive Petroleomics: Measurement of the Total Acid Number by Electrospray Fourier Transform Mass Spectrometry and Chemometric Analysis

Boniek G. Vaz,*,[†,‡] Patrícia V. Abdelnur,[§] Werickson F. C. Rocha,[∥] Alexandre O. Gomes,[⊥] and Rosana C. L. Pereira*,[⊥]

[†]Pontifical Catholic of Rio de Janeiro, Rio de Janeiro (RJ) 22451-900, Brazil

[‡]Chemistry Institute, Federal University of Goiás, Goiânia, (GO) 74001-970, Brazil

[§]Embrapa Agroenergia, Brasília, Distrito Federal (DF) 70770-901, Brazil

[∥]National Institute of Metrology, Quality and Technology (Inmetro), Directorate of Industrial and Scientific Metrology (DIMCI), Chemical Metrology Division (DQUIM), , Xerém, Duque de Caxias (RJ), 25250-020, Brazil

[⊥]CENPES, Petróleo Brasileiro S.A. (Petrobras), Rio de Janeiro, Rio de Janeiro (RJ) 28999-999, Brazil

ABSTRACT: Crude oil samples are uniquely complex because of the number of compounds present that can only be resolved using Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS). The FT-ICR MS technique has been redefined for examining the composition of crude oil and its products, which has led to a new field called "petroleomics". The chemical composition ultimately determines the chemical and physical properties and the behavior of petroleum and its products. "Petroleomics" predicts the properties and behavior of petroleum using its composition to solve production and processing problems. This paper correlates the chemical composition of crude oil with the total acid number (TAN), which enables the development of prediction models using partial least squares (PLS) and support vector machines (SVMs) as alternative multivariate calibration methods that allow for the application of FT-ICR MS analysis in direct measurements. The prediction models using PLS and SVM demonstrated low prediction errors and superior performance in relation to the univariate method. These results support the development of robust models to predict crude oil properties based on the vast quantity of information provided by FT-ICR MS using PLS and SVM as multivariate calibration procedures.

1. INTRODUCTION

Crude oil, which is currently considered the most complex mixture in nature,[1] has challenged the analytical and petrochemical community for decades to unravel its complexity and describe its individual constituents on a molecular level. Such characterization is vital to understanding the functionality and the various properties of crude oil, which require the identification of tens of thousands of chemical components in a typical oil sample. Among the mass spectrometry techniques used for the analysis of crude oil, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) is most able to achieve the peak capacity needed to resolve the individual components with a minimal sample preparation.[2] Various ionization methods can be coupled to a FT-ICR mass spectrometer, allowing the user to generate ions from their samples using different methods, each of which has advantages and disadvantages. The use of electrospray ionization (ESI),[3] for example, is widespread; however, this technique only detects strongly polar or ionic species; therefore, it is best suited to studying the acidic or basic components of petroleum. In addition, the high resolution and accuracy provided by FT-ICR MS allows for unique elemental compositions ($C_cH_hN_nO_oS_s$) to be determined.[4] Typically, more than 20 000 chemically distinct components can be detected and analyzed from a single sample.[5] On the basis of these data, the sample can be further characterized according to the distribution of the heteroatom classes or the degree of aromaticity.[6]

The ability of FT-ICR MS to evaluate crude oil components has led to the term "petroleomics", which refers to the principle that the properties and behavior of the organic components of petroleum and its derivatives and products can be correlated (and ultimately predicted) through sufficiently complete characterizations.[7] The petroleomic MS characterization of crude oils has highlighted the compositional trends to elucidate important crude oil properties. A fundamental goal of petroleomics is to link such detailed crude oil compositions to its properties. However, to our knowledge, no systematic studies have yet been published on how to predict crude oil properties using the spectral information obtained from FT-ICR MS.[8]

To relate the measured spectra to specific parameters, uni- and multivariate calibration are often used, which are especially useful with parameters that are difficult to measure directly.[9] Many properties of crude oil and its products can only be determined through laborious means using uni- or multivariate

Special Issue: 13th International Conference on Petroleum Phase Behavior and Fouling

Received: September 17, 2012
Revised: January 23, 2013
Published: January 27, 2013

calibration, and these parameters can be derived much more quickly via indirect measurements using methods such as mass spectrometry. In addition, the use of mass spectrometry can reduce waste, minimize the consumption of raw materials and energy, and diminish the environmental impact.[10] These factors are especially important with respect to the multitrillion (U.S.)-dollar petroleum industry.

Various methods combined with multivariate calibration have been applied to quantify the properties of crude oil.[11] Methods based on near-infrared (NIR), Fourier-transform infrared (FTIR), and Raman spectroscopies[12] have been used to quantify the saturates, aromatics, resins, and asphaltenes (SARA),[13] the sulfur and nitrogen content,[14] and the total acid number (TAN)[15] in crude oil samples. Although these methods provide reliable results, pretreatment steps and elaborate methodologies render them time-consuming and somewhat limited for wide screening because few components are monitored: information related to the chemical compositions of the samples and the compounds responsible for differentiation between the varieties is normally unavailable or is poorly described.

Given its high resolution and accurate measurement capabilities, FT-ICR MS can provide the resolution and accuracy necessary for crude oil analysis. On the other hand, instrumental constraints such as duty cycles and fundamental constraints of ionization suppression difficult quantitative analysis using this technique. However, ESI FT-ICR MS combined with partial least-squares (PLS) multivariate calibration technique was used sucessfully as a fast method to quantify blends of *robusta* and *arabica* coffee.[16]

In this paper, we describe the combination of an information-rich analytical technique, ESI FT-ICR MS, with efficient, modern mathematical regression tools, such as uni- and multivariate methods, including PLS regression and support vector machine (SVM), to create an accurate and robust method for predicting crude oil properties. As an example of this combination, we predicted the TAN of crude oil samples based on the normalized relative abundance of O_2 compounds detected via ESI(−) FT-ICR MS. The accuracy and robustness of the SVM, PLS, and univariate regression are compared.

The TAN is an important parameter measured during crude oil assays that affects refinery optimization, corrosion management, and the safe refining of high-TAN crudes.[17] These measurements normally require considerable crude oil quantities and cannot be performed on small samples or when the crude oil contains large quantities of water. The presented method allows for the measurement of the TAN in such situations, and additional crude oil physical−chemical properties can be extracted from the FT-ICR MS analysis. This method eliminates many steps and reduces the number of analyses (a single unique analysis may provide many properties) and the analytical cost.

1.1. Theoretical Considerations: Univariate Calibration.

In routine analytical work, classical univariate calibration remains the most common calibration method. Typically, the calibration data are fitted using ordinary least squares (OLS).[18] Ordinary linear regression predicts the unknown quantity (the random response variable) as a linear combination of a set of observed values (predictors), which implies that a constant change in a predictor can lead to a constant change in the response variable (i.e., a linear-response model). This approach is appropriate when the response variable has a normal distribution (i.e., when a response variable can vary indefinitely in either direction with no fixed "zero value"). If the basic conditions for the use of a least-squares fit are not fulfilled or if strongly deviating calibration points appear, the OLS method fails; the estimated parameters are biased and are therefore not representative of the relationship between x and y. The default alternative has been to ignore the data characteristics and to apply an OLS method.

However, alternatives, such as the generalized linear model (GLM), a flexible generalization of ordinary linear regression, allow for response variables that have abnormal distributions. The generalized linear model alleviates the limitations of the OLS by allowing for response variables with arbitrary distributions (rather than only normal distributions) and allowing for the arbitrary function of the response variables (the link function) to vary linearly with the predicted values (rather than assuming that the response itself must vary linearly). This model appears suitable for correlating the ESI FT-ICR MS data and response variables with some crude oil proprieties (e.g., TAN), particularly when such properties of a given crude oil sample do not have normal distributions. Additional details on the generalized linear model can be found in the work by Dobson and Barnett.[18]

1.2. Theoretical Considerations: Multivariate Calibration.

1.2.1. PLS Method. The PLS approach is one way to resolve the regression problem (i.e., how to model one or several dependent variables or responses, **Y**, via a set of predictor variables, **X**), which is among the most common data/analytical problems in science and technology.[19] Examples in petroleomics include the relation of **Y**, the properties of crude oil samples, to **X**, their composition as determined using FT-ICR MS. The PLS model has been discussed in detail in the literature;[20] thus, only a brief description is presented here. PLS can be used to find the fundamental relationship between two matrices (**X** and **Y**) and can be a latent variable approach to modeling the covariance structures in these two spaces. A PLS model attempts to determine the multidimensional direction in the **X** space that explains the maximum multidimensional variance direction in the **Y** space. A PLS regression is particularly appropriate when the predictor matrix has more variables than observations and when multicollinearity exists among the **X** values. Standard regression, however, fails in these cases.[21]

1.2.2. SVMs. SVMs are a group of learning techniques that can be applied to classification or regression.[22] This powerful methodology solves nonlinear classification, function estimation, and density estimation problems and yields prediction functions that are expanded through a subset of support vectors. The SVMs can generalize complicated gray-level structures with few support vectors and, thus, provide a new mechanism for image compression. The SVM algorithm is based on the statistical learning theory and the Vapnik−Chervonenkis (VC) dimension.[23]

2. EXPERIMENTAL SECTION

2.1. Samples. A total of 27 heavy crude oil samples from different Brazilian basins with TANs that ranged from 0.05 to 4.80 mg of KOH/g were used in this study. These samples were codified as show in Table 1.

2.2. ESI FT-ICR MS. Samples (approximately 4 mg) were dissolved in 10 mL of toluene. Then, 0.5 mL of this solution was transfer to a 1 mL vial and diluted with 0.5 mL of methanol that contained 0.2% ammonium hydroxide. All solvents and additives were of high-performance liquid chromatography (HPLC) grade, were purchased from Sigma-Aldrich, and were used as received. The general ESI

Table 1. TANs of 27 Crude Oil Samples Analyzed in This Work

sample	TAN (mg of KOH/g)
S 1	1.92
S 3	0.82
S 7	0.70
S 9	1.24
S 10	0.41
S 12	2.69
S 15	0.71
S 18	0.89
S 19A	1.23
S 20	1.19
S 21	1.22
S 22	0.74
S 24	4.80
S 26	0.51
S 27	1.24
S 32	0.41
S 34	2.12
S 36	1.24
S 38	0.49
S 39	3.35
S 40	0.06
S 46	0.33
S 49	0.06
S 50	0.23
S 53	0.30
S 54	0.50
S F1	2.35

conditions were a capillary voltage of 3.10 kV, a tube lens of −100 V, and a flow rate of 3 μL min^{-1}.

Ultrahigh-resolution MS was performed with a ThermoScientific 7.2 T ESI FT-ICR mass spectrometer (ThermoScientific, Bremen, Germany). A scan range of m/z 200−1000 was used, and 100 microscans were collected in each run. The average resolving power (R_p) was 400 000 at m/z 400, and R_p was calculated as $m/\Delta m_{50\%}$ (i.e., the m/z value divided by the full width at half maximum). The time-domain data (ICR signal or transient signal) were acquired over 700 ms. The molecular-weight distribution for each sample was first verified using linear ion trap (LTQ, ThermoScientific, Bremen, Germany) analysis to ensure the validity of the molecular-weight distribution based on the FT-ICR MS.

To approximate the total concentration of O_2-containing species, stearic acid ($C_{18}H_{36}O_2$, neutral monoisotopic mass = 284.270 98 Da; Sigma Aldrich, St. Louis, MO) was used as an internal standard. The samples were prepared as follows: 200 mL of methanol that contained 2% (v/v) NH_4OH was combined with 200 μL of crude oil (4 mg of oil/10 mL of toluene) and 10 μL of 0.266 μmol/L stearic acid. A total of 50 time-domain transients were collected and co-added.[23] The mass spectrum with the internal standard was then used to semi-quantitatively determine the total naphthenic acid concentration in the unspiked samples (e.g., the mass spectrum generated from 100 time-domain transients). The relative abundance for the m/z 283.263 16 ions in the unspiked sample was normalized relative to the deprotonated internal standard $[C_{18}H_{35}O_2]^-$ ion with the same mass, which allowed for the concentration of $[C_{18}H_{35}O_2]^-$ (in micromolar) in the unspiked sample to be determined. This concentration was then scaled to the total relative abundance of the O_2 class to yield the total naphthenic acid concentration in the crude oil (in micromolar).

In addition to external calibration, an internal recalibration was applied to the peak list (using Composer, Sierra Analytics, Modesto, CA) prior to the final peak assignments. A set of theoretical homologous series for a specific heteroatom class (the most abundance class for each ion mode) was selected as the internal calibrant because the most abundance class components have low errors and high average peak intensities.

2.3. Formula Assignment. For each spectrum, an automated analysis was used to assign the formulas to the peaks with a signal-to-noise ratio (S/N) greater than 3. The elements allowed were ^{12}C, ^{1}H, ^{16}O, ^{14}N, ^{32}S, and ^{13}C. The maximum allowed formula error was 1 ppm, and the mass limit for empirically assigning the elemental formulas was 500 Da. The formulas with atomic masses greater than

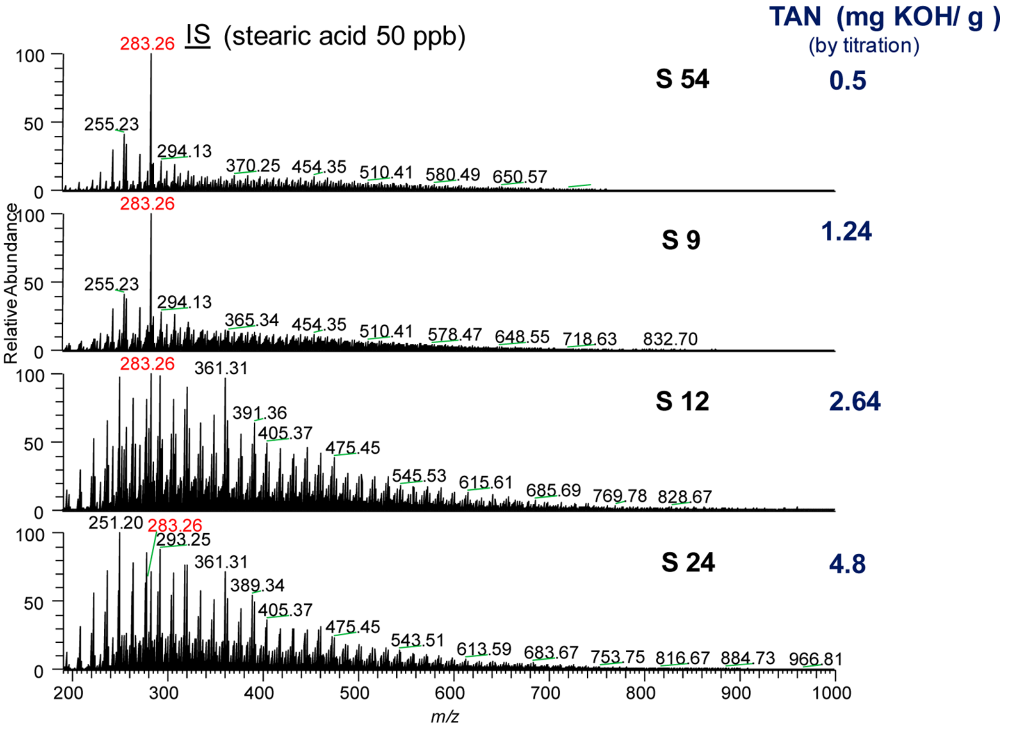

Figure 1. Negative-ion ESI FT-ICR mass spectra of some crude oils with an internal standard. The ESI MS patterns vary with crude oil acidity.

500 Da were assigned through the detection of a homologous series. If no chemical formula matched a m/z value within the allowed error, the peak was not included in the list of elemental formulas. For each elemental composition, $C_cH_hN_nO_oS_s$, the heteroatom class, the type [double bond equivalents (DBE) = the number of rings plus the double bonds involving carbon], and the carbon number, c, were tabulated to generate the relative abundance distributions of the heteroatom class and the graphical DBE versus the carbon number images.

2.4. Univariate Calibration. A generalized linear model was constructed to quantify the TAN in a heavy crude oil sample using the software R.

2.5. Chemometric Analysis. The PLS and SVM models were constructed to quantify the TAN in heavy crude oil samples. The Matlab (version R2007b) and PLS Toolbox (version 6.2) software packages from Eigenvector Research were used to build the PLS and SVM models. To quantify the TAN, 17 samples were used for calibrations, which resulted in a $\mathbf{X}_{\text{calibration}}$ matrix (17 × 300) and a $\mathbf{y}_{\text{calibration}}$ vector (17 × 1), whereas 10 samples were employed for validation, which resulted in a $x_{\text{validation}}$ matrix (10 × 300) and a $\mathbf{y}_{\text{validation}}$ vector (10 × 1). Several preprocessing techniques and their combinations were tested. The best results were obtained from the mean center. The Kennard−Stone algorithm was used to separate the samples in these groups.[24] Notably, these samples were used to construct the calibration and validation phases of the univariate model.

3. RESULTS AND DISCUSSION

3.1. ESI FT-ICR MS Analysis. Negative ESI selectively ionizes the acidic components (carboxylic acids and neutral nitrogen compounds) in a hydrocarbon crude oil matrix. The carboxylic acids are preferentially ionized by an order of magnitude over the neutral nitrogen species (pyrollic benzalogs).[25] The preferential ionization of the acidic species via negative ESI renders this the desired technique because these acidic components, specifically the carboxylic acids, are believed to be the key contributors to the TAN.

Figure 1 illustrates the negative ESI mass spectra for the four representative crude oils spiked with the internal standard (note that, as the TAN increases, the signal of m/z 283 decreases). All mass spectra exhibit an average mass resolving power of 350 000 < $m/\Delta m_{50\%}$ < 400 000. The distribution of the m/z values differs for all four crude oils, 200 < m/z < 800 for the negative ions, despite the differences in the TANs. As the TAN increases, the abundance, primarily the odd ions, increases according to a Gaussian distribution, which indicates that the molecular-weight distribution plays a role in the TAN of the crude oils.

Classes of compounds that contain a carboxylic acid group have been proposed as the main contributors to the TAN of a crude oil.[17] Figure 2 shows a plot of the abundance of the O_2-containing acids relative to all species versus the TAN; the plot shows a clear correlation between the relative abundance of O_2 compounds and the TAN. Although these acids are the primary contributors to the overall acidity of the crude oils, other non-basic nitrogen compounds, phenols and nonpolar sulfur components (not detected by ESI FT-ICR MS), also contribute to the TAN of crude oils.[17a]

The compositional differences within the O_2 class can also be monitored via contour plots as a function of the carbon number versus the DBE, as illustrated in Figure 3. A pronounced increase in the complexity is evident in the spread in both the carbon number and the DBE with an increasing TAN.

3.2. Univariate Model for TAN Prediction. To test the use of the univariate model for TAN prediction, we used the model proposed by Qian et al.[26] The hypothesis of this model is based on a proportionality of the ESI signal with the quantity of acid in the sample, which, in turn, is related to the KOH needed to neutralize the acid. The TAN measurement via ESI is based on the semi-quantification of all O_2 species (i.e., the O_2 compound concentration of the unspiked samples was estimated on the basis of the abundance of stearic acid in the spiked samples). Uniform response factors were assumed for all acid molecules in the TAN calculation according to eq 1

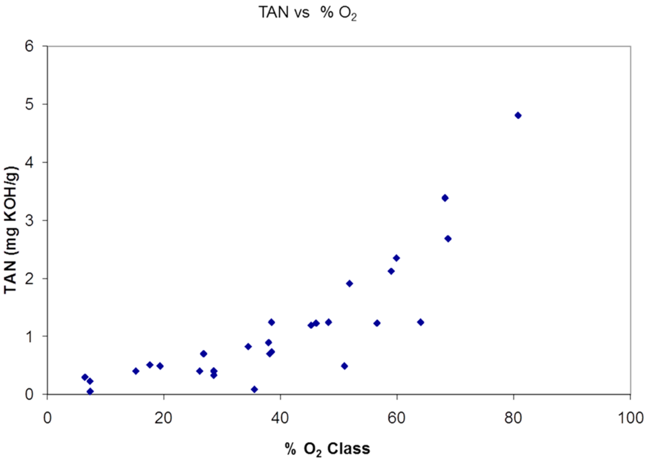

Figure 2. Correlation of the percentage of the O_2 class and TAN values.

$$\text{TAN (mg of KOH/g)} = \frac{c(56.1/W)(M_S/R_S)}{\sum R_A}$$

$$y = c \times b \tag{1}$$

where c is a constant instrument factor, W is the weight of the sample (in grams), M_S is the concentration of stearic acid (in millimoles), R_S is the ESI response of the stearic acid, and R_A is the response of the acid molecules in the sample.

The fundamental basis for this univariate model is the constant c, which is obtained through the relationship between the TAN (measured via titration) and the constant b. The TAN estimated by ESI MS is subsequently obtained as the product of the constants c and b.

The constant c can typically be obtained via an ordinary least-squares method. However, to use this method, the dependent variable (TAN) must have a normal distribution, which does not occur in this case (see Figure 4a). We performed the Shapiro−Wilk test,[27] which rejected the hypothesis of normality for the variable y (TAN). Given that the dependent variable has a positively skewed distribution, the generalized linear model obtained by γ regression (Figure 4b) was used as represented by eq 2.

$$y = 0.15292 + 0.00036b_i \tag{2}$$

Figure 5 compares the TANs determined via ESI FT-ICR MS to those obtained via titration for a series of crude oils. To verify the agreement between these results, two variables (TAN measured via titration versus TAN measured via ESI FT-ICR MS) were compared using a regression line and t test, as recommended by Miller and Miller.[28]

The significant correlation between the two variables, i.e., the H_0 = zero correlation, was calculated as follows:

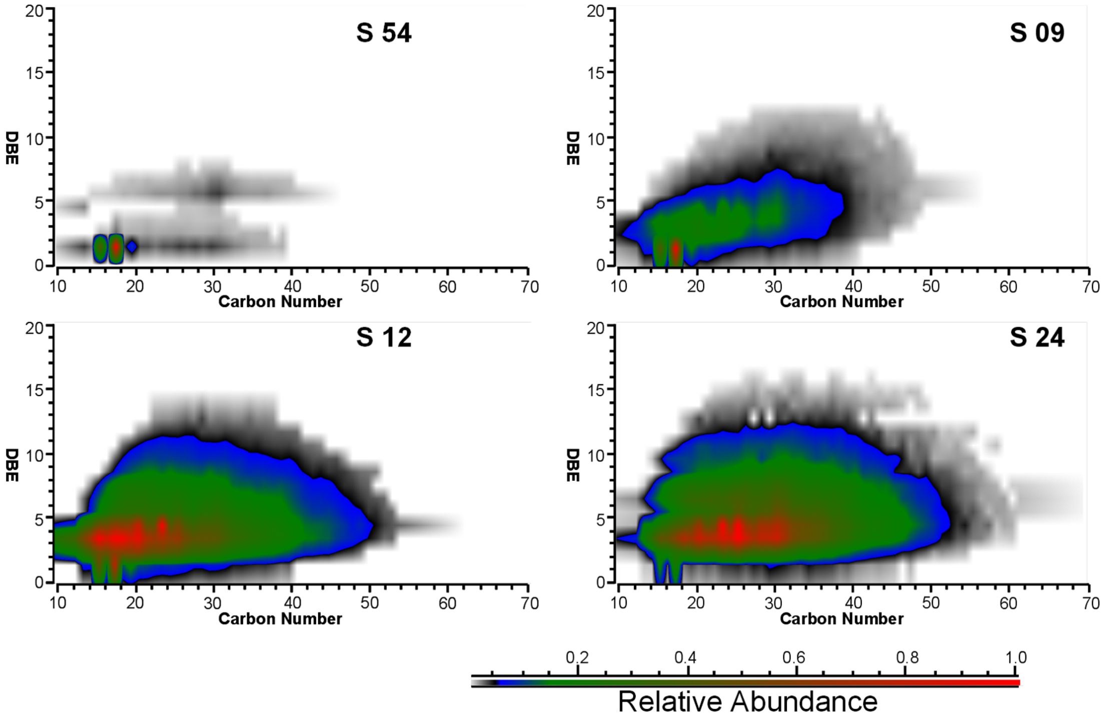

Figure 3. Plots of DBE versus carbon number for the O_2 class from four crude oils.

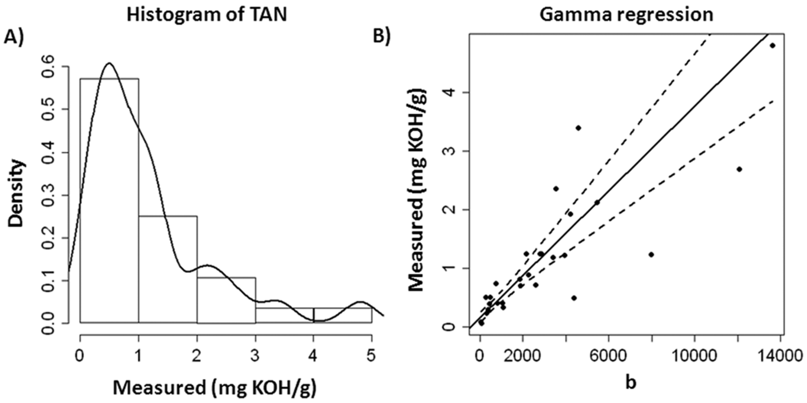

Figure 4. Distribution of the measured TAN values and the regression line obtained using the GLM method.

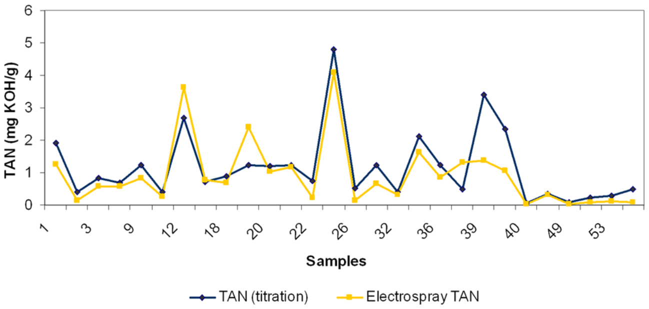

Figure 5. TAN distribution of 27 crude oil samples determined via ESI FT-ICR MS and compared to a nonaqueous titration assay.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{3}$$

where t is the critical value of the t distribution with $n - 2$ degrees of freedom, n is the number of data points in the regression line, and r is the correlation coefficient.

The calculated value of t was compared to the tabulated value at the desired significance level using a two-sided t test and $n - 2$ degrees of freedom. The null hypothesis in this case states that no correlation exists between the TAN measured via ESI FT-ICR MS and that measured via titration. If the calculated value of t is greater than the tabulated value, the null hypothesis is rejected, and we conclude in such cases that a significant correlation does exist. The value found for the t test was 8.32, and the critical value of t was 2.05 ($P = 0.05$). We thus concluded that the two assay methods do not provide significantly different values for the TAN in crude oil samples, which suggests that the assumption of a uniform response factor for the various acid types is reasonable.

3.3. Multivariate Analysis for TAN Prediction. The root-mean-square error of cross-validation (RMSECV) was used to determine the optimum number of latent variables (LVs) for the PLS model. Six LVs were used to build the PLS model, which provided a lower RMSECV and explained the 97.61% variance of **X** and the 94.34% variance of **Y**.

To perform the SVM model, a proper Kernel function must be selected and its optimal parameters must be determined to build the best model, i.e., the model with a lower RMSECV value. In this work, we used the kernel function of the Gaussian radial basis function and optimized the parameters ν, the cost (C), and γ, as previously reported.[29] Suitable values for the parameters ν, C, and γ were obtained after several trials. We compared the root-mean-square error of calibration (RMSEC) and RMSECV (a low value is good) to select suitable parameters. In the SVM algorithm, the parameters obtained were $C = 100$, $\nu = 0.2$, and $\gamma = 0.01$. A leave-one-out cross-validation was used to generate the model with the SVM for the input data set. Figure 6 shows the cross-validation optimization.

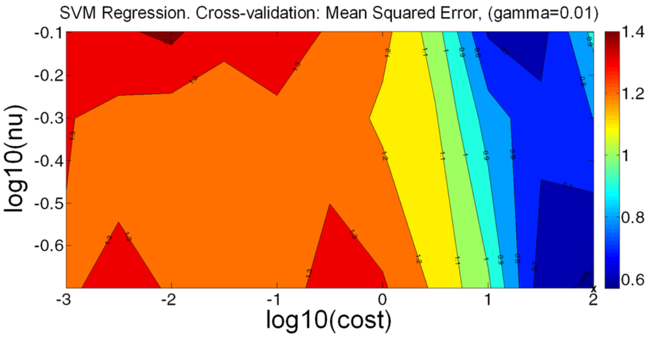

Figure 6. Contour plot of the cross-validation accuracy for the SVM regression.

Figure 7 shows the predicted values (for the ESI FT-ICR MS assay) according to the PLS (panels A and B of Figure 7) and the SVM (panels C and D of Figure 7) models against the measured values (i.e., those from the titration). In all plots, the validation samples are represented by white circles, whereas the calibration samples are indicated by black circles. The distribution pattern for the majority of data points close to the 45° solid line demonstrates a good agreement between the predicted and measured values. The percent errors for these fits are shown in Figure 7. In this figure, the residuals present a random distribution, which indicates a suitable fit. The calculation of the residuals was performed using eq 4

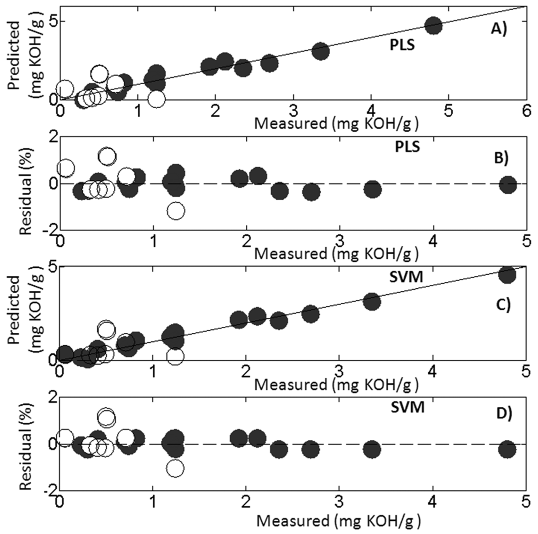

Figure 7. Plot of the predicted versus the reference values for the TAN obtained by PLS and SVM. Calibration (●) and validation (○) samples.

$$\text{residual (\%)} = \left(\frac{y_{\text{ref}} - y_{\text{pred}}}{y_{\text{ref}}}\right) \times 100 \tag{4}$$

where y_{ref} contains the reference values for the TAN and y_{pred} is the value predicted by the model. The errors of all of the predicted samples were less than 2%. The major errors were observed for samples that contained a low acid concentration. This observation may be rationalized as follows: non-basic nitrogen compounds (i.e., carbazoles and their benzologues) and phenolic compounds, which are weak acids, were not used in the models. These compounds contribute to the TAN in samples with low TAN values.

3.4. General Remarks.

Table 2 presents three statistics often used to compare the performances of calibration models: root-mean-square error of calibration (RMSEC), the root-mean-square error of prediction (RMSEP), and Pearson's correlation coefficient between the real and predicted concentrations (rcal). Both errors are based on the calculated root-mean-squared error (RMSE) as follows:

Table 2. Comparison of the Performance Results for the Calibration Models

	PLS	SVM	univariate
RMSEC (mg of KOH/g)[a]	0.29	0.20	2.99
RMSEP (mg of KOH/g)[b]	0.77	0.68	2.05
r_{cal}[c]	0.97	0.98	0.87

[a] Root-mean-square error of calibration for values of TAN. [b] Root-mean-square error of prediction for values of TAN. [c] Pearson's correlation coefficient between the real and predicted concentrations (calibration).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n}} \tag{5}$$

where $\hat{y}_i$ is the predicted value, Y_i the measured value, and n is the number of samples. The RMSEC and RMSEP differ in the determination of $\hat{y}_i$. For details on the calculation of the parameters see Brereton.[30]

The RMSEC was used to evaluate the error of the proposed calibration models, and the RMSEP was used to evaluate the prediction ability of the different models and to select the best model.

Pearson's correlation coefficient between the real and predicted concentrations (r_{cal}) was calculated for the calibration set, which was calculated according to eq 6, where $\overline{y}_i$ is the mean of the reference measurements for all samples in the training set.[31]

$$r = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}} \tag{6}$$

A comparison of the results for the multivariate models (PLS and SVM) to those for the univariate model (Table 2) reveals that the RMSEC and RMSEP values obtained using the multivariate models are better, with smaller error, than those obtained using the univariate model. In addition, these values were similar for both the PLS and SVM models, which makes both of them suitable for quantifying the TAN values.

The similarity between these two multivariate models was confirmed using a F-test[32] with a 95% confidence level (P = 0.05) considering the null hypothesis that the determination of TAN values by the two models (PLS and SVM) yields no significant difference in errors. For the F-test, the following expression was used:

$$F = \frac{(\mathrm{RMSEP}_i)^2}{(\mathrm{RMSEP}_j)^2} \qquad (7)$$

where RMSEP is the root-mean-square error of prediction (validation) and the subscripts i and j represent the models with the largest and smallest RMSEP values, respectively. The degree of freedom in the F test was 9 for both models, and the F-test result was 1.26. The calculated F value was less than 3.18, which was the critical F value (with a confidence level of 95%). These results reconfirmed the equivalence between the PLS and SVM models.

The F test was also performed to compare each of the multivariate models to the univariate model, as shown below. The F-test results were greater than the critical F value of 3.18 (with a confidence level of 95%) and confirmed that both multivariate models are more effective than the univariate model.

$$F = \frac{(\mathrm{RMSEP}_{\mathrm{univariate}})^2}{(\mathrm{RMSEP}_{\mathrm{PLS}})^2} = \frac{(2.05)^2}{(0.77)^2} = 7.01$$

$$F = \frac{(\mathrm{RMSEP}_{\mathrm{univariate}})^2}{(\mathrm{RMSEP}_{\mathrm{SVM}})^2} = \frac{(2.05)^2}{(0.68)^2} = 9.01$$

4. CONCLUSION

This paper proposes the use of uni- and multivariate calibration methods to predict crude oil properties based on the information provided by ESI FT-ICR MS analysis, as illustrated through the TAN prediction. Because of their ability to provide good predictions, PLS and SVM are promising techniques for petroleomic studies and for solving multivariate calibration problems for the oil industry. These methods estimate the parameters via indirect yet fast and reliable measurements, such as ESI FT-ICR mass spectra, and thereby offer improvements over the standard physical approaches for the determination of the parameters, which are laborious and can be time-consuming.

Multivariate calibration methods offer important advantages that lead to global (and often unique) models for simplifying calculations; this paper demonstrates the performance of such methods using well-known statistics tests. In comparison to the previously applied univariate modeling method, the multivariate calibration methods best solve this problem.

We hope that this study provides a clear application of the role of PLS and SVM regression in chemometrics and multivariate data analysis for predictive petroleomics and that the possibilities and obstacles to their application in both the analytical and industrial communities have been made evident.

We believe that our results will help future chemometric and petroleomic investigations. The results presented herein can help to achieve rapid and accurate analysis of TAN values and can be applied to predict other properties of crude oil. The use of FT-ICR MS in fuel and biofuel analyses can be enhanced through the application of the methods of multivariate data analysis, including SVMs, artificial neural networks, and other machine-learning techniques.

AUTHOR INFORMATION

Corresponding Author

*Telephone: +55-62-3521-1016 R261 (B.G.V.); +55-21-2162-6175 (R.C.L.P.). E-mail: bonigontijo@yahoo.com.br (B.G.V.); rosanacardoso@petrobras.com.br (R.C.L.P.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Petrobras for permission to publish this work.

REFERENCES

(1) Hsu, C. S.; Hendrickson, C. L.; Rodgers, R. P.; McKenna, A. M.; Marshall, A. G. *J. Mass Spectrom.* **2011**, *46*, 337−343.

(2) (a) Saab, J.; Mokbel, I.; Razzouk, A. C.; Ainous, N.; Zydowicz, N.; Jose, J. *Energy Fuels* **2005**, *19*, 525−531. (b) Islas-Flores, C. A.; Buenrostro-Gonzalez, E.; Lira-Galeana, C. *Energy Fuels* **2005**, *19*, 2080−2088. (c) Hughey, C. A.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2002**, *74*, 4145−4149. (d) Qian, K.; Robbins, W. K.; Hughey, C. A.; Cooper, H. J.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2001**, *15*, 1505−1511. (e) Wu, Z.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2004**, *76*, 2511−2516. (f) Barrow, M. P.; Headley, J. V.; Peru, K. M.; Derrick, P. J. *J. Chromatogr., A* **2004**, *1058*, 51−59.

(3) Zhan, D.; Fenn, J. B. *Int. J. Mass Spectrom.* **2000**, *194*, 197−208.

(4) (a) Hsu, C. S *Energy Fuels* **2010**, *26*, 1169−1177. (b) Hsu, C. S. *Energy Fuels* **2010**, *24*, 4097−4098.

(5) Rodgers, R. P.; McKenna, A. M. *Anal. Chem.* **2011**, *83*, 4665−4687.

(6) (a) Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37*, 53−59. (b) Rodgers, R. P.; Schaub, T. M.; Marshall, A. G. *Anal. Chem.* **2005**, *77*, 0A−27A. (c) Marshall, A. G.; Rodgers, R. P. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18090−18095.

(7) Hsu, C. S.; Lobodin, V. V.; Rodgers, R. P.; McKenna, A. M.; Marshall, A. G. *Energy Fuels* **2011**, *25*, 2174−2178.

(8) Hur, M.; Yeo, I.; Kim, E.; No, M.; Koh, J.; Cho, Y. J.; Lee, J. W.; Kim, S. *Energy Fuels* **2010**, *24*, 5524−5532.

(9) Manly, B. F. J. *Multivariate Statistical Methods: A Primer*, 3rd ed.; Chapman and Hall/CRC: Boca Raton, FL, 2004.

(10) Næs, T.; Isaksson, T.; Fearn, T.; Davies, T. *A User-Friendly Guide to Multivariate Calibration and Classification*; NIR Publications: Chichester, U.K., 2002.

(11) (a) Sun, F.; Ma, W.; Xu, L.; Zhu, Y.; Liu, L.; Peng, C.; Wang, L.; Kuang, H.; Xu, C. *TrAC, Trends Anal. Chem.* **2010**, *29*, 1239−1249. (b) Shao, Y.-N.; He, Y.; Bao, Y.-D. *Spectrosc. Spectral Anal.* **2008**, *28*, 602−605. (c) Ala-Korpela, M.; Hiltune, Y.; Bell, J. D. *NMR Biomed.* **2005**, *8*, 235−244.

(12) Balabin, R. M.; Zafieva, R. Z.; Lomakina, E. I. *Anal. Chim. Acta* **2010**, *671*, 27−35.

(13) Laxalde, J.; Ruckebush, C.; Devos, O.; Caillol, N.; Wahl, F.; Duponchel, L. *Anal. Chim. Acta* **2011**, *705*, 227−234.

(14) (a) Müller, A. L. H.; Picoloto, R. S.; Mello, P. A.; Ferrão, M. F.; Santos, M. F. P.; Guimarães, R. C. L.; Müller, E. I.; Flores, E. M. M. *Spectrochim. Acta, Part A* **2012**, *89*, 82−87. (b) Li, J.; Chu, X.; Tian, S.; Lu, W. *China Pet. Process. Petrochem. Technol.* **2011**, *4*, 1−7.

(15) Dong, J.; Van De Voort, F. R.; Ismail, A. A.; Akochi-Koble, E.; Pinchuk, D. *Lubr. Eng.* **2000**, *56*, 12−20.

(16) Garret, R.; Vaz, B. G.; Hovell, A. M. C.; Eberlin, M. N.; Rezende, C. M. *J. Agric. Food Chem.* **2012**, *60*, 4253−4258.

(17) (a) Wu, Z.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2004**, *18*, 1424−1428. (b) Tomczyk, N. A.; Winans, R. E.; Shinn, J. H.; Robinson, R. C. *Energy Fuels* **2001**, *15*, 1498−1504. (c) Barrow, M. P.; McDonnel, L. A.; Feng, X.; Walker, J.; Derrick, P. J. *Anal. Chem.* **2003**, *75*, 860−866.

(18) Dobson, A. J.; Barnett, A. G. *Introduction to Generalized Linear Models*, 3rd ed.; Chapman and Hall/CRC: Boca Raton, FL, 2008.

(19) (a) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1−17. (b) Brereton, R. G. *Analyst* **2000**, *125*, 2125−2154. (c) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley and Sons: New York, 1989.

(20) (a) Rocha, W. F.; Sabin, G. P.; Março, P. H.; Poppi, R. J. *Chemom. Intell. Lab. Syst.* **2011**, *106*, 198−204. (b) Rocha, W. F. C.; Rosa, A. L.; Martins, J. A.; Poppi, R. J. *J. Braz. Chem. Soc.* **2010**, *21*, 1929−1936. (c) Rocha, W. F.; Nogueira, R.; Vaz, B. G. *J. Chemometr.* **2012**, *26*, 456−461. (d) Rocha, W. F.; Vaz, B. G.; Sarmanho, G. F.; Leal, L. H. C.; Borges, C. N.; Silva, V. F. *Anal. Lett.* **2012**, *45*, 1−14.

(21) Ivanciuc, O. *Rev. Comput. Chem.* **2007**, *23*, 291−400.

(22) (a) Vapnik, V.; Lerner, A. *Autom. Remote Control* **1963**, *24*, 774−780. (b) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.

(23) Hughey, C. A.; Minardi, C. S.; Galasso-Roth, S. A.; Paspalof, G. B.; Mapolelo, M. M.; Rodgers, R. P.; Marshall, A. G.; Ruderman, D. L. *Rapid Commun. Mass Spectrom.* **2008**, *23*, 3968−3976.

(24) Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137−148.

(25) Sterner, J. L.; Johnston, G. R.; Ridge, D. P. *J. Mass Spectrom.* **2000**, *35*, 385−391. (b) King, R.; Bonfiglio, R.; Fernendez-Metzler, C.; Miller-Stein, C.; Olah, T. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 942−950. (c) Enke, C. G. *Anal. Chem.* **1997**, *69*, 4885−4893. (d) Wu, Z.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2004**, *76*, 2511−2516.

(26) Qian, K.; Edwards, K. E.; Dechert, G. J.; Jaffe, S. B.; Green, L. A.; Olmstead, W. N. *Anal. Chem.* **2008**, *80*, 849−855.

(27) Shapiro, S. S.; Wilk, M. B. *Biometrika* **1965**, *52*, 591−611.

(28) Miller, J. N.; Miller, J. C. *Statistics and Chemometrics for Analytical Chemistry*, 6th ed.; Pearson: Harlow, U.K., 2010.

(29) Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R.; O'Sullivan, S. D. *PLS_Toolbox 6.2 for Use with MATLAB*; Eigenvector Research, Inc.: Wenatchee, WA, March, 2011.

(30) Brereton, R. G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*; John Wiley and Sons: Chichester, U.K., 2003; Chemical Analysis Series.

(31) Chen, Q.; Zhao, J.; Liu, M.; Cai, J.; Liu, J. *J. Pharm. Biomed. Anal.* **2008**, *46*, 568−573.

(32) Danzer, K. *Analytical Chemistry: Theoretical and Metrological Fundamentals*; Springer: Berlin, Germany, 2007.