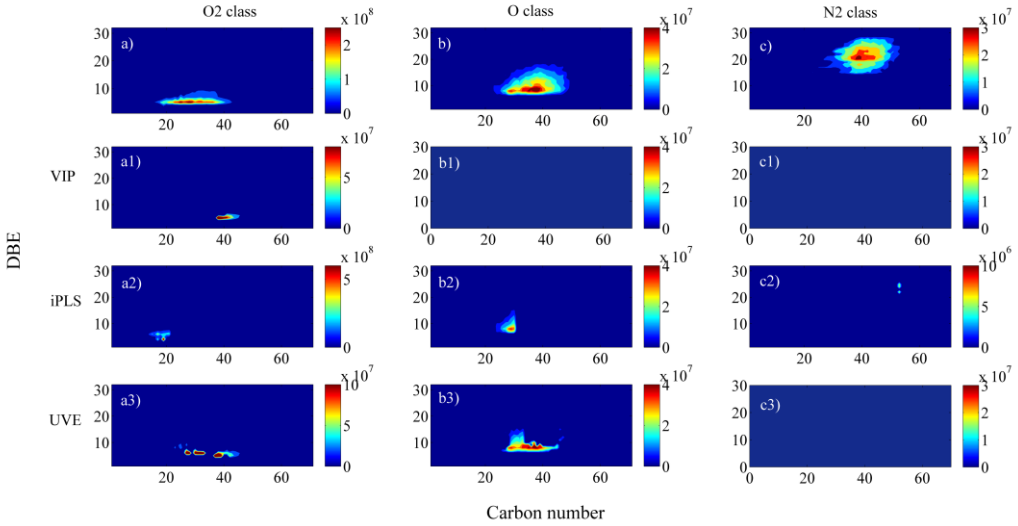
Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) was coupled to a Partial Least Squares (PLS) regression and variable selection methods to estimate the total acid number (TAN) of Brazilian crude oil samples



Petroleomics by Electrospray Ionization FT-ICR Mass Spectrometry Coupled to Partial Least Squares with Variable Selection Methods: Prediction of the Total Acid Number of Crude Oils

Luciana A. Terra[a], Paulo R. Filgueiras[a], Lílian V. Tose[b], Wanderson Romão[b,c], Douglas D. de Souza[d], Eustáquio V. R. de Castro[b], Mirela S. L. de Oliveira[e], Júlio C. M. Dias[e] and Ronei J. Poppi[a*]

[a] Institute of Chemistry, University of Campinas, Campinas, SP, Brazil.

[b] Petroleomic and Forensic Chemistry Laboratory, Department of Chemistry, Federal University of Espírito Santo, 29075-910, Vitória, ES, Brazil.

[c] Federal Institute of Education, Science and Tecnology of Espírito Santo, 29106-010, Vila Velha, ES, Brazil.

[d] Institute of Physics "Gleb Wataghin", University of Campinas, Campinas, SP, Brazil.

[e] CENPES/PETROBRAS, Rio de Janeiro, RJ, Brazil.

[*] Corresponding author. Tel./fax: +55 19 35212134.

E-mail address: ronei@iqm.unicamp.br (R. J. Poppi).

Abstract

Negative-ion mode electrospray ionization, ESI(-), with Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) was coupled to a Partial Least Squares (PLS) regression and variable selection methods to estimate the total acid number (TAN) of Brazilian crude oil samples. Generally, ESI(-)-FT-ICR mass spectra presents a power of resolution of ca. 500.000 and a mass accuracy less than 1 ppm, producing a data matrix containing over 5700 variables per sample. These variables correspond to heteroatom-containing species detected as deprotonated molecules, $[M-H]^-$ ions, which are identified primarily as naphthenic acids, phenols and carbazole analog species. The TAN values for all samples ranged from 0.06 to 3.61 mg of KOH g^{-1}. To facilitate the spectral interpretation, three methods of variable selection were studied: variable importance in the projection (VIP), interval partial least squares (iPLS) and elimination of uninformative variables (UVE). The UVE method seems to be more appropriate for selecting important variables, reducing the dimension of the variables to 183 and producing a root mean square error of prediction of 0.32 mg of KOH g^{-1}. By reducing the size of the data, it was possible to relate the selected variables with their corresponding molecular formulas, thus identifying the main chemical species responsible for the TAN values.

Keywords: ESI(-)FT-ICR MS; Petroleomic; Total acid number; UVE-PLS.

1. Introduction

Petroleomics is defined as a field of petrol sciences able to elucidate the chemical composition of constituents present in crude oil due to its physical and chemical properties and reactivity. In general, crude oil is composed mainly of carbon (80 to 90%), hydrogen (10 to 15%), sulfur (up to 5%), oxygen (up to 4%), nitrogen (up to 2%) and traces of other elements (e.g., nickel and vanadium). The composition of the oil is classified in terms of the proportion of hydrocarbons and polar aromatic compounds.[1]

Oxygen-containing compound classes (naphthenic acids, O2 class; phenols, O1 class), are responsible for some undesirable properties, such as acidity (caused mainly by the presence of naphthenic acids), coloring, odors (phenols), the formation of emulsions and corrosion.[2] Among the main corrosive species, the naphthenic acids are evidenced in acidic crude oils, although they represent less than 3 wt %.[3] Naphthenic acids are defined as organic acids with the general formula R-$(CH2)_n$-COOH, where R is a radical including one or more cyclopentane or cyclohexane rings (Figure 1**)**. Due to naphthenic corrosivity effects and their biological marker role in geochemistry, many studies were focused on the identification of the naphthenic acids structures present in different crude oils.[4] This identification proved to be very difficult because naphthenic acids form complex mixtures. Some references mentioned that a single crude oil sample contains approximately 1500 different organic acids that are identified with molecular weights ranging from 200 to 700 Da.[3,5] More recent works on some crude oils identified naphthenic acid with a mass range of 115-1500 Da and a carbon content of C20 to C80.

The concentration of naphthenic acids in oils was one of the first tasks performed in the naphthenic corrosion studies. Currently, naphthenic concentrations are measured by titrating them with an alcoholic solution of potassium hydroxide (KOH), being

expressed by the total acid number (TAN) that represents the milligrams of KOH used to neutralize all of the acidic species in 1 g of oil sample. Crude oils with TAN > 0.5 mg of KOH/g may cause severe corrosion problems to refinery operations.[6-8] However, the value of the TAN is not directly correlated to the corrosivity of naphthenic acids. The TAN value depends upon the size and structure of the naphthenic acids and their interaction with other compounds present in the crude oil (sulfites, carbon dioxide, etc.).[6]

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) offers the highest available mass resolution, mass resolving power and mass accuracy, which enable the analysis of complex petroleum mixtures on a molecular level.[9] High-resolution MS data have demonstrated that it is possible to discriminate many different compounds[10,11] because of the different ionization efficiencies of the crude oil constituents.[12] Accurate mass measurements[13,14] allow unambiguous elemental composition ($C_cH_hN_nO_oS_s$) assignment and DBE (double bond equivalents), facilitating material classification by the heteroatom content and the degree of aromaticity.[15,16] Naphthenic acids can be analyzed by negative-ion electrospray ionization, ESI(-), coupled to FT-ICR MS, being detected in the form of a deprotonated molecule, $[M - H]^-$.

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. To relate the measured spectra to specific parameters, uni- and multi-variate calibration are often used, which are especially useful with parameters that are difficult to measure directly.[17,18] In addition, the use of mass spectrometry can reduce waste, minimizing the

consumption of raw materials and energy, thereby diminishing the environmental impact.

The aim of this work was the utilization of ESI(-)-FT-ICR MS technique in conjunction to PLS regression to find a relationship between the mass spectra and the TAN value. After that, by using variable selection methods, the dimension of the MS data was reduced and, most important, the main chemical species responsible for the relationship was identified. Then, it was possible to link the crude oil composition with the TAN parameter that is the fundamental goal of petroleomics.

2. Data Analysis

2.1. Partial Least Squares

In the partial least squares (PLS) approach, the matrix of instrumental responses (X) is related to the vector of property of interest (**y**) by the liner relationship presented in Equation 1:[19]

$$\mathbf{y} = \mathbf{Xb} \quad (1)$$

The arrays **X** and **y** are decomposed into latent variables, similar to the principal component analysis (PCA)[20], as represented by Equations 2 and 3,

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E} \quad (2)$$

$$\mathbf{y} = \mathbf{Tq}^t + \mathbf{f} \quad (3)$$

where **T** is the matrix of scores, $\mathbf{P}^t$ and $\mathbf{q}^t$ are the loadings, and **E** and **f** are the residues. The scores **T** are estimated from the weight coefficient **W** that is obtained to minimize the vector **f** and to maximize the relationship between **X** and y given by Equation 1,

$$\mathbf{T} = \mathbf{XW} \quad (4)$$

The regression coefficients used to relate **X** and **y**, are calculated according to Equation 5:

$$\mathbf{b} = \mathbf{W}\left(\mathbf{p}^t\mathbf{W}\right)^{-1}\left(\mathbf{T}^t\mathbf{T}\right)^{-1}\mathbf{T}^t\mathbf{y} \quad (5)$$

2.2. Variables Selection Methods

The variable selection methods have been used to produce simpler, robust and interpretable models. In this paper, the variables selection methods tested were: uninformative variable elimination (UVE), variable importance in projection (VIP) and interval partial least squares (iPLS).

2.2.1. Uninformative variable elimination (UVE)

The UVE proposed by Centner[21] is a method of variable selection based on the reliability analysis of regression coefficients (**b**) that represent the contribution of each variable to the established model, which is calculated through a leave-one-out validation method. The principle of UVE is to add noise (artificial variables) to the matrix of instrumental responses and develop a PLS model for the data set containing the experimental and the artificial variables. The reliability criterion c_j (for each j variable) is calculated based on the ratio of the regression coefficient $mean(b_j)$ and $std(b_j)$, which are the mean and the standard deviation, respectively, of the regression coefficients (Equation 6).

$$c_j = \frac{mean(b_j)}{std(b_j)}, \qquad \text{for} \quad j = 1,..., 2p \quad (6)$$

where p is the number of variables of the instrumental responses. To estimate a cutoff, i.e., the limit for the included and excluded variables, a criterion of informative or uninformative variable is formulated and presented in Equation 7:

$$cutoff = k * \max(abs(c_{noise})) \quad (7)$$

where k is an arbitrary value (normally 2), c_{nose} are the values for the artificial variables, and $\max(abs(c_{noise}))$ is maximum absolute value of the reliability criterion.[22]

2.2.2. *Variable importance in projection (VIP)*

VIP is a combined measure of how much a variable contributes to describe the two datasets as PLS regression.[23] The idea is to accumulate the importance of each variable *j* being reflected by weight w_h from each latent variable. The VIP measure v_j is defined by the Equation 8,

$$v_j = \sqrt{p \sum_{h=1}^{H} \left[SS_h \left(w_{hj} / \|w_h\|^2 \right) \right] / \sum_{h=1}^{H} (SS_h)} \quad (8)$$

where SS_h is the sum of squares described by the *h* latent variables. Hence, the v_j is a measure of the contribution of each variable according to the variance described by each PLS component, where $\left(w_{hj} / \|w_h\|^2 \right)$ represents the importance of the variable *j*. The $v_j <$ 1 (average of v) condition indicates a non-important variable that could probably be removed.

2.2.3. *Interval Partial Least Squares (iPLS)*

The interval Partial Least Squares (iPLS) is a method introduced by Norgaard[24], in which the matrix of instrumental response is divided in a defined number of intervals, and a PLS model is developed for each one of them. The sub-interval (or more than one sub-interval) that presented the smallest cross-validated error was selected. The iPLS can be an effective tool to determine the importance of different parts of a data set.

3. Experimental

3.1. Reagents and samples

Anhydrous propanol, toluene and potassium hydroxide, each with a purity higher than 99.5%, were purchase from Vetec, Brazil and used for the TAN determinations. Ammonium hydroxide (NH_4OH) and sodium trifluoroacetate (NaTFA) were purchased from Sigma–Aldrich, USA and used for the ESI(-)-FT-ICR MS measurements. In this study, thirty four crude oil samples from sedimentary basin of the Brazilian coast were used.

3.2. TAN determination

TAN measurements were performed according to the standard ASTM method (ASTM D664-09)[25] using a potentiometer (Metrohm Analytical Instruments and Accessories, USA) with a reproducibility of pH ± 0.005 mg KOH g-1. In 250 mL beakers, the crude oil samples (5 ± 0.5 g) were weighed, dissolved in 125 mL of water/anhydrous propan-2-ol/toluene (0.5/49.5/50 % in volume), and then titrated with a 0.1 mol L^{-1} alcoholic KOH solution. Silver/Silver chloride (Ag/AgCl) reference

electrode built into the same electrode body and a magnetic stirrer were used during the titration. The results are expressed as milligrams of potassium hydroxide per gram of sample required to titrate a sample in a solvent to a specified end point (see Table 1). The TAN values ranged from 0.06 to 3.61 mg KOH g^{-1}.

3.3. ESI(-)-FT-ICR MS measurements

Petroleum samples were analyzed by ESI(-). Briefly, the crude oil samples were diluted to ≈ 1.2 mg mL^{-1} in 50:50 (v/v) toluene/methanol (which contained 0.1% w/v of NH_4OH for ESI(-)). The resulting solution was directly infused at a flow rate of 5 µl min^{-1} into the ESI(-) source. The mass spectrometer (model 9.4 T Solarix, Bruker Daltonics, Bremen, Germany) was set to operate over a mass range of *m/z* 200-1300. The ESI(-) source conditions were as follows: a nebulizer gas pressure of 1.0 bar, a capillary voltage of 3.0-3.5 kV and a transfer capillary temperature of 250 °C. The ions are accumulated in the hexapolar collision cell over a time period of 0.15-0.20 s, followed by transport to the analyzer cell (ICR) through the multipole ion guide system (another hexapole). Each spectrum was acquired by accumulating 200 scans of time-domain transient signals in four mega-point time-domain data sets. The front and back trapping voltages in the ICR cell were - 0.60 V and - 0.65 V, respectively, for ESI(-). All mass spectra were externally calibrated using a NaTFA solution (*m/z* from 200-1200) after they were internally recalibrated using a set of the most abundant homologous alkylated compounds for each sample. A resolving power ($m/\Delta m_{50\%}$ ≈ 450 000, in which $\Delta m_{50\%}$ is the full peak width at half-maximum peak height) of *m/z* 400 and a mass accuracy of < 1 ppm provided the unambiguous molecular formula assignments for singly charged molecular ions. The mass spectra were acquired and

processed using a custom algorithm developed specifically for petroleum data processing, Composer software (Sierra Analytics, Pasadena, CA, USA). The MS data were processed, and the elemental compositions of the compounds were determined by measuring the *m*/*z* values. To help visualize and interpret the MS data, a typical plot was constructed, such as a heteroatomic-containing compounds diagram and a carbon number *versus* the double bond equivalents (DBE) plot, where DBE is defined as the number of rings plus the number of double bonds in a molecular structure. The aromaticity of a petroleum component can be deduced directly from its DBE value according to Equation 9,

$$DBE = c - h/2 + n/2 + 1 \quad (9)$$

where *c*, *h* and *n* are the numbers of carbon, hydrogen and nitrogen atoms, respectively, in the molecular formula.

3.4. Models development

After the mass attribution by the Composer software, the data set was transformed to a data matrix in the software Matlab. The optimization of the PLS models was performed using a k-fold cross-validation, and the best results were obtained using auto-scaling as a preprocessing step. In this study, 25 samples for calibration and nine samples for prediction were used. To facilitate the spectral interpretation, three different variables selection methods were tested: UVE, VIP and iPLS. The best method was chosen based on the error of prediction (RMSEP) determined using Equation 10, the number of selected variables and the region of the image corresponding to the original image obtained from the DBE vs. carbon number

plots. All of the PLS models were development in the software PLS Toolbox version 6.7 for Matlab.

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (10)$$

where $\hat{y}$ is the predicted TAN value by the PLS model and y is the reference TAN value given by ASTM D664-09.

4. Results and discussion

4.1. ESI(-) FT-ICR MS

Figures 2a-c displays the mass spectra of three typical Brazilian crude oils with different TANs: S1 (Figure 2a, TAN = 3.61 mg KOH g^{-1}), S4 (Figure 2b, TAN = 1.90 mg KOH g^{-1}) and S21 (Figure 2c, TAN = 0.17 mg KOH g^{-1}). The ESI(-)-FT-ICR MS technique was able to identify 5703 variables per sample (the number of molecular formula assignments from the Composer software). In general, the ESI(-)-FT-ICR mass spectra have Gaussian profiles from *m/z* 200 to 800 with an average molar mass distribution (M_w) centered from 520 (sample S4, 2b) to 550 Da (sample 2c). Heteroatom-containing species were detected as deprotonated molecules, $[M-H]^-$ ions, corresponding primarily to naphthenic acids, phenols and carbazole-analog species.

The naphthenic acid species concentration changes as a function of the TAN values. The enlarged area around *m/z* 417 (Figures 2a-c) highlights the decrease of the intensity of the ion $[C_{28}H_{50}O_2 - H]^-$, which has an *m/z* of 417.3738 and DBE of 4; whereas the ions $[C_{29}H_{38}O_2 - H]^-$ and $[C_{30}H_{42}O - H]^-$ of *m/z* 417.2799 and 417.3164, and DBEs of 11 and 10, respectively, increases in function of TAN.

Generally, petroleum samples have chemical compositions that differ significantly from one another. One way to display the similarities or differences between the signal patterns of crude oil samples is to build certain types of plots, such as plots of the relative abundances of different classes of compounds. In the present study, for the class profile diagrams shown in Figure 3, the relative amounts of each class were calculated by summing the abundances in one compound class and dividing by the total abundance of all species. The relative amounts of N, N2, NO, NO2, O and O2 class compounds present in the samples S01, S04 and S021 are shown in Figure 3. For sample S01 of higher acidity (TAN = 3.61 mg KOH g^{-1}), the O2 class, composed primarily of naphthenic acids,[26] was the most abundant class. The N class (carbazole analog species) was the second most abundant, followed by the O and NO2 classes (analogous to phenols and carbazoles with one carboxylic acid group or two hydroxylic groups). For sample S04 of intermediary TAN value, (TAN = 1.90 mg KOH g^{-1}), an increase of the relative abundance of non-basic nitrogen-containing compound classes is observed (N, N2, NO and NO2 classes), with the majority presence of phenol analogous compounds, O class, also evident for sample S21 of TAN = 0.17 mg KOH g^{-1}. As consequence of the relationship observed between the TAN values and the relative abundance of the O2 class, a plot of the relative abundance of the O_2-containing acids species *versus* the TAN value was built for the 34 samples analyzed, as shown in Figure 4, where a clear relation is observed similar to reported in literature.[18] Although these acids are the primary contributors to the overall acidity of the crude oils, other non-basic nitrogen compounds, phenols and non-polar sulfur components (not detected by ESI-FT-ICR MS), also contribute to the TAN value of the crude oils.

4.2. Variable selection methods

Figure 5a shows the original mass spectra and Figure 5b shows the spectra separated by classes, where the classes are organized in alphabetic order. This data reorganization was performed to facilitate the data analysis by chemometric methods and the results interpretation. Analyzing Figure 5, the need of data preprocessing is identified for the correction of the main differences between the intensity of the molecular ions. These differences were corrected by using auto-scaling as the preprocessing step, due to its performance in removing unwilling variation.

Figures 6, 7 and 8 show the procedure for variable selection by UVE, VIP and iPLS, respectively. For the UVE, a confidence level of 99% was defined, with six latent variables in the model, whereas for the VIP, VIP > 5 was chosen as the threshold, with five latent variables. For iPLS, the spectra were split in intervals of size of 100 variables, and four intervals with five latent variables were selected to obtain the best results of the RMSEP. It is important to know that the O class is 3731 to 4346 and the O2 class is 4347 to 5000 compound numbers. Note that in all figures, the majority of variables selected were between these intervals, being the most important to be used to predict the TAN value in the rearranged spectrum (Figure 5b).

Subsequently, a new PLS model was developed with the different variable selection methods, and the results are presented in Table 2, including the number of latent variables and the number of variables selected along with the respective calibration and validation errors. The F-test is found to exhibit a similar accuracy for all of the methods because $F_{exp} < F_{crit}$ with a confidence level of 95 %. Additionally, these errors were consistent with those obtained when all of the variables were used. However, now the model presents the advantage of the reduced number of variables generating more parsimonious models. Another important observation in Table 2 is that the number of

selected variables is similar between the VIP and UVE methods (the VIP and UVE methods reduced the amount of 5703 variables to 159 and 183 variables, respectively).

The choice of the most representative variable selection method can be based on the relationship of the selected variables with their respective molecular formula, thus identifying the species responsible for the TAN parameter (with the carbon number ranging from C_{15} to C_{69}, and the DBEs from 3 to 8). Therefore, it is possible to plot DBE versus the carbon number using the selected variables, as presented in Figure 9, for the O2 (a) O (b) and N2 (c) classes from the three chemometric methods used: VIP (a1,b1,c1), iPLS (a2,b2,c2) and UVE (a3,b3,c3). For this plot, it was chosen a sample with TAN of 1.18 mg KOH/g, an intermediated acidity value. For others samples, with different TAN values, different figures could be obtained. It is possible to see, by the observation of the images in Figure 9 that the region selected by UVE method for O2 and O classes corresponds to the correct range of carbon number in comparison with the image with all variables. This observation demonstrates that the UVE method reduced the number of variables with the same error of prediction of the other tested methods, while retaining the chemical information; therefore, it can be considered to be the best method for variables reduction.

The predicted TAN values obtained by UVE-PLS compared to the measured ones by ASTM method are shown in Figure 10. Note the clear correlation between the value of TAN obtained from ESI(-)-FT-ICR MS data and that obtained from conventional method. The RMSEP of 0.32 mg KOH/g (for nine validation samples) is less than the result obtained by the other studies[18] of 0.77 mg KOH/g (for 10 validation samples).

5. Conclusions

ESI(-)FT-ICR MS coupled to multivariate calibration produces a powerful analytical tool to predict the values of TAN for different crude oils. This approach combines the reliable information supplied from ESI(-) FT-ICR MS data with the possibility of variables reduction and the multivariate calculation supplied by chemometric methods, such as PLS.

Generally, the ESI(-)-FT-ICR mass spectra produced a data matrix containing a total number of variables of 5703 for the different crude oils studied. Among the three variable selection methods studied (variable importance in the projection (VIP), interval partial least squares (iPLS) and elimination of uninformative variables (UVE)), the UVE was found to be more appropriate to select the important variables for the model, reducing the dimension of the variables from 5703 to 183. This result was confirmed from the DBE *versus* carbon number plot, where maximum distributions for the abundant species are in good agreement. The RMSEP obtained was 0.32 mg of KOH g^{-1}, which is consistent with the error when using all variables. By reducing the size of the data, it was possible to relate the selected compounds with their corresponding molecular formula, thus identifying the relationship with the TAN of the oil.

In future works, other physical and chemical properties of crude oils will be predicted, such as density, viscosity, SARA method (saturate, aromatic, resin and asphaltene) and elemental analyses (percentage of carbon, hydrogen, nitrogen and oxygen), from non-polar compounds ionization, being then applied other ionization sources, such as laser desorption/ionization and atmospheric pressure photoionization.

Acknowledgments

This research was generously funded by PETROBRAS/CENPES, CNPq, CAPES, FAPESP and FINEP.

References

1. B. P. Tissot and D. H. Welte, *Petroleum Formation and Occurrence*, Berlin, Springer-Verlag, 1984.

2. J. G. Speight; *Handbook of Petroleum Analysis*, John Wiley and Sons, Inc.: New York, 2001.

3. M. P. Barrow, J. V. Headley, , K. M. Peru, P. J. Derrick, *Energy Fuels*, 2009, **23**, 2592-2599.

4. R. Piehl, *Mater. Perform*, 1988, **27**, 37-43.

5. L. B. Moura, R. F. Guimarães, H. F. G. Abreu, H. C. Miranda, S. S. M. Tavares, *Material Research*, 2012, **15**, 277-284.

6. Q. Shi, S. Zhao, Z. Xu, K. H. Chung, Y. Zhang, C. Xu, *Energy Fuels,* 2010; **24**, 4005–4011.

7. E. Slavcheva, B. Shone, A. Turnbull, *Corros. J.*, 1999, **34**, 125-131.

8. C. S. Hsu, G. J. Dechert, W. K. Robbins, E. K. Fukuda, *Energy Fuels*, 1999, **14**, 217-223.

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

10. A. Gaspar, W. Schrader, *Rapid Commun Mass Spectrom.*, 2012, **26**, 1047-1052.

11. A. M. McKenna, J. M. Purcell, R. P. Rodgers, A. G. Marshall, *Energy Fuels*, 2010, **24**, 2929-2938.

12. G. C. Klein, A. Angstrom, R. P. Rodgers, A. G. Marshall, *Energy Fuels*, 2006, **20**, 668-672.

13. C. S. Hsu, C. L. Hendrickson, R. P. Rodgers, A. M. McKenna, A. G. Marshall, *J. Mass Spectrom*., 2011, **46**, 337-343.

14. J. J. Savory, N. K. Kaiser, A. M. McKenna, F. Xian, G. T. Blakney, R. P. Rodgers, C. L. Hendrickson, A. G. Marshall, *Anal. Chem.*, 2011, **83**, 1732-1736.

15. R. L. Cunico, E. Y. Sheu, O. C. Mullins, *Pet. Sci. Technol.,* 2004, **22**, 787-798.

16. J. M. Purcell, I. Merdrignac, R. P. Rodgers, A. G. Marshall, T. Gauthier, I. Guibard, *Energy Fuels*, 2010, **24**, 2257-2265.

17. M. Benassi, A. Berisha, W. Romão, E. Babayev, A. Rompp, B. Spengler, *Rapid Commun Mass Spectrom.,* 2013, **27**, 825-834.

18. G. V. Boniek; V. A. Patrícia, F. C. R. Werickson, O. G. Alexandre; C. L. Rosana, Pereira, *Energy Fuels*, 2013, **27**, 1873-1880.

19. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 1986, **185**, 1-17.

20. S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37-52.

21. V. Centner, D. Massart, O. E. Noord, S.Jong, B.M. Vandeginste, C. Sterna, Anal. Chem., 1996, 68, 3851-3858.

22. Z. Xiaobo, Z. Jiewen, M. J.W. Povey, M.Holmes, M. Hanpin, *Anal. Chim. Acta.*, 2010, **667**, 14-32.

23. C. M. Andersen and R. Bro, *Journal of chemometrics*, **24**, pp.728-737, 2010.

24. L. Norgaard, A. Saudland, J.Wagner, J. Nielsen, L.Munck, S. Engelsen, *Applied Spectroscopy*, 2000, **54**, 413-419.

25. ASTM D664-09, Standard test Method for Acid Number of Petroleum Products by Potentiometric Titration, ASTM International, West Conshohocken, PA, 2006.

26. K. A. P. Colati, G. P. Dalmaschio, E. V. R. Castro, A. O. Gomes, B. G. Vaz, W. Romão, *Fuel,* 2013, **108**, 647-655.

Table 1. TAN values for the crude oil samples.

Samples	TAN (mg-KOH / g)	Samples	TAN (mg-KOH / g)
S 01	3.61	S 19	0.19
S 02	3.09	S 20	0.18
S 03	2.64	S 21	0.17
S 04	1.90	S 22	0.17
S 05	1.18	S 23	0.15
S 06	0.89	S 24	0.14
S 07	0.72	S 25	0.13
S 08	0.70	S 26	0.11
S 09	0.65	S 27	0.10
S 10	0.58	S 28	0.06
S 11	0.51	S 29	0.06
S 12	0.42	S 30	0.06
S 13	0.45	S 31	0.06
S 14	0.35	S 32	0.04
S 15	0.32	S 33	0.04
S 16	0.29	S 34	0.14
S 17	0.27		
S 18	0.24		

Table 2. RMSEP for the variable selection methods.

Variable Selection Method	**Number latent variables**	**Number select variables**	**RMSEC** (mg-KOH/g)	**RMSECV** (mg-KOH/g)	**RMSEP** (mg-KOH/g)	$*R_{cal}^2$
None	6	5703	0.12	1.05	0.34	0.89
VIP	5	159	0.13	0.80	0.35	0.98
iPLS	6	400	0.34	0.56	0.38	0.87
UVE	6	183	0.20	0.79	0.32	0.93

* Determination coefficient for calibration samples set.

Captions for the Figures

Figure 1. Some chemical structures of the naphthenic acids present in crude oil.

Figure 2. ESI(-)-FT-ICR MS for the crude oil samples S01(a)**,** S04 (b), e S21 (c). The insert shows that the intensity of ion $[C_{28}H_{50}O_2 - H]^-$ of *m/z* 417.3738 and DBE of four decreases, whereas the ions $[C_{29}H_{38}O_2 - H]^-$ and $[C_{30}H_{42}O - H]^-$ of *m/z* 417.2799 and 417.3164, and DBEs of 11 and 10, respectively, increases in function of TAN.

Figure 3. Heteroatom-containing compound class distribution from the ESI(-)-FT-ICR MS data of the crude oil samples S01 (high acidity), S04 (intermediate acidity) and S21 (low acidity).

Figure 4. Percentage of the O2 class *versus* the TAN values for 34 crude oil samples.

Figure 5. ESI(-) FT-ICR mass spectra of 34 crude oils samples. Original mass spectrum (a) and mass spectrum rearranged by classes (b).

Figure 6. UVE results. t values for experimental (1-5703) and artificial random (5704-11406) variables (a) and selected variables (b). The cutoff level at 0.99 is indicated by the dashed line.

Figure 7. VIP variables of the PLS model.

Figure 8. Illustration of the results from iPLS. The columns denote the RMSECV obtained from each of the intervals. The horizontal line denotes the RMSECV of the full-spectrum model.

Figure 9. DBE *versus* carbon number for O2 class (a), O class (b) and N2 class (c) using all variables and for variable selection methods: VIP (a1,b1,c1), iPLS (a2,b2,c2) and UVE (a3,b3,c3).

Figure 10. Plot of predicted by UVE-PLS *versus* the measurement by the ASTM method for the TAN values.

Figure 1

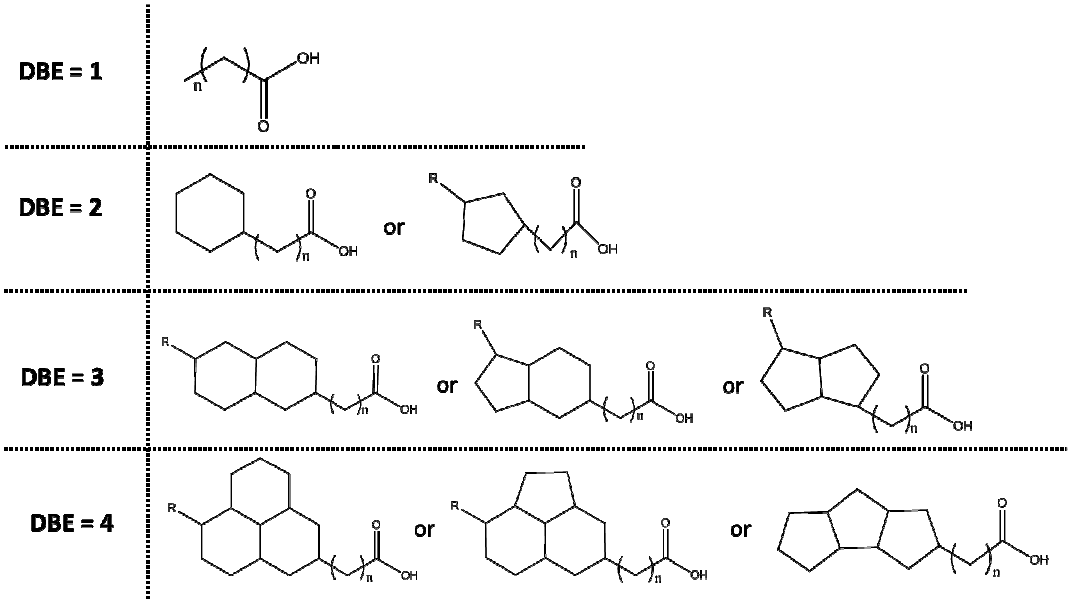


Figure 2

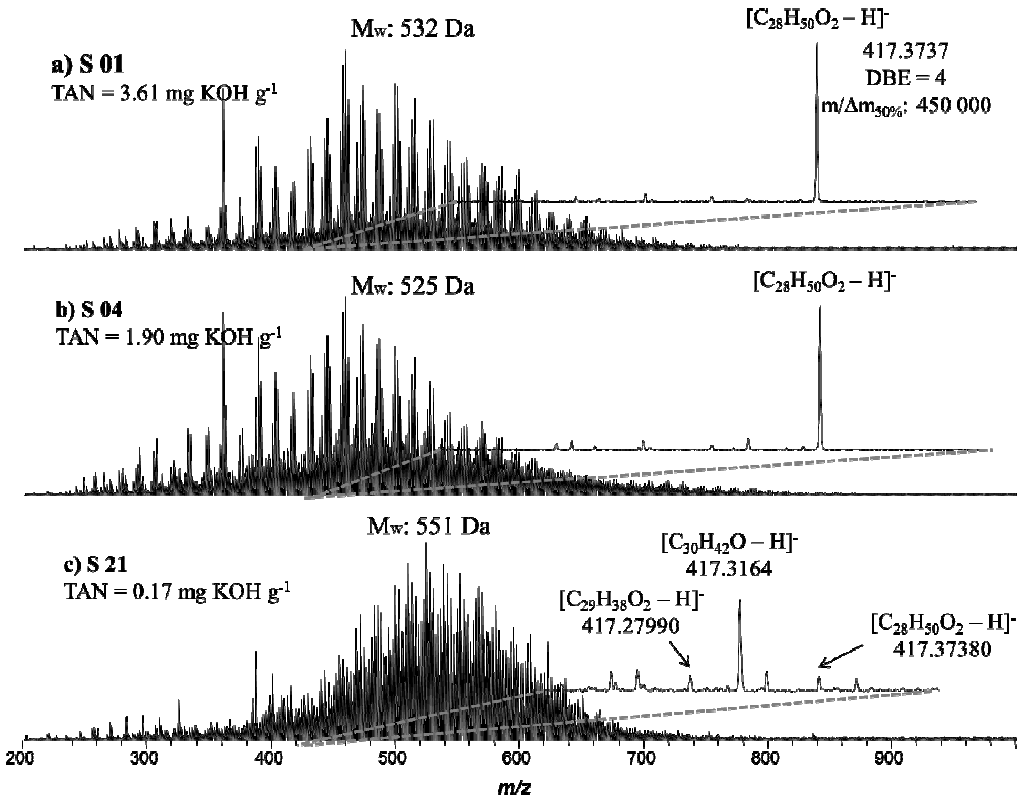


Figure 3

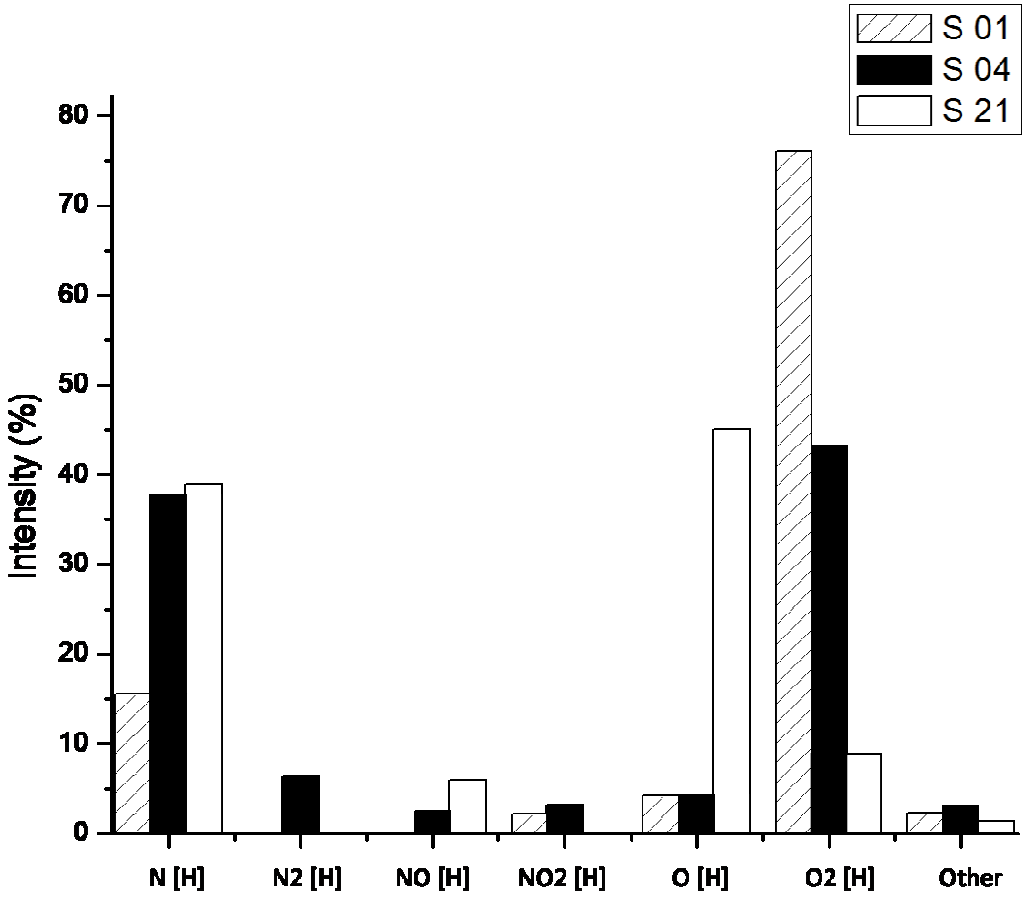


Figure 4

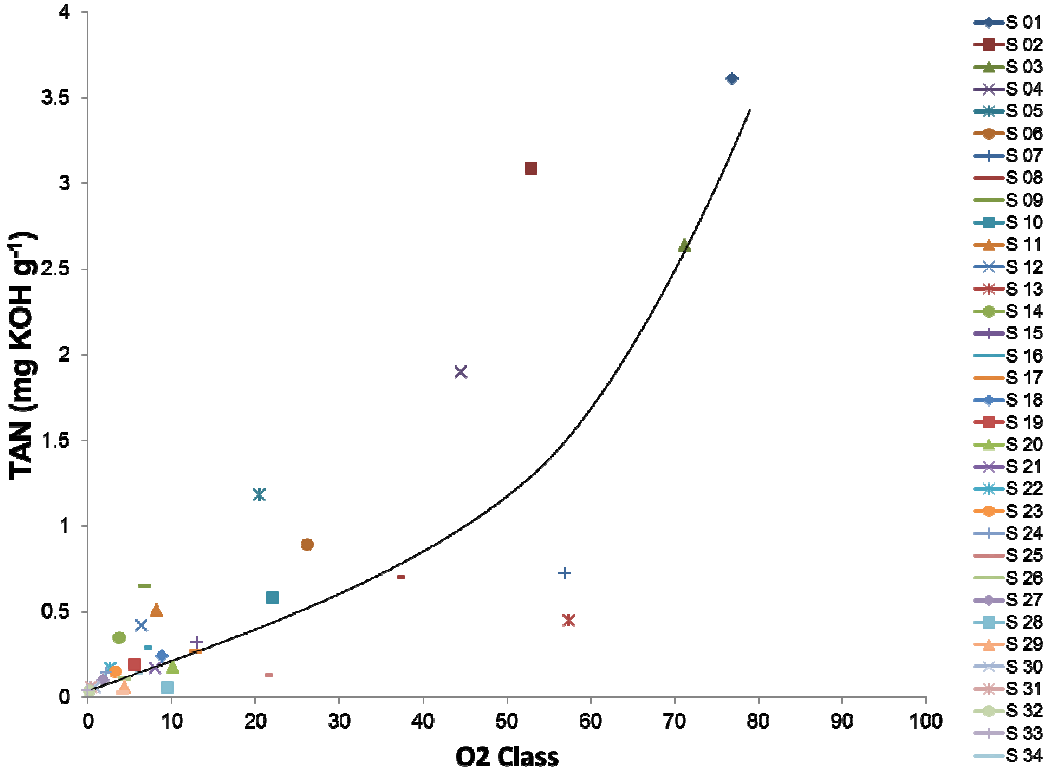


Figure 5

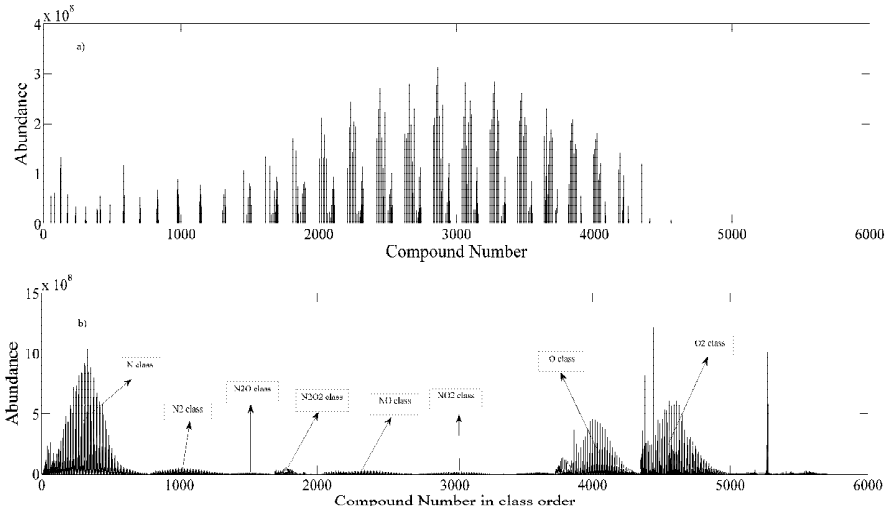


Figure 6

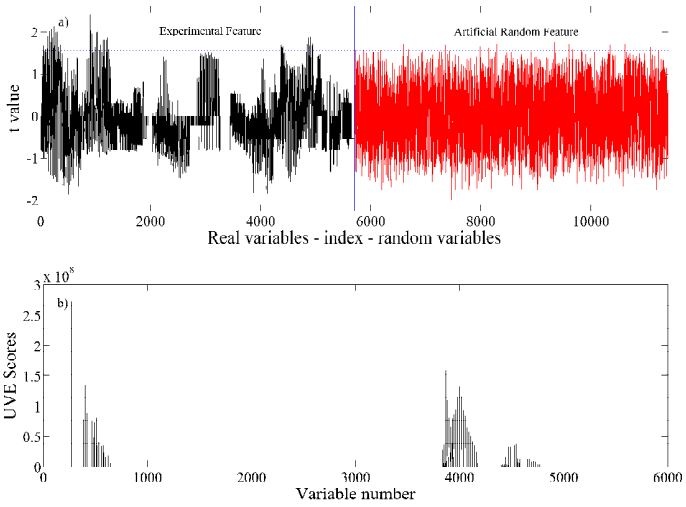


Figure 7

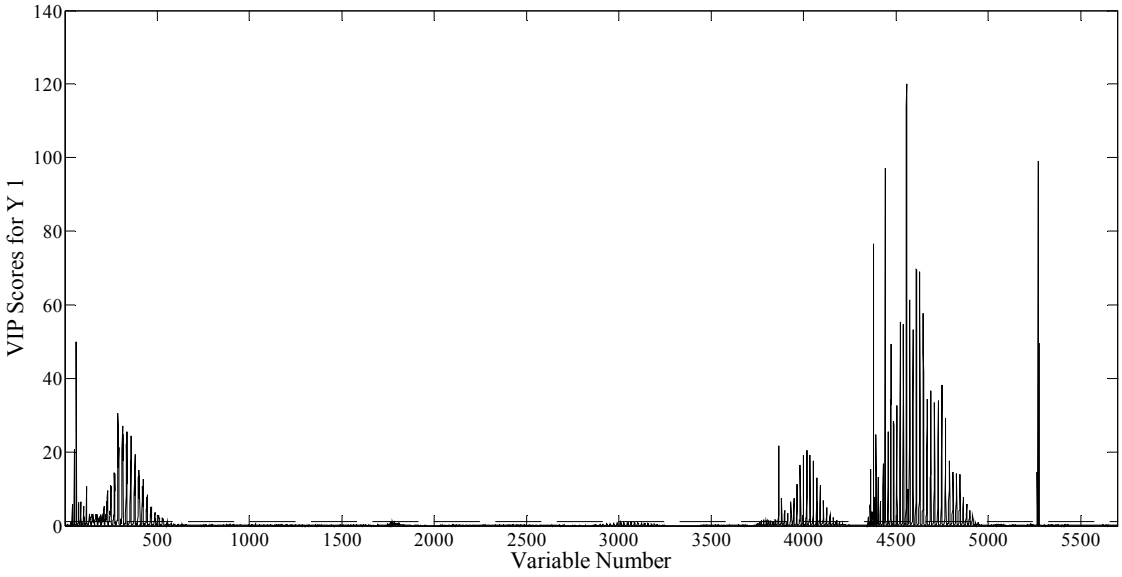


Figure 8

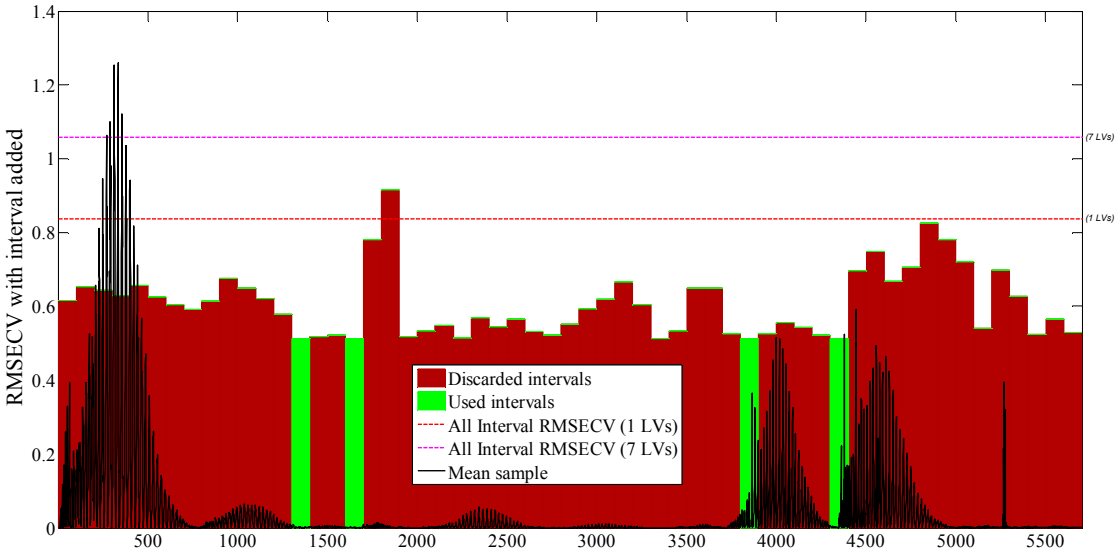


Figure 9

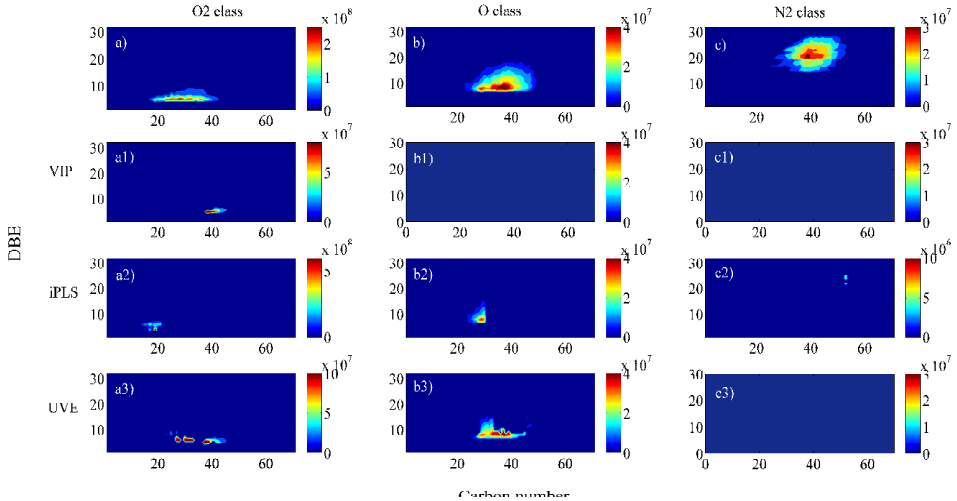


Figure 10