Full Length Article

Comprehensive and multidimensional tools for crude oil property prediction and petrochemical industry refinery inferences

Daniella L. Vale[a,*], Paula F. de Aguiar[a], Lize Mirela S.L. de Oliveira[b], Gabriela Vanini[a], Vinicius B. Pereira[a], Larissa O. Alexandre[a], Giovani S.C. da Silva[b], Luiz André Mendes[b], Alexandre O. Gomes[b], Débora A. Azevedo[a,*]

[a] Universidade Federal do Rio de Janeiro, Instituto de Química, Ilha do Fundão, Rio de Janeiro, RJ 21941-909, Brazil
[b] Petrobras/Cenpes, Ilha do Fundão, Rio de Janeiro, RJ 21941-909, Brazil

GRAPHICAL ABSTRACT
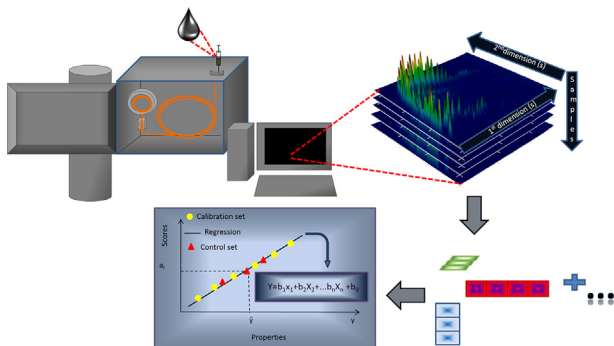
ABSTRACT

The association of comprehensive bidimensional gas chromatography with time-of-flight mass spectrometry (GC × GC-TOFMS) with high-order chemometric, N-way partial least squares (N-PLS), is an analytical innovation for the characterization of complex samples such as crude oil. The N-PLS method was applied to calibrate third-order data for sets of crude oil samples using whole oil comprehensive bidimensional chromatograms. The calibration model for API gravity had a bias equal to -5.8×10^{-3} and R^2_{Cal} of 0.9808 and WAT model had coefficient of determination for calibration model equal to 0.9436 and a bias of 8.7×10^{-3}. The results obtained by the decomposition of 11 components for API gravity were 99.79% for the X data and 98.08% for the Y data. The root mean square error for calibration (RMSEC) was equal to 0.81 and 1.01, while the root mean square error for prediction (RMSEP) was equal to 1.96 and 1.97 for the API gravity model and WAT, respectively and the explained variance obtained by decomposition in 9 components for WAT was 99.90% for the X data and 94.36% for the Y data. In the calibration models, all the errors for each sample were below 3.0 and 2.5 for °API and WAT, respectively. For the prediction set that was used to validate the model, the errors for each sample were below 3.0 and 3.2 for °API and WAT, respectively. The data indicates improvements for the correlation of petroleomic properties, thus allowing for the simultaneous prediction of certain properties instead of traditional analyses for each property when making inferences in the refining process. This application allows automation of the responses generated using crude oil samples without the need for pretreatment or fractionation steps; in addition, only one drop of each sample is required. This analytical application leads to cost reductions compared

* Corresponding authors.
    E-mail address: daniellavale@iq.ufrj.br (D.L. Vale).

to the numerous forms of traditional analyses currently used as well as lower sample and time consumption, thus allowing for environmentally cleaner characterization.

1. Introduction

Fossil fuels are currently the most important source of the world's energy resources. The quality of the oils influences the characteristics of the products to be obtained from them. Crude oil properties vary dramatically, which implies complex characterization. Crude oil may contain more than 100,000 different organic compounds that can influence these characteristics. This variation in physicochemical parameters results in the need to make operational decisions regarding refining processes. Conducting measurements using standard methods requires considerable time expenditures, large sample quantities, and the use of experts and specialized instrumentation, all of which result in higher economic costs. There is thus an need for methods that can predict various physical and chemical properties in a single analysis in order to estimate future crude oil cuts and refinement conditions, thus yielding economic benefits [1–7]. In this context, chemometric assessment is an important tool for application to modern analyses in the oil industry [8]. There is also a need for an analytical method that could promote comprehensive characterization aided by a chemometric tool able to generate a prediction model for various physicochemical properties [1,2,4,6].

Comprehensive bidimensional gas chromatography with time-of-flight mass spectrometry (GC × GC-TOFMS) may be applied to promote better knowledge of compositional information. This is a valuable analytical technique for analyzing complex samples such as petroleum because it presents higher peak capacity and chromatographic resolution and is capable of processing data in the order of 200 μs [6,9–11]. Associated with the greater power of differentiation of chemical species in a complex sample is the necessity of applying the proper treatment of large amounts of data generated to facilitate differentiation, and therefore the prediction of properties [12–14].

The GC × GC raw data generates a large and complex array of information called second-order data. When several samples are related to one another, they generate a data cube formed by stacking arrays [15]. Correlations with this type of data are often complex, which causes difficulties in making behavioral inferences without the use of multivariate analysis. It is necessary to use higher-order chemometric tools or multiway tools to organize this second-order data [16–18].

Although several higher-order techniques and restrictions may be used for second-order data, each has advantages and limitations; it is necessary to evaluate the suitability of the model, depending on the response modeled. PARAllel FACtor analysis (PARAFAC), PARAllel FACtor analysis 2 (PARAFAC2), and N-way partial least squares (N-PLS) analysis are the main iterative algorithms used for multivariate analysis of higher-order data. This type of algorithms is the most appropriate for generating a predictive model according to the analytical data generated by GC × GC in order to infer the operational parameters of a refinery. Each of these tools has characteristics such as trilinear data and "second-order advantage." Depending on the data set and the aim of the study, more than one tool might be suitable for modeling [12,15,18–24].

Several studies in the literature have described predictions of oil sample properties associated with analytical techniques and multivariate regression tools. Meléndez et al. developed the prediction of SARA (saturates, aromatics, resins, and asphaltenes) analysis of Colombian crude oils and obtained errors of between 1.3 and 3.7 [25]. Abbas et al. utilized Fourier transform infrared–attenuated reflectance (FTIR–ATR) associated with PLS analysis to predict API gravity and aliphatic/aromatic ratios and obtained the square error for prediction (SEP) of 1.66 [26]. In their study, Filgueiras et al. predicted

API gravity, kinematic viscosity, and water contents in Brazilian crude oil samples using infrared spectroscopy with attenuated total reflectance (FT-IR/ATR) with support vector regression (SVR) and PLS regression. The validation of the models showed RMSEP values of below 0.4 for API gravity and below 0.3 and 0.4 for kinematic viscosity and water contents, respectively [27]. Terra et al. used Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) coupled with PLS to estimate total acid number (TAN); the authors obtained RMSEP values below 0.4 for the crude oil samples [28]. Filgueiras et al. also estimated the temperature equivalent to 10% (T10%), 50% (T50%), and 90% (T90%) of distilled volume in crude oils using ^{1}H nuclear magnetic resonance (NMR) and SVR [2]. Duarte et al. used PLS associated with ^{1}H NMR spectroscopy to estimate petroleum sample properties and obtained RMSEP values equal to 0.8, 0.598% w/w, 3.8 °C, and 0.009% w/w for API gravity, carbon residue (CR), WAT, and basic organic nitrogen (BON), respectively [4]. Fakayode et al. determined the boiling point of petrochemicals by GC and PLS [29]. Other studies have used chemometric models for other purposes, such as toxicity identification in fresh and weathered crude oils [30]. Tomren et al. proved the potential of multivariate previsions by obtaining the density of crude oils by N-PLS with IR and GC data [31]. The same researchers compared prevision of asphaltene and acid number using PLS for GC data and FTIR data [7]. Godoy et al. presented a method for the differentiation of gasoline according to its geographical origin using comprehensive two-dimensional gas chromatography-flame ionization detection (GC × GC-FID) combined with multivariate analysis [32]. Besides that, several Brazilian commercial gasoline physicochemical parameters were predicted using PLS with GC data by Flumignan et al. [33]. Other studies focused on structural characteristics of petroleum asphaltenes using fluorescence emission spectroscopy and multivariate curve resolution alternating least-squares (MCR-ALS), besides quantification of polycyclic aromatic hydrocarbons by Parastar et al. [34,35].

Only Godoy et al., however, have described the application of an analytical technique using comprehensive and high-order information such as GC × GC associated with a multiway tool (namely PARAFAC) to model the properties of fuel samples. Although gasoline fuel is simpler than crude oil, this strategy may be used to develop new and significant analytical methodologies to solve various petroleomic problems. The potential of GC × GC-TOFMS for obtaining comprehensive characterization and third-order information associated with the N-way tool can improve petroleomic information and provide many properties, utilizing only one drop of each sample. The development of GC × GC-TOFMS associated with N-PLS is an analytically innovative method for characterizing complex samples such as crude oil.

As previously cited, GC × GC data is considered as third order when samples' data are overlapped, constructing a data cube or an array. This array can be decomposed to acquire high order methods or unfolding and later decomposed. The most appropriate algorithms used in literature and their advantages can be described in Table 1. The multivariate curve resolution-alternating least squares MCR-ALS is largely applied in GC and GC × GC data. However, the latter can only be apply in tree-way data if it was unfolding into bilinear two-way data matrix, like other methods based in unfolding process as unfolded-PLS (U-PLS) [24]. Likewise, the tools such as PARAFAC, PARAFAC2 and N-PLS can be directly applied in tree way data, such as GC × GC data, granting greater power of prediction and interpretability to the model [36]. Furthermore, N-PLS has the advantage of constructing a regression model directly using the X and Y scores.

The aim of this work is thus to develop a GC × GC-TOFMS method that would have the advantages of requiring minimal sample