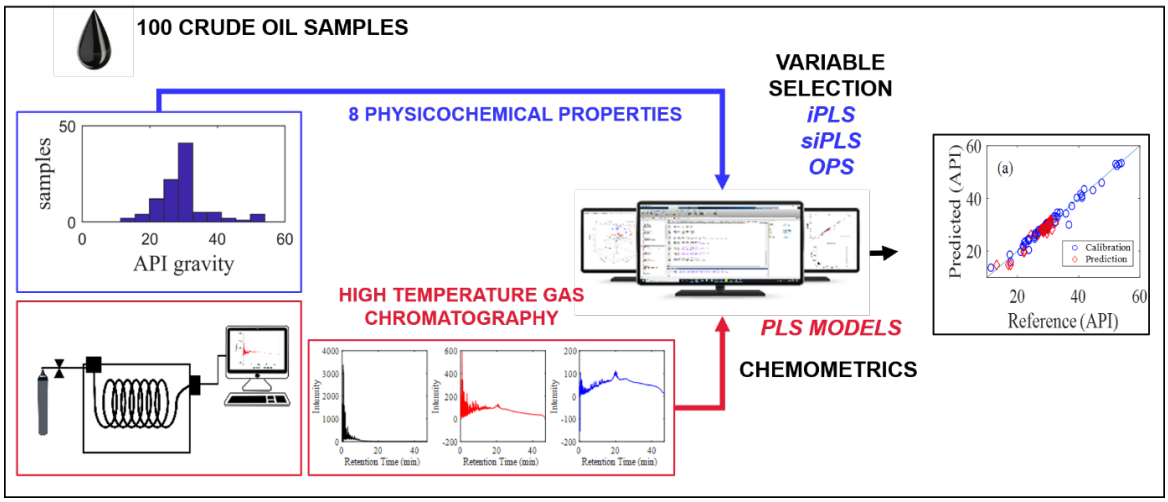
ISSN 2359-4799
Volume 10 / Número 1 / Ano 2024 – p. 01-27
DOI: 10.36524/ric.v10i1.2516

REVISTA Ifesciência



VARIABLE SELECTION METHODS APPLIED IN HTGC DATA TO DETERMINE PHYSICOCHEMICAL PROPERTIES OF CRUDE OILS

Ellisson Henrique de Paulo[1,2]*, Francine Dalapícola dos Santos[1], Gabriely Silveira Folli[1] Márcia Helena Cassago Nascimento[1], Mariana Kuster Moro[1], Pedro Henrique Pereira da Cunha[1], Samantha Ribeiro Campos da Silva[1], Eustáquio Vinícius Ribeiro de Castro[1], Alvaro Cunha Neto[1], Paulo Roberto Filgueiras[1], Marco Flôres Ferrão[2,3].

[1]Department of Chemistry, Center of Exact Sciences, Federal University of Espírito Santo – UFES, Goiabeiras, Vitória – ES, Brazil, Zip Code: 29075-910.

[2]Department of Inorganic Chemistry, Institute of Chemistry, Federal University of Rio Grande do Sul – UFRGS, Agronomia, Porto Alegre – RS, Brazil, Zip Code: 90650-001.

[3]National Institute of Science and Technology in Bioanalytics – INCT-Bio, Cidade Universitária Zeferino Vaz, Campinas – SP, Brazil, Zip Code: 13083-970.

*ellisson.hp@gmail.com

Article submitted in 06/11/2023, accepted in 16/02/2024 and published in 25/03/2024

ORCID – Ellisson H. de Paulo: https://orcid.org/0000-0001-9960-846X
ORCID – Francine D. dos Santos: https://orcid.org/0000-0002-3673-9427
ORCID – Gabriely S. Folli https://orcid.org/0000-0003-0665-7540
ORCID – Márcia H. C. Nascimento: https://orcid.org/0000-0001-5252-586X
ORCID – Mariana K. Moro: https://orcid.org/0000-0001-7726-802X
ORCID – Pedro H. P. da Cunha: https://orcid.org/0000-0003-1850-4664
ORCID – Eustáquio V. R. de Castro: https://orcid.org/0000-0002-7888-8076
ORCID – Alvaro Cunha Neto: https://orcid.org/0000-0002-1814-6214
ORCID – Paulo R. Filgueiras: https://orcid.org/0000-0003-2617-1601
ORCID – Marco F. Ferrão: https://orcid.org/0000-0002-3332-0540

Abstract: High-temperature gas chromatography (HTGC) is an analytical technique employed in the petroleum industry for component separation. By incorporating chemometrics, HTGC data can be effectively utilized to predict various properties of crude oil. However, HTGC chromatograms generate a substantial number of variables, some of which may lack pertinent chemical information. Consequently, employing variable selection methods becomes crucial to reduce the number of variables and enhance the predictive capability of calibration models. In this study, the interval partial least squares (iPLS), synergy interval partial least squares (siPLS), and ordered predictors selection (OPS) methods were applied for variable selection to construct linear regression models. The main objective was to investigate the potential of these methods in predicting eight properties of crude oil: American Petroleum Institute (API) gravity, standardized kinematic viscosity at 50 °C (VISp), flash point (FP), Reid vapor pressure (RVP), micro carbon residue (MCR), saturates (SAT), aromatics (ARO), and polar (POL) content. While all variable selection methods yielded satisfactory results, the OPS-PLS regression models consistently exhibited the best performance in estimating these properties, achieving root mean squared error of prediction (RMSEP) values of 1.244 for API, 0.029 for VISp, 15.356 °C for FP, 0.324 kPa for RVP, 0.629 wt% for MCR, 3.691 wt% for SAT, 2.939 wt% for ARO, and 3.374 wt% for POL. Variable selection demonstrated remarkable effectiveness, significantly improving the accuracy of the models, and allowing for the creation of concise models with a focused set of variables.

Keywords: Variables selection; crude oil; HTGC; PLS; OPS.

1 INTRODUCTION

Gas chromatography (GC) has emerged as a widely utilized technique in the petroleum industry for analyzing crude oil (BLOMBERG; SCHOENMAKERS; BRINKMAN, 2002). It offers several advantages, such as high sensitivity, efficient column performance, speedy analysis, and compatibility with complementary methods like mass spectrometry (POLLO et al., 2021; ZENG et al., 2012). High-temperature gas chromatography (HTGC) enables the separation of compounds at temperatures as high as 720 °C, in accordance with the ASTM D7169 standard method (ASTM D7169-16, 2016). This distinctive characteristic makes HTGC highly suitable for estimating the true boiling point (TBP) curve through simulated distillation (SIMDIS) analysis. Unlike traditional TBP analysis, which necessitates 18 liters of crude oil and takes 24 hours to complete, SIMDIS utilizing HTGC only requires 5 mL of the sample and can be accomplished within a mere 20 minutes. Consequently, HTGC presents significant savings in terms of time, sample volume, equipment, and labor when compared to the conventional TBP method (ASTM D7169-16, 2016; AUSTRICH; BUENROSTRO-GONZALEZ; LIRA-GALEANA, 2015; DE ANDRADE FERREIRA; DE AQUINO NETO, 2005; ESPIÑOSA-PEN; FIGUEROA-GOMEZ; JIME´NEZ-CRUZ, 2004; ZENG et al., 2012)

While GC is a well-established technique for fuel analysis, its application in the field of chemometrics for crude oil and its derivatives remains relatively limited (CHUA et al., 2020; DASZYKOWSKI; WALCZAK, 2006; LI et al., 2019). However, there have been notable early endeavors to explore this approach. In 1987, Telnaes et al. utilized principal component analysis (PCA) in conjunction with GC to analyze the distribution of phenanthrene in 36 crude oil samples (TELNAES et al., 1987). Similarly, Hupp et al. (2008) employed PCA and Pearson product moment correlation (PPMC) using gas chromatography-mass spectrometry (GC-MS) to differentiate 25 diesel samples, with a specific focus on aromatic compounds that exhibited significant discriminatory power (HUPP et al., 2008). Additionally, they identified the chemical components that contributed the most to the observed variance (HUPP et al., 2008). These studies demonstrate the potential of combining GC

with chemometrics for exploratory analysis and classification tasks. Moreover, this approach can be extended to quantification problems through the utilization of regression methods.

In a similar vein, Nascimento et al. (2018) employed HTGC in conjunction with detailed hydrocarbon analysis (DHA) to estimate the true boiling point (TBP) curve and predict flash point and Reid vapor pressure using partial least squares (PLS) models (NASCIMENTO et al., 2018). By combining these chromatographic methods, the researchers were able to develop predictive models that closely resembled the outcomes of the standard method. This success can be attributed to the fact that each method offered a complementary elution range compared to the other, enhancing the overall predictive capabilities (NASCIMENTO et al., 2018).

The combination of chromatography with chemometrics methodologies offers an alternative approach to analyzing chromatograms. This approach enables the extraction of a wide range of chemical information within a relatively short time and with reduced sample consumption when compared to standard physicochemical analysis methods. In this way, Medina et al. (MORALES-MEDINA; GUZMÁN, 2012), Rodrigues et al. (RODRIGUES et al., 2018), and Rocha et al. (ROCHA; SHEEN, 2019) used GC and HTGC to estimate physicochemical properties of biodiesel, crude oil, and its derivatives. Some physicochemical properties, such as saturates and aromatics content are directly related to chromatograms, whose cause-effect relationship is intimately explained by HTGC, which, in turn, makes chemometrics modeling from HTGC data effective and reliable (MERDRIGNAC, I. ESPINAT, 2007; RODRIGUES et al., 2018).

Nevertheless, a single HTGC chromatogram of a crude oil sample can generate an overwhelming number of variables, often exceeding 4,000. In such cases, employing variable selection methods becomes crucial to streamline computational processing, enhance accuracy, and facilitate the interpretation of prediction models (DE ARAÚJO GOMES et al., 2022). By employing these methods, the aim is to select the most relevant information that is highly correlated with the property of interest (DE ARAÚJO GOMES et al., 2022). This approach helps to reduce the complexity of the dataset, enabling more efficient analysis and improving the overall performance of the prediction models. Variable selection methods like genetic algorithms (BALLABIO et al., 2008; GUO et al., 2002), forward selection (BALLABIO et al., 2008), and LASSO (MA et al., 2018) are used to manage the vast amount of chromatographic data. Interval PLS (iPLS) (PEREIRA RAINHA et al., 2019; VIEIRA et al., 2019), synergy interval PLS (siPLS) (PEREIRA RAINHA et al., 2019; VIEIRA et al., 2019), and ordered predictors selection (OPS) (RIBEIRO et al., 2012; RIBEIRO; FERREIRA; SALVA, 2011; ROQUE et al., 2019) are emerging as effective variable selection approaches.

This study focuses on utilizing PLS regression and variable selection methods, namely iPLS, siPLS, and OPS-PLS on HTGC dataset. The objective is to predict various physicochemical properties of crude oil samples, including American Petroleum Institute (API) gravity, standardized kinematic viscosity at 50 °C (VISp), Reid vapor pressure (RVP), flash point (FP), micro carbon residue (MCR), saturates (SAT), aromatics (ARO), and polar (POL) content.

2 THEORETICAL BACKGROUNDS

Variable selection plays a fundamental role in simplifying models, improving their interpretability, and predictive performance. These methods allow for the identification of the most relevant variables while discarding those that contribute little or have insignificant impact on the analysis's objective. This not only conserves

computational resources but also reduces the risk of overfitting, which can occur when models are excessively tuned to irrelevant variables. The choice of the appropriate method depends on the type of data, the modeling algorithm, and the analysis's goal, making variable selection a critical step in data preparation and the construction of statistical and machine learning models.

The variable selection in chromatographic data is reported in the literature, such as genetic algorithm (GA) (ZHANG et al., 2018), forward selection (FS) (BALLABIO et al., 2008), variable importance in projection (VIP) (FARRÉS et al., 2015; PARK et al., 2013), selectivity ratio (SR) (FARRÉS et al., 2015), and least absolute shrinkage and selection operator (LASSO) (MA et al., 2018) for various purposes. To the best of our knowledge there are few works with application of variable selection in CG data for crude oil.

Other methods such as interval partial least squares (iPLS) and synergy interval partial least squares (siPLS) are already becoming the main algorithms for variable selection, based on selecting the intervals that generate the most accurate models (DE PAULO et al., 2022; PEREIRA RAINHA et al., 2019). In this way, a smaller set of variables is selected and used for regression (DE ARAÚJO GOMES et al., 2022; MEHMOOD; SÆBØ; LILAND, 2020). However, selecting intervals can often include variables without chemical significance or noise, producing less accurate models. To overcome this, it can be select discrete variables, rather than intervals, in the entire range of the chromatogram.

Ordered predictors selection (OPS) is a variable selection algorithm developed by Teófilo et al. in 2008 (TEÓFILO; MARTINS; FERREIRA, 2009). The OPS resize the original data matrix in descending order of importance. The variables are conditioned to a vector that carries information about the property of interest and shows which variables are the most important for the property (TEÓFILO; MARTINS; FERREIRA, 2009). Ribeiro et al. applied PLS regression on GC data to estimate sensory attributes of Arabica coffee and OPS algorithm to improve the prediction by selecting peaks for some compounds (RIBEIRO; FERREIRA; SALVA, 2011). Besides that, OPS have been widely applied in data set from QSAR (quantitative structure-activity relationship) (ROQUE et al., 2019; TEÓFILO; MARTINS; FERREIRA, 2009), nuclear magnetic resonance (DE PAULO et al., 2020, 2022, 2023; ROQUE et al., 2019), Raman (ROQUE et al., 2019; TEÓFILO; MARTINS; FERREIRA, 2009), infrared (CALIARI et al., 2017; FERREIRA et al., 2018; ROQUE et al., 2019; TEÓFILO; MARTINS; FERREIRA, 2009) and ultraviolet spectroscopy (ROQUE et al., 2019; ROQUE; DIAS; TEÓFILO, 2017), X-ray fluorescence (ROQUE et al., 2019; TEÓFILO; MARTINS; FERREIRA, 2009) and mass spectrometry (ROQUE et al., 2019; TEÓFILO; MARTINS; FERREIRA, 2009), voltammetry (ROQUE et al., 2019; TEÓFILO; MARTINS; FERREIRA, 2009), and GC (RIBEIRO et al., 2012; ROQUE et al., 2019; TEÓFILO; MARTINS; FERREIRA, 2009) in pharmaceutical, food and fuel areas.

3 METHODOLOGICAL PROCESSES/MATERIALS AND METHODS

3.1. Physicochemical analysis.

In this paper we used 100 crude oils samples from Brazilian coast sedimentary basin. API gravity was determined by the standard method ISO 12185 ("ISO 12185. Crude petroleum and petroleum products – determination of density – oscillating U-tube method.", 1996) following Equation 1, where ρ is the specific gravity of the sample.

$$\text{API} = \frac{141{,}5}{\rho} - 131{,}5 \qquad (1)$$

Kinematic viscosity (KVIS) at 50 °C was measured by ASTM D7042 (ASTM D7042., 2013) standard method and the VIS_p was obtained by log treatment on

KVIS according to Equation 2 (DIAS; AGUIAR, 2011).

$$\mathrm{VIS}_p = log(log(\mathrm{KVIS} + 0.7)) \quad (2)$$

RVP was obtained according to ASTM D323 (ASTM D323-15A, 2015). FP was determined following ISO 13736 (ISO 13736, 2006). MCR was measured following ASTM D4530 (ASTM 4530, 2015) standard method. SAT, ARO, and POL content were classified by ASTM D2549-02 (ASTM D2549, 2002) modified as described in previous works (FILGUEIRAS et al., 2016; RODRIGUES et al., 2018) using supercritical fluid chromatography/thin layer chromatography-flame ionization detector (SFC/TLC-FID).

3.2. Chromatographic analysis.

The HTGC analysis followed the ASTM D7169 (ASTM D7169-16, 2016) with extension of calibration (C_5-C_{120}) of *n*-paraffins. A chromatograph from Agilent Technologies, model 6890N was used. The equipment presented automatic injection system by programmable temperature; metallic capillary column, coated internally with polydimethylsiloxane of 5 m x 0.53 mm in internal diameter and 0.09-0.15 µm of stationary phase thickness; and flame ionization detector (FID). The assay was carried out in the followed chromatographic conditions: initial oven temperature at −20 °C, with programing from 10 ºC·min^{-1} to 430 °C, maintaining this temperature by 2 min; injector temperature at 430 °C and detector temperature at 435 °C; helium as carrier gas with a flow rate of 20 mL·min^{-1}.

All samples were diluted in carbon dissulfide (2 wt%) and injected in the column with a ramp of 50°C–430°C at a rate of 15°C·min^{-1}. The mixture of C_5-C_{28} light *n*-paraffins standard and the mixture of C_{30}-C_{120} heavy *n*-paraffins standard were used for the retention times calibration (Analytical Controls). The chromatograms were obtained in quadruplicate and the data was processed by Agilent Technologies' GC ChemStation software.

3.3. Data analysis.

The chromatographic data were used to build the matrix X, while physicochemical analysis provided the y vectors. The Icoshift (TOMASI; SAVORANI; ENGELSEN, 2011) algorithm was used to align chromatograms. Samples were split into 70% for calibration set and 30% for prediction set by Kennard-Stone algorithm (KENNARD; STONE, 1969). Before that, data set was preprocessed using one of these methods: normalization (NORM), mean centering (CENTER), autoscaling (AUTO), first order derivative (DERIV), and standard normal variation (SNV) methods (RINNAN; BERG; ENGELSEN, 2009). All chemometrics steps were carried out in the software MATLAB® R2013a (The Mathworks, Natick, USA).

3.3.1. Variables selection methods.

The iPLS, siPLS, and OPS algorithms were used to select variables, reducing the chromatographic matrix. The intervals with the lower error in both iPLS and siPLS modeling were selected to build a regression model. In each model the optimal number of latent variables (LVs), which minimize the root mean squared error of cross-validation (RMSECV), was selected (DE PAULO et al., 2022; PEREIRA RAINHA et al., 2019).

For OPS, it was needed to define some parameters to optimize the algorithm. First, the initial number of latent variables (hOPS) was used to build an informative vector that was used to sort the set of variables (MARTINS; FERREIRA, 2013; TEÓFILO; MARTINS; FERREIRA, 2009). The vectors used in this paper were the regression vector, obtained when a first PLS model is made with all variables set following Equation 3, where y is the dependent variable, *i.e.*, physicochemical data, X is the independent variable (retention time from HTGC chromatogram) and b_{REG} is the regression coefficient.

$$y = X \cdot b_{REG} \quad (3)$$

The correlation vector was built by the correlation between a variable in matrix X and its corresponding variable in vector y, measured following the Equation 4, where r is the correlation coefficient, I is the number of samples, X_a^t and y_a with subscript a, are the autoscaled matrix and vector for independent and dependent variables, respectively.

$$r = (X_a^t \cdot y_a)/(I-1) \quad (4)$$

Finally, the product vector is made by the product between regression and correlation vectors. This vector carries a lot of information from data set and was the major vector used to resize the X matrix. After obtaining the vector, OPS algorithm needs two subsets named window and increment. The first one is the initial number of variables in the current matrix and the second one is the set of variables that will be added to the window by the OPS algorithm. The percentage of variables that will be analyzed by the algorithm and the number of variables removed in cross-validation step are chosen together (MARTINS; FERREIRA, 2013; TEÓFILO; MARTINS; FERREIRA, 2009; VALE et al., 2018).

The informative vector is compared to original data set and according to intensity of the vector signal, the variables are ordered by descending importance (TEÓFILO; MARTINS; FERREIRA, 2009). Thus, the window and increments subsets are determined and PLS model is built to estimate cross-validation parameters. PLS regression is performed until all increment subsets are analyzed. The new subset is chosen based on the PLS model with the lowest RMSECV values (TEÓFILO; MARTINS; FERREIRA, 2009).

3.3.2. PLS modeling.

The intervals selected by iPLS and siPLS, and the new data set chosen by the OPS algorithm, were used to build PLS regression models. For comparison purposes, full chromatograms were also used to build models.

To avoid overfitting to the calibration data, the cross-validation method k-fold was applied during calibration with PLS modeling (LILAND; STEFANSSON; INDAHL, 2020). Thus, LVs were selected for modeling optimization. After that, the evaluation parameters were calculated for the built PLS models. To evaluate which model presented the best adjust and prediction capacity, the root mean squared error of calibration (RMSEC) and prediction (RMSEP) were used (OLIVIERI, 2015). The parameters were calculated according to Equations 5 and 6, where y_i is the property reference value, $\hat{y}_i$ is the property value predicted by the model, $\bar{y}_i$, is the mean reference value, $ncal$ is the number of samples used for calibration, and $npred$ is the number of samples used for prediction. Besides, coefficients of determination (R^2) were calculated using Equation 7 for calibration and prediction as well (OLIVIERI, 2014).

$$RMSEC = \sqrt{\sum_{i=1}^{ncal} \frac{(y_i - \hat{y}_i)^2}{ncal - LV - 1}} \quad (5)$$

$$RMSEP = \sqrt{\sum_{i=1}^{npred} \frac{(y_i - \hat{y}_i)^2}{npred}} \quad (6)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \quad (7)$$

4 RESULTS AND DISCUSSION

The methodology of multivariate data analysis combined to HTGC chromatograms allows the identification of crude oil features with high accuracy. It can dramatically reduce standard crude oil characterization methods by saving time and volume.

4.1. Physicochemical properties.

Figure 1 shows the range of each property for all samples. These oil samples presented API gravity ranging from 11.4 to 54.0 API as can be seen in Figure 1a. Many samples had API gravity higher than 31, which characterizes light oils, according to the API classification (SPEIGHT, 2015). Some of them showed API between 22 and 31, which characterizes intermediary oils. A

few samples are classified as heavy crude oils (API lower than 22).

The VIS_p showed values with high distribution (Figure 1b). This property mainly affects oil handling and transportation. Viscosity, as well as API gravity, is one of the main physicochemical properties for assessing oil quality in industry (SPEIGHT, 2015). Due to its relationship with temperature, it is necessary to improve the technology and equipment used in transportation of oil to avoid flow issues in different temperatures. Thus, the process of estimating the property in any temperature enables decision-making process regarding transportation problems (RIAZI, 2007; SPEIGHT, 2015).

The distribution of RVP values of our samples (Figure 1c) is characteristic of Brazilian oils (less than 70 kPa) (BRAZILIAN PETROLEUM, NATURAL GAS; (ANP), 2016). Some samples have RVP below 38 kPa, characteristic value of aviation gasoline (38 to 49 kPa). Other samples ranged from 45 to 54 kPa, values found in commercial automotive gasoline (45 kPa to 69 kPa). No sample presented RVP above 70 kPa, characteristic value of condensed gas.

Flash point (FP) (Figure 1d) of a hydrocarbon or a fuel is defined as the lowest temperature at which its vapor pressure is sufficient to produce the needed vapor for spontaneous ignition with the air and an external heat source, such as a spark or a flame (RIAZI, 2007; SPEIGHT, 2015). FP is related to the volatility of a fuel and, therefore, the presence of light and volatile components. FP indicates the maximum temperature that it can be stored without serious fire hazard. It is directly related to the safe storage and handling of such crude oil products (RIAZI, 2007; SPEIGHT, 2015).

Another important property is SAP, usually used for petroleum assessment. As can be seen Figure 1e, saturated components are the most abundant (up to 60%), being composed of normal chain (paraffins), branched (isoparaffins) and cyclic (naphthenic) hydrocarbons (RIAZI, 2007; SPEIGHT, 2015). ARO, in turn, represented around 30% of the oil composition and it is formed by single and polyaromatic carbon rings in structures (Figure 1f) (RIAZI, 2007; SPEIGHT, 2015). POL components represent around 10% of the oil composition, as can be seen in Figure 1g (RIAZI, 2007; SPEIGHT, 2015). This class is predominantly polar, due to heteroatoms, such as O, N, and S, presented in this structure (RIAZI, 2007; SPEIGHT, 2015).

Knowing oil's composition, it is possible to determine its quality and economic value. Oils composed mainly by saturates and aromatic hydrocarbons have higher economic value, because they are easier to refine, since they generally have low molecular weight. On the other hand, oils with high polar content are usually more undervalued as they present greater challenges during refining for industry. SAP classification is very important in stability studies during oil transportation, because precipitation of organic compounds in refinery pipelines is related to the proportion and interaction between these classes (MERDRIGNAC, I. ESPINAT, 2007; RIAZI, 2007; SPEIGHT, 2015). The physicochemical properties of oils vary considerably depending on the constituent substances of each class. MCR indicates the number of lubricant oils that can be produced in the refining process. Furthermore, this parameter also indicates the possibility of deposit formation in injectors and engines by the residue generated during the combustion of a fuel (DUARTE et al., 2016). In our sample set, most samples presented MCR between 0 and 5 wt%, and few samples between 10 and 15 wt% (Figure 1h).

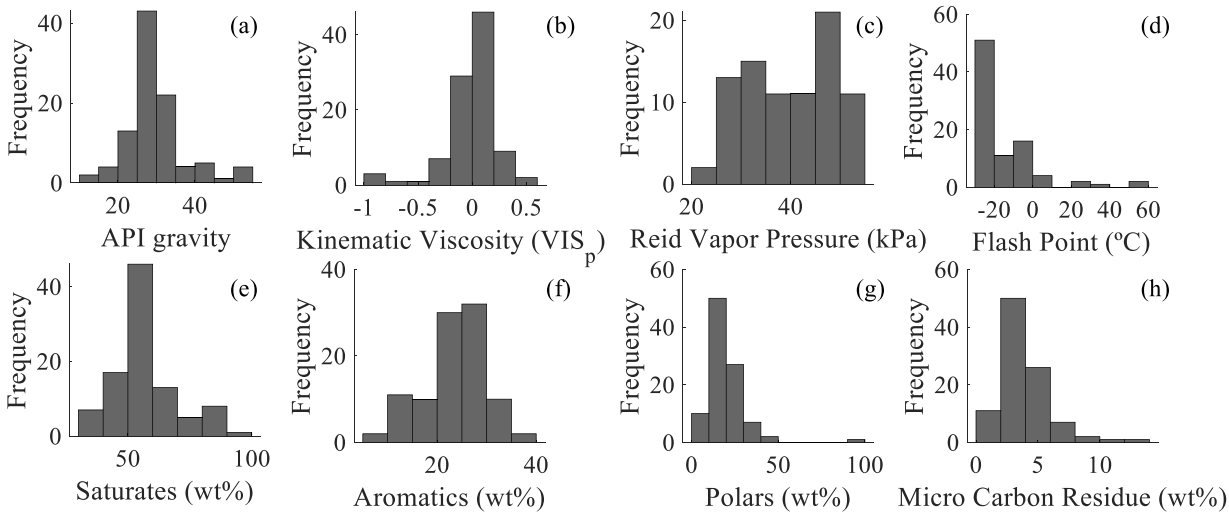


Figure 1. Histograms of sample distribution for API gravity (a), VIS_p (b), RVP (c), FP (d) SAT (e), ARO (f), POL (g), and MCR (d).

Figure S1 shows the correlation between the studied properties. API gravity is directly correlated with SAT, which was already expected, since the higher SAT, the higher the paraffin content and, therefore, the lighter the oil. API is also direct, but less, correlated with RVP. All the other properties are indirectly correlated with API, especially MCR, ARO, and POL. Both MCR and FP are inversely correlated with SAT and directly correlated with ARO and POL. The values of the correlation coefficients between each property can be found in table S1 of the supplementary material.

4.2. HTGC.

Figure S2 shows HTGC chromatograms for a light, an intermediary, and a heavy oil sample. The major difference between the chromatograms is in the first minutes of retention, where we can see the predominance of *n*-alkanes peaks for the light and intermediate samples, which does not occur for the heavy sample. From light to heavy samples, there is a decrease of n-alkanes quantity, which reduces the peak intensity in the first minutes. The profile shown of the heavy oil chromatogram is the profile of biodegraded oil, whose main characteristic is the lower peak intensity of *n*-alkanes (LARTER et al., 2012; SILVA et al., 2020).

4.3. Variables selected.

Figure S3 shows HTGC chromatogram versus OPS vector graphics. The regression and correlation vectors were used to build a product vector. This last one shows similarities to regression vector which makes it possible that the most importance on it comes from regression vector.

The peaks with higher intensity are mostly in the first thousand variables and some other ones after two thousand. Those chromatogram regions are primarily due to light and intermediary *n*-paraffin with chains ranging up to n-C_{30}, approximately (RIAZI, 2007). Figure 2A shows the sets of variables selected in OPS algorithm for API gravity. OPS selected 450 variables between 0 to 9 minutes (min), 11 to 12 min, 14 to 20 min and in 46 min, which corresponds retention time to compounds n-C_5 to n-C_{18}, n-C_{22} to n-C_{24}, n-C_{30} to n-C_{48}, and above n-C_{100} respectively (ASTM D7169-16, 2016). These regions are represented by light, intermediary and heavy compounds, respectively. The iPLS model selected the retention times from 0 to 2 min (n-C_5 to n-C_9), 7 to 9 min (n-C_{16} to n-C_{18}), 14 to 16 min (n-C_{28} to n-C_{34}), and 26 to 28 min (n-C_{78} to n-C_{94}), totalizing 940 variables. siPLS selected the variables from 0 to 18 minutes (n-C_5 to n-C_{40}), totalizing 1,880 variables (ASTM D7169-16, 2016).

For VIS_p, OPS selected 4,230 variables throughout the entire retention time range (from 0 to 46 min), as can be seen in Figure 2B. A total of 940 variables were selected by iPLS from 0 to 4 min, 7 to 9 min, and 23 to 25 min from *n*-C_5 to *n*-C_{12}, *n*-C_{15} to *n*-C_{18}, and *n*-C_{62} to *n*-C_{72}, respectively. Thus, siPLS selected 1,880 variables in retention time from 0 to 18 min (*n*-C_5 to *n*-C_{40}). For FP, iPLS selected compounds of type *n*-C_5 to *n*-C_{18} (0 to 9 min), siPLS selected compounds of type *n*-C_{12} to above *n*-C_{100+} (4 to 37 min), and OPS selected compounds of type *n*-C_5 to *n*-C_{100+} (0 to 31 min and 38 to 44 min) (Figure 2C). It selected a total of 940 variables for iPLS, 3,290 for siPLS, and 550 for OPS (ASTM D7169-16, 2016).

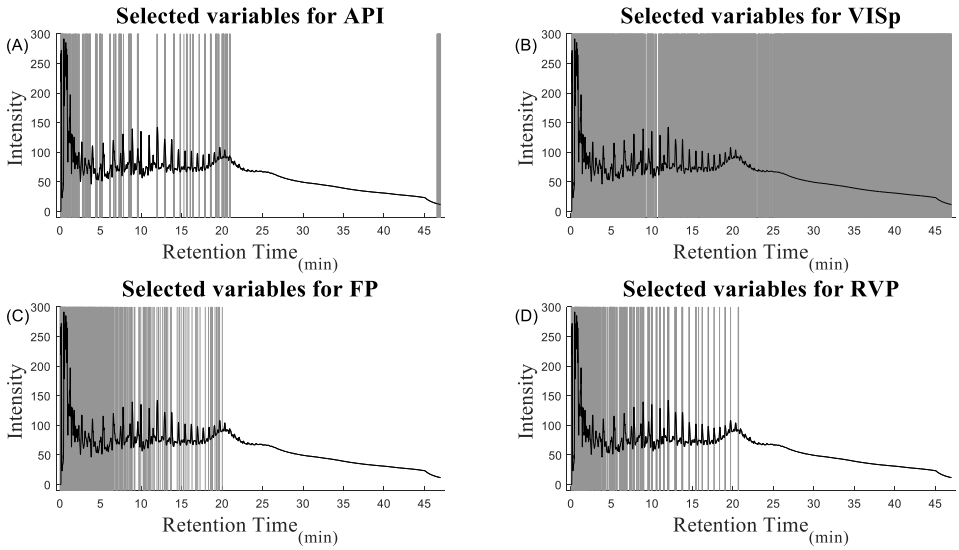


Figure 2. Sets of variables selected in OPS algorithm for API (a), VIS_p (b), FP (c), and RVP (d).

RVP was related to compounds from 0 to 4 min (*n*-C_5 to *n*-C_{12}) and 23 to 25 min (*n*-C_{62} to *n*-C_{72},) using iPLS. The siPLS model selected 23 to 42 min (*n*-C_{62} to *n*-C_{100+}), 0 to 20 min, that is, *n*-C_5 to *n*-C_{48} (Figure 2D). It was selected a total of 705 variables for iPLS, 1,880 for siPLS, and 400 for OPS (ASTM D7169-16, 2016).

Figure 3A, 3B, 3C, and 3D show the variables selected by OPS algorithm, respectively, for SAT, ARO, POL, and MCR. Variables carrying information of saturated compounds were selected in retention times by iPLS between 0 and 4 min, 7 and 9 min and between 23 and 28 min (*n*-C_5 to *n*-C_{12}, *n*-C_{16} to *n*-C_{18} and *n*-C_{62} to *n*-C_{94}, respectively). siPLS selected variables from 0 to 14 min (*n*-C_5 to *n*-C_{30}) and OPS selected variables from 0 to 15 min and18 to 20 min (*n*-C_5 to *n*-C_{32}, *n*-C_{40} to *n*-C_{48}, respectively) (Figure 3A). It was selected a total of 1,175 variables for iPLS, 1,410 for siPLS, and 450 for OPS. The selected regions are related to hydrocarbons formed by chains lower than C_{50}, demonstrating that the class is mainly conditioned to these compounds. According to Zeng et al. branched *n*-alkanes have a low boiling point compared to equivalent but normal chain *n*-alkanes (ZENG et al., 2012). Thus, the selection of some variables at lower boiling points indicates greater importance of branched *n*-alkanes for chemometrics modeling (ASTM D7169-16, 2016).

Retention times related to aromatics compounds (Figure 3B) were selected by iPLS from 0 to 28 min (*n*-C_5 to *n*-C_{94}), by siPLS from 0 to 18 min (*n*-C_5 to *n*-C_{40}), and by OPS from 0 to 41 min (*n*-C_5 to *n*-C_{100+}), totalizing 2,820, 1,880, and 250 variables, respectively. To estimate POL (Figure 3C), 3,995 variables were selected by iPLS from 0 to 9 min and from 16 to 47 min (*n*-C_5 to *n*-C_{18} and *n*-C_{34} to *n*-C_{100+}, respectively), 3,290 variables were selected by siPLS from 4 to 47 min (*n*-C_{12} to *n*-C_{100+}), and 250 variables were selected by OPS at 7 min and

from 28 to 43 min (n-C_{16}, n-C_{62} to n- C_{100+}). For MCR (Figure 3D), iPLS selected regions from 0 to 2 min, 7 to 11 min and 23 to 25 min (n-C_5 to n-C_8, n-C_{16} to n-C_{22} and n-C_{62} to n-C_{94}), siPLS selected regions from 23 to 32 min (n-C_{62} to n- C_{100+}), and OPS selected from 0 to 21 min and at 46 min (n-C_5 to n-C_{52} and n- C_{100+}). iPLS, siPLS and OPS selected a total of 940, 940, and 600 variables for MCR, respectively (ASTM D7169-16, 2016).

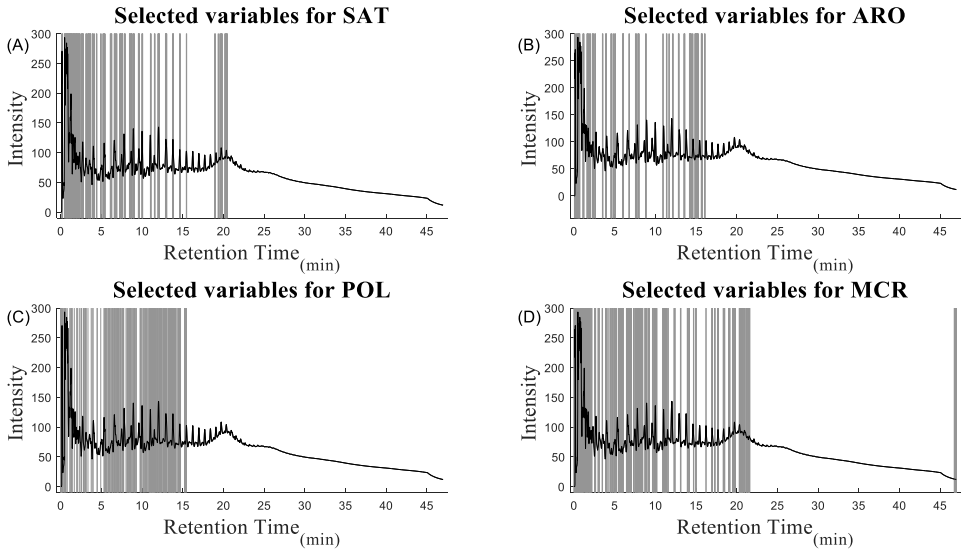


Figure 3. Sets of variables selected in OPS algorithm for SAT (a), ARO (b), POL (c), and MCR (d).

4.4. Regression models.

iPLS, siPLS, OPS-PLS, and PLS models with the full chromatogram were built to predict crude oil physicochemical properties. The main parameters for each model are shown in Table 1. The number of latent variables ranged from 3 to 8, while autoscaling, normalization and SNV methods predominated in the data preprocessing methods. The higher the R^2 value (closer to 1) and the lower the RMSEC and RMSEP values, the higher the model quality.

Table 1. Statistical parameters of PLS, iPLS, siPLS, and OPS-PLS models for API gravity, VIS_p, RVP, FP, SAT, ARO, POL, and MCR.

		Parameters						
Property	Model	Variable	Pretreat[a]	LV[b]	RMSEC (wt%)	RMSEP (wt%)	R^2_c	R^2_p
API	PLS	4,700	AUTO	6	1.71	1.15	0.96	0.94
	iPLS	940	AUTO	6	1.59	1.34	0.97	0.93
	siPLS	1,880	AUTO	8	1.37	1.48	0.98	0.91
	OPS-PLS	450	AUTO	7	1.41	1.24	0.97	0.93
VIS_p	PLS	4,700	SNV	4	0.07	0.06	0.94	0.89
	iPLS	940	SNV	3	0.09	0.08	0.87	0.84
	siPLS	1,880	AUTO	6	0.05	0.03	0.96	0.93
	OPS-PLS	4,230	AUTO	5	0.05	0.03	0.96	0.94
RVP	PLS	4,700	NORM	4	0.41	0.37	0.99	0.99
	iPLS	705	NORM	8	0.24	0.32	0.99	0.99

	siPLS	1,880	AUTO	3	0.49	0.48	0.99	0.99
	OPS-PLS	400	NORM	5	0.31	0.32	0.99	0.99
FP	PLS	4,700	AUTO	7	7.24	16.34	0.82	0.72
	iPLS	940	NORM	6	8.57	16.71	0.75	0.65
	siPLS	3,290	AUTO	7	7.34	16.99	0.82	0.68
	OPS-PLS	550	AUTO	7	7.99	15.36	0.79	0.79
SAT	PLS	4,700	SNV	3	6.01	3.73	0.82	0.73
	iPLS	1,175	SNV	3	6.93	3.78	0.76	0.73
	siPLS	1,410	SNV	3	7.17	4.77	0.75	0.58
	OPS-PLS	450	NORM	3	6.3	3.69	0.8	0.76
ARO	PLS	4,700	SNV	3	4	3.53	0.61	0.7
	iPLS	2,820	SNV	4	3.99	3.57	0.62	0.73
	siPLS	1,880	SNV	3	4.27	3.72	0.56	0.71
	OPS-PLS	250	SNV	7	3.72	2.94	0.69	0.8
POL	PLS	4,700	SNV	4	4.22	5.15	0.78	0.64
	iPLS	3,995	SNV	5	4	5.39	0.8	0.6
	siPLS	3,290	SNV	3	5.23	4.03	0.64	0.79
	OPS-PLS	250	SNV	7	4.14	3.37	0.79	0.86
MCR	PLS	4,700	NORM	5	1.06	0.7	0.78	0.85
	iPLS	940	NORM	6	0.9	0.79	0.84	0.82
	siPLS	940	NORM	6	1.2	0.8	0.72	0.83
	OPS-PLS	600	NORM	3	1.17	0.63	0.72	0.88

Source: The authors.

For API gravity, the PLS model from the full chromatogram provided the lowest RMSEP (1.15 API), however, OPS showed comparable results (RMSEP 1.24 API), using a smaller number of variables. The iPLS and siPLS models also presented RMSEP values closer to OPS and PLS. Medina et al. predicted the API gravity in crude oil using CG data and PLS regression (MORALES-MEDINA; GUZMÁN, 2012). They reported an R^2_p equal to 0.82 and RMSEP equal to 1.4 API. Rodrigues et al. also used HTCG to estimate API gravity, obtaining a R^2_p of 0.951 and a RMSEP of 1.7 (RODRIGUES et al., 2018).

Medina et al. also predicted kinematic viscosity, obtaining a R^2_p of 0.89 and a RMSEP of 2.6 $mm^2 \cdot s^{-1}$ (MORALES-MEDINA; GUZMÁN, 2012). Rodrigues et al. reported a RMSEP of 0.31 $mm^2 \cdot s^{-1}$ and a R^2_p of 0.911 for kinematic viscosity at 50 °C (RODRIGUES et al., 2018). In this study, we estimated this property with a RMSEP equal to 0.029 and a R^2_p equal to 0.94, using the OPS-PLS method. For VIS_P modeling, OPS-PLS selected almost all chromatographic variables (4,230 variables) but provided the best model among all obtained models.

For RVP, the lowest values of RMSEP were achieved using iPLS and OPS-PLS methods (0.322 kPa and 0.324 kPa, respectively) as well as a R^2_p of 0.96 in both cases. Nascimento et al. used PLS in HTGC to predict RVP, reporting a R^2_p of 0.99 and a RMSEP of 0.4 kPa (NASCIMENTO et al., 2018). This suggests that using several variables about ten times smaller (400 variables) can produce comparable results to using full data set (4,700).

For modeling the FP property, Nascimento et al. applied PLS in HTGC and DHA data and applied data fusion strategy

(NASCIMENTO et al., 2018). The authors reported R^2_p values of 0.53, 0.69, 0.73, 0.82 and 0.89, as well as RMSEP values of 8.0 °C, 17.2 °C, 12.4 °C, 11.6 °C and 5.3 °C from DHA_1, $HTGC_2$, $HTGC_1$, DHA_2, and data fusion model, respectively. Here, our best model, OPS-PLS, showed RMSEP of 15.356 °C and R^2_p of 0.78 using only 550 variables. between the two chromatography techniques can improve the ability to predict FP. The advantage of data fusion in this case lies in the fact that low chain compounds can evaporate during sample preparation before HTGC chromatograms register them. Adding DHA data, by data fusion strategies, can improve results, since DHA increases the resolution up to *n*-C_{14} compounds allowing the addition of information on lower chain compounds.

Rodrigues et al. also predicted CR with a RMSEP of 0.83 wt% and a R^2 of 0.768 (RODRIGUES et al., 2018).We obtained a RMSEP value of 0.629 wt% and a R^2 of 0.88 for MCR. This clearly indicates a growth in both linearity and accuracy of the prediction model as the variables decrease in selection with the OPS. The other methods, PLS-full, iPLS, and siPLS showed RMSEP values above 0.7 wt% and R^2 above 0.81.

SAT was also predicted by Rodrigues et al. with a RMSEP of 6.76 wt% and a R^2_p of 0.692 (RODRIGUES et al., 2018). In this study, we obtained a RMSEP of 3.691 wt% and a R^2_p of 0.759, using a small data set of 450 variables in OPS-PLS. Rodrigues et al. reached a RMSEP of 4.05 wt% and a R^2_p of 0.505 for ARO while we obtained a RMSEP and R^2_p values of 2.939 wt% and 0.796, respectively (RODRIGUES et al., 2018). There are no studies reporting the estimate of POL using HTGC in oil, perhaps due to the difficulty of explaining it, since the measure is indirect. Filgueiras et al. determined POL in crude oil using nuclear magnetic resonance (^{13}C NMR) with PLS and SVR (support vector regression) associated with the genetic algorithm (GA) (FILGUEIRAS et al., 2016). The GA-PLS model presented RMSEP equal to 4.0 wt% and R^2_p to 0.778, while GA-SVR model presented RMSEP equal to 3.7 wt% and R^2_p to 0.774. For POL prediction, OPS-PLS provided the best model, using 250 variables, obtaining an RMSEP value of 3.374 wt% and R^2_p of 0.86. Although ^{13}C-NMR can provide important structural information of crude oil compounds, the use of HTGC and variable selection proved to generate comparable results for predicting this property (MERDRIGNAC, I. ESPINAT, 2007; RIAZI, 2007; SPEIGHT, 2015).

Figure 4 and 5 shows graphics with the properties values from the reference method (ASTM) versus values predicted by the OPS-PLS models. As stated earlier, all properties presented high R^2 values, which demonstrates the ability of the proposed method. Most of the samples fitted well to the model and a few of them did not. The outlier detection was carried out through the evaluation of residues and no one outlier was detected. In general, iPLS and siPLS models showed great applicability to estimate most of the proposed properties. These variable selection methods have been constantly described in literature and are great tools for reducing the data set and obtaining models with better predictive ability than using the entire chromatographic or spectral information. However, the selected variables must be truly representative for the property of interest in the case of iPLS model or they must have high synergism with each other in the case of siPLS model. OPS algorithm presented the best models for predicting properties, while the other ones showed comparable or less effective models than OPS-PLS. This may be because the OPS algorithm rearranges the chromatographic matrix according to the individual importance of each variable for the property.

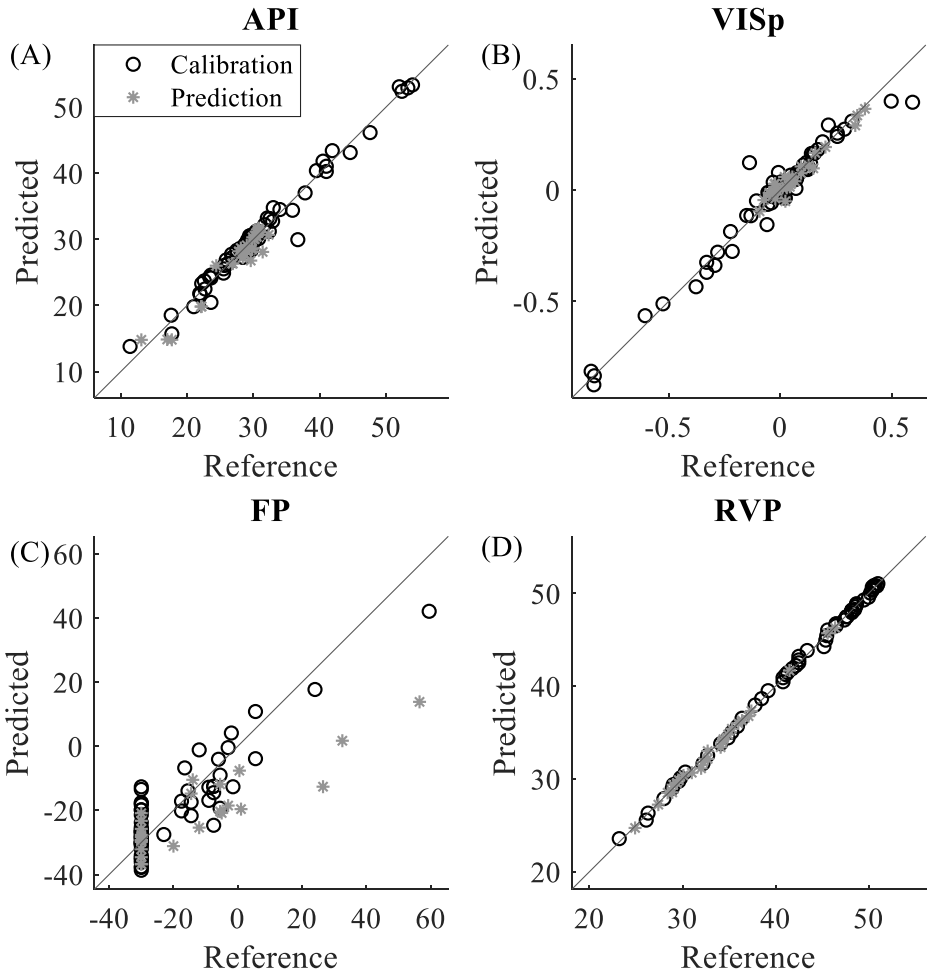


Figure 4. Graph of the OPS-PLS regression models for API (A), VIS_p (B), FP (C) in °C, and RVP (D) in kPa.

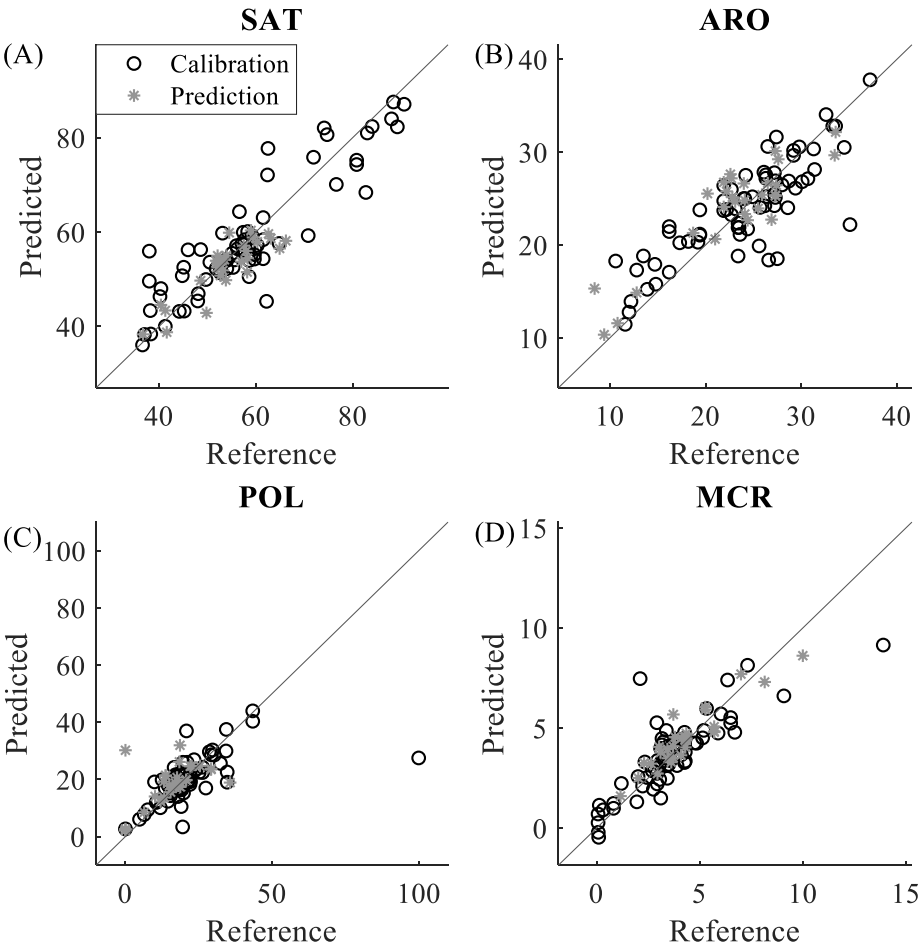


Figure 5. Graph of the OPS-PLS regression models for SAT (A), ARO (B), POL (C), and MCR (D) with values in wt%.

5 CONCLUSIONS

In this study, we employed PLS, iPLS, siPLS, and OPS-PLS models to predict eight physicochemical properties of crude oil. The OPS-PLS models demonstrated superior performance in accurately estimating standardized kinematic viscosity (RMSEP of 0.029), flash point (RMSEP of 15.356 °C), Reid vapor pressure (RMSEP of 0.324 kPa), micro carbon residue (RMSEP of 0.629 wt%), saturates (RMSEP of 3.691 wt%), aromatics (RMSEP of 2.939 wt%), and polar content (RMSEP of 3.374 wt%). For API gravity, the PLS-full model exhibited the lowest RMSEP (1.15), although the OPS-PLS model (RMSEP of 1.24) yielded comparable results while utilizing a smaller number of variables. Additionally, iPLS for RVP (RMSEP of 0.322 kPa) showed similar performance to OPS-PLS for RVP. Notably, for all properties, we were able to identify the selected peaks in the chromatograms, providing insights into the relevant compounds associated with each retention time. This suggests that the OPS algorithm effectively identifies and selects the most significant regions for all properties, thereby improving the predictive capacity of the models.

ACKNOWLODGEMENTS

The authors would like to thank the Laboratório de Pesquisa e Desenvolvimento de Metodologias para Análises de Petróleos (LABPETRO) of UFES and Centro de Pesquisas e Desenvolvimento Leopoldo Américo Miguez de Mello (CENPES) of Petrobras for providing the samples and physicochemical analyses. This work was supported by FAPES (Fundação de Amparo

à Pesquisa e Inovação do Espírito Santo - 442/2021; 691/2022 P: 2022-2DRM4; 1036/2022 P: 2022-VZ8G9; 343/2023 P 2023-6SJG7), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - 001), and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - 409700/2022-3;305459/2020-1).

REFERENCES

ASTM 4530. Standard Test Method for Determination of Carbon Residue (Micro Method). ASTM International: West Conshohocken, PA. **ASTM International.** 2020.

ASTM D323-15A. Standard Test Method for Vapor Pressure of Petroleum Products (Reid Method). ASTM International: West Conshohocken, PA. **ASTM International.** 2020.

ASTM D2549. Standard Method for Separation of Representative Aromatics and Nonaromatics Fractions of High-Boiling Oils by Elution Chromatography. **ASTM International:** West Conshohocken, PA. ASTM International. 2022.

ASTM D7042.21a. Standard Test Method for Dynamic Viscosity and Density of Liquids by Stabinger Viscometer (and the Calculation of Kinematic Viscosity). **ASTM International:** West Conshohocken, PA. ASTM International. 2021.

ASTM D7169-23. Standard Test Method for Boiling Point Distribution of Samples with Residues Such as Crude Oils and Atmospheric and Vacuum Residues by High Temperature Gas Chromatography. **ASTM International:** West Conshohocken, PA. ASTM International. 2023.

AUSTRICH, A. J.; BUENROSTRO-GONZALEZ, E.; LIRA-GALEANA, C. ASTM D-5307 and ASTM D-7169 SIMDIS standards: A comparison and correlation of methods. **Petroleum Science and Technology**, v. 33, n. 6, p. 657–663, 2015.

BALLABIO, D. et al. Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques. **Journal of Chemometrics**, v. 22, n. 8, p. 457–463, 2008.

BLOMBERG, J.; SCHOENMAKERS, P. J.; BRINKMAN, U. A. T. Gas chromatographic methods for oil analysis. **Journal of Chromatography A**, v. 972, n. 2, p. 137–173, 2002.

BRAZILIAN AGENCY OF PETROLEUM, NATURAL GAS, AND BIOFUELS (ANP). Ordinance n.206 of 29/08/2000.

CALIARI, Í. P. et al. Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods. **Carbohydrate Polymers**, v. 158, p. 20–28, 2017.

CHUA, C. C. et al. Enhanced analysis of weathered crude oils by gas chromatography-flame ionization detection, gas chromatography-mass spectrometry diagnostic ratios, and multivariate statistics. v. 1634, 2020.

DASZYKOWSKI, M.; WALCZAK, B. Use and abuse of chemometrics in chromatography. **TrAC - Trends in Analytical Chemistry**, v. 25, n. 11, p. 1081–1096, 2006.

DE ANDRADE FERREIRA, A.; DE AQUINO NETO, F. R. a Destilação Simulada Na Indústria Do Petróleo. **Quimica Nova**, v. 28, n. 3, p. 478–482, 2005.

DE ARAÚJO GOMES, A. et al. Variable selection in the chemometric treatment of food data: A tutorial review. **Food Chemistry**, v. 370, 15 fev. 2022.

DE PAULO, E. H. et al. Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties. **Fuel**, v. 279, 1 nov. 2020.

DE PAULO, E. H. et al. Determination of gross calorific value in crude oil by variable selection methods applied to 13C NMR spectroscopy. **Fuel**, v. 311, 1 mar. 2022.

DE PAULO, E. H. et al. Study of coffee sensory attributes by ordered predictors selection applied to 1H NMR spectroscopy. **Microchemical Journal**, v. 190, n. January 2023.

DIAS, J. C. M.; AGUIAR, P. F. a Statistical Method for Acceptance of Crude Oil Viscosity-Temperature Curves. **Brazilian Journal of Petroleum and Gas**, v. 5, n. 1, p. 019–024, 2011.

DUARTE, L. M. et al. Determination of some physicochemical properties in Brazilian crude oil by 1H NMR spectroscopy associated to chemometric approach. **Fuel**, v. 181, p. 660–669, 2016.

ESPIÑOSA-PEN, M.; FIGUEROA-GOMEZ, Y.; JIME´NEZ-CRUZ, F. Simulated Distillation Yield Curves in Heavy Crude Oils: A Comparison of Precision between ASTM D-5307 and ASTM D-2892 Physical Distillation. **Energy & Fuels**, v. 18, n. 6, p. 1832–1840, 2004.

FARRÉS, M. et al. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. **Journal of Chemometrics**, v. 29, n. 10, p. 528–536, 1 out. 2015.

FERREIRA, G. W. D. et al. Temporal decomposition sampling and chemical characterization of eucalyptus harvest residues using NIR spectroscopy and chemometric methods. **Talanta**, v. 188, n. May, p. 168–177, 2018.

FILGUEIRAS, P. R. et al. Determination of Saturates, Aromatics, and Polars in Crude Oil by 13C NMR and Support Vector Regression with Variable Selection by Genetic Algorithm. **Energy and Fuels**, v. 30, n. 3, p. 1972–1978, 2016.

GUO, Q. et al. Feature selection in principal component analysis of analytical data. **Chemometrics and Intelligent Laboratory Systems**, v. 61, n. 1–2, p. 123–132, 2002.

HUPP, A. M. et al. Chemometric analysis of diesel fuel for forensic and environmental applications. **Analytica Chimica Acta**, v. 606, n. 2, p. 159–171, 2008.

ISO 12185. Crude petroleum and petroleum products – determination of density – oscillating U-tube method. Geneva, SwitzerlandInternational Organization for Standardization, 1996.

ISO 13736. International Standard International Standard. **61010-1 © Iec:2001**, v. 2006, p. 13, 2006.

KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. **Technometric**, v. 11, n. 1, p. 137–148, 1969.

LARTER, S. et al. A practical biodegradation scale for use in reservoir geochemical studies of biodegraded oils. **Organic Geochemistry**, v. 45, p. 66–76, 2012.

LI, W. et al. Analysis of light weight fractions of coal-based crude oil by gas chromatography combined with mass spectroscopy and fl ame ionization detection. **Fuel**, v. 241, n. October 2018, p. 392–401, 2019.

LILAND, K. H.; STEFANSSON, P.; INDAHL, U. G. Much faster cross-validation in PLSR-modelling by avoiding redundant calculations. **Journal of Chemometrics**, v. 34, n. 3, p. 1–11, 2020.

MA, S. et al. Discrimination of Acori Tatarinowii Rhizoma from two habitats based on GC-MS fingerprinting and LASSO-PLS-DA. **Journal of Central South University**, v. 25, n. 5, p. 1063–1075, 2018.

MARTINS, J. P. A; FERREIRA, M. M. C. Qsar Modeling: a New Open-Source Computational Package To Generate and Validate Qsar Models. **Quimica Nova**, v. 36, n. 4, p. 554–560, 2013.

MEHMOOD, T.; SÆBØ, S.; LILAND, K. H. Comparison of variable selection methods in partial least squares regression. **Journal of Chemometrics**, v. 34, n. 6, 1 jun. 2020.

MERDRIGNAC, I. ESPINAT, D. Phisicochemical characterization of petroleum fractions: The state of the art. **Oil & Gas Science and Technology**, v. 62, n. 1, p. 7–32, 2007.

MORALES-MEDINA, G.; GUZMÁN, A. Prediction of density and viscosity of colombian crude oils from chromatographic data. **CTyF - Ciencia, Tecnologia y Futuro**, v. 4, n. 5, p. 57–73, 2012.

NASCIMENTO, M. H. C. et al. Determination of flash point and Reid vapor pressure in petroleum from HTGC and DHA associated with chemometrics. **Fuel**, v. 234, n. July, p. 643–649, 2018.

OLIVIERI, A. C. Analytical figures of merit: From univariate to multiway calibration. **Chemical Reviews**, v. 114, n. 10, p. 5358–5378, 2014.

OLIVIERI, A. C. Practical guidelines for reporting results in single- and multi-component analytical calibration: A tutorial. **Analytica Chimica Acta**, v. 868, p. 10–22, 2015.

PARK, H. E. et al. Gas chromatography/mass spectrometry-based metabolic profiling and differentiation of Ginseng roots according to cultivation age using variable selection. **Journal of AOAC International**, v. 96, n. 6, p. 1266–1272, 2013.

PEREIRA RAINHA, K. et al. Determination of API Gravity and Total and Basic Nitrogen Content by Mid- and Near-Infrared Spectroscopy in Crude Oil with Multivariate Regression and Variable Selection Tools. **Analytical Letters**, v. 52, n. 18, p. 2914–2930, 12 dez. 2019.

POLLO, B. J. et al. Trends in Analytical Chemistry Chemometrics, Comprehensive Two-Dimensional gas chromatography and “omics ” sciences : Basic tools and recent applications. **Trends in Analytical Chemistry**, v. 134, p. 116111, 2021.

RIAZI, M. **Characterization and Properties of Petroleum Fractions**. 1 st ed. West Conshohocken, PA: ASTM International, 2007.

RIBEIRO, J. S. et al. Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics. **Talanta**, v. 101, p. 253–260, 2012.

RIBEIRO, J. S.; FERREIRA, M. M. C.; SALVA, T. J. G. Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy. **Talanta**, v. 83, n. 5, p. 1352–1358, 2011.

RINNAN, Å.; BERG, F. VAN DEN; ENGELSEN, S. B. Review of the most common pre-processing techniques for near-infrared spectra. **TrAC - Trends in Analytical Chemistry**, v. 28, n. 10, p. 1201–1222, 2009.

ROCHA, W. F. DE C.; SHEEN, D. A. Determination of physicochemical properties of petroleum derivatives and biodiesel using GC/MS and chemometric methods with uncertainty estimation. **Fuel**, v. 243, n. July 2018, p. 413–422, 2019.

RODRIGUES, É. V. A. et al. Determination of crude oil physicochemical properties by high-temperature gas chromatography associated with multivariate calibration. **Fuel**, v. 220, n. November 2017, p. 389–395, 2018.

ROQUE, J. V. et al. Comprehensive new approaches for variable selection using ordered predictors selection. **Analytica Chimica Acta**, v. 1075, p. 57–70, 10 out. 2019.

ROQUE, J. V.; DIAS, L. A. S.; TEÓFILO, R. F. Multivariate calibration to determine phorbol esters in seeds of Jatropha curcas L. using near infrared and ultraviolet spectroscopies. **Journal of the Brazilian Chemical Society**, v. 28, n. 8, p. 1506–1516, 1 ago. 2017.

SILVA, S. R. C. et al. Preparation of a Nitrogen Oil Compound Fraction by Modified Gel Silica Column Chromatography. **Energy and Fuels**, v. 34, n. 5, p. 5652–5664, 2020.

SPEIGHT, J. G. **Handbook of petroleum product analysis (2nd edition)**. 2015. v. 53

TELNAES, N. et al. Interpretation of multivariate data: Relationship between phenanthrenes in crude oils. **Chemometrics and Intelligent Laboratory Systems**, v. 2, n. 1–3, p. 149–153, 1987.

TEÓFILO, R. F.; MARTINS, J. P. A.; FERREIRA, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. **Journal of Chemometrics**, v. 23, n. 1, p. 32–48, 2009.

TOMASI, G.; SAVORANI, F.; ENGELSEN, S. B. Icoshift: An effective tool for the alignment of chromatographic data. **Journal of Chromatography A**, v. 1218, n. 43, p. 7832–7840, 2011.

VALE, D. L. et al. Comprehensive and multidimensional tools for crude oil property prediction and petrochemical industry refinery inferences. **Fuel**, v. 223, n. September 2017, p. 188–197, 2018.

VIEIRA, A. P. et al. Determination of physicochemical properties of petroleum using 1H NMR spectroscopy combined with multivariate calibration. **Fuel**, v. 253, p. 320–326, 1 out. 2019.

ZENG, H. et al. Gas Chromatograph Applications in Petroleum Hydrocarbon Fluids. **Advanced Gas Chromatography - Progress in Agricultural, Biomedical and Industrial Applications**, n. May 2014, 2012.

ZHANG, Z. et al. Variable selection in Logistic regression model with genetic algorithm. **Annals of Translational**

Medicine, v. 6, n. 3, p. 45–45, fev. 2018.

SUPPLEMENTARY MATERIAL

Figure S1. Correlation graph of the physicochemical properties.

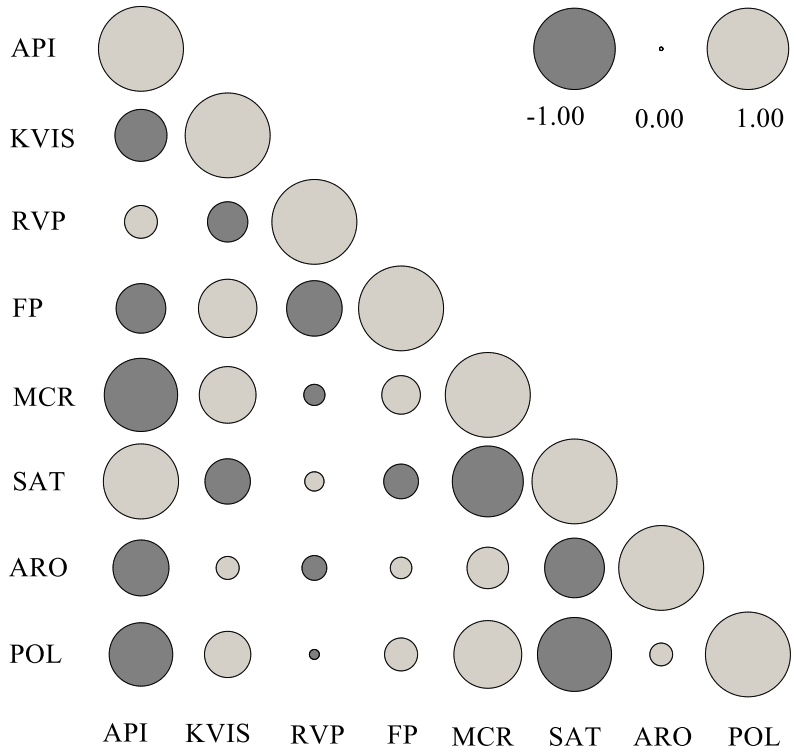


Figure S2. Examples of HTGC chromatograms for a light (a), intermediary (b), and a heavy (c) oil sample.

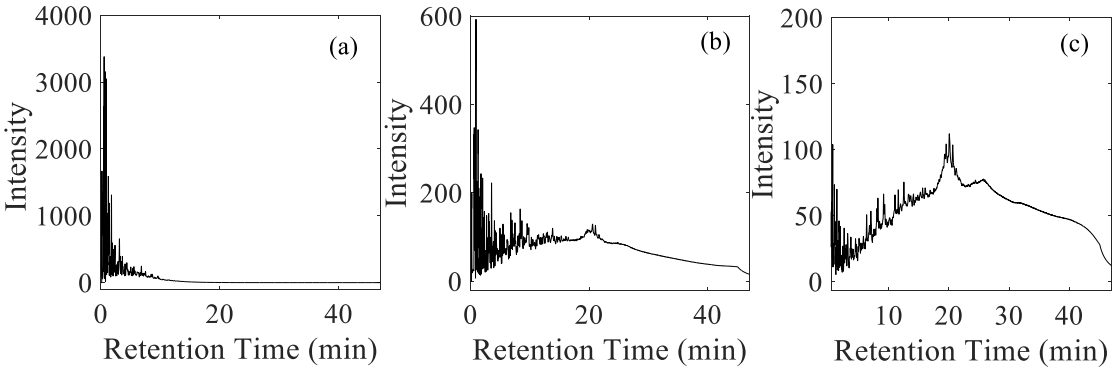


Figure S3. HTGC chromatogram *vs* OPS vector graphics.

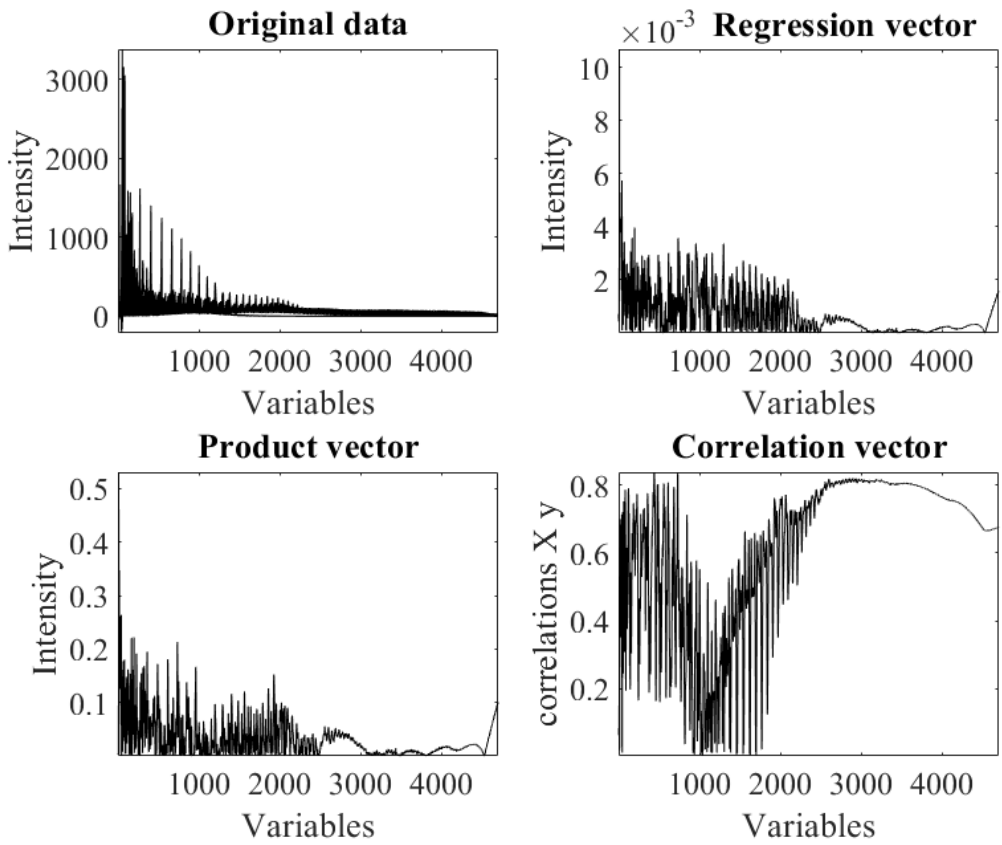


Figure S4. Graphics of reference vs predicted values of the PLS, iPLS, siPLS, OPS-PLS models for API, VIS, RVP, and FP.

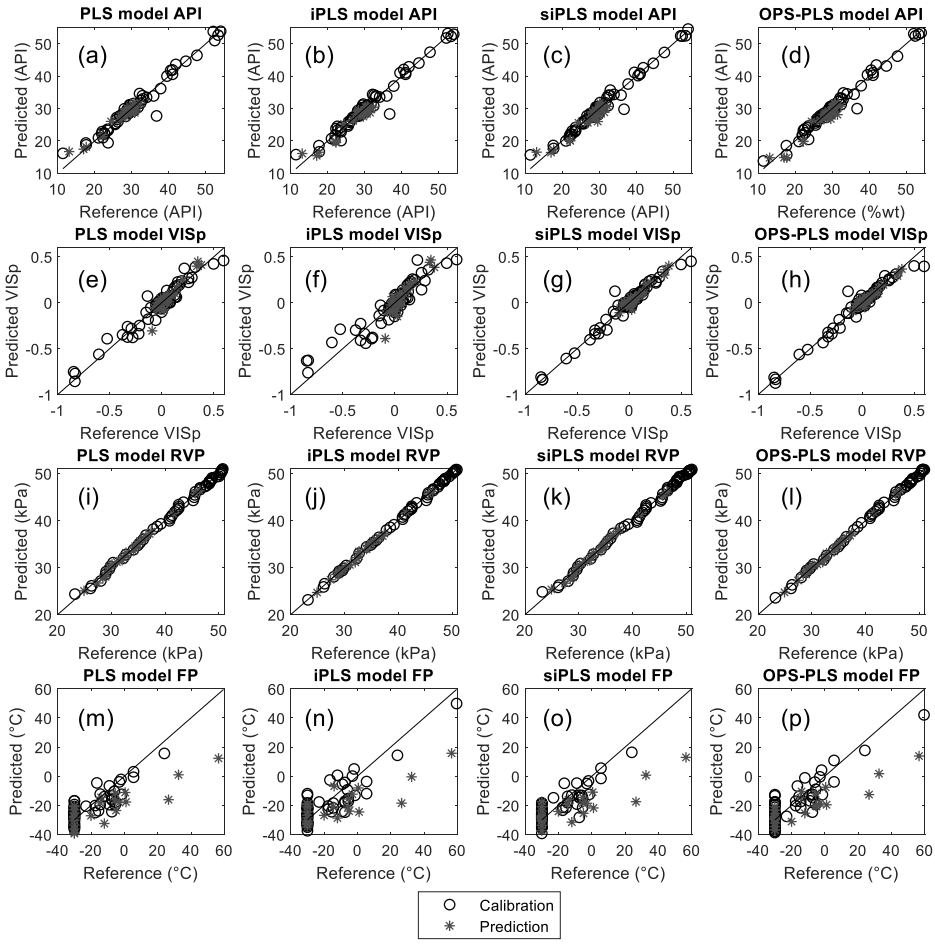


Figure S5. Graphics of reference vs predicted values of the PLS, iPLS, siPLS, OPS-PLS models for SAT, ARO, POL, and MCR.

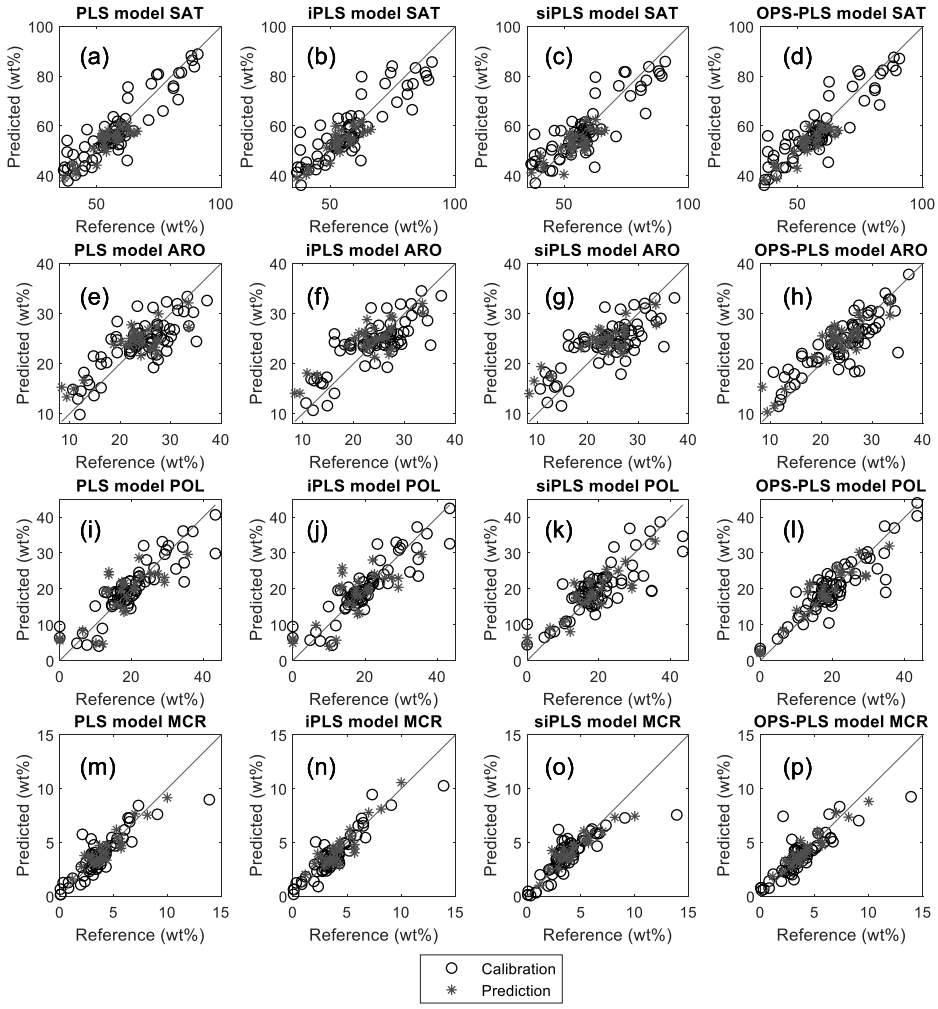


Figure S6. Variable selection plots of the iPLS, siPLS, OPS-PLS models for API, VIS, RVP, and FP.

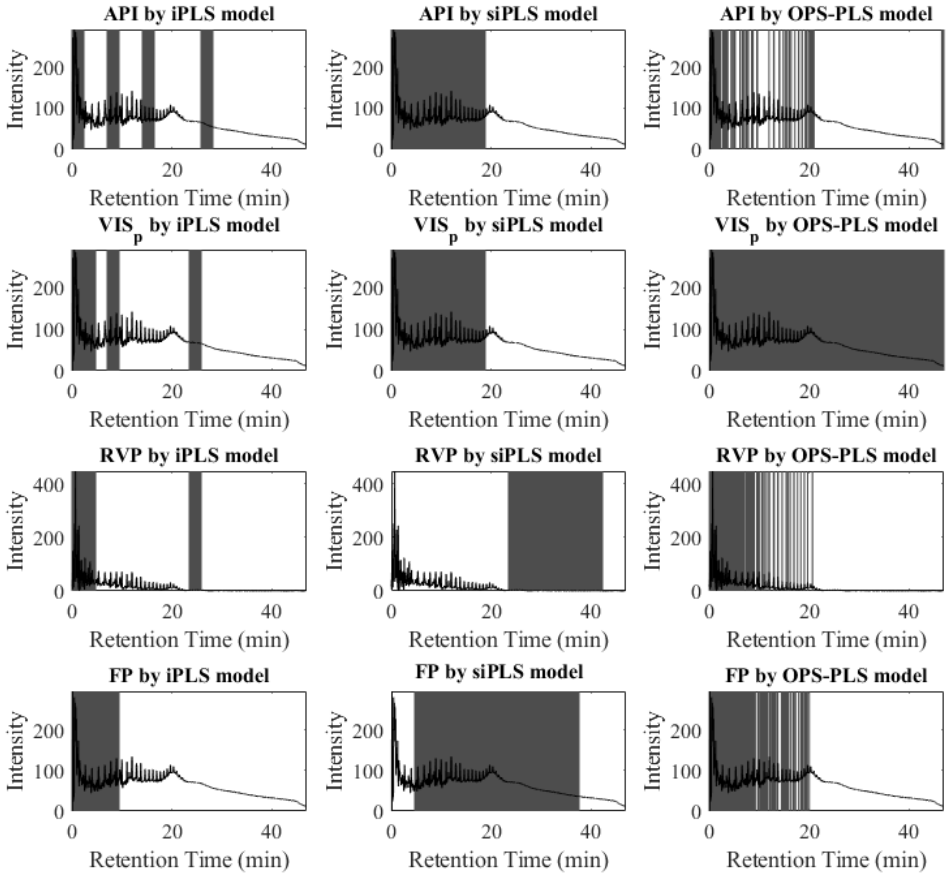


Figure S7. Variable selection plots of the iPLS, siPLS, OPS-PLS models for SAT, ARO, POL, and MCR.

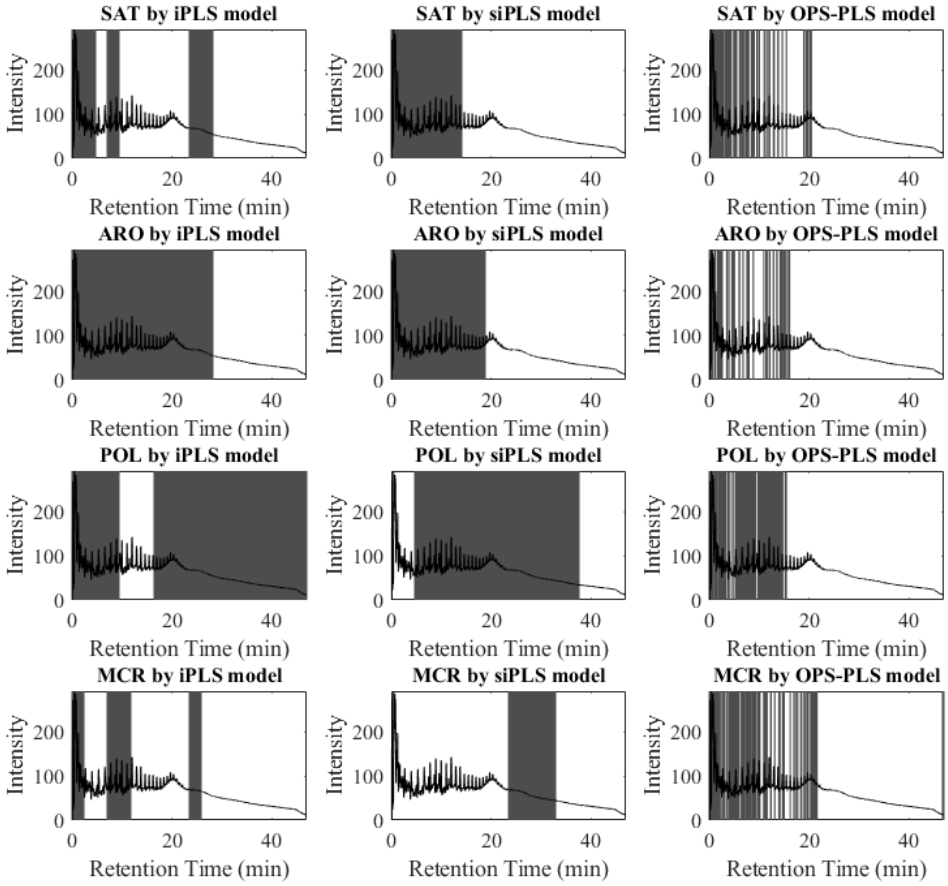


Table S1. Coefficients of correlation between the properties studied.

	API	KVIS	RVP	FP	MCR	SAT	ARO	POL
API	1.00	-0.60	0.36	-0.57	-0.86	0.88	-0.64	-0.74
KVIS	-0.60	1.00	-0.45	0.67	0.65	-0.51	0.24	0.53
RVP	0.36	-0.45	1.00	-0.64	-0.22	0.19	-0.26	-0.08
FP	-0.57	0.67	-0,.64	1.00	0.43	-0.38	0.22	0.36
MCR	-0.86	0.65	-0.22	0.43	1.00	-0.83	0.47	0.79
SAT	0.88	-0.51	0.19	-0.38	-0.83	1.00	-0.69	-0.87
ARO	-0.64	0.24	-0.26	0.22	0.47	-0.69	1.00	0.24
POL	-0.74	0.53	-0.08	0.36	0.79	-0.87	0.24	1.00