Infrared spectroscopy and multivariate methods as a tool for identification and quantification of fuels and lubricant oils in soil

Maurílio Gustavo Nespeca · Gabriel Baroffaldi Piassalonga · José Eduardo de Oliveira

Received: 15 September 2017 / Accepted: 28 December 2017 / Published online: 9 January 2018

Abstract Environmental contamination caused by leakage of fuels and lubricant oils at gas stations is of great concern due to the presence of carcinogenic compounds in the composition of gasoline, diesel, and mineral lubricant oils. Chromatographic methods or non-selective infrared methods are usually used to assess soil contamination, which makes environmental monitoring costly or not appropriate. In this perspective, the present work proposes a methodology to identify the type of contaminant (gasoline, diesel, or lubricant oil) and, subsequently, to quantify the contaminant concentration using attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy and multivariate methods. Firstly, gasoline, diesel, and lubricating oil samples were acquired from gas stations and analyzed by gas chromatography to determine the total petroleum hydrocarbon (TPH) fractions (gasoline range organics, diesel range organics, and oil range organics). Then, solutions of these contaminants in hexane were prepared in the concentration range of about 5–10,000 mg kg^{-1}. The infrared spectra of the solutions were obtained and used for the development of the pattern recognition model and the calibration models. The partial least square discriminant analysis (PLS-DA) model could correctly classify 100% of the samples of each type of contaminant and presented selectivity equal to 1.00, which provides a suitable method for the identification of the source of contamination. The PLS regression models were developed using multivariate filters, such as orthogonal signal correction (OSC) and general least square weighting (GLSW), and selection variable by genetic algorithm (GA). The validation of the models resulted in correlation coefficients above 0.96 and root-mean-square error of prediction values below the maximum permissible contamination limit (1000 mg kg^{-1}). The methodology was validated through the addition of fuels and lubricating oil in soil samples and quantification of the TPH fractions through the developed models after the extraction of the analytes by the EPA 3550 method adapted by the authors. The recovery percentage of the analytes was within the acceptance limits of ASTM D7678 (70–130%), except for one sample (69% of recovery). Therefore, the methodology proposed here provides faster and less costly analyses than the chromatographic methods and it is adequate for the environmental monitoring of soil contamination by gas stations.

Keywords Soil contamination · Fuel leakage · Infrared spectroscopy · Partial least square · Genetic algorithm · Multivariate filters

Introduction

Despite the development of technologies for the use of renewable fuels, today's society is still highly dependent on petroleum-derived fuels (BP 2016). Due to the

M. G. Nespeca (✉) · G. B. Piassalonga · J. E. de Oliveira
Center for Monitoring and Research of the Quality of Fuels, Biofuels, Crude Oil, and Derivatives (Cempeqc), Institute of Chemistry, São Paulo State University (UNESP), Prof. Francisco Degni 55, Araraquara, SP Zip Code 14800-060, Brazil
e-mail: mauriliogn@gmail.com

dangerous nature of these fuels to the environment and health, the Environmental Company of the State of São Paulo (Cetesb), in Brazil, configures the gas stations as risky establishments. Although gasoline and diesel fuel are mainly composed of saturated hydrocarbons, they pose a high risk to human health, since the concentration of aromatic hydrocarbons, substances with carcinogenic properties, can reach up to 35% of the gasoline volume and 11% of the diesel volume (Todd et al. 1999; Y. Wang et al. 2016). According to Cetesb reports, gas stations are the main responsible for environmental contamination in the state of São Paulo, representing approximately 74% of the 5376 contaminated areas registered by the company (COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO 2016a).

In view of the environmental risk, gas station licensing depends on the environmental analysis to verify fuel contamination in soil and groundwater. The verification is performed by the analysis of total petroleum hydrocarbons (TPH), in which the concentration cannot exceed 1000 mg kg^{-1} in the soil and 600 μg L^{-1} in the groundwater according to Brazilian regulatory (COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO 2016b). Given the high complexity of petroleum product composition, TPH are divided into fractions according to the number of carbons in chain. Hydrocarbons containing 6 to 10 atoms of carbon are attributed to the gasoline range organics (GRO), 10 to 28 carbons to the diesel range organics (DRO), and up to 28 carbons to the oil range organics (ORO) (US ENVIRONMENTAL PROTECTION AGENCY 2007a).

Conventionally, TPH analysis is performed by gas chromatography (GC) or mid-infrared (MIR) spectroscopy. For methods based on MIR, TPH is defined as any substance extractable by a solvent, which is not removed by the cleanup step and can be detected by MIR at specified wavelengths (Weisman 1998). Although the conventional MIR methods provide rapid, precise, and adequate screening information, they do not provide information of the contaminant type that is present in the environmental matrix, which may be crucial to identify the source of leakage and perform a rapid correction.

The first established MIR method for TPH analysis was the Environmental Protection Agency (EPA) method 418.1, in which the quantification of hydrocarbons is based on the wavenumber of the carbon-hydrogen bond of the CH_3 group (2950 cm^{-1}) (US ENVIRONMENTAL PROTECTION AGENCY 1978). Later, the EPA removed the method because of the environmental risks of Freon or tetrachloromethane used as extraction solvent; however, the method was used for a long time in some countries (Schwartz et al. 2012). The EPA method 418.1 was adapted to the method 8440, in which the hydrocarbons are extracted by supercritical carbon dioxide in solid samples; however, it is a time-consuming and non-selective method and, moreover, is not applicable for volatile petroleum fractions such as gasoline (US ENVIRONMENTAL PROTECTION AGENCY 1996a). As an alternative, the ASTM (American Society for Testing and Materials) developed a similar method for determination of TPH in water using cyclohexane as solvent (ASTM D7678), but the method is also not specific (AMERICAN SOCIETY FOR TESTING AND MATERIALS 2011). The limitation of these methods lies in the univariate modeling based only on the wavenumber related to the stretching of the CH_3 group (EPA methods) or angular deformation of the CH_2 group (ASTM method).

In recent years, several studies have shown the possibility of measuring soil contaminants through multivariate analysis of fingerprints in visible-near (vis-NIR, 27,778–12,821–4000 cm^{-1}) and mid (MIR, 4000–400 cm^{-1}) infrared ranges (Horta et al. 2015; Okparanma and Mouazen 2013; Vershinin and Petrov 2016; Webster et al. 2016; Workman and Weyer 2007). The diffuse reflectance spectroscopy has been widely used to estimate soil properties, such as total carbon, sand and clay contents, pH, and total nitrogen, due to the lack of sample preparation, rapid analysis, and portability provided by the technique. However, recent studies showed a low accuracy for TPH analysis using MIR diffuse reflectance due to the granulometric differences between soil types, particle aggregates, and heterogeneous distribution of TPH in the sample (S. Forrester et al. 2010; S. T. Forrester et al. 2013; Webster et al. 2016).

One way to improve the accuracy of the TPH quantification is to bypass the soil interferences through extraction of the analytes with a solvent, as is done in EPA method 418.1 and ASTM D7678. Unlike these methods based on univariate calibration, multivariate analysis of MIR spectra allows the use of several solvents for extraction since the contaminant quantification is based on fingerprint analysis. Although sample preparation makes the method laborious and time-consuming, the use of attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy provides a fast analysis (less than 1 min per sample) with a small volume of sample (1 mL or less) (Pejcic et al. 2013). Thus, the improvement in method accuracy will compensate the sample preparation step if the extraction method is adequate. Vershinin and Petrov, 2016, demonstrated that ATR-FTIR allows a fast and accurate analysis for the determination of hydrocarbon mixture in residual water; however, the authors used tetrachloromethane as the extraction solvent.

Pattern recognition methods associated with FTIR spectroscopy have been extensively used in several areas, including for fuel classification (Ballabio and Consonni 2013; Da Silva et al. 2014; Worley et al. 2013); however, the identification of fuel or lubricating oil type present in the environmental media has not yet been explored by these methods. Although gasoline, diesel, and lubricating oil are mainly hydrocarbons, these products have different spectra in the infrared region due to different carbon chains and additive in composition. The spectral characteristics of each product allow the identification of the origin of the contamination in the environmental media using methods such as partial least square discriminant analysis (PLS-DA).

If the contaminant is identified by a classification method, the quantification can be performed by a multivariate calibration method such as partial least square (PLS) developed specifically for the type of contaminant. Although PLS models are very useful to resolve various calibration problems, the predictive ability of the model is highly influenced by the background interferences and spectrum noise (Bosch-Reig et al. 2017; Zhang et al. 2009). To reduce the instrumental and matrix influences, calibration data are often preprocessed prior to data analysis. The baseline shifting is usually corrected by applying the first or second derivative, or by polynomials that correct the displacement based on a standard spectrum (Burns and Ciurczak 2009; Gemperline 2006). When the analytes are present in a similar solvent, e.g., gasoline solubilized in hexane, the selectivity of the method is compromised by the overlap of the bands inherent to the analyte by the solvent bands. Nonetheless, the selectivity can be improved using multivariate filters, such as orthogonal signal correction (OSC), generalized least square weighting (GLSW), and external parameter orthogonalization (EPO), to remove signals from background and interferences through identification of some unwanted covariance structure (Eigenvector Research 2013; Laghi et al. 2011; Roudier et al. 2017; Zhang et al. 2009). These multivariate filters use samples with similar *Y*-block (concentration of analyte) values to identify the sources of variance in the *X*-block (spectra) to downweight and generate a multivariate regression with more captured variance in the *Y*-block and, usually, using less latent variables (Eigenvector Research 2013; Wold et al. 1998). Another way to improve the selectivity and, consequently, the predictive ability of the chemometric model is by the selection of variables (Vohland et al. 2014). The removal of variables, in which the noise dominates over the information related to the analyte, often leads to better accuracy and performance of the analytical method and it is a technique widely accepted (Mehmood et al. 2012; Xiaobo et al. 2010). The selection of variables can be performed based on the spectral knowledge (manual approach) or through algorithms that seek to minimize the prediction error of the model such as genetic algorithm (GA). This algorithm is a popular heuristic optimization technique that employs a probabilistic, non-local search process that manipulates binary strings (chromosomes) with the coded experimental variables (genes) (Xiaobo et al. 2010). Mixtures with almost identical spectra have been successfully calibrated using GA in addition to the better understanding of the chemical system provided by the algorithm (Vohland et al. 2014; L. Wang et al. 2015; Xiaobo et al. 2010).

Considering the large number of contaminations in the environment caused by fuel leaks in gas stations, this study aimed to develop an analytical method to identify the type of contaminant present in environmental media and quantify the TPH fractions (GRO, DRO, and ORO) using ATR-FTIR spectroscopy associated with PLS-DA for identification and GA-PLS for quantification. The extraction of the contaminants and the preparation of the calibration samples were carried out using non-halogenated solvent and seeking to make the analytical method simple and practical for a rapid assessment of soil contaminated by petroleum-based fuels and lubricants.

Materials and methods

Fuel and lubricant oil samples

The fuel samples were collected at gas stations located in the state of São Paulo, Brazil. The samples collected were: one common-type gasoline (without additives), one additive-type gasoline, one S10 diesel fuel, and one S500 diesel fuel. Besides the fuel samples, two samples of lubricant oil for light vehicles were purchased: one mineral oil with SAE 50 viscosity grade and one semi-synthetic oil with SAE 15W50 viscosity grade.

The collected samples were analyzed by gas chromatography to quantify the GRO and DRO fractions and subtract the biofuel content present in the samples. In Brazil, the addition of biofuels in gasoline and diesel is mandatory and the concentration of biofuels varies according to Brazilian legislation. Currently, the anhydrous ethanol must be present in gasoline at 27% (*v*/*v*) and biodiesel in diesel at 8% (*v*/*v*). The analysis was performed by a gas chromatograph with flame ionization detector (GC-FID), model Trace GC Ultra (Thermo Fisher Scientific), equipped with split/splitless injector and Triplus autosampler. The separation of GRO compounds was performed according to ASTM D5769 (AMERICAN SOCIETY FOR TESTING AND MATERIALS 2015) using a Thermo Scientific TR-1 capillary column of dimensions 60 m × 0.25 mm i.d. and 1.0 μm of stationary phase thickness (dimethylpolysiloxane). The DRO compounds were separated by an Agilent HP-1 capillary column (100 m × 0.25 mm × 0.5 μm) using a chromatographic method adapted from ASTM D6209 (AMERICAN SOCIETY FOR TESTING AND MATERIALS 2013). The quantification of the GRO and DRO fractions was performed by external calibration according to the EPA 8015C method and the retention times were determined using the certified reference materials GRO Mix (Supelco) and TPH Mix (RTC).

Solution preparation

Hexane was chosen as the extraction solvent because of its high affinity with nonpolar compounds and lower toxicity relative to the halogenated solvents used in EPA 3510, 3540, and 3550 methods (US ENVIRONMENTAL PROTECTION AGENCY 1996b, 1996c, 2007b, 2010). The common-type gasoline, the diesel S10, and the mineral lubricant oil were used to prepare 108 solutions (36 solutions for each type of contaminant in hexane P.A. 99%) in the concentration range of 5–10,000 mg kg^{-1}, approximately. The fraction concentration of each solution was subsequently corrected by the values determined in the chromatographic analysis.

Spectrum acquisition

The infrared spectra were acquired by a Nicolet 6700 FTIR (Thermo Scientific, Waltham, USA) spectrometer using 32 scans, 4 cm^{-1} resolution, and spectral range of 4000–650 cm^{-1}. A Smart ARK ATR accessory of ZnSe crystal with 45° of incidence was used for sampling. The conditions of temperature and relative humidity during the analysis were, respectively, 22.7 °C ± 0.1 °C and 49% ± 2%.

Chemometric analysis

The chemometric analysis was performed using MATLAB software (version R2013a) and PLS toolbox (version 7.9.3). The first step was to separate 66% of the solutions for model development (calibration set) and 34% for model validation (test set). The identification of contaminant type was performed by the PLS-DA method and the quantification of the GRO, DRO, and ORO fractions by the PLS regression method. The number of latent variables (LV) was chosen based on root mean square errors of calibration (RMSEC), cross-validation (RMSECV), and root-mean-square error of prediction (RMSEP) values to avoid model overfitting (Hawkins 2004).

To provide the best predictive ability to the model, different preprocessing data were evaluated. The tested preprocessing data were: mean center, first and second derivatives, smoothing, standard normal variate (SNV), multiplicative scatter correction (MSC), orthogonal signal correction (OSC), and generalized least squares weighting (GLSW). After verification of the most appropriate preprocessing for each model, the genetic algorithm (GA) was applied for variable selection. Since the GA from the PLS toolbox is limited to 200 generations, the algorithm was executed twice for each model. Both executions were performed with population size of 128 models, initial terms of 30%, mutation rate of 0.5%, double crossover, and PLS regression method. The first execution was performed using window width with four variables and the second with one variable. The optimization in the model provided by GA was statistically evaluated through an *F*-test (Eq. 1):

$$F = \frac{RMSEP_1^2}{\mathrm{RMSEP}_2^2} \tag{1}$$

where RMSEP$_1$ > RMSEP$_2$ (Rocha et al. 2012). If the calculated F value was greater than the F tabulated, the use of GA significantly optimized the prediction ability of the model.

The efficiency of the calibration models was evaluated by the values of RMSEC and RMSEP, determination coefficients (R^2), bias of the test set, and correlation coefficients (r). The equations of these figures of merit are widely described in the literature and can be found in detail in ASTM E1655 (AMERICAN SOCIETY FOR TESTING AND MATERIALS 2012).

Method validation

The analytical method developed for the quantification of TPH fractions was validated by analysis of soil samples spiked with gasoline, diesel, and lubricant oil. Three validation samples for each contaminant type were prepared using the fuel and lubricant oil samples that were not used in the calibration and test sets. The soil sample was collected at the northern state of Paraná, Brazil (23°18′60″S 51°39′55″W). It was classified as red latosol and presented 18.4% of clay, 3.9% of silt, and 77.7% of coarse sand in the granulometric analysis.

The extraction of the contaminants was performed according to the EPA 3550 (US ENVIRONMENTAL PROTECTION AGENCY 2007b) adapted by the authors. The adaptations include 4 g of samples, 22 mL headspace vials, 30 s of stirring before sonication, a polypropylene syringe filled with $NaSO_4$ and silica gel for cleanup, and filtration by a 5-μm porosity filter connected to the syringe. After the extraction, the solution was taken to the spectrometer for infrared spectrum acquisition and the TPH fractions were quantified using the developed models.

Results and discussion

Characterization of the fuel samples

The analytical method was developed using commercial fuel samples rather than analytical standards to better represent the reality. Therefore, it was essential to characterize the fuel samples to avoid false positives caused by the presence of biofuels in the composition. The common-type and additive-type gasoline samples presented GRO content, respectively, 61.92 ± 1.81% (*w*/*w*) and 78.60 ± 6.45% (*w*/*w*); and the S10 and S500 diesel samples presented DRO content, respectively, 75.56 ± 1.99% (*w*/*w*) and 73.19 ± 3.36% (*w*/*w*). The determined values of GRO and DRO contents were used to correct the concentrations of the samples prepared in hexane.

Spectral features

Hexane is an extractive solvent that exhibits high interaction with the hydrocarbons of gasoline, diesel, and lubricating oils, which provides a high partition coefficient; however, the use of this solvent in the conventional analyses of TPH presents some obstacles. In univariate MIR methods, for example, the major obstacle is the high spectral similarity between hexane and the extracted analytes. The spectra of the hexane P.A. and contaminant solutions at high concentration (about 10,000 mg kg^{-1}) are shown in Fig. 1.

The choice of a specific wavelength for quantification of analytes becomes difficult due to the great spectral similarity; however, the standard deviation of each spectral variable reveals regions where there is a greater difference between the solvent and the classes of contaminant (Fig. 2). The main differences between the classes were related to the vibrations of the CH group, more specifically to the axial deformation of CH_2 at 2889 cm^{-1} and CH_3 at 2917 cm^{-1}, and to the angular deformation of CH_2 in 1465 cm^{-1} and CH_3 in 1375 cm^{-1} (Silverstein et al. 2005). This fact presents two reasons: first, these bands are the most intense; therefore, the variations of absorbance are more sensitive and generated greater standard deviation; secondly, fuels and lubricant oils are complex mixtures of hydrocarbons and, consequently, proportion of CH, CH_2, and CH_3 groups may vary considerably according to the contaminant composition. Low-intensity bands in Fig. 1 could be observed in the standard deviation plot in Fig. 2. The deviations close to 3500 cm^{-1} and 1700 cm^{-1} were related, respectively, to the hydroxyl from the anhydrous ethanol present in the gasoline and to the carbonyl from the biodiesel present in the diesel oil. In addition, the deviation in the fingerprint region (1000–650 cm^{-1}) was further indicative of the spectral difference between the solutions and the pure solvent. Although the abovementioned absorbance variations do not permit univariate modeling, the identification and quantification of the interest groups can be performed by multivariate techniques.

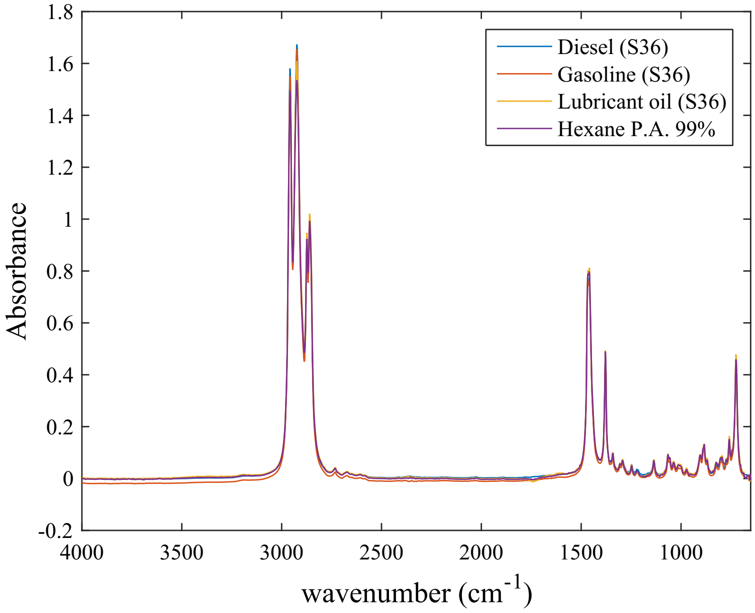

Fig. 1 Spectra of hexane P.A. and contaminant solutions at 10,000 mg kg^{-1}

Identification of the type of contaminant

The classification of the type of contaminant is valuable information in the environmental monitoring at gas stations since it makes possible the identification of the contamination source. PLS-DA is a pattern recognition method based on the PLS regression method; however, coded classes are used as the vector **y** to develop the prediction model instead of a property of interest (Brereton and Lloyd 2014). This multivariate method maximizes the spectral (*X*-block) covariance with the classes through the decomposition of *x* variables into latent variables (LVs) (Rajalahti et al. 2009).

Firstly, the PLS-DA model was developed with the full spectral range and mean centered data. The resulting PLS-DA model required many latent variables (10 LVs), which can be justified by the great spectral similarity between the classes and several variables with

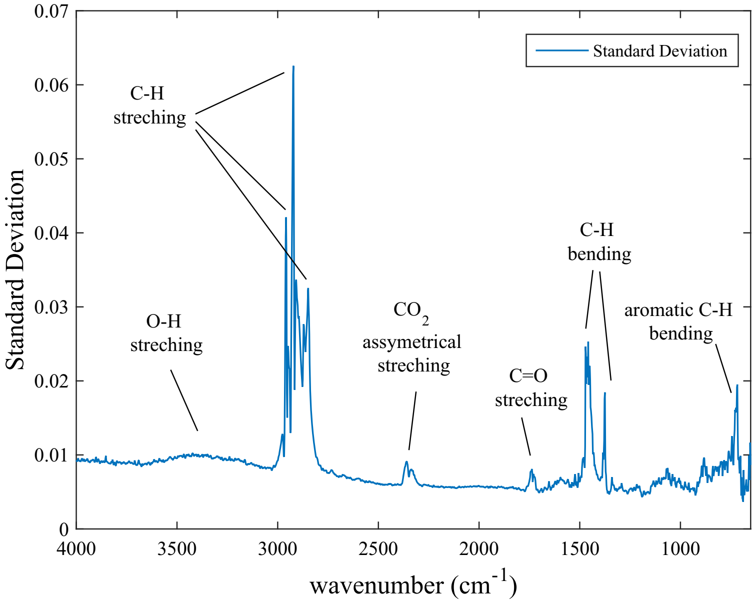

Fig. 2 Standard deviation calculated from hexane and contaminant solution spectra. The main spectral differences between the classes

large variation and small correlation with the **y** vector, so the covariance with the response is usually spread across several PLS components (Rajalahti et al. 2009). Therefore, the number of variables uncorrelated to the **y** vector was reduced by the exclusion of the range 2500–2000 cm^{-1} since this region presents variations from the absorption of carbon dioxide present in the atmosphere (Fig. 2). Afterwards, the GLSW preprocessing was applied with the objective to down-weight the variables that were poorly correlated with the response. The use of GLSW and the exclusion of the 2500–2000 cm^{-1} band provided a PLS-DA model with lower RMSE values and higher correlation coefficients using less latent variables (6 LVs). The analytical parameters of the final PLS-DA model are shown in Table 1.

The application of the model can be evaluated through the values of sensitivity (true positive rate) and selectivity (true negative rate). Sensitivity and selectivity equal to 100% indicate that the model correctly classified all samples without any false positive. Therefore, the PLS-DA model correctly classified 100% of the calibration and validation samples and presented correlation coefficients above 0.95. The prediction of the blank samples showed lower RMSE values and higher correlation coefficients, so the discrimination between uncontaminated and contaminated samples presented the best fit and, therefore, higher analytical reliability. Figure 3 is the graphical representation of class separation across the discrimination threshold. The samples above the threshold belong to the class predicted by the model. Thus, the better discrimination of the blank samples and the absence of false positives and false negatives in the classification of each class are notable.

Table 1 Parameters of the PLS-DA model

Parameter	Blank	Gasoline	Diesel	Lubricant oil
RMSEC	0.04	0.09	0.10	0.08
RMSEP	0.06	0.16	0.14	0.14
r (cal)	0.98	0.96	0.95	0.97
r (pred)	0.97	0.95	0.96	0.96
Sensitivity (cal)	100%	100%	100%	100%
Specificity (cal)	100%	100%	100%	100%
Sensitivity (pred)	100%	100%	100%	100%
Specificity (pred)	100%	100%	100%	100%

Quantification of GRO, DRO, and ORO

Unlike the conventional infrared methods for the quantification of TPH (EPA 418.1, EPA 8440, and ASTM D7678), the methodology proposed in this work makes possible the quantification of the total petroleum hydrocarbon fractions—GRO, DRO, and ORO—after the identification of the type of contaminant by the PLS-DA model. The PLS regression method was chosen for the development of calibration models since it is widely used for the analysis of complex mixtures, especially related to spectroscopic techniques (Nespeca et al. 2017; Yin et al. 2016).

Spectral data are usually treated with preprocessing such as first and second derivatives, SNV, and MSC to correct baseline displacement, especially SNV and MSC to correct deviations caused by different particle sizes and diffuse radiation scattering, and smoothing to reduce instrumental noise (Gemperline 2006). Although the mathematical transformations mentioned have generated PLS models with better predictive abilities, the multivariate filters (OSC and GLSW) provided the lowest values of RMSE and higher correlation coefficients. The main statistical parameters of the PLS regression models with full spectra and variables selected by GA are presented in Table 2. Except for the ORO prediction model, the models with the full spectra had good linearity ($R^2 > 0.91$) and correlation coefficients above 0.95. The variable selection through GA provided models with better prediction results for all TPH fractions, that is, models with lower RMSEC and RMSEP values, better correlations between the reference values and the values predicted by the model ($r > 0.96$), higher linearity ($R^2 > 0.92$), and less bias. In Fig. 4, the plots of reference versus predicted values were indications of the good linearity of the calibration (black circles) and validation (red triangles) sets. Although the prediction model for ORO showed a positive bias for the validation set, the high positive residues were related to samples with measured concentration of ORO above 1000 mg kg^{-1}; therefore, the model presented no risk of false positives for samples below the limit of contamination established by the legislation.

Since the selection of variables by GA can make the computational processing time-consuming, approximately 4 h for each model with the settings used in this work, an *F*-test was performed to verify if the reduction of RMSEP values was statistically significant. The *F* value was calculated by Eq. 1 and compared with the *F*

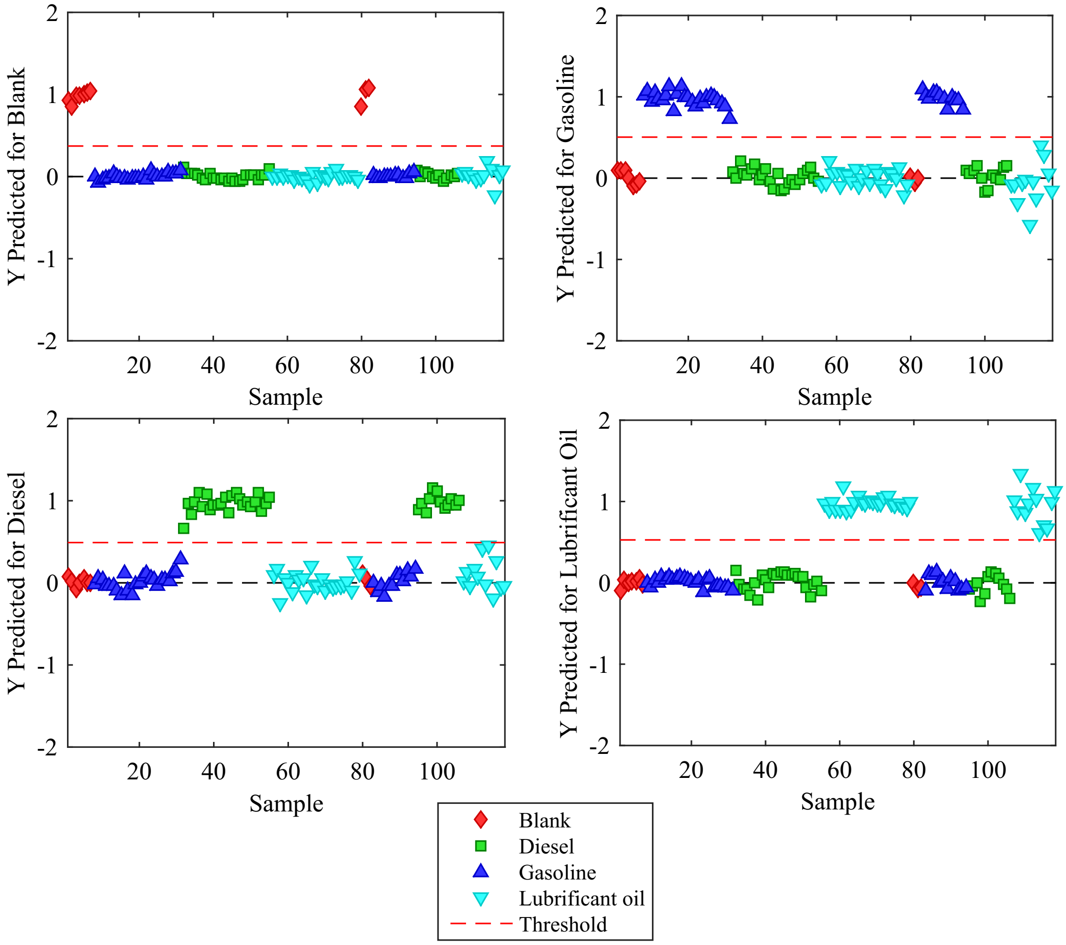

Fig. 3 Discrimination plot for each class. The samples classified above the threshold (red dashed line) were correctly assigned to the respective class

Table 2 Main statistical parameters of the developed PLS regression models

	GRO		DRO		ORO	
	Full spectra	GA	Full spectra	GA	Full spectra	GA
#Variables	1738	62	1738	63	1738	66
Preprocessing	OSC	OSC	OSC	OSC	GLS + MC + 2^{a}der + smooth	GLS
#LVs	2	6	5	12	4	10
RMSEC	63	85	212	21	632	262
RMSEP	451	136	837	217	1915	938
Bias (pred)	− 37	40	143	− 10	701	462
R^2 (cal)	0,9988	0,9978	0,9942	0,9999	0,9562	0,9925
R^2 (pred)	0,9383	0,9957	0,9123	0,9941	0,6340	0,9285
r (cal)	0,9994	0,9989	0,9971	1,0000	0,9778	0,9962
r (pred)	0,9687	0,9979	0,9551	0,9971	0,7963	0,9636

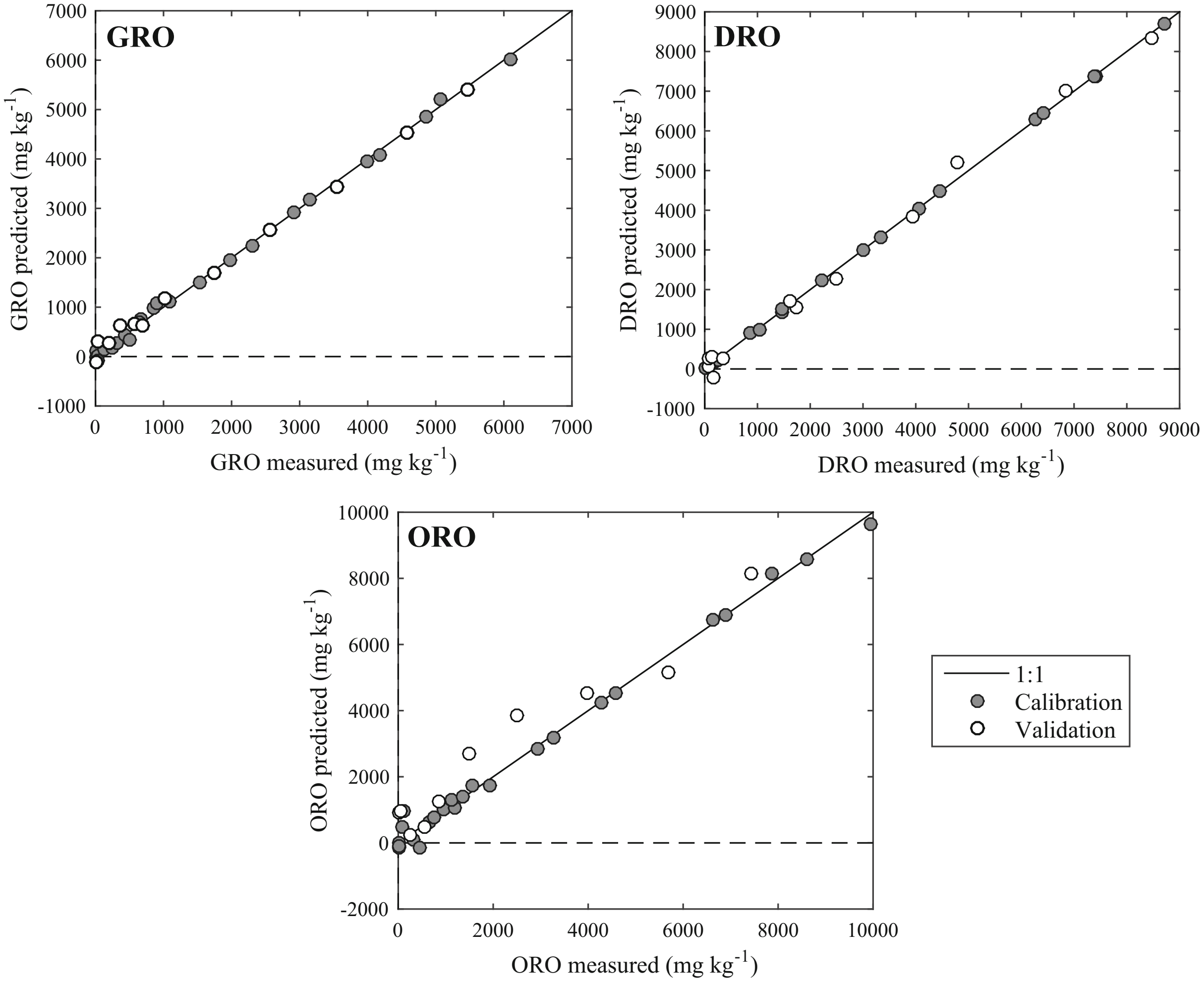

Fig. 4 Plots of reference versus predicted values from the GA-PLS models

tabulated for 11° freedom and 95% confidence level. According to the values presented in Table 3, the genetic algorithm provided a significant improvement in the prediction ability of the TPH fraction models.

Results of the method validation

The extraction process described in item 2.5 was carried out using the soil samples spiked with the contaminants to verify the application of the developed method. The acceptance limits of 70–130% for the percent recovery were based on ASTM D7678. Except for the sample "soil/diesel-2", all samples had recovery value within the accepted limits (Table 4). Therefore, hexane was a suitable solvent for the extraction of the TPH fractions in soil by the adapted EPA method 3550. In addition, the use of sealed vials in the extraction attenuated the loss of volatile compounds and provided recovery values above 69%.

Table 3 *F*-test results to evaluate the GA variable selection

	GRO	DRO	ORO
F calculated	10.94	14.95	4.16
F tabulated	2.82		

Although the method proposed in this work requires a sample preparation step, the extraction by vortex followed by sonication was simple and fast when compared to other extraction techniques such as Sohxlet, which requires a large volume of solvent and time of extraction. Since several samples could be sonicated simultaneously, the total time of extraction and clean up was about 5 min per sample.

Table 4 Results from method validation using soil samples spiked with contaminants

Sample	Concentration (mg kg^{-1})		
	Theoretical	Determined	Recovery (%)
Soil/gasoline-1	2009	2598	129
Soil/gasoline-2	2784	2234	80
Soil/gasoline-3	3599	2668	74
Soil/diesel-1	1173	1419	121
Soil/diesel-2	3608	2472	69
Soil/diesel-3	5193	6345	122
Soil/lubricant-1	2516	2026	81
Soil/lubricant-2	6179	6490	105
Soil/lubricant-3	5985	5339	89

Extraction of TPH fractions in groundwater can also be performed using hexane; however, the specified maximum limit of 600 μg L^{-1} is too low to quantify TPH without preconcentrating the sample. Since hexane has a boiling point close to the gasoline and diesel compounds, it is not possible to concentrate the sample without loss of volatile compounds. Therefore, the method would be limited to contaminations in groundwater at high concentrations.

Conclusion

In this work, we present a methodology for identification and quantification of contaminants, such as gasoline, diesel, and lubricant oil, in soil samples by extraction with hexane and subsequent ATR/FTIR analysis associated with multivariate methods. The PLS-DA model was sensitive and selective with no false positive or negative. The regression models to quantify the GRO, DRO, and ORO fractions presented high correlation coefficients ($r > 0.96$) and sufficient accuracy (RMSE values) to quantify values below the maximum limit of contamination (1000 mg kg^{-1}). The use of multivariate filters (OSC and GLSW) provided better fitness to the PLS models and the selection of variables by the genetic algorithm significantly reduced the values of the prediction errors, which was essential for the prediction ORO fraction. Extraction of contaminants from the spiked samples using the adapted EPA 3550 method and quantification of the fractions by the GA-PLS models resulted in recovery values between 69 and 129%; therefore, the determined concentrations were within the limits established by ASTM D7678, except for one soil sample spiked with diesel. Therefore, the methodology proposed here is adequate for the monitoring of soil contamination caused by gas stations and, in addition, provides faster and less costly analyses than the chromatographic methods and more selective quantification than mid-infrared methods currently used in environmental monitoring.

Acknowledgements The authors would like to thank Cempeqc for providing samples and equipment for analyses.

Funding information The authors would like to thank Capes for providing academic scholarship.

References

AMERICAN SOCIETY FOR TESTING AND MATERIALS. (2011). ASTM D7678—standard test method for total petroleum hydrocarbons (TPH) in water and wastewater with solvent extraction using mid-IR laser. *Transportation*, (C), 1–9. https://doi.org/10.1520/D7678

AMERICAN SOCIETY FOR TESTING AND MATERIALS. (2012). ASTM E1655—standard practices for infrared multivariate quantitative analysis. *ASTM International*, *5*(Reapproved 2012), 29. https://doi.org/10.1520/E1655-05R12.2

AMERICAN SOCIETY FOR TESTING AND MATERIALS. (2013). ASTMD 6209—standard test method for determination of gaseous and particulate polycyclic aromatic hydrocarbons in ambient air (collection on sorbent-backed filters with gas chromatographic/mass spectrometric analysis), 1–14. https://doi.org/10.1520/D6209-13.2.

AMERICAN SOCIETY FOR TESTING AND MATERIALS. (2015). ASTM D5769. Standard test method for determination of benzene, toluene, and total aromatics in finished gasolines by gas chromatography/mass spectrometry., *5*(October), 1–7. https://doi.org/10.1128/AEM.68.6.2660

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods, 5*(16), 3790. https://doi.org/10.1039/c3ay40582f.

Bosch-Reig, F., Gimeno-Adelantado, J. V., Bosch-Mossi, F., & Domnech-Carb, A. (2017). Quantification of minerals from ATR-FTIR spectra with spectral interferences using the MRC method. *Spectrochimica Acta-Part A: Molecular and Biomolecular Spectroscopy, 181*, 7–12. https://doi.org/10.1016/j.saa.2017.02.012.

BP. (2016). *Statistical review of world energy. BP Statistical Review of World Energy*, 65ª ed (p. 48). London. https://www.bp.com/content/dam/bp/pdf/energy-economics/statistical-review-2016/bp-statistical-review-of-world-energy-2016-full-report.pdf.

Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics, 28*(4), 213–225. https://doi.org/10.1002/cem.2609.

Burns, D. A., & Ciurczak, E. W. (2009). Handbook of near-infrared analysis, 3rd ed. *Analytical and Bioanalytical Chemistry., 124*(19), 5603–5604. https://doi.org/10.1021/ja015320c.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO. (2016a). *Texto explicativo: relação de áreas contaminadas e reabilitadas no Estado de São Paulo*. Cetesb. http://cetesb.sp.gov.br/areas-contaminadas/wp-content/uploads/sites/17/2013/11/Texto-explicativo-2016.pdf.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO. (2016b). SISTEMA DE LICENCIAMENTO DE POSTOS IV - Procedimento para Identificação de Passivos Ambientais em Estabelecimentos com Sistema de Armazenamento Subterrâneo de Combustíveis (SASC).

Da Silva, M. P. F., Brito, L. R. E., Honorato, F. A., Paim, A. P. S., Pasquini, C., & Pimentel, M. F. (2014). Classification of gasoline as with or without dispersant and detergent additives using infrared spectroscopy and multivariate classification. *Fuel, 116*, 151–157. https://doi.org/10.1016/j.fuel.2013.07.110.

Eigenvector Research. (2013). Advanced preprocessing: multivariate filtering. http://wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Multivariate_Filtering

Forrester, S., Janik, L., & Mclaughlin, M. (2010). An infrared spectroscopic test for total petroleum hydrocarbon (TPH) contamination in soils. *Proc. 19th World Congress of Soil Science, 1–6 August 2010, Brisbane, Australia*, (August), 13–16.

Forrester, S. T., Janik, L. J., McLaughlin, M. J., Soriano-Disla, J. M., Stewart, R., & Dearman, B. (2013). Total petroleum hydrocarbon concentration prediction in soils using diffuse reflectance infrared spectroscopy. *Soil Science Society of America Journal, 77*(2), 450. https://doi.org/10.2136/sssaj2012.0201.

Gemperline, P. (2006). Practical guide to chemometrics, Second Edition. Book (Chemo). https://doi.org/10.1201/9781420018301.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences, 44*(1), 1–12. https://doi.org/10.1021/ci0342472.

Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T. F. A., McBratney, A. B., Pallasser, R., & Pozza, L. (2015). Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: a prospective review. *Geoderma, 241–242*, 180–209. https://doi.org/10.1016/j.geoderma.2014.11.024.

Laghi, L., Versari, A., Parpinello, G. P., Nakaji, D. Y., & Boulton, R. B. (2011). FTIR spectroscopy and direct orthogonal signal correction preprocessing applied to selected phenolic compounds in red wines. *Food Analytical Methods, 4*(4), 619–625. https://doi.org/10.1007/s12161-011-9240-2.

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems, 118*, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010.

Nespeca, M. G., Rodrigues, C. V., Santana, K. O., Maintinguer, S. I., & de Oliveira, J. E. (2017). Determination of alcohols and volatile organic acids in anaerobic bioreactors for H 2 production by near infrared spectroscopy. *International Journal of Hydrogen Energy, 42*(32), 20480–20493. https://doi.org/10.1016/j.ijhydene.2017.07.044.

Okparanma, R. N., & Mouazen, A. M. (2013). Determination of total petroleum hydrocarbon (TPH) and polycyclic aromatic hydrocarbon (PAH) in soils: a review of spectroscopic and nonspectroscopic techniques. *Applied Spectroscopy Reviews, 48*(6), 458–486. https://doi.org/10.1080/05704928.2012.736048.

Pejcic, B., Boyd, L., Myers, M., Ross, A., Raichlin, Y., Katzir, A., Lu, R., & Mizaikoff, B. (2013). Direct quantification of aromatic hydrocarbons in geochemical fluids with a mid-infrared attenuated total reflection sensor. *Organic Geochemistry, 55*, 63–71. https://doi.org/10.1016/j.orggeochem.2012.11.011.

Rajalahti, T., Arneberg, R., Berven, F. S., Myhr, K. M., Ulvik, R. J., & Kvalheim, O. M. (2009). Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems, 95*(1), 35–48. https://doi.org/10.1016/j.chemolab.2008.08.004.

Rocha, W. F. C., Vaz, B. G., Sarmanho, G. F., Leal, L. H. C., Nogueira, R., Silva, V. F., & Borges, C. N. (2012). Chemometric techniques applied for classification and quantification of binary biodiesel/diesel blends. *Analytical Letters, 45*(16), 2398–2411. https://doi.org/10.1080/00032719.2012.686135.

Roudier, P., Hedley, C. B., Lobsey, C. R., Viscarra Rossel, R. A., & Leroux, C. (2017). Evaluation of two methods to eliminate the effect of water from soil vis–NIR spectra for predictions of organic carbon. *Geoderma, 296*, 98–107. https://doi.org/10.1016/j.geoderma.2017.02.014.

Schwartz, G., Ben-Dor, E., & Eshel, G. (2012). Quantitative analysis of total petroleum hydrocarbons in soils: comparison between reflectance spectroscopy and solvent extraction by 3 certified laboratories. *Applied and Environmental Soil Science, 2012*, 1–11. https://doi.org/10.1155/2012/751956.

Silverstein, M. R., Webster, F. X., & Kiemle, D. J. (2005). Spectrometric identification of organic compounds-7th Ed. State University of New York.

Todd, G. D., Chessin, R. L., & Colman, J. (1999). *Toxicological profile for total petroleum hydrocarbons (TPH). Agency for Toxic Substances and Disease Registry.* Atlanta: Agency for Toxic Substances and Disease Registry.

US ENVIRONMENTAL PROTECTION AGENCY. (1978). Method 418.1: petroleum hydrocarbons (spectrophotometric, infrared).

US ENVIRONMENTAL PROTECTION AGENCY. (1996a). *Method 8440—total recoverable petroleum hydrocarbons by infrared spectrophotometry.* Washington DC: EPA Methods.

US ENVIRONMENTAL PROTECTION AGENCY. (1996b). Method 3510C: separatory funnel liquid-liquid extraction. *Test Methods for Evaluating Solid Waste, Physical/Chemical Methods*, (December), 1–8.

US ENVIRONMENTAL PROTECTION AGENCY. (1996c). *Method 3540C: Soxhlet extraction*. Washington DC: US Environmental Protection Agency. https://doi.org/10.1017/CBO9781107415324.004.

US ENVIRONMENTAL PROTECTION AGENCY. (2007a). Method 8015C: nonhalogenated organics by gas chromatography. Washington DC.

US ENVIRONMENTAL PROTECTION AGENCY. (2007b). *Method 3550: ultrasonic extraction*. Washington DC: EPA Methods.

US ENVIRONMENTAL PROTECTION AGENCY. (2010). Method 1664: n-hexane extractable material (HEM; oil and grease) and silica gel treated n-hexane extractable material (SGT-HEM; non-polar material) by extraction and gravimetry. Washington DC.

Vershinin, V. I., & Petrov, S. V. (2016). The estimation of total petroleum hydrocarbons content in waste water by IR spectrometry with multivariate calibrations. *Talanta, 148*, 163–169. https://doi.org/10.1016/j.talanta.2015.10.076.

Vohland, M., Ludwig, M., Thiele-Bruhn, S., & Ludwig, B. (2014). Determination of soil properties with visible to near- and mid-infrared spectroscopy: effects of spectral variable selection. *Geoderma, 223–225*(1), 88–96. https://doi.org/10.1016/j.geoderma.2014.01.013.

Wang, L., Liu, E., Cheng, Y., Bekele, D. N., Lamb, D., Chen, Z., Megharaj, M., & Naidu, R. (2015). Novel methodologies for automatically and simultaneously determining BTEX components using FTIR spectra. *Talanta, 144*, 1104–1110. https://doi.org/10.1016/j.talanta.2015.07.044.

Wang, Y., Rong, Z., Qin, Y., Peng, J., Li, M., Lei, J., et al. (2016). The impact of fuel compositions on the particulate emissions of direct injection gasoline engine. *Fuel, 166*, 543–552. https://doi.org/10.1016/j.fuel.2015.11.019.

Webster, G. T., Soriano-Disla, J. M., Kirk, J., Janik, L. J., Forrester, S. T., McLaughlin, M. J., & Stewart, R. J. (2016). Rapid prediction of total petroleum hydrocarbons in soil using a hand-held mid-infrared field instrument. *Talanta, 160*, 410–416. https://doi.org/10.1016/j.talanta.2016.07.044.

Weisman, W. H. (1998). *Analysis of petroleum hydrocarbons in environmental media*. (W. Weisman, Ed.)*Total Petroleum Hydrocarbon Criteria Working Group Series* (Vol. 1). Amherst: Total Petroleum Hydrocarbon Criteria Working Group Series. http://www.qros.co.uk/Total Petroleum Hydrocarbon Criteria Working Group Series Volume 1 Analysis of petroleum hydrocarbons in environmental media.pdf

Wold, S., Antti, H., Lindgren, F., & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems, 44*(1–2), 175–185. https://doi.org/10.1016/S0169-7439(98)00109-9.

Workman, J., & Weyer, L. (2007). Practical guide to interpretive near-infrared spectroscopy. *CRC Press., 47*(25), 4628–4629. https://doi.org/10.1002/anie.200885575.

Worley, B., Halouska, S., & Powers, R. (2013). Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry, 433*(2), 102–104. https://doi.org/10.1016/j.ab.2012.10.011.

Xiaobo, Z., Jiewen, Z., Povey, M. J. W., Holmes, M., & Hanpin, M. (2010). Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta, 667*(1–2), 14–32. https://doi.org/10.1016/j.aca.2010.03.048.

Yin, M., Tang, S., & Tong, M. (2016). Identification of edible oils using terahertz spectroscopy combined with genetic algorithm and partial least squares discriminant analysis. *Analytical Methods, 8*(13), 2794–2798. https://doi.org/10.1039/C6AY00259E.

Zhang, M. L., Sheng, G. P., Mu, Y., Li, W. H., Yu, H. Q., Harada, H., & Li, Y. Y. (2009). Rapid and accurate determination of VFAs and ethanol in the effluent of an anaerobic H2-producing bioreactor using near-infrared spectroscopy. *Water Research, 43*(7), 1823–1830. https://doi.org/10.1016/j.watres.2009.01.018.