Determination of Density and Viscosity of Crude Oil Samples from FTIR Data using Multivariate Regression, Variable Selection and Classification

Shahla Alizadeh*. Souvik Ta. Ajay K. Ray***. Lakshminarayanan, S.******

** Department of Chemical and Biochemical Engineering, Western University, Canada N6A 5B9
(e-mail: salizad7@uwo.ca)*

*** Department of Chemical and Biochemical Engineering, Western University, Canada N6A 5B9
(e-mail: sta4@uwo.ca)*

**** Department of Chemical and Biochemical Engineering, Western University, Canada N6A 5B9
(e-mail: aray6@uwo.ca)*

***** Department of Chemical and Biochemical Engineering, Western University, Canada N6A 5B9
Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore
(e- mail: lsamaved@uwo.ca)*

Abstract: The use of Fourier Transform Infrared (FTIR) spectroscopy for quantification of crude oil properties was investigated using chemometric methods. Sample sets consisting of crude oil from seven different Canadian fields were analyzed. Different methods such as PLS, PCA, iPLS, and PLS-GA were used for model building and the results were compared. Evaluation of the models was conducted by determination of the coefficient of determination (R^2) and cross validation error. The best results for quantification of density and viscosity were obtained by partial least squares (PLS) regression on FTIR data. Data analysis on the total sample set of 82 samples yielded a prediction error (root mean square error of cross validation) of 4.5 x 10^{-5} and 0.33 respectively for density and viscosity. Improvement in prediction accuracy of viscosity was obtained by using Decision tree classification on samples before applying PLS regression.

Keywords: Chemometric tools, Partial least squares, FTIR, Crude oil, Viscosity, Density, Classification, Multivariate regression, Spectroscopy.

1. INTRODUCTION

An accurate evaluation of physical properties for crude oil is essential for addressing many reservoir engineering and process operational problems. Since petroleum is a complex mixture of organic compounds, its quality is mostly evaluated by physicochemical properties. These properties are ideally determined experimentally on actual fluid samples via elaborate laboratory procedures. Crude oil from different fields and wells come with different characteristics, hence determination of their physicochemical properties is quite valuable to production specialists as well as reservoir engineers. Over the last few decades, several correlations have been developed to estimate the crude oil properties. However, these correlations may be useful only in regional geological provinces and may not provide satisfactory results when applied to crude oils from other regions (Hanafy. H, 1997). Among oil properties, density and viscosity are two important parameters in crude oil specification – they are normally measured in the laboratory, with procedures that can last about hours and cost hundreds of dollars for viscosity measurement of each sample, especially when dealing with heavy oils, though density evaluation is quite faster and more inexpensive.

Owing to rapid and significant advances in the fields of multivariate statistics and machine learning techniques, it is now possible to estimate many properties of interest (that are difficult or costly to measure) using other measurements that are relatively simpler, faster, and less costly. Analytical procedures which are less dependent on sample size (Rocha et al., 2016) are now available for mapping out the relationship between the easily available measurements such as spectra and the difficult to obtain measurements such as the viscosity of crude oil.

Multivariate calibration methods, specifically partial least squares (PLS) regression has become a standard tool in chemometrics and used in chemistry and engineering (Wold et al., 2001) for model building in the above context. Techniques such as PLS allow the treatment of complex data from a mathematical and statistical point of view by correlating instrumental measurements and values with a corresponding property of interest.

In recent years, infrared spectroscopy has shown to be a promising tool in qualitative analysis of petroleum, diesel and biodiesel. Santos et al., 2005 analyzed diesel samples by FTIR and FT-Raman spectroscopy using PLS and artificial neural networks (ANN). Filguerias et. al., 2014 estimated the API gravity, kinematic viscosity and water content in petroleum using ATR-FTIR spectroscopy measurements. Rocha et al., 2016 used mid-infrared and near-infrared spectroscopy and PLS to determine sulfur content in Brazilian petroleum fractions.

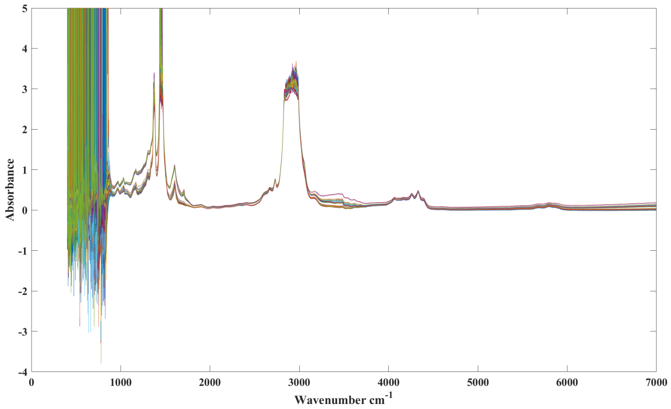

Figure 1. FTIR spectra of eighty-two crude oil samples in the range of 400 to 7000 cm^{-1}

This study focuses on the usage of a combination of infrared spectroscopy data with chemometric methods that have been successfully applied by researchers and industry practitioners in many applications. A reliable, fast and low-cost method is developed for determination of density and viscosity of Canadian crude oil samples based on decision tree classification and linear PLS modeling.

2. MATERIALS AND METHODS

Eighty-two crude oil samples obtained from seven different Canadian oil fields were supplied by a petroleum company in Canada. They obtained the FTIR spectra corresponding to these samples using a Thermo Fisher FTIR microscope. They also have measured several physicochemical properties of these crude oil samples using appropriate analytical instruments. The company wants to develop robust models that can provide accurate estimates of the physicochemical properties utilizing FTIR data only thus avoiding the need for elaborate, expensive and time-consuming laboratory procedures. In this work, we focus only on the estimation of density and viscosity of crude oil samples from their FTIR data.

3. CHEMOMETRIC METHODS

3.1 Principal Component Analysis (PCA)

Principal Components Analysis (PCA) is a very popular multivariate statistical technique for analyzing data samples where there is a high collinearity between the variables. PCA can be used to create a new set of orthogonal variable space by creating optimal linear combinations of the original variables. In the presence of collinear variables, the new orthogonal variables (called principal components) form a lower dimensional subspace that captures most of the variability present in the original data space. Thus, PCA is also a dimensionality reduction method that facilitates a low-dimensional view of the original data set with minimal or no loss of information content of the original dataset. Mathematically, PCA decomposes the appropriately scaled original data matrix (X; dimension n x c where n is the number of samples and c is the number of variables) into two matrices namely scores matrix (T) and loadings matrix (P) such that $X = TP^{'}$. T (dimension n x c) represents the new set of orthogonal variables (note: T = X P) and is referred to as the *scores matrix*. The columns of the *loadings matrix* P represent the weights given to the original X variables to form the principal components (latent variables). Thus, elements of the P matrix tell us how much an original X variable is loaded into the corresponding principal component (latent variable). Samples that are more alike and have close T values will cluster together in the scores plot (where one column of T is plotted against another). Likewise, variables that are highly correlated tend to cluster together in the loadings plot (where one column of P is plotted against another). Thus, the scores and loadings plot have useful diagnostic value when analyzing data sets with no a priori information making PCA a versatile exploratory data analysis and visualization tool. In practice, however, one recognizes that the measurement matrix X has noise and redundancies and therefore X is decomposed as $X = T_k P_k^{'} + E_k$ where $T_k P_k^{'}$ represents the signal portion of X and E_k represents the noise or redundancies in the system (i.e., residual space). Symbol k represents the number of latent variables retained to avoid both underfitting and overfitting of the data. The value of k is determined via methods such as eigenvalue 1 criterion, percentage variance captured or by empirical methods such as multi-fold cross validation (Wold, S. and Sjöström, M., 1998). In this work, the optimal number of principal components k was obtained via cross validation.

Once the PCA model is constructed, statistical measures such as Hotelling's T^2 and squared prediction error (SPE) can be constructed for the signal and residual spaces respectively. These measures serve to identify outliers and abnormalities in the process data in the spirit of statistical process control (i.e., like control charts).

3.2 Principal Component Regression (PCR)

Oftentimes, the need is to relate information related to two sets of variables i.e., construction of a mapping relationship (e.g. a linear regression model) between the process variables X and the quality variables Y. If the X space is comprised of correlated variables, the construction of a multiple linear regression model using the least squares formulation becomes difficult owing to ill-conditioning of the X space which makes its inversion numerically unstable. One potential way to construct the regression model in such situations is to get a PCA model of the X space (as described above) and using the latent variables (T space) as regressors and obtaining a multiple linear regression model between T and Y. Thus, when a new sample is obtained in the X-space, it is projected into the lower dimensional orthogonal T-space and then the linear regression model is used to get the estimated Y values. Thus, PCR is able to circumvent the ill-conditioning problem associated with the collinear X space and also permit its lower dimensional representation and provide a regression model with lesser number of parameters. (Keithley et al., 2009)

3.3 Partial Least Squares (PLS)

PLS represents one of the most commonly used methods for multivariate calibration. PLS determines the mapping between the input space X (say FTIR measurements) and the output space Y (say properties such as density and kinematic viscosity) by identifying maximally related latent variables (using covariance measure) of the X and Y spaces (called the outer model) and relating them via a univariate regression model component by component (called the inner model). Conceptually, PLS constructs a multivariate regression model relating the X and Y spaces (each of which may have highly correlated variables) by decomposing it into several univariate regression problems and finally putting them together to get the multivariate regression model. In effect, the suitably scaled X and Y matrices are decomposed as:

$$X = T_k P_k^{'} + E_k \qquad (1)$$

and

$$Y = U_k Q_k^{'} + F_k \qquad (2)$$

where T and U are the scores matrices of the X and Y spaces respectively and represent the latent variables of the X and Y spaces (i.e. optimal linear combinations of the original variables). Matrices P and Q represent the loadings matrices of the X and Y spaces respectively and k represents the optimal number of dimensions in the PLS model that is typically obtained via cross validation. $T_k P_k^{'}$ and $U_k Q_k^{'}$ represent the signal component of the X and Y spaces while E_k and F_k represent the residual (noise and redundancies) component of the two spaces respectively.

The inner model relating each dimension (column) of the T and U spaces are obtained via simple linear regression as:

$$u_i = b_i t_i + \varepsilon_i \qquad (3)$$

for i = 1, 2, …, k.

Equations (1) and (2) represent the PLS outer model and equation (3) represents the PLS inner model (Mohammadi et al., 2019).

3.4 Decision Trees (DT)

Classification (of class-labelled data) using DT has been successfully deployed for a wide range of applications across domains such as engineering, medicine, and business. The DT is composed of a root node which consist of the entire data, a set of internal nodes (splits), and a set of terminal nodes (leaves). DT is a classification procedure that recursively partitions a data set into smaller subdivisions based on optimizing a measure such as entropy or purity defined at each branch (or node) in the tree. The splitting is continued until the terminal nodes are of sufficient purity or the complexity of the tree (its depth and width) is acceptable. Typically, a binary split is applied to a parent node by choosing the best “split variable” and a value that is chosen for the best “split variable” so that the total purity of the two child nodes is better than the purity of the parent node. A class label is assigned to each terminal node. When a new sample is obtained, the rules can be checked starting from the root node all the way into one of the terminal nodes and the sample is classified as belonging to the class label assigned to the terminal node (Friedl et al., 1997).

Despite the popularity and effectiveness of the DT method, its use in the classification of crude oil samples based on spectroscopy data has not been widely discussed. In this paper, we employ the decision tree classification method for classifying the Canadian crude oil samples and using the result to determine the viscosity based on a multivariate regression model.

4. RESULTS AND DISCUSSION

The Canadian oil company provided us with 103 crude oil samples, which was reduced to 82 during the preprocessing stage due to some outlier detection, duplicity and missing crucial data. The remaining 82 crude oil samples investigated in this study have density (at 25 ºC) ranging between 0.8 to 0.92 g/cm^3, which means that the API gravity ranges between 21.84 to 44.84 ºAPI and includes both heavy/medium and light oils. Figure 1 shows the FTIR spectra obtained for these samples, from where it can be observed that the FTIR spectra have considerable noise in the region between 400 and 860 cm^{-1} as compared to the other wavelength regions. This prompted us to omit the spectral information between 400 and 860 cm^{-1} in some aspects of this work. However, after more inspection we found that information in the aforementioned interval was important in predicting viscosity values.

4.1 Models for Density

Several chemometric methods such as PCR, PLS, PLS-GA, iPLS and PCA were used to predict the density. Considering the accepted prediction error to be in the range of 0.15, all these methods were providing reliable results.

Principal Components Regression (PCR) was used to relate the FTIR spectra to the density values. The optimum number of principal components was found by iterating our algorithm and changing the number of principal components to be included in our regression. Figure 2 shows it was found that the best results were obtained using 10 principal components for the X (FTIR spectra) space. This optimal model returned a root mean squared error cross validation (RMSECV) value of 4.5 x 10^{-5} and R-squared value (between the actual and model predicted densities) of 0.9795. The results of the predictions of density were deemed to be very accurate and valuable for industry implementation according to our industry partner. Hence, we are not discussing density prediction results in more detail here. Besides being useful, the density predictions from the PCR model were useful as an additional input in the viscosity prediction model as will be explained below.

4.2 Models for Viscosity

Kinematic viscosity of the 82 studied samples were in the range of 2.84 to 178 mm2/s (cSt) at 25ºC and it was aimed to predict them with error values below 0.15. First of all, we tried to find out a global model between the FTIR spectra and viscosity values for all the 82 samples using different chemometric methods. Shown in table 1, even the best results of these models were not acceptable. Figure 3 shows that there is a strong linear relationship between the density and viscosity values, hence in the second approach we tried to use the predicted values of density as the sole input variable to estimate viscosity of samples. Despite this strong linear relationship between density and viscosity values of corresponding samples, using density as a single input did not work out well. Although it was providing us with a good match between current data with acceptable RMSECV values, the problem of overfitting turned out to make huge errors in the prediction of new samples' viscosities due to R^2 values in range of 0.2. As discussed in section 4.1, the predictions for density were quite accurate which emboldened us to consider the predicted values of density as an input value in combination with the spectral data to predict the viscosity of the crude oil samples as the third approach. And the final approach, which was not a global one, consisted of first identifying whether the sample belongs to heavy or light oils based on the FTIR records, then predicting the viscosity. All above approaches and some results are briefly discussed here and indicated in table 1. All the results of discussed models were evaluated by cross validation.

As can be seen in Table 1, PCR models were constructed using the full spectra data as well as by augmenting the full spectra data with the predicted density values. The results and the stability of the models were verified by cross validating them over over 100000 different subsets including 50 samples of the dataset using random subsampling to create the different test-train data splits. The PCR model coupled with density as an input gave the best possible result on full spectra models.

Two approaches that incorporated density predictions as an additional input were tried. (i) The first one was to construct the PCR model on a combined dataset where both the spectra and the predicted density formed the X space and (ii) where PCA was first applied on the spectral data and the regression was done using an input space that contained the spectra related latent variables and the predicted density as an additional input. The latter approach gave considerably better results. The RMSECV turned out to be 201.1 and 160.5 respectively for the two approaches. The R^2 metric for the first and second approaches were found to be 0.938 and 0.9507 respectively. Overall, with the incorporation of predicted density as an input variable, the RMSECV in viscosity prediction dropped by 41.5% (from 274.41 to 160.54). This shows the utility of making predicted density as an input variable in the viscosity prediction model. PLS models were constructed using the full spectra and full spectra augmented with predicted density values (rows 3 and 4 in Table 1). While these PLS models (with k = 20 obtained by cross-validation) substantially reduced the MSE compared to those obtained with PCR, there was only a slight improvement in MSE values when predicted density was added as an input variable. (from 52.9 to 49.2).

In many applications (including those involving spectral data), it is known that variable selection in conjunction with regression gives better results. Many methods such as stepwise regression, genetic algorithm (GA) and interval PLS (iPLS) may be employed for variable selection. First, PLS was combined with GA (PLS-GA) and was implemented using the *PLS-Genetic Algorithm Toolbox* by Leardi (2000). In Leardi (2000), it was suggested that the number of original input variables be limited to less than 200 to reduce the risk of overfitting. Hence, we grouped every 36 neighbor spectra data (absorbance values) and used their average as a variable, in this way the total number of 6842 variables were reduced to 190. This modeling strategy led to a model that had 14 latent variables and RMSECV value of 21.45 that was substantially lower than that obtained with full spectra and density prediction as input variables. If the predicted density was included as the 191st variable, a PLS model with 7 latent variables emerged as the best model with RMSECV of 11.25. Yet another method of variable selection, interval-PLS (iPLS) was tried (Norgaard et al., 2000). The iToolbox available in MATLAB (Leardi and Norgaard, 2004) was used to predict the viscosity using the full FTIR spectra. The spectra were split into 30 wavelength intervals and PLS models were determined for each interval. With this procedure the optimal number of PLS components was found to be 4 with a RMSECV value of 26.14. Thus, variable selection procedures in conjunction with PLS helps in better prediction of viscosity values as does the inclusion of predicted density values.

In Figure 3, the measured density and viscosity values for the 82 samples are plotted. It is apparent that there are two major clusters of crude oil samples. Light oils which have lower values of density (API higher than 31.1 °API) and viscosity, and heavy oils, having higher values of viscosity and density (API lower than 31.1 °API). These two classes are labeled as class 1 and class 2 in this study. A decision tree model was built for the dataset to classify the high and low viscosity classes before passing them through the appropriate regression model to predict the viscosity. When the decision tree was constructed for the data set, it was seen that the spectral measurements at two specific wavelengths (corresponding to variable numbers 506 and 2610) are enough to classify the samples as belonging to either class with 100% accuracy (Figure 4).These selected variables and their neighbors were in agreement with the often selected variables by other methods such as PLS-GA and iPLS. Separate PLS models were then constructed for each of the two classes using only spectral data. As seen from the last two rows of Table 1, the two models with 10 and 19 PLS dimensions respectively resulted in considerably improved MSE values for the two classes (0.33 and 0.11 respectively) even without inclusion of the density estimates. Figure 5 shows the scatter plots for actual and model-predicted viscosities for samples belonging to the two classes. The agreement for heavy oils is very good while it is quite acceptable for light oil samples.

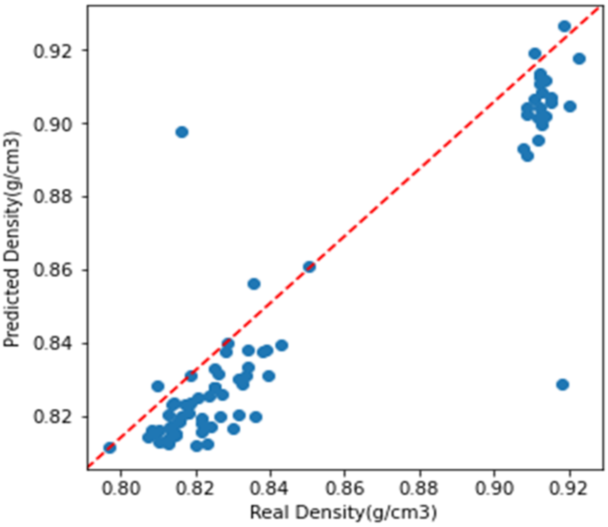

Figure 2. PCR results for prediction of density of 82 crude oil samples based on the FTIR data

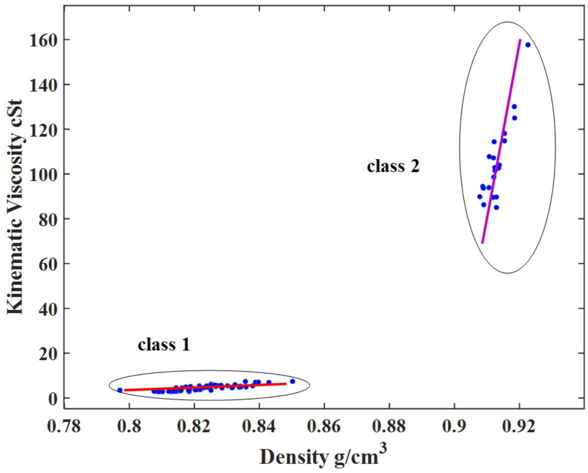

Figure 3. Scatter plot of kinematic viscosity vs density of 82 crude oil samples and their linear relationship

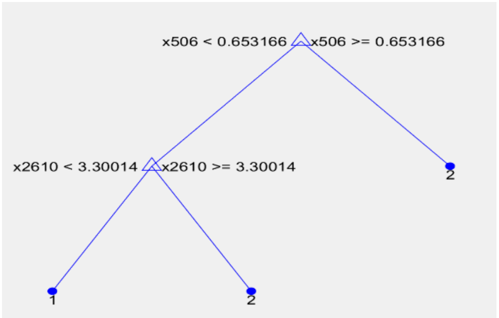

Figure 4. Decision tree result for classification of 82 crude oil samples

In summary, partitioning the data set into different classes and building different PLS models for the classes had the greatest impact on prediction accuracy. Variable selection using GA or a method such as interval PLS (iPLS) was also found to result in improved prediction accuracy.

Table 1. Summary of PCR and PLS Models for Viscosity Prediction for 82 Crude Oil Samples

Model	Input Variables	Number of Components	RMSECV
PCR	Full spectra	10	274.41
PCR	Full spectra & predicted Density	11	160.54
PLS	Full spectra	20	52.9
PLS	Full spectra & predicted Density	20	49.2
PLS-GA	Full spectra pre-processed to 190 variables	14	21.45
PLS-GA	Full spectra pre-processed to 190 variables & predicted Density	7	11.25
iPLS	Full spectra	4	26.14
PLS	Light Oils (Class 1)	10	0.33
PLS	Heavy Oils (Class 2)	19	0.11

6. CONCLUSIONS

Standard chemometric techniques such as Principal Components Regression (PCR) and Partial Least Squares

(PLS) augmented by prior classification using Decision Trees (DT) and variable selection procedures were found to predict to sufficient accuracy (as required for industrial deployment) the density and viscosity values from FTIR data. The best results were obtained by firstly classifying the samples into a specific class and then using the PLS regression model developed for that class. Presently, we are developing models to estimate the levels of industrially relevant contaminants in crude oil from FTIR data.

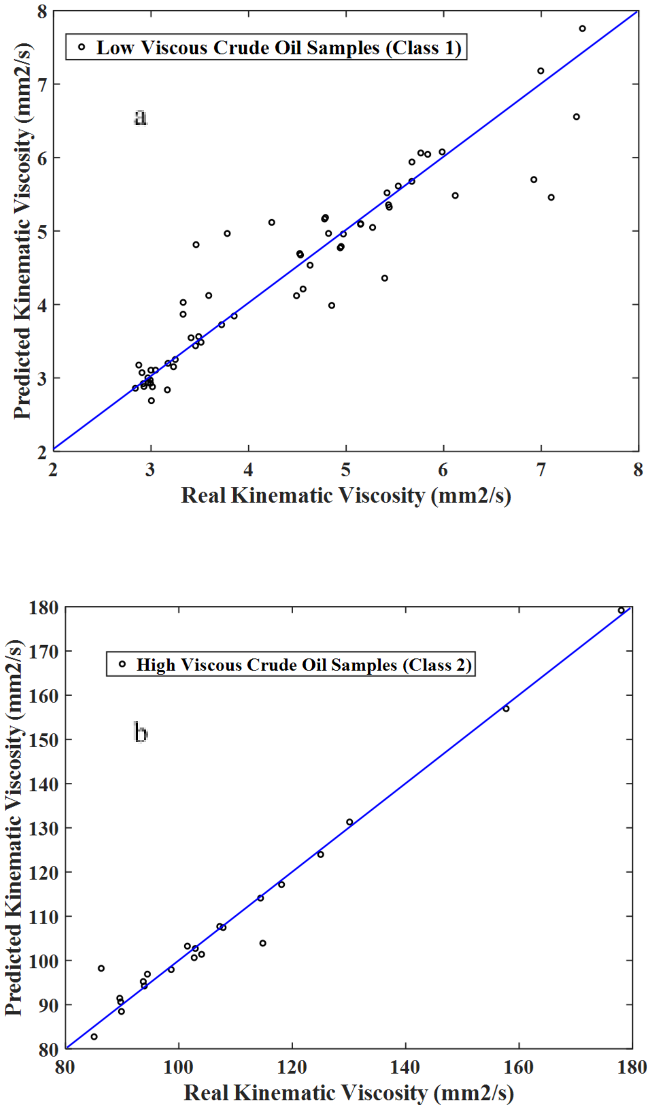

Figure 5. Predicted kinematic viscosities of crude oil samples vs. real kinematic viscosities a) Light crude oil samples (class 1) b) Heavy crude oil samples (class 2)

REFERENCES

Filgueiras, P.R., Sad, C.M., Loureiro, A.R., Santos, M.F., Castro, E.V., Dias, J.C. and Poppi, R.J., 2014. Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. *Fuel*, *116*, pp. 123-130.

Friedl, M.A. and Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, *61*(3), pp. 399-409.

Hanafy, H.H., Macary, S.M., ElNady, Y.M., Bayomi, A.A. and El Batanony, M.H., 1997, March. A new approach for predicting the crude oil properties. In *SPE Production Operations Symposium*. OnePetro.

Keithley, R.B., Wightman, R.M. and Heien, M.L., 2009. Multivariate concentration determination using principal component regression with residual analysis. *TrAC Trends in Analytical Chemistry*, *28*(9), pp.1127-1136.

Leardi, R., 2000. Application of genetic algorithm–PLS for feature selection in spectral data sets. Journal of Chemometrics: A Journal of the Chemometrics Society, 14(5‐6), pp. 643-655.

Leardi R., and Nørgaard L., 2004. Sequential application of backward interval PLS and Genetic Algorithms for the selection of relevant spectral regions, *Journal of Chemometrics,* 18(11), pp. 486-497.

Mohammadi, M., Khorrami, M.K. and Ghasemzadeh, H., 2019. ATR-FTIR spectroscopy and chemometric techniques for determination of polymer solution viscosity in the presence of SiO2 nanoparticle and salinity. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *220*, p. 117049.

Norgaard, L. et al. (2000) Interval Partial Least-Squares Regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. Applied Spectroscopy. [Online] 54 (3), pp. 413-419.

Rocha, J.T., Oliveira, L.M., Dias, J.C., Pinto, U.B., Marques, M.D.L.S., Oliveira, B.P., Filgueiras, P.R., Castro, E.V. and de Oliveira, M.A., 2016. Sulfur determination in brazilian petroleum fractions by mid-infrared and near-infrared spectroscopy and partial least squares associated with variable selection methods. *Energy & Fuels*, *30*(1), pp. 698-705.

Santos Jr, V.O., Oliveira, F.C.C., Lima, D.G., Petry, A.C., Garcia, E., Suarez, P.A. and Rubim, J.C., 2005. A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis. *Analytica Chimica Acta*, *547*(2), pp. 188-196.

Wold, S. and Sjöström, M., 1998. Chemometrics, present and future success. *Chemometrics and Intelligent Laboratory Systems*, *44*(1), 3–14.

Wold, S., Sjöström, M. and Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems, 58*(2), pp. 109-130.