A Simple Statistical Model for Predicting Crude Oil API Values*

[1]S. A. Marfo and [1]C. B. Bavoh
[1]University of Mines and Technology, Box 237, Tarkwa, Ghana

Marfo, S. A., Bavoh, C. B. (2023), "A Simple Statistical Model for Predicting Crude Oil API Values", *Ghana Mining Journal*, Vol. 23, No. 1, pp. 22-26.

Abstract

American Petroleum Institute (API) gravity value is the main indicator of crude oil quality and marketing value; hence, it must be simply and accurately determined. Existing crude oil API prediction models are complex and time-consuming because they use lots of parametric properties for predictions. Herein, we propose a simple two-variable (aromatic and naphthene content) statistical model for predicting crude oil API values. The statistical model in this study was developed using multiple linear regression techniques on about 80 crude oil samples from different locations. The study shows that the developed model in this work could accurately predict crude oil API gravity values with a standard error of 3.14 and a correlation matrix of 0.92. Also, the model confirmed that the use of crude oil aromatic and naphthene content could accurately describe its API values. The model could predict crude oil API better than some API models in literature by 38% - 62%. The findings in this work provide a simple and fast method of determining crude oil API for crude marketing and inspection.

Keywords: Crude oil; Multiple linear regression; API; Aromatics; Naphthene

1 Introduction

Generally, crude oil is the main source of the fuels used to power most economies in the world (Demirbas et al., 2015). Crude oil in its normal form is valueless until it is refined into useful fractions. However, the presence of heavy hydrocarbon content and "hetero" atoms such as nitrogen, sulphur, and oxygen mostly reduce the refinability of the crude (Alzarieni et al., 2021). These properties of crude importantly affect the crude oil market value. The value of crude in the market is controlled by its API gravity values as proposed by the American Petroleum Institute. This defines the quality of the crude as light (°API > 31.1), medium (22.3 < °API < 31.1), heavy (10 < °API < 22.3) and extra heavy (°API < 10)(Goel et al., 2017). Aside the use of API to determine the price of crude oil, it also has a direct impact on the quantum of investment and energy consumption of the refinery. Therefore, an accurate estimation of crude oil API values is important to refineries' operational efficiency.

Crude API is mostly measured using the American Society for Testing and Materials (ASTM) methods D287 and D1298 (Goel et al., 2017). The ASTM method is time-consuming, lacks quick results obtainability, and requires expensive equipment when used for online crude oil monitoring. It is therefore important to replace the experimental methods with empirical models. Several authors have proposed mathematics for predicting crude oil API values using other crude oil-associated properties such as viscosity, aromatic, naphthalic, and saturate content. Also, the use of crude oil spectroscopic data for predicting API has been reported. Abbas *et al.*, (2012) employed FTIR data to predict crude oil API values using the Partial Least-Square (PLS) analysis. Also, the use of PLS regression on NIR spectroscopy data to predict crude oil API densities was reported by Hidajat and Chong, (2000). Aside the use of spectroscopic data in predicting crude oil API, SARA (saturate, aromatic, resin and asphaltene) fractions are other parameters some authors use to model the API densities of crude oils. Four SARA fractions were employed as inputs to develop an API prediction model by Fan et al., (2002). A nonlinear SARA data-based API model was also proposed by Goel *et al.*, (2017). Despite the acceptable level of crude oil API prediction accuracies from the existing models, these methods are not that simple with respect to the number and quality of the variables used for the predictions.

Recent crude API models are based on artificial intelligence techniques by employing 'nonlinear' modelling techniques. Goel *et al.*, (2017) used the artificial neural network (ANN) model, support vector regression (SVR), and genetic programming (GP) to predict crude oil APIs using four SARA analysis parameters. Aside their work, Lozano *et al.*, (2017); de Paulo *et al.*, (2020) have also used genetic algorithms and Partial Least Squares to predict the API properties of crude oils. Though the use of machine learning for API predictions is encouraging, current machine learning models employ lots of variables for their predictions which makes them complex and sophisticated. This complex nature and accuracy levels usually define the rigorous nature of machine learning prediction techniques. However, in the context of predicting crude properties during process monitoring, the complexity and demand for lots of variables are potential drawbacks to be considered in machine learning models in terms of cost and fast API gravities prediction characterisation (Abbas et al., 2012; Correa Pabón and Souza Filho, 2019; Goel et al., 2017; Guzmán-Osorio et al., 2020; Pantoja et al., 2011). Also, machine learning algorithms'

*Manuscript received March 12, 2023
Revised version accepted June 25, 2023
https://doi.org/10.4314/gm.v22i1.1

predictions are hidden and cannot be simply denoted by equations that can be used readily. Hence, the need for a simple API prediction equation or model with few suitable variables is necessary for academic and industrial purposes.

In this work, we established the dependence of crude oil API on the naphthenic and aromatic properties in various crudes around the world. These variables were chosen as suitable properties that could describe crude oil API behaviour. A simple multiple regression model was developed to predict the crude oil API gravity using the crude oil's naphthenic and aromatic properties. The findings in this work give further evidence and confirm the behaviour of crude oil API in relation to their naphthenic and aromatic properties. It further introduces a simple and easy two-variable API gravity prediction model for industrial and academic use.

2 Resources and Methods Used

2.1 Database

The use of quality and accuracy is the key to developing accurate statistical models between the variables. Thus, this work was conducted by carefully selecting quality crude oil process monitoring data that could accurately predict oil API gravities. In this work, the API, aromatic and naphthene content of over 80 different crude oils around the world was extracted from their crude oil assay data and used for the model development as shown in Table 1. The extracted data were pretreated and used for the model development.

Table 1. Summary of crude oil assay data used in this work.

Variables	Mean	Max	Min
Aromatics (vol%)	30.40	55.72	7.40
Naphthenes (Vol%)	33.71	44.04	17.26
API	36.61	73.10	19.08

Max (Maximum); Min (Minimum)

2.2 Model development

Before the model development, the data visualisation analysis was conducted using Predictive Analytics Software (JMP) Pro 16. The data visualisation process focused on determining the correlation matrix, skewness, and collinearity of the data. Fig. 1 shows the adopted model development flow chart. Multiple linear regression (MLR) was used to develop API gravity prediction model in this study. The MLR model theory in this work constructs a baseline regression that aims at creating a linear relationship between crude oil properties (aromatics and naphthene contents) and the response variables (crude oil API) (Goel et al., 2017). The MLR model then provides a simple equation as a linear function of the crude API gravity. The MLR model is most suitable for establishing simple and easy-to-use predictive models. The MLR equation of the API prediction in this study is expressed as:

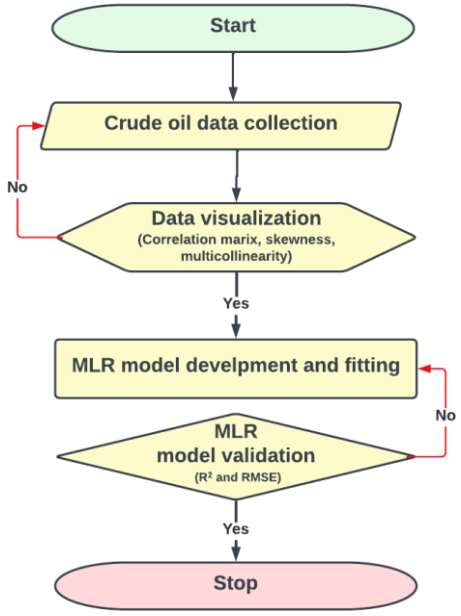

Fig. 1 MLR model development flowchart.

$$API = k_{1_{API}}a + k_{2_{API}}n + k_{3_{API}} \quad (1)$$

where a and n are the crude oil aromatics and naphthenes contents. $k_{1_{API}}, k_{2_{API}}, and\ k_{3_{API}}$ are the model constants for API. The model coefficients were optimised using JMP Pro 16. The API model performance was evaluated using root mean squared error (RSME) and the determination coefficient R^2. The specific formulas of the above two evaluation indices are defined below as.

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

$$R^2 = 1 - \frac{\sum_i(\hat{y}-y_i)^2}{\sum_i(\overline{y_i}-y_i)^2} \quad (3)$$

where $\hat{y}_i$ is the MLR model predicted values, y_i denotes the API experimental values, and m represents the number of data sets.

3 Results and Discussion

An initial data visualisation and statistical correlations were conducted to ensure the absence of collinearity for the API MLR model development. The dependent variables are crude oil API, while the independent variables are the aromatics and naphthenes content. The estimated correlation

matrix for the studied variables is presented in Table 2 and Fig. 2. The observed correlations among the independent variables (aromatics and naphthenes) are very weak (R < 0.1). The weak correlation matrix suggests the absence of collinearity between the independent variables. Aside the absence of collinearity in the data, the histogram and scatter plots in Fig. 2 exhibit an acceptable skewness and distribution.

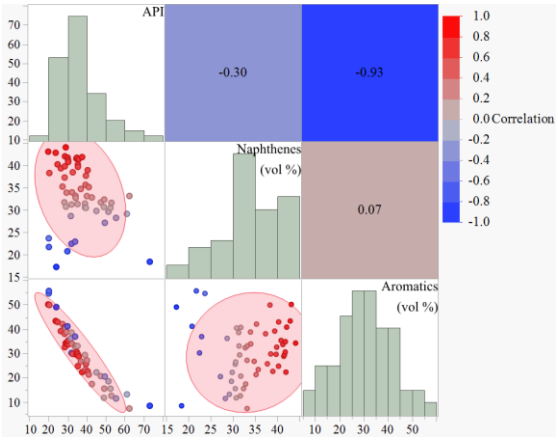

Fig. 2. Histograms, correlation matrix heat maps, and scatter plots of the variables used in this work

Table 2. Correlation matrix for API gravity and modelled variables used in this work.

Variables	API	Naphthenes (vol %)	Aromatics (vol %)
API	1.0000	-0.2962	-0.9339
Naphthenes (vol %)	-0.2962	1.0000	0.0746
Aromatics (vol %)	-0.9339	0.0746	1.0000

The MLR model was used in this study to provide a simple linear two-parameter function for predicting crude oil API gravity. The developed model in this study could effectively predict crude oil API values accurately as shown in Fig. 3. The standard error and correlation coefficient for the prediction of API was 3.14 and 0.92, respectively. The API model coefficients parameters for Equation 1 are shown in Table 3. The model coefficients show that both aromatics and naphthenes content of crude oil negatively affects the crude oil API gravity. The crude oil's aromatic content has the most dominating impact on determining the crude's API. Crude oil aromatic content API predictor strength is 2 times than the naphthene content. This implies that crude oil with more aromatic content would exhibit poor crude oil API values. This further agrees with the findings by Stasiuk and Snowdon, (1997) who confirmed that crude oil's aromatic content highly affects its API value. The API model coefficients further agree with the ternary classification of crude oil based on their aromatic, naphthenic and paraffinic content (Behrenbruch and Dedigama, 2007). However, during the model development, the paraffinic data of crude oils was excluded due to collinearity correlations with the aromatic and naphthenic variables. Thus, the aromatic and naphthenic properties of the crude have a strong ability to predict crude oil API by representing the crude oils paraffinic content as a constant value ($k_{3_{API}}$). Hence, if one assumes a pure paraffinic crude oil system where there are little or no aromatic and naphthene contents, the API value of paraffinic crude oil system would be 76.5504. This value is in the same range with maximum API gravity value (75), representing the API gravity for a lease condensate (Kawthar et al., 2021). Which mainly consist of natural gas liquids (NGLs) such as ethane, propane and butane.

Table 3. Coefficients for Equations 1

Equation parameters	API
$k_{1_{API}}$	-0.8888
$k_{2_{API}}$	-0.3831
$k_{3_{API}}$	76.5504
RMSE	3.1350
R^2	0.9209

The performance of the developed model for predicting crude oil was compared with some existing API predictive models in literature that used different crude oil properties as their model variables. Thus, neglecting the models' optimisation performances, the impact of the choice of variable type on crude oil API prediction could be determined. The five models used for comparison in this work are MLR, genetic programming (GP), support-vector regression (SVR), Fan and Buckley (FB), and modified FB. All the models were used as machine learning techniques and thus their prediction performances are expected to be accurate owing to the use of more variables (4) than this current study. The models used saturates, aromatics, resins, and asphaltenes as the variable for predicting crude API gravity. Table 4 presents the model performance comparison between this study and those of Goel *et al.*, (2017). Surprisingly, the simple two-parameter MLR proposed in this work outperforms all the five machine learning models reported by Goel *et al.*, (2017). In Table 4, the correlation matrix and standard error for the proposed model in this work are about 6% - 26% and 38% - 62% better than the models in literature, respectively.

Table 4. Comparison of the performance of the proposed model in this work with conventional API predictive models in literature

Model	Type of Model	No. Parameters	Parameters	R^2	RMSE	Ref.
MLP	Statistical Method	2	Aromatics and naphthenes	0.92	3.12	This work
MLP	ML	4	Saturates, aromatics, resins, and asphaltenes	0.86	5.19	(Goel et al., 2017)
GP	ML	4	Saturates, aromatics, resins, and asphaltenes	0.84	5.54	(Goel et al., 2017)
SVR	ML	4	Saturates, aromatics, resins, and asphaltenes	0.87	5.00	(Goel et al., 2017)
FB	ML	4	Saturates, aromatics, resins, and asphaltenes	0.73	8.17	(Goel et al., 2017)
Modified FB	ML	4	Saturates, aromatics, resins, and asphaltenes	0.82	5.81	(Goel et al., 2017)

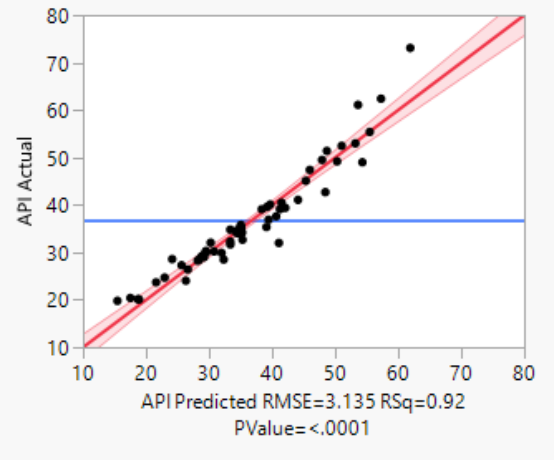

Fig. 3 The proposed simple two-variable MLR model accuracy for crude oil API model

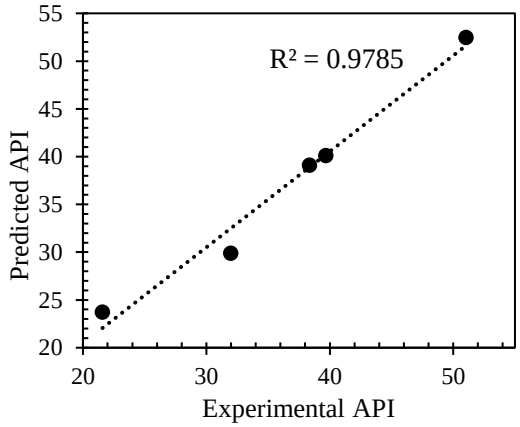

Fig. 4 Experimental and predicted API gravities of the blind data set

It must be stated that the comparative analysis conducted in this work is not intended to undermine the optimisation capabilities of machine learning optimisation techniques. Rather, the bases of comparison in this study were to compare other API models with our simplified model. The performance of the model in this work could be due to the choice of model parameters that highly describes the crude API properties compared with those used in the machine learning models. We suspect that the use of machine learning to optimise the parameters in our model would yield much better API predictions.

The developed model in this work was further tested on blind data sets from different locations to appreciate its prediction capabilities. In Fig. 4, it is observed that the predicted date agrees with the blind experimental data set with a correlation matrix of 0.9785.

4 Conclusion

In this study, a simple crude oil API value prediction model was developed using the MLR technique. The developed model in this work exhibited suitable crude oil API prediction accurately with a standard error of 3.12. The model parameters highly affected the API values and describe the chemical structure of crude oils. Specifically, crude oil aromatics and naphthene content linearly relates to the crude oil's API gravities. Our model performed better than existing crude oil API models reported in literature. The finding in this work is useful for predicting crude oil API values accurately and easily for quality inspection and marketing purposes.

Acknowledgement

The authors are grateful to the University of Mines and Technology, Tarkwa for providing facilities for this work.

References

Abbas O, Rebufa C, Dupuy N, et al. (2012), "PLS regression on spectroscopic data for the prediction of crude oil quality: API gravity and aliphatic/aromatic ratio", *Fuel*, Vol. 98, pp. 5-14.

Alzarieni KZ, Zhang Y, Niyonsaba E, et al. (2021),

"Determination of the Chemical Compositions of Condensate-like Oils with Different API Gravities by Using the Distillation, Precipitation, Fractionation Mass Spectrometry (DPF MS) Method", *Energy and Fuels*, Vol. 35, No. 10, pp. 8646–8656.

Behrenbruch P and Dedigama T (2007), "Classification and characterisation of crude oils based on distillation properties", Vol. 57, No. 1-2, pp. 166–180.

Correa Pabón RE and Souza Filho CR de (2019), "Crude oil spectral signatures and empirical models to derive API gravity", *Fuel*, Vol. 237, pp. 1119–1131.

de Paulo EH, Folli GS, Nascimento MHC, et al. (2020), "Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties", *Fuel*, Vol. 279, pp. 118462.

Demirbas A, Alidrisi H and Balubaid MA (2015), "API gravity, sulfur content, and desulfurization of crude oil", *Petroleum Science and Technology* Vol. 33, No. 1, pp. 93–101.

Fan T, Wang J and Buckley JS (2002) Evaluating Crude Oils by SARA Analysis. *Proceedings - SPE Symposium on Improved Oil Recovery*: 883–889. DOI: 10.2118/75228-ms.

Goel P, Saurabh K, Patil-Shinde V, et al. (2017), "Prediction of °aPI values of crude oils by use of saturates/aromatics/resins/asphaltenes analysis: Computational-intelligence-based models", *SPE Journal*, Vol. 22, No. 3, pp. 817–853..

Guzmán-Osorio FJ, Adams RH, Domínguez-Rodríguez VI, et al. (2020), "Alternative method for determining API degrees of petroleum in contaminated soil by FTIR", *Egyptian Journal of Petroleum*, Vol. 29, No. 1, pp. 39–44.

Hidajat K and Chong SM (2000), "Quality characterisation of crude oils by partial least square calibration of NIR spectral profiles", *Journal of Near Infrared Spectroscopy*, Vol. 8, No. 1, pp. 53–59.

Kawthar A, Zhang Y, Niyonsaba E, et al. (2021), "Determination of the Chemical Compositions of Condensate-like Oils with Di ff erent API Gravities by Using the Distillation , Precipitation , Fractionation Mass Spectrometry (DPF MS) Method", *energy & fuel*, Vol. 35, pp. 8646–8656.

Lozano PDC, Orrego-Ruiz JA, Cabanzo Hernández R, et al. (2017), "APPI(+)-FTICR mass spectrometry coupled to partial least squares with genetic algorithm variable selection for prediction of API gravity and CCR of crude oil and vacuum residues", *Fuel*, Vol. 193, pp. 39–44.

Pantoja PA, López-Gejo J, Le Roux GAC, et al. (2011), "Prediction of crude oil properties and chemical composition by means of steady-state and time-resolved fluorescence", *Energy and Fuels*, Vol. 25, No. 8, pp. 3598–3604.

Stasiuk LD and Snowdon LR (1997), "Fluorescence micro-spectrometry of synthetic and natural hydrocarbon fluid inclusions: Crude oil chemistry, density and application to petroleum migration", *Applied Geochemistry*, Vol. 12, No. 3, pp. 229–241.

Authors

Engr Solomon Adjei Marfo is a Senior Lecturer at the Chemical and Petrochemical Engineering Department of University of Mines and Technology, Tarkwa, Ghana. He holds PhD in Petroleum Engineering from the University of Port Harcourt, Nigeria, MEng Degree in Mining (Petroleum Engineering) from the University of Belgrade, Serbia and BSc in Chemical Engineering from the Kwame Nkrumah University of Science and Technology, KNUST, Kumasi, Ghana. He is a member of the Ghana Institution of Engineering and Technology (PE.IET GH), a member of the Society of Petroleum Engineers (SPE), a Registered Environmental Specialist (RES) with the National Registry of Environmental Professionals (NREP) of USA. His research interests include Fuel an Lubricant Quality Analysis, Crude Oil Analysis, Application of Artificial Intelligence in Petroleum, Petrochemical, Refining Engineering.

Cornelius Borecho Bavoh is a Lecturer at the Chemical and Petrochemical Engineering at the University of Mines and Technology, Tarkwa, Ghana. He holds the degrees of MSc and PhD Universiti Tecknologi PETRONAS, Malaysia. He is a member of the Society of Petroleum Engineers and the Malaysia Board of Technologist. His research and consultancy work cover projects related to Drilling fliuds, Gas hydrate, and CO_2 capture and separation, Fuel characterisation, etc.