ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.434

Correlation Model Development for Saybolt Colour of Condensates and Light Crude Oils

Cheng Seong Khor[1*], Nur Nelly Sofia Nurazrin[1], Fatimah Mohd Hanafi[1], Fatin Nadhirah Asallehan[1], Nur Zawani Rosman[1], Jia Jia Leam[1], Sarat C. Dass[1,3], Shahrul Azman Zainal Abidin[4] and Farah Syamim Anuar[4]

[1]*Chemical Engineering, Universiti Teknologi PETRONAS (UTP),*
Bandar Seri Iskandar, Perak, Malaysia

[2]*Centre for Process Systems Engineering, Universiti Teknologi PETRONAS (UTP),*
Bandar Seri Iskandar, Perak, Malaysia

[3]*School of Mathematical and Computer Sciences, Heriot-Watt University Malaysia Campus,*
Putrajaya, Malaysia

[4]*PETRONAS Group Technical Solutions, Process Simulation and Optimization Group,*
Kuala Lumpur, Malaysia

Saybolt colour or number is a measured physical property of petroleum condensates and light crude oils which can be used as a quality indicator. As an alternative approach to the laboratory-based colour measurement method, this work aims to determine the influential physical properties in predicting Saybolt colour by applying a regression modelling approach. Data available on Saybolt colour and several physical properties are obtained from assay reports for condensates and light crude oils of Malaysian oil and gas fields. Other unavailable but potentially influential properties are estimated using a commercial process simulation software, iCON. The properties identified as explanatory variables in this study are refractive index, kinematic viscosity at 40°C, and characterization factor. This machine learning problem gives rise to applying multiple linear regression techniques based on a backward elimination approach in developing a correlation to predict Saybolt colour using the identified key properties of characterization factor, kinematic viscosity at 40°C, and refractive index.

Keywords: saybolt number; condensates; modelling; linear regression; liquid refractive index; kinematic viscosity

I. INTRODUCTION

Saybolt colour or number is a measurement scale used mainly for refined oils such as light petroleum products. It can serve as a quality indicator, e.g., of contaminants presence (Andrews *et al.,* 2001), which can influence feedstock selection decision for refinery processing to ensure product specifications are met. Colour can be a physical property used in this way, especially if it is readily observable. In the petroleum industry, colour is measured depending on requirements according to available methods or scales such as ASTM, Hazen, Rosin (or Gardner), and Lovibond (Speight, 2001).

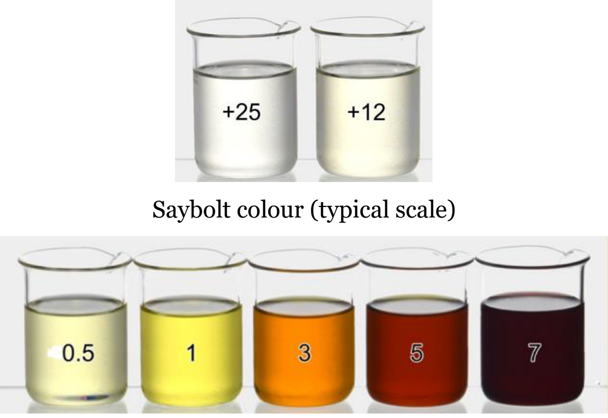
The two scales used to define petroleum colour particularly for products (with their respective standard test methods) are: (1) ASTM colour scale using the ASTM D 1500 method (ASTM International 2008), and (2) Saybolt colour scale using the ASTM D 156 Saybolt chromometer method (ASTM International 2003). As shown in Figure 1, the ASTM colour scale is used to quantify a broad range of petroleum products defined by numbers ranging from 0.5 (lightest) to 8 (darkest). Lighter petroleum colour with less than 0.5 on the

*Corresponding author's e-mail: chengseong.khor@utp.edu.my; khorchengseong@gmail.com

ASTM scale (typically for refined products) is graded using the Saybolt scale ranging from 30 (lightest) to –16 (darkest). Condensates, which have high API gravity and low density (Schlumberger 2018), possesses lighter shade and thus registers positive Saybolt numbers closer to values around 30.

A difference between the ASTM and Saybolt scales lies in the opposing magnitude used to describe the shade of a sample, e.g., ASTM scale uses smaller values for lighter coloured materials while Saybolt uses larger values. The colour value is determined by matching the sample with a specific set of standard scale either by direct or indirect visualization. The latter (indirect visualization) is aided with an instrument in matching output with the selected scale, e.g., the height of a column of a sample using Saybolt chromometer is adjusted until the colour match with that of standard. However, a detailed comparison between these two scales is beyond the scope of this work.

Several physical properties have been reported in the literature to be influential to petroleum colour, as summarized in Table 1. However, there is still a lack of such a model describing a relation between petroleum colour and its physical properties. Historically, petroleum feedstock quality is evaluated by measurement of bulk physical properties as it is readily determined, quick, and economical. The typical physical properties measured are specific density (or API gravity), refractive index, and viscosity. Among the physical properties reported to affect petroleum colour are; refractive index, surface tension, and specific dispersion (Diller *et al.,* 1943; Lykken and Rae, 1949; Speight, 2001; Rodriguez *et al.,* 2017).



Saybolt colour (typical scale)

ASTM colour (illustrative typical scale)

Figure 1. Colour spectrum (partial) for ASTM and Saybolt colour scales (Kemtrak 2019)

Table 1. Summary of physical properties related to petroleum colour reported in past work

Work	Physical Property	Remark
Diller *et al.,* (1943)	Refractive index, surface tension, specific dispersion	Saybolt colour is linearly related to the refractive index.
Lykken & Rae (1949)	Optical density	Refractive index is used as an indicator of optical density.
Speight (2001)	Composition, acid or basic nature	Total acid number is typically available from assay reports.
Rodriguez *et al.* (2017)	Composition of dodecane (C9 paraffins)	Dodecane composition is mostly available from assay reports (through detailed hydrocarbon analysis).

Analysis of colour by visual inspection is susceptible to low accuracy as variation, and thus inconsistency may arise across multiple observers as colour can be influenced by individual perspective. Several strategies have been developed to overcome the subjectivity of relying on the human eye in improving colour determination accuracy, mainly by modifying existing instruments without developing new colour measurement method (Rodriguez *et al.,* 2017).

As part of the recent digitalization trend as established through the smart manufacturing initiative or Industry 4.0 drive (Saudagar *et al.,* 2019), there is interest to automate colour determination instead of relying on experimental- or instrumentation-based technique, which is subject to cost, time, and accuracy issues. In this regard, developing a mathematical correlation to estimate petroleum colour based on physical properties data typically reported in assay reports (e.g., density and kinematic viscosity) offers promise. To the best of our knowledge, there is no correlation developed for the automated colour determination of petroleum feedstock or products towards assessing their quality. Methods involving eye visualization or laboratory analysis remain largely used in the industry.

This work focuses on the Saybolt colour scale in correlating it with physical properties identified as potentially influential variables, particularly in the case of light crude oils and condensates. We conduct statistical analysis coupled with using a commercial process simulator towards formulating suitable linear regression models to describe the relationship between the physical properties as factors and Saybolt colour as the response. The work contributes by postulating such correlation as an automated way to determine Saybolt colour that can be an alternative to conventional laboratory analysis, which may be expensive, slow, and inaccurate.

II. PROBLEM STATEMENT

Several challenges in providing fast and cost-effective yet reliably accurate petroleum colour measurement motivates us to develop an alternative method. First, colour is currently determined using laboratory analysis by direct or indirect visual inspection. The method incurs time and cost besides possibly generating inaccurate results due to the reasons. Second, a reliable colour estimation can assist refinery operators in assessing feedstock quality in avoiding possible operational problems and inability to meet processing targets such as product specifications. Third, the colour measurement must be done timely because petroleum colour can age over time and become unstable. Petroleum colour tends to darken due to the oxidization of unstable components present such as olefins. Consequently, colour measurement at a certain lapsed time might not reflect the actual quality (Speight, 2001; Rodriguez *et al.,* 2017). Thus, an alternative approach of developing correlation-based models empirically provides an alternative in determining petroleum colour instantly, overcomes the subjectivity of relying on human eye besides obviating time-consuming and cost-incurring laboratory analysis.

III. REGRESSION MODELING FRAMEWORK

This work adopts a semi-empirical modelling framework which integrates data-driven correlation development with physical modelling based on thermodynamic property estimation. As shown in the procedural flowchart in Figure 2 starting with data extraction, we proceed to conduct regression modelling at two stages: first is to predict unknown properties using flowsheet simulation tools, then to relate the identified potentially influential properties to Saybolt colour.

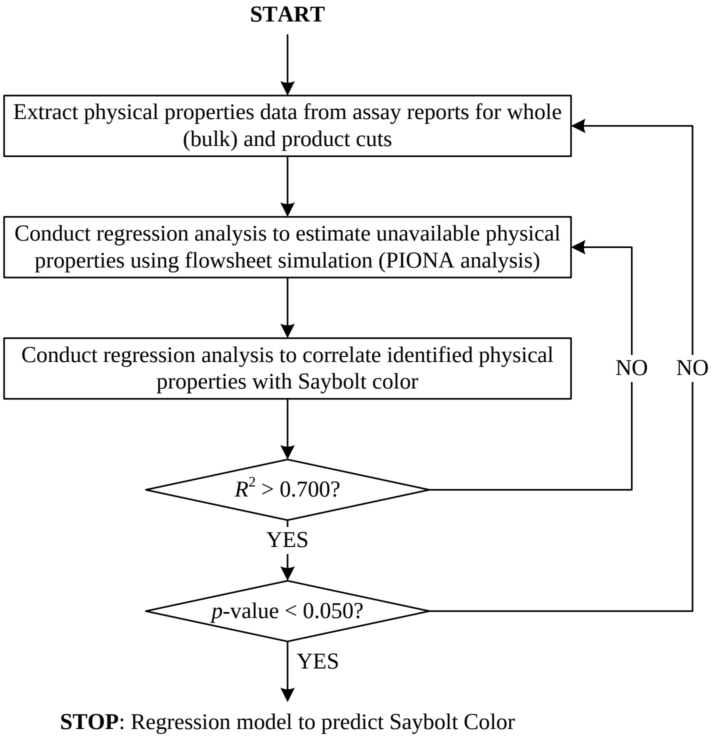

START
Extract physical properties data from assay reports for whole (bulk) and product cuts
Conduct regression analysis to estimate unavailable physical properties using flowsheet simulation (PIONA analysis)
Conduct regression analysis to correlate identified physical properties with Saybolt color
NO
NO
R^2 > 0.700?
YES
p-value < 0.050?
YES
STOP: Regression model to predict Saybolt Color

Figure 2. Regression modelling framework adopted in this work

A. Data Extraction

Physical properties of condensates and light crude oils are extracted from assay reports made available by the industrial collaborator on this work. The data are systematically recorded; the properties include density, boiling point, vapour pressure, kinematic viscosity (at temperatures –20°C, 20°C, 40°C, 50°C, 70°C, 75°C, and 100°C), and characterization factor (also called UOP or Watson K factor) (Gary *et al.,* 2007), which are identified as potentially influential factors to determine Saybolt colour as based on preliminary screening with insight from subject matter experts. Note that we consider properties data for the whole petroleum (i.e., bulk properties) as well as their product fractions or cuts at various boiling point ranges (such as C5–70°C, 70–90°C, 90–140°C, 140–155°C until 450–520°C). Part of the data is used as inputs to a flowsheet simulation package to estimate other physical properties not provided in the assay reports, as explained next.

B. Unknown Property Estimation

To estimate certain physical properties such as liquid refractive index and Reid vapour pressure which are not available from the assay reports but deemed potentially influential for Saybolt colour prediction, we use a proprietary in-house process simulation software of PETRONAS called iCON (Virtual Materials Group, 2017). Within iCON, a predictive tool (called PIONA Slate) is invoked to postulate a compositional makeup with hydrocarbon pseudo-component characterization comprising molecular structural groups according to the PIONA (n-paraffin, isoparaffins, olefins, naphthenes, and aromatics classification. Each PIONA group for a certain number of carbon atoms with different boiling points may exhibit distinct thermodynamic properties which govern the physical properties determination. A parameter estimation procedure (called Oil Source) is then used to estimate the composition of the component slate which optimally matches the assay data available on distillation (mainly true boiling point) and physical properties. Finally, we determine the desired physical property value by executing a black-box modelling tool (called Special Property).

C. Pairwise Variable Analysis

Before performing regression analysis, we develop pairwise scatterplots between all the variables in the dataset, i.e., comprising Saybolt colour and the three physical properties considered, namely liquid refractive index (RI), kinematic viscosity at 40°C (KV40), and characterization factor (KF). Each of the individual plots, as shown in Figure 3, depicts the relation (or non-relation) between the row variable and the column variable. For example, the individual plot on the first column of the second row represents a possible relation between Saybolt colour and characterization factor. The scatterplots show that there is no linear relation observed between Saybolt colour and any of the three factors nor is there linear relation between any of the three factors. Therefore, in our further analysis, we have considered higher-order terms of each variable and their possible interactions. Linear regression is used to assess the importance of these higher-order terms. This methodology is explained in detail in Section 4.

D. Regression Analysis

Regression analysis is used to determine the relationship between Saybolt colour and the physical properties. The independent variable or model response is Saybolt colour (or number) of a condensate type based on its certain physical properties as dependent variables or model predictors. We consider simple and multiple linear regression models to describe the correlation between Saybolt colour and the potentially influencing physical properties. The former (simple linear regression model) admits only one predictor property while the latter can handle more than one predictor property in which two and three such influential variables are considered appropriate and thus investigated for this problem using a backward elimination approach (Montgomery *et al.,* 2012).

The regression analysis is performed with the aid of an Excel spreadsheet (installed with an add-in called Data Analysis that features a regression tool); any similar software can also be used for this purpose. To verify the results obtained, we use a specific statistical package in the form of the open-source package of R (version 3.3.1) (R Core Team 2016). The value of the coefficient of determination R^2 allows us to determine whether the predictor and regressor variables have a strong or weak relation.

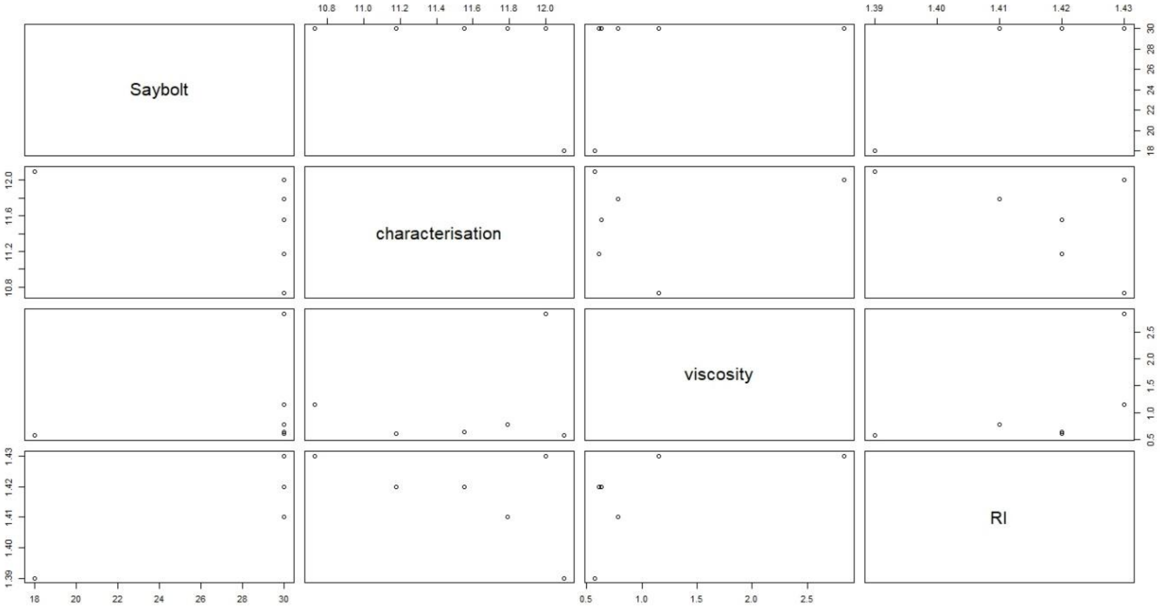


Figure 3. Scatter plots of Saybolt colour versus the three factors considered: characterization factor (KF), kinematic viscosity at 40°C (KV40), and liquid refractive index (RI).

The value is indicative particularly for a single regressor variable in developing a simple linear regression model. A value of R^2 closer to one indicates a higher percentage of data that fits a proposed regression model. A p-value of less than 0.05 (for a 5% level of significance) indicates that there exists a statistically-significant linear relation between the response and the factors considered (Lind *et al.,* 2001).

It is noteworthy that the value of R^2 tends to increase as more factors are added to the regression model. Thus, R^2 does not penalize for model complexity—a few important features, together with many spurious factors result in a higher R^2 value. In other words, R^2 does not discount spurious factors that do not affect the regression in a meaningful way. To address this drawback of R^2, we use the adjusted R^2 measure, which penalizes factors that are added but unimportant in the regression model (Draper and Smith, 1998).

IV. REGRESSION MODEL DEVELOPMENT

A. Raw Data

This work considers data on physical properties of both bulk and product cuts (or fractions) of condensate and light crude oil types from Malaysian oil and gas fields mainly in offshore Sabah and Sarawak (e.g., Kimanis, Marjoram, Bintulu, and Kasawari). The underlying assumption is that physical property values are independent and identically distributed (i.i.d.). Due to commercial confidentiality, the data is not reported here.

B. Multiple Linear Regression Model

We use a multiple linear regression model as a basis for the model building in which the generalized formulation is given by:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_i x_{ij} + \cdots + \varepsilon_j \qquad (1)$$

where for data point j, x_{ij} is the value of the independent variable i (i.e., predictor or regressor), y_j is the observed value of the dependent variable, β_0 is intercept, β_i is slope coefficient associated with independent variable i, and ε_j is random error. As summarized in Table 2, we develop four-second order variants of a multiple linear regression model, with and without considering interaction involving the factors for two or three of such independent variables to predict Saybolt colour. An interaction term is represented by the multiplication of a pair of variables such as $\mathrm{RI} \cdot \mathrm{KV40}$ and $\mathrm{RI} \cdot \mathrm{KF}$.

The best result obtained among the models considered is that for a second-order two-variable without interaction model; the other models tend to lead to a problem of parameters overfitting. For the former, the adjusted R^2 value is 0.9903 with an F-statistic of 590.7 corresponding to a p-value of 2.2×10^{-16} at a 95% confidence level, which indicates that 99.03% of the total sum of squares of the Saybolt colour response can be accounted for by the following correlation:

$$SN = -2.234\times10^4 + (3.120\times10^4)RI + 13.71KV40 - 1.088\times10^4 RI^2 - (3.098\times10^4)KV40^2 \quad (2)$$

Table 2. Summary of regression models developed for identified physical properties.

Model Type	**Regressor**	**Adjusted R^2**	**p-value**
Second-order two-variable without interaction	$\mathrm{RI}, \mathrm{KV40}, \mathrm{RI}^2, \mathrm{KV40}^2$	0.9903	2.2×10^{-16}
Second-order two-variable with two-way interaction	$\mathrm{RI}, \mathrm{KV40}, \mathrm{RI}^2, \mathrm{KV40}^2, \mathrm{RI}\cdot\mathrm{KV40}$	1 (overfitting)	0.0
Second-order three-variable without interaction	$\mathrm{RI}, \mathrm{KV40}, \mathrm{KF}, \mathrm{RI}^2, \mathrm{KV40}^2, \mathrm{KF}^2$	1 (overfitting)	0.0
Second-order three-variable with two-way interaction	$\mathrm{RI}, \mathrm{KV40}, \mathrm{KF}, \mathrm{RI}^2, \mathrm{KV40}^2, \mathrm{KF}^2, \mathrm{RI}\cdot\mathrm{KV40}, \mathrm{RI}\cdot\mathrm{KF}, \mathrm{KV40}\cdot\mathrm{KF}$	1 (overfitting)	0.0

V. CONCLUDING REMARKS

The results obtained from multiple linear regression modelling indicates that a statistically significant relationship exists between Saybolt colour and condensate physical properties comprising liquid refractive index and kinematic viscosity at 40°C. Future work involves exploring other potentially influential physical properties (e.g., contaminants content such as sulphur) (Speight, 2015) and systematic regression modelling strategies (e.g., such as stepwise regression) (Harrell, 2001).

VI. ACKNOWLEDGEMENT

We are grateful for the financial support provided from YUTP grant no. 015LC0-166 to carry out this study.

VII. REFERENCES

Andrews, RJ, Beck, G, Castelijns, K, Chen, A, Cribbs, ME, Fadnes, FH, Irvine-Fortescue, J, Williams, S, Hashem, M, Jamaluddin, A, Korkjian, A, Sass, B, Mullins, OC, Rylander, E, Dusen, AV 2001, 'Quantifying contamination using color of crude and condensate', *Oilfield Review*, vol. 13, pp. 24-43.

ASTM International 2003, *Standard Test Method for Saybolt Color of Petroleum Products (Saybolt Chromometer Method)*, ASTM International, United States.

ASTM International 2008, *Standard Test Method for ASTM Color of Petroleum Products (ASTM Color Scale)*, ASTM International, United States.

Diller, IM, Dean, JC, DeGray, RJ, Wilson, JW 1943, Color index: Light color petroleum products, in *Industrial and Engineering Chemistry*, Brooklyn, NY: Socony-Vacuum Oil Co., Inc.

Draper, NR, Smith, H 1998, *Applied Regression Analysis*. Third ed. Wiley-Interscience, New York, Original edition, pp. 246.

Gary, JH, Handwerk, GE, Kaiser, MJ 2007, *Petroleum Refining: Technology and Economics*. Fifth ed. New York: Marcel Dekker.

Harrell, J, Frank E. 2001, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Kemtrak 2019, ASTM D 1500 Color Scale, viewed July 11 2019, <https://www.kemtrak.com/application/astm-d-1500-color-scale/>.

Lind, DA, Marchal, WG, Mason, RD 2001, *Statistical Techniques in Business and Economics*. New York: McGraw-Hill516.

Lykken, R, Rae, J 1949, 'Determination of relative color density of liquids: A rapid photoelectric method', *Analytical Chemistry*, vol. 21, no. 7, pp. 787-793.

Montgomery, DC, Peck, EA, Vining, GG 2012, *Introduction to Linear Regression Analysis*. Hoboken, New Jersey: Wiley.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rodriguez, JD, Comstock, M, Auz, B, Olmstead, T. 2017, A spectroscopic method of determining color of petroleum products using CIELab color space with LED illumination, in *Photonic Instrumentation Engineering IV*, edited by YG Soskind and C Olson.

Saudagar, M, Ye, M, Al-Otaibi, S, Al-Jarba, K 2019, Smart Manufacturing: Hope or Hype? *Chemical Engineering Progress* (June).

Schlumberger 2019 Condensate. Schlumberger 2018, viewed 17 January 2019, <https://www.glossary.oilfield.slb.com/Terms/c/condensate.aspx>.

Speight, JG 2001, *Handbook of Petroleum Analysis*. New York: Wiley.

Speight, JG 2015, *Handbook of Petroleum Product Analysis*: New York: Wiley.

Virtual Materials Group 2017, Oil Source, iCON Version 10.0 user manual.