Journal of Petroleum Exploration and Production Technology
https://doi.org/10.1007/s13202-020-01031-y

ORIGINAL PAPER-PRODUCTION ENGINEERING

Saybolt color prediction for condensates and light crude oils

Jia Jia Leam[1] · Cheng Seong Khor[1,2] · Sarat C. Dass[3]

Received: 11 June 2020 / Accepted: 20 October 2020

Abstract

Saybolt color determination is one of the techniques used to evaluate the quality of petroleum products as an indicator of the degree of refinement. As color is a property readily observed by operators, conventional procedures require operators to determine Saybolt color either through direct visual observation or through Saybolt chromometers. These methods are subjective due to the variability in perception of colors across different observers and may be influenced by external factors such as the level of illuminance. Digital oil color analyzers, on the other hand, cost almost four times as much as Saybolt chromometers. An alternative approach to color measurement is to develop a correlation model between Saybolt color with the physical and chemical properties of condensates and light crude oils from Malaysian oil and gas fields. This work applies several multiple linear regression techniques (such as stepwise regression) performed both manually and using the *R* software (version 3.6.1) to obtain statistically significant results. The step, regsubsets and glmulti functions from *R* are explored to develop the correlation model which predicts Saybolt color using only identified key properties, overcoming the possible drawbacks associated with conventional laboratory analysis. The models developed through these different techniques are analyzed and compared based on criteria indicated through the coefficient of multiple determination, R^2 and *F*-tests to infer on suitable regression approaches. Results obtained from these regression methods for models with and without interaction terms report deviations of less than 5% for 75% of the samples used for validation.

Keywords Forward selection · Backward elimination · Bidirectional elimination · *R* · Machine learning · Glmulti

Introduction

Color observations of petroleum products are standardized through two international standards developed by the American Society for Testing and Materials (ATSM), namely ASTM D 156 and ASTM D 1500. The two standards cover different ranges of color. Highly refined petroleum products use the ASTM D 156 scale, also known as the Saybolt color scale which ranges from −16 (darkest) to +30 (lightest) (ASTM International 2003). For colors darker than −16 of the ASTM D 156 scale, the ASTM D 1500 scale is used, ranging from 0.5 (lightest) to 8 (darkest) (ASTM International 2008). Petroleum products for which colors fall outside of the established range are deemed contaminated. Conventional ways of measuring color are through direct and indirect visual observation. Direct observation involves comparing the color of oil samples directly with color standards, whereas indirect observation utilizes chromometers (Khor et al. 2020).

The measurement of Saybolt color using a Saybolt chromometer is carried out in the presence of a constant light source. The oil level in the sample tube is adjusted in a way so that the short-wavelength (violet) portion of the light energy reaching the eye is equal to that passing through the standard disk and the empty tube. Since surface tension, refractive index and specific dispersion of oil determine the angle at which light hits the wall from the oil surface, these attributes directly affect Saybolt color (Diller et al. 1943).

The American Petroleum Institute (API) gravity expresses the density of petroleum liquids in comparison to water where high API gravity represents low density. While condensates and light oils have low viscosity and high API

✉ Cheng Seong Khor
chengseong.khor@utp.edu.my; khorchengseong@gmail.com

[1] Chemical Engineering Department, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak Darul Ridzuan, Malaysia

[2] Centre for Systems Engineering, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak Darul Ridzuan, Malaysia

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

Published online: 12 November 2020

gravity, degraded oils are heavier and more viscous. As crude oils get heavier, API gravity decreases, the absorption spectra moves to the red region, and fluorescence emissions become weaker (Hagemann and Hollerbach 1986). Furthermore, different types of hydrocarbons behave differently: Aromatics absorb visible or near-infrared light, while aliphatic compounds are only excited by high ultraviolet light. Hence, light hydrocarbons are colorless as they do not absorb light in the visible spectrum. Heavier or degraded crude oils with high concentrations of complex aromatic molecules are distinctively darker since they absorb light effectively in the visible light region (Steffens et al. 2011).

Regression modeling has been applied in the petroleum industry to develop correlations and pose models to predict physical properties; see, for example, Tomren and Barth (2014) and Douglas et al. (2018). The former work (Tomren and Barth 2014) involves formulating partial least squares calibration models to estimate properties such as viscosity, acid number and asphaltene content of crude oils and condensates based on information from gas chromatography (GC) and infrared spectroscopy (IR). However, the applicability range of the models might be limited and not readily extended to a wide range of petroleum sources. The latter work (Douglas et al. 2018) aims to predict hydrocarbon concentrations in contaminated soil in which different regression techniques are compared. Due to nonlinearity of soil spectral responses, higher prediction accuracy is observed using the random forest machine learning technique compared to partial least squares regression.

A main contribution of this work is to develop a Saybolt color correlation model for devising a fast and potentially cost-efficient method of estimating the color compared to laboratory-based measurements of the same. To the best of our knowledge, there is no correlation model developed for an automated determination of Saybolt color since the practice remains dependent on laboratory analysis. Arguably, the novelty of the paper lies in its attempt to develop such an empirical correlation model for Saybolt color to support measurement of this physical property as a standard quality indicator in the oil and gas industry. This color property has become more prominent in recent years due to increased interest in petroleum condensates as refinery feedstock resulting from shale gas extraction activities (IHS Markit 2018). With previous studies reporting the correlation between color with petroleum product properties, this work aims to demonstrate that regression modeling and analysis can be used to develop such a correlation model for predictive purpose.

Problem statement

It is reported that direct visual observations used for color determination of petroleum are highly subjective due to the variability in color perception across different observers (Rodriguez et al. 2017). On the other hand, measurements using Saybolt chromometers are affected by environmental factors: Varying illuminance levels can be obtained from different light sources such as fluorescent lamps and halogen lamps. Moreover, compounds such as olefins in crude oils and condensates are prone to oxidation, thus resulting in darkening and aging of samples which affect Saybolt color analysis (Rodriguez et al. 2017; Speight 2015). Digital oil color analyzers, on the other hand, cost almost four times as much as Saybolt chromometers (Clarkson Laboratory and Supply Inc 2019; IndiaMart 2019). Hence, an alternative approach is to rely on mathematical models for determining Saybolt color.

Best subset regression methodology

The aim of this paper is to develop best regression models for Saybolt color based on four properties of oil samples, namely (1) refractive index (R), (2) density (D), (3) kinematic viscosity at 75 °C (V_1) and (4) kinematic viscosity at 100 °C (V_2). To achieve this, we utilize the methodology of best subset regression where multiple statistical hypotheses tests are performed either to add or to remove regressor terms from a full model. The full model consists of all possible regressor terms constructed from considering all possible powers and interactions (i.e., main- and higher-order interactions) among the original four attributes $\{R, D, V_1, V_2\}$. The general mathematical formulation of a full model based on a response variable y and $m=4$ explanatory variables, x_1, x_2, $\cdots$, x_m, is given by:

$$y = \beta_0 + \sum_{p=1}^{M}\sum_{i=1}^{m}\beta_{ip}x_i^p + \sum_{p,q=1,p\neq q}^{M}\sum_{i,j=1,i<j}^{m}\beta_{ijpq}x_i^p x_j^q + \varepsilon, \quad (1)$$

where y is the observed Saybolt color; x_i^p is the regressor i ($i \in \{R, D, V_1, V_2\}$) raised to the power p with $1 \le p \le M$ where M is the highest power considered; x_j^q is similar to x_i^p with $j \in \{R, D, V_1, V_2\}$ and q such that $1 \le q \le M$; β_0 is the constant intercept term; β_{ip} and β_{ijpq} are, respectively, the regression coefficient corresponding to x_i^p and $x_i^p x_j^q$; and ε is the random error assumed to arise from a normal distribution with mean zero and constant, but unknown variance σ^2. Best subset regression analysis is performed using a dataset of oil samples of size n, $\left(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)}, y^{(k)}\right), k = 1, 2, \ldots, n$. We develop two main full models in this paper, both with $M=2$, without and with pairwise interactions between the explanatory variables.

To arrive at the best subset regression model, we implement the stepwise regression technique which adds or removes regressors from the current model one at a time (Montgomery et al. 2012) and tests for the significance of

the added/removed term. The technique can be classified into forward selection, backward elimination and bidirectional elimination methods (Rawlings et al. 2006). In forward selection, the initial model starts with zero regressors. Subsequently, regressor terms from the full model in Eq. (1) are fitted into the current model, and the regressor with the best correlation with Saybolt color is selected for inclusion in the current regression model. The forward selection procedural flowchart is shown in Fig. 1. Backward elimination works in the opposite direction where a regressor is removed from the full model (1) if the corresponding test of significance for this regressor falls below a pre-specified threshold as shown in Fig. 2. A combination of the forward and backward methods constitutes the bidirectional elimination method. To perform statistical tests on the added/removed terms, two different F-tests, namely the partial and overall F-tests, are used to evaluate the significance (Draper and Smith 1998). The flowcharts of procedures for both bidirectional elimination methods are presented in Figs. 3 and 4.

We perform the forward, backward and bidirectional elimination methods by manually creating columns in Excel that represent all the regressor terms in Eq. (1). The second approach that we implement is more automated and similar hypotheses tests are carried out via functions accessed

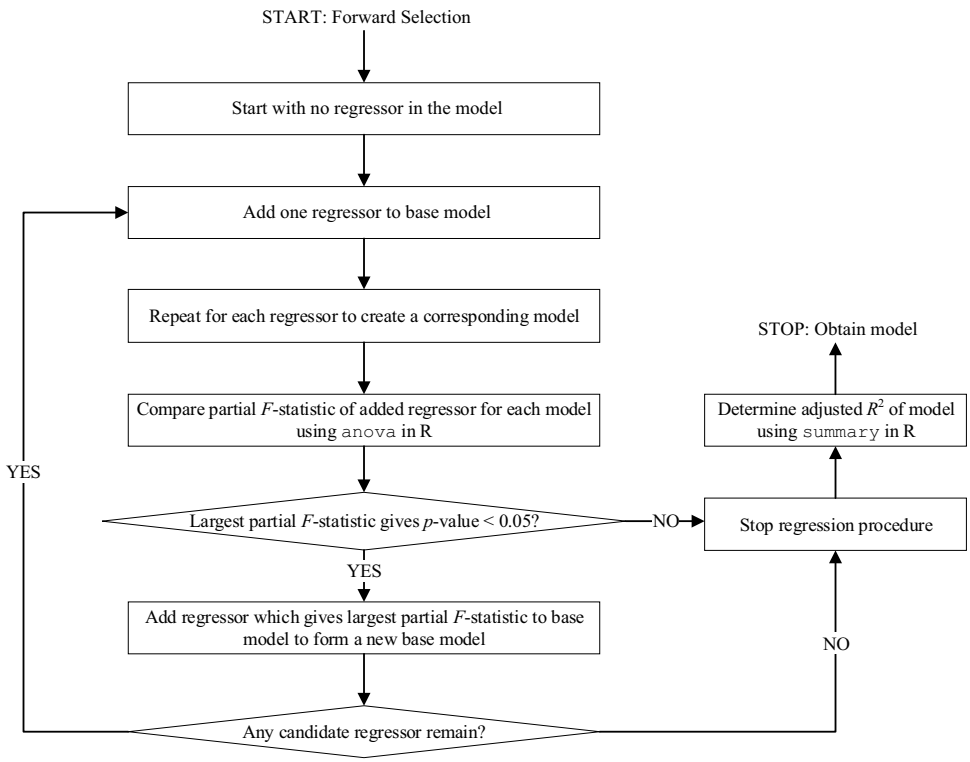

Fig. 1 Flowchart of forward selection procedure

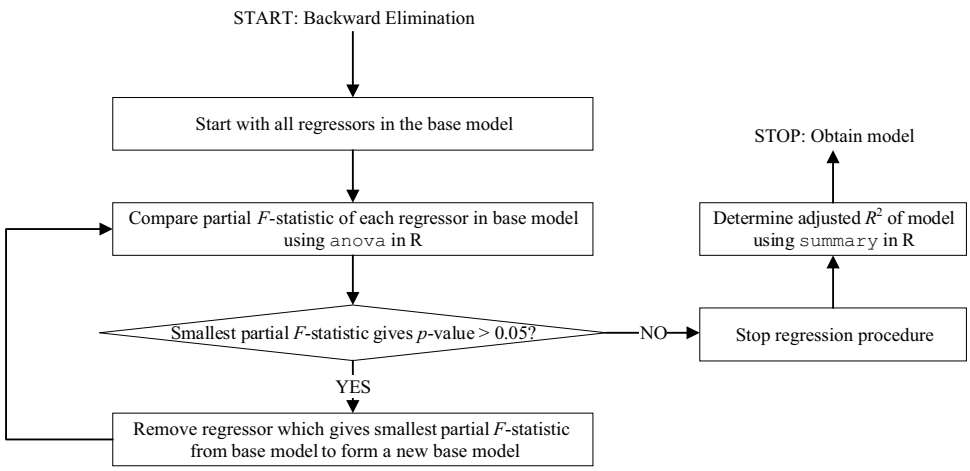

Fig. 2 Flowchart of backward elimination procedure

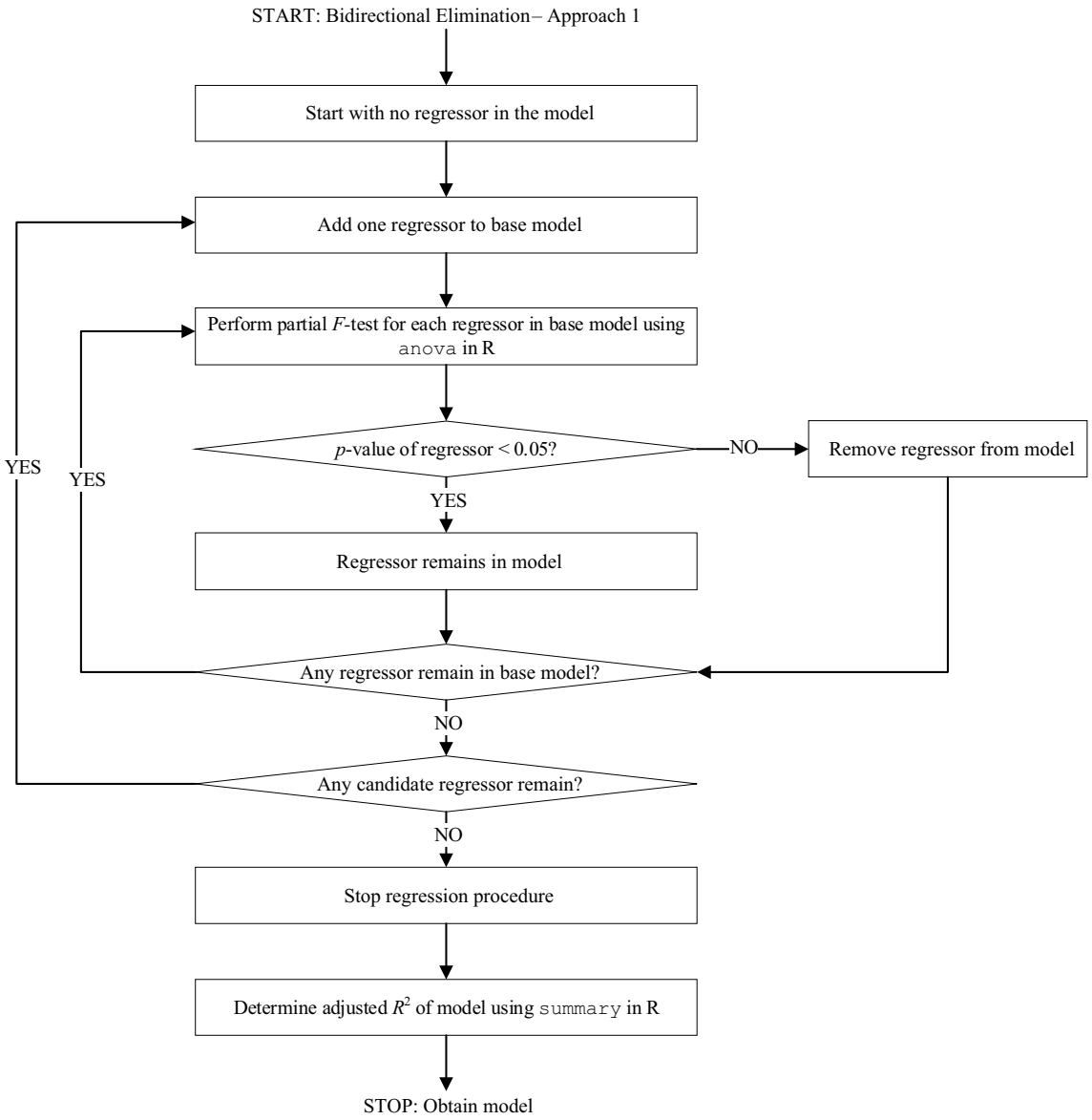

Fig. 3 Flowchart of bidirectional elimination procedure using partial *F*-test

through statistical packages in *R* (version 3.6.1) (R Core Team 2013). The *R* function step from the **stats** package uses the Akaike information criteria (AIC) to estimate the relative quality of model fit to the dataset for a given set of regressors. Procedures for forward selection and backward elimination using step are shown in Figs. 5 and 6, respectively. On the other hand, the regsubsets function from the **leaps** package evaluates all possible models for a given set of regressors and returns the model with the highest adjusted R^2 (Lumley 2014). Another set of functions (or classes) used in this work is from the **glmulti** package which automatically considers all possible generalized linear models arising from all possible subsets of a given collection of regressors from the full model. As an exhaustive screening tool, **glmulti** ranks the subset regression models according to a specific information criterion: It gives three choices, namely AIC, Bayesian information criterion (BIC) and corrected AIC (AICc). The first-ranked (i.e., highest ranked) model has the lowest value of such information criterion (Calcagno and de Mazancourt 2010).

Results and analysis

Regression modeling based on physical properties

The four properties of refractive index (*R*), density (*D*), kinematic viscosity at 75 °C (V_1) and kinematic viscosity at

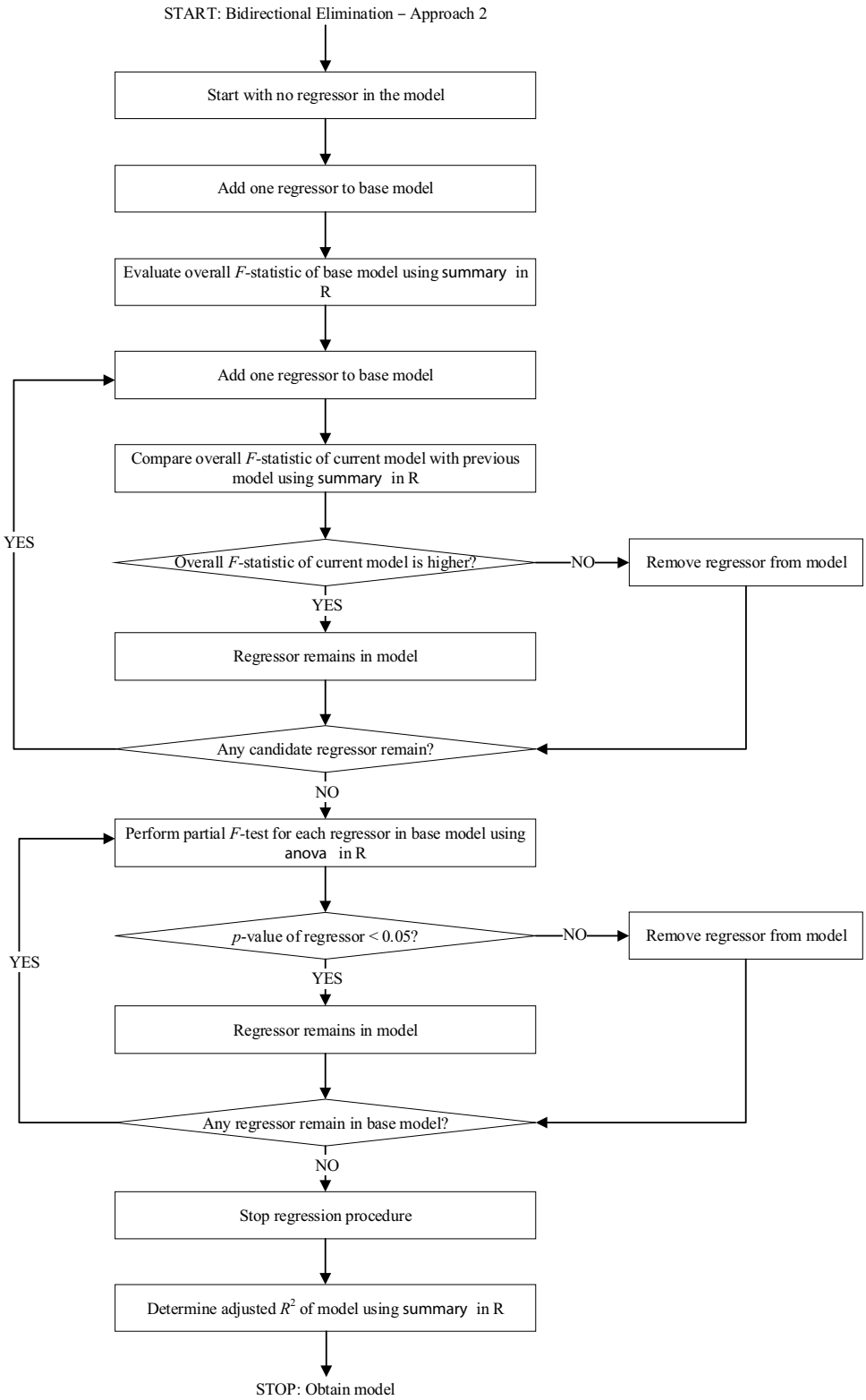

Fig. 4 Flowchart of bidirectional elimination procedure using overall *F*-test

100 °C (V_2), as well as Saybolt color measurements, are obtained from assay reports for the whole (i.e., bulk) and product fractions (i.e., cuts) of condensates and light crude oils from Malaysian oil and gas fields located mainly in offshore Sabah and Sarawak (e.g., of the types named Kimanis, Marjoram, Bintulu and Kawasari). The dataset consists of $n = 15$ samples. The scatterplots obtained for analyzing pairwise variable relationships, as shown in Fig. 7, indicate

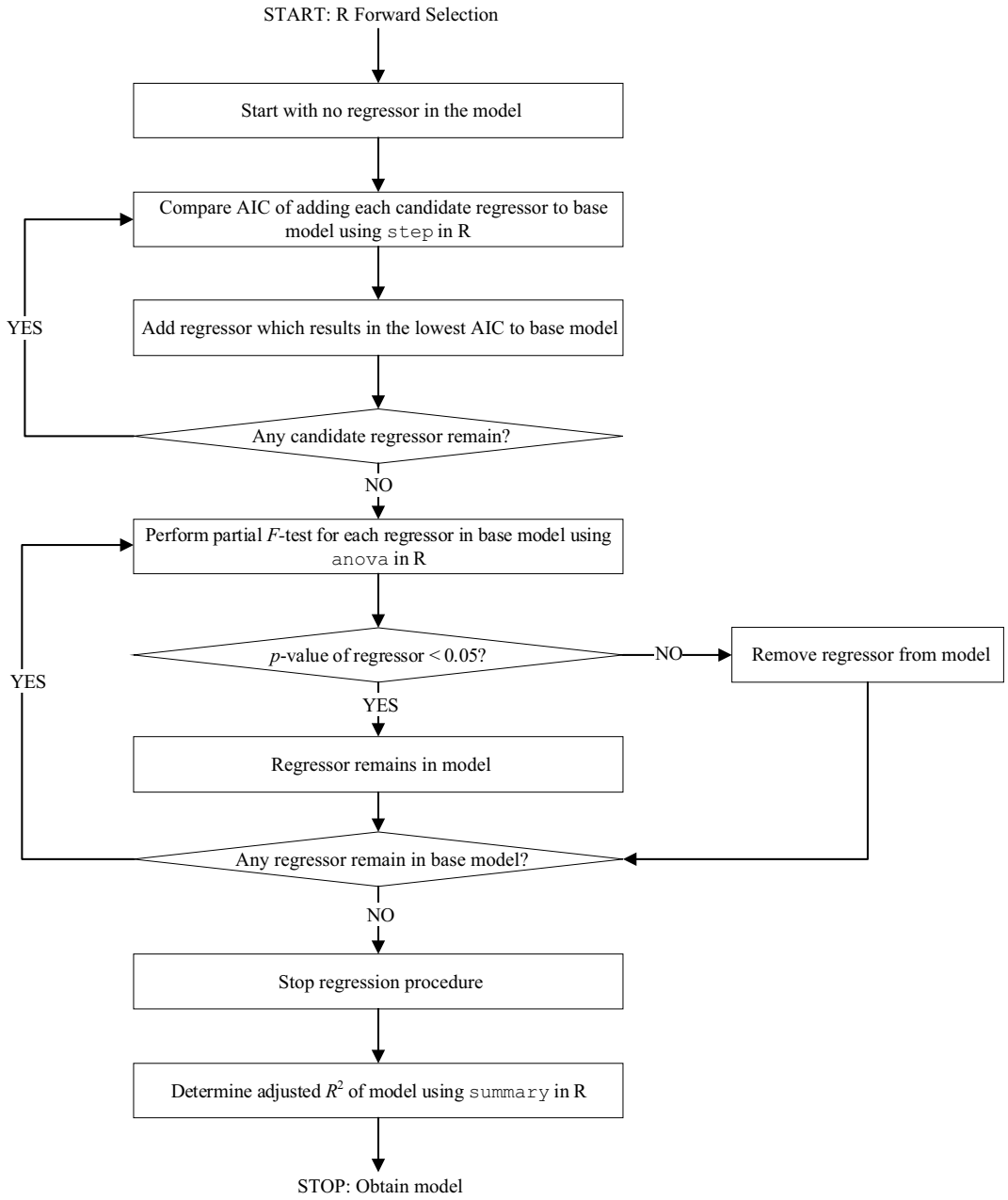

Fig. 5 Flowchart of forward selection procedure performed using step function in *R*

the absence of linear trends between Saybolt color and the four variables. These scatterplots suggest that there is a high degree of nonlinear relationships between Saybolt color and the potential regressors. Henceforth, we consider higher-order powers and interaction terms for the four variables. The catterplots of R *vs*. D and V_1 *vs*. V_2 show strong linear relationships (high collinearity). Thus, we expect the best subset regression model in our experiments to select either R or D, but not both, and similarly for V_1 and V_2.

The two full models considered with $M=2$ have the following explicit forms:

$$S = \beta_0 + \beta_1 R + \beta_2 D + \beta_3 V_1 + \beta_4 V_2 + \beta_5 R^2 + \beta_6 D^2 + \beta_7 V_1^2 + \beta_8 V_2^2 + \varepsilon \tag{2}$$

as the full model without interaction terms and

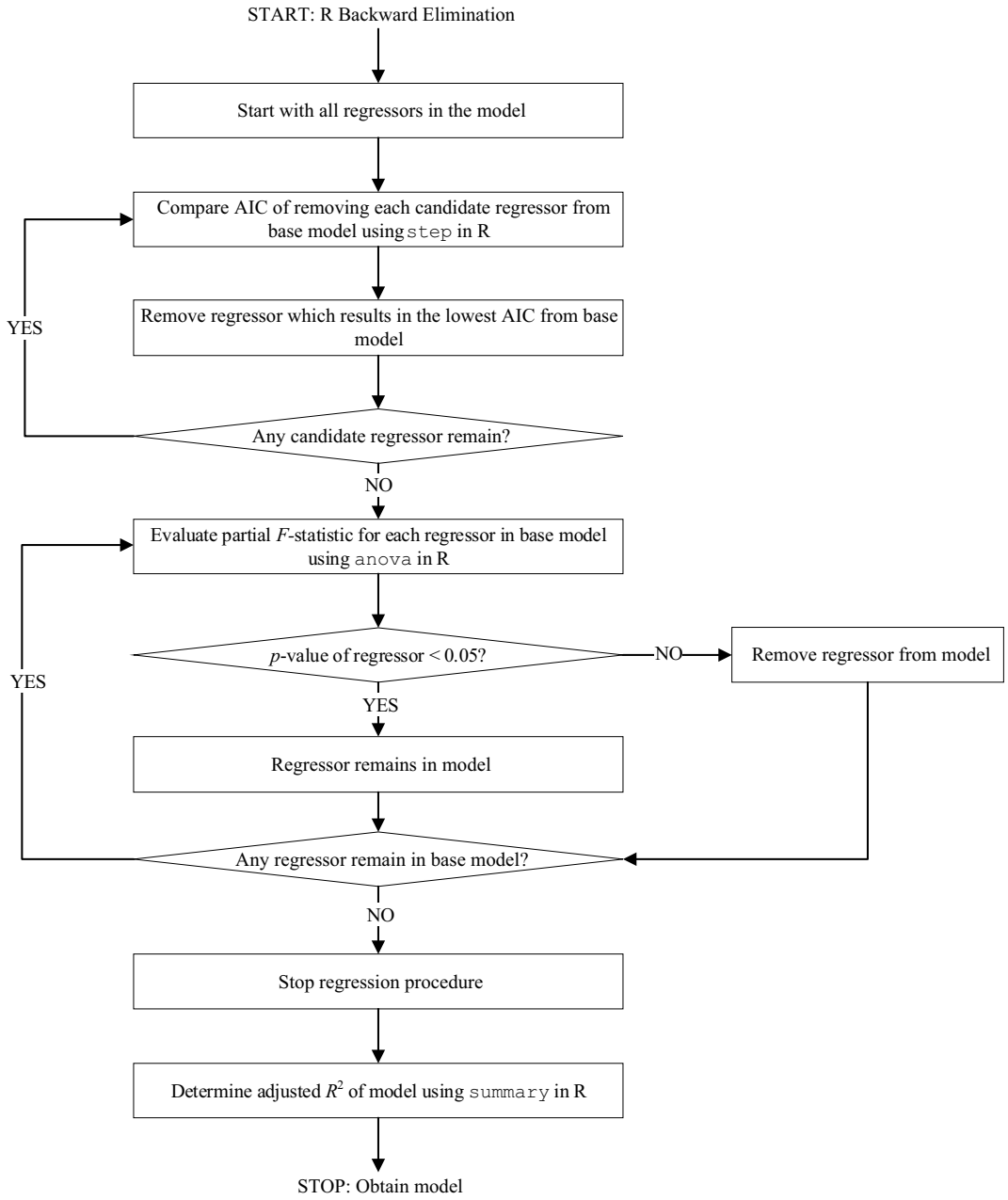

Fig. 6 Flowchart of backward elimination procedure performed using step function in *R*

$$S = \beta_0 + \beta_1 R + \beta_2 D + \beta_3 V_1 + \beta_4 V_2 + \beta_5 R^2 + \beta_6 D^2 + \beta_7 V_1^2 + \beta_8 V_2^2 + \beta_9 RD + \beta_{10} RV_1 + \beta_{11} RV_2 + \beta_{12} DV_1 + \beta_{13} DV_2 + \beta_{14} V_1 V_2 + \varepsilon \quad (3)$$

as the model with all pairwise interaction terms. Note that higher powers of explanatory variables are deliberately missing when considering the interaction terms in Eq. (3). This is because of the small sample size ($n = 15$) of the dataset used in this study. The full model (3) with 15 unknown parameters has zero residual degrees of freedom and hence cannot be tested for statistical significance. We consider the full model in (3) only as a strict upper bound for all pairwise interaction terms to be considered in the subset regression models. We find that the best regression model typically

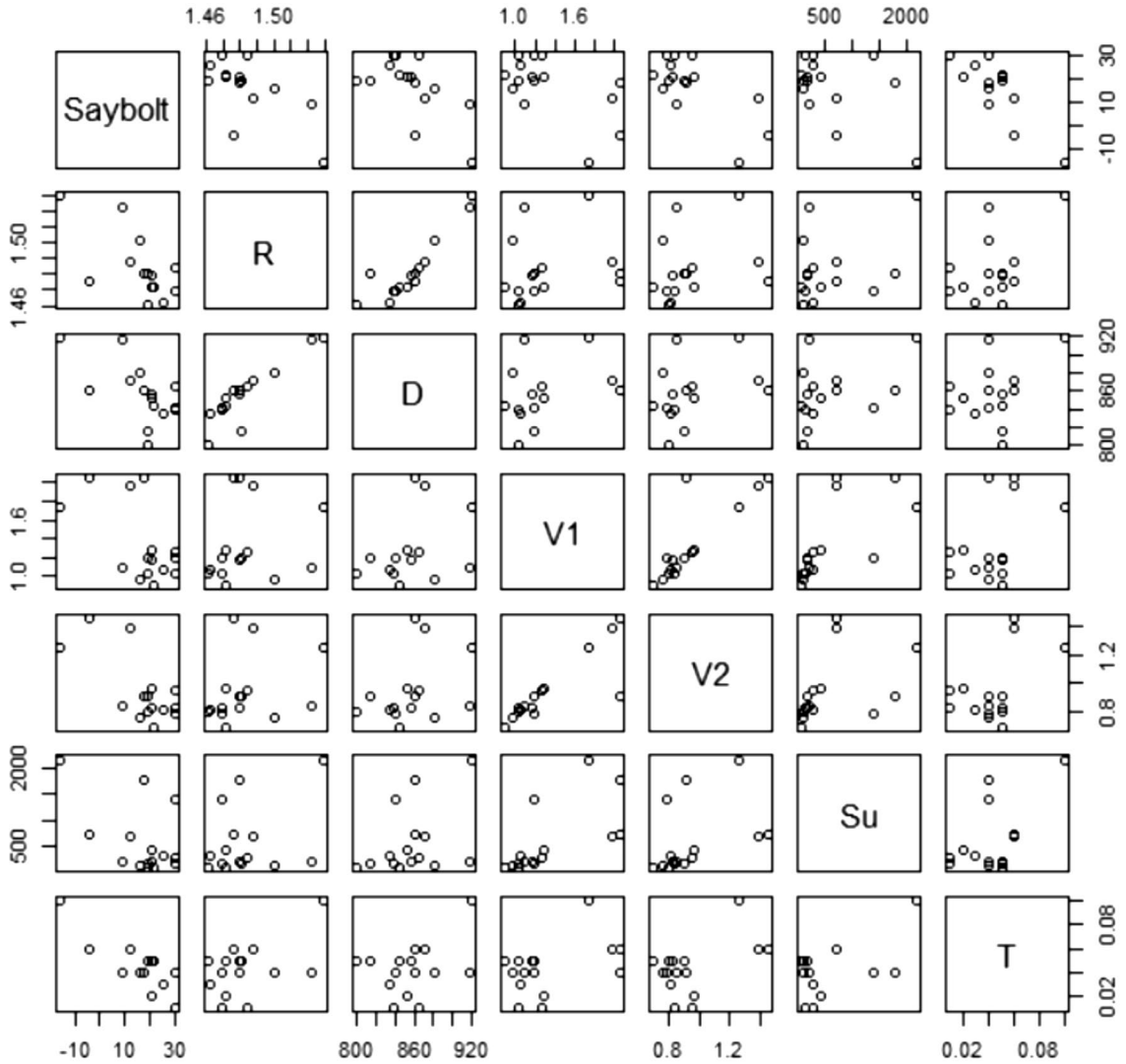

Fig. 7 Scatterplots of Saybolt color versus refractive index (*R*), density (*D*), kinematic viscosity at 75 °C (V_1), kinematic viscosity at 100 °C (V_2), sulfur content (*Su*) and total acid number (*T*)

Table 1 Evaluation of best models with and without pairwise interaction developed using stepwise regression

Regression technique	Regressor	Adjusted R^2	Overall *F*	*p* value
Second-order model without pairwise interaction				
Forward selection	R^2, V_2^2	0.7135	18.43	0.00022
Backward elimination	R, V_1	0.6050	11.72	0.00151
Bidirectional elimination based on partial *F*-test	R, V_1	0.6050	11.72	0.00151
Bidirectional elimination based on overall *F*-test	R, V_1	0.6050	11.72	0.00151
Second-order model with pairwise interaction				
Forward selection	DV_2, R^2	0.6923	16.75	0.00034
Backward elimination	–	–	–	–
Bidirectional elimination based on partial *F*-test	R, V_1	0.6050	11.72	0.00151
Bidirectional elimination based on overall *F*-test	R, V_1	0.6050	11.72	0.00151

Table 2 Evaluation of best models with and without pairwise interaction developed using *R* functions

R function	Regressor	Adjusted R^2	Overall F	p value
Second-order model without pairwise interaction				
Step (forward selection)	R^2, V_2^2	0.7135	18.43	0.00022
Step (backward elimination)	R, V_1^2	0.6139	12.13	0.00131
Regsubsets	R, V_1^2	0.6139	12.13	0.00131
Second-order model with pairwise interaction				
Step (forward selection)	DV_2, R^2	0.6923	16.75	0.00034
Step (backward elimination)	–	–	–	–
Regsubsets	R, V_1	0.6050	11.72	0.00151

includes only one or two of the pairwise interaction terms this model.

The stepwise regression results are summarized in Table 1 for both types of full models (2) and (3). Note that in the case of the latter model, a zero degree of freedom does not permit backward elimination to be applied (Hu 2016). For both types of full models (2) and (3), the overall F-statistic and adjusted R^2 values obtained using the forward selection procedure are higher than those obtained using the other techniques. This outcome is supported by Berk (1978) and Dempster et al. (1977): Forward selection produces model subsets with smaller residual variances compared to backward and bidirectional eliminations because only regressors which improve the model significantly are added. In the case where sample size is small, backward elimination which starts by fitting all candidate regressors into a model can result in overfitting with huge rounding errors (Draper and Smith 1998).

Best subset regression results using the *R* functions step and regsubsets are summarized in Table 2. Backward elimination using step for the full model (2) gives a better subset regression model (see Table 2) compared to manual backward elimination (see corresponding entry in Table 1) as indicated by a higher adjusted R^2 and lower p value. The same result is obtained using regsubsets. In this case, the initial model produced by regsubsets gives an adjusted R^2 value of 0.8177, but after the removal of insignificant regressors through partial F-tests, the adjusted R^2 value drops to 0.6139. This poses a problem when using regsubsets: There is no guarantee of obtaining a model with all statistically significant terms despite a high adjusted R^2 value (Kassambara 2018). As for the full model (3), regsubsets displays the same problem where the adjusted R^2 value drops from 0.9303 to 0.6050 after removing statistically insignificant terms.

Table 3 Regressors for models developed based on AIC using the first-level iterations of glmulti for physical properties

Rank	Regressor	AIC	Adjusted R^2	Overall F	p value
1	$R, R^2, D^2, V_1^2, V_2, V_2^2$	99.96	0.8177	11.47	0.00148
2	$R, R^2, D, V_1^2, V_2, V_2^2$	100.00	0.8172	11.43	0.00150
3	$R, R^2, D^2, V_1, V_2, V_2^2$	100.08	0.8162	11.36	0.00153
4	$R, R^2, D, V_2, V_2, V_2^2$	100.12	0.8157	11.33	0.00155
5	$R, R^2, D^2, V_1, V_1^2, V_2^2$	100.80	0.8072	10.77	0.00184

Table 4 Regressors for models developed based on BIC using the first-level iterations of glmulti for physical properties

Rank	Regressor	BIC	Adjusted R^2	Overall F	p value
1	R, R^2, V_2^2	105.02	0.7812	17.66	0.00016
2	$R, R^2, D^2, V_1^2, V_2, V_2^2$	105.62	0.8177	11.47	0.00148
3	$R, R^2, D, V_1^2, V_2, V_2^2$	105.66	0.8172	11.43	0.00150
4	D, D^2, V_2^2	105.70	0.7710	16.71	0.00021
5	$R, R^2, D^2, V_1, V_2, V_2^2$	105.75	0.8162	11.36	0.00153

Table 5 Regressors for models developed based on AICc using the first-level iterations of glmulti for physical properties

Rank	Regressor	AICc	Adjusted R^2	Overall F	p value
1	R, R^2, V_2^2	108.14	0.7812	17.66	0.00016
2	R^2, V_2^2	108.82	0.7135	18.43	0.00022
3	D, D^2, V_2^2	108.82	0.7710	16.71	0.00021
4	R, V_2^2	108.94	0.7113	18.25	0.00023
5	R, R^2, V_2	109.42	0.7617	15.91	0.00026

Comparing all entries in Tables 1 and 2, the best subset regression model based on the full model (2) is obtained via forward selection either performed manually or using the step function. The explicit model fit is given by:

$$S = 280.687 - 113.875R^2 - 14.102V_2^2. \quad (4)$$

This best subset regression model has an adjusted R^2 value of 0.7135 with an F-statistic value of 18.43 and a corresponding p value of 0.00022.

As an extension of the *R* functions utilized earlier, the **glmulti** package enables outputs in the form of multi-model results in which models are ranked according to specific information criteria such as AIC, BIC and AICc. This work considers only applying the glmulti function to fitting regressor terms from the full models (2) and (3), i.e., without and with interaction terms, respectively. Models without pairwise interaction are obtained using the first-level iterations where we report the top-five models ranked according to

Table 6 Regressors for models developed based on BIC using the second-level iterations of glmulti for physical properties

Rank	Regressor	BIC	Adjusted R^2	Overall F	p value
1	$R, V_2^2, V_2^2R, V_2^2R^2$	103.69	0.8160	16.53	0.00021
2	$R^2, V_2^2, V_2^2R, V_2^2R^2$	103.73	0.8156	16.48	0.00021
3	$V_2^2, R^3, V_2^2R, V_2^2R^2$	103.77	0.8151	16.43	0.00021
4	R^2, V_2^2, R^3	105.01	0.7813	17.67	0.00016
5	R, V_2^2, R^3	105.01	0.7812	17.67	0.00016

Table 7 Regressors for models developed based on AICc using the second-level iterations of glmulti for physical properties

Rank	Regressor	AICc	Adjusted R^2	Overall F	p value
1	$V_2^2R, V_2^2R^2$	107.80	0.7324	20.16	0.00015
2	$V_2^2, V_2^2R^2$	107.87	0.7312	20.04	0.00015
3	V_2^2, V_2^2R	107.95	0.7298	19.90	0.00015
4	R^2, V_2^2, R^3	108.13	0.7813	17.67	0.00016
5	$R, , R^3$	108.13	0.7812	17.67	0.00016

increasing AIC, BIC and AICc values as given in Tables 3, 4 and 5, respectively.

We find that the best subset regression model obtained previously using forward selection in terms of regressors R^2 and V_2^2 (as shown in Eq. (4)) is returned as the top-two model via the AICc criterion (refer to Table 5). AICc implements a correction factor on AIC to prevent overfitting for small sample sizes (Burnham and Anderson 2002; Cavanaugh 1997); hence, it is a more appropriate measure compared to AIC in this case. Comparing all three criteria, the best model developed is found as the first-ranked model via AIC and the second-ranked model via BIC. These two models are similar with the highest adjusted R^2 value of 0.8177 and corresponding p value of 0.00148, and they are obtained as:

$$S = -2.203 \times 10^4 + 2.994 \times 10^4 R - 1.022 \times 10^4 R^2 + 1.030 \times 10^{-4} D^2 - 3.298 V_1^2 + 1.504 \times 10^2 V_2 - 7.825 \times 10 V_2^2. \quad (5)$$

We then proceed with the second-level glmulti iterations using the main effects obtained from the best-ranked model determined using AIC, BIC and AICc to check whether further model improvement can be made. Since the second-level iterations include pairwise interaction terms, they do not converge for the best-ranked model obtained via AIC (which is a model with six main effects, namely R, R^2, D^2, V_1^2, V_2 and V_2^2) as additional degrees of freedom do not exist for hypothesis testing once interaction effects are included. Results for the second-level iterations based on BIC and AICc are presented in Table 6 and Table 7 in which both criteria develop models with the same main effects, namely R, R^2 and V_2^2, as shown in Tables 4 and 5, respectively.

From these results, the best-ranked model has a highest adjusted R^2 of 0.8160 with main effects of R and V_2^2 and interaction effects of V_2^2R and $V_2^2R^2$. Despite the higher adjusted R^2 as compared to the first-level iterations, the glmulti function fits interaction terms into the model without guaranteeing the inclusion of their corresponding main effects, thus violating the model hierarchy or heredity that prescribes including interaction terms in a model only if all corresponding main effect terms are present (Coulton and Chow

Table 8 Regressors for models developed based on AIC using the first-level iterations of glmulti for physical properties after the removal of insignificant regressors using ANOVA

Rank	Regressor	AIC	Adjusted R^2	Overall F	p value
1	R, V_1^2	109.30	0.6139	12.13	0.00131
2	R, V_1^2	109.30	0.6139	12.13	0.00131
3	R, V_1	109.64	0.6050	11.72	0.00151
4	R, V_1	109.64	0.6050	11.72	0.00151
5	R, V_1	109.64	0.6050	11.72	0.00151

Table 9 Regressors for models developed based on BIC using the first-level iterations of glmulti for physical properties after the removal of insignificant regressors using ANOVA

Rank	Regressor	BIC	Adjusted R^2	Overall F	p value
1	R, V_2^2	107.77	0.7113	18.25	0.00023
2	R, V_1^2	112.13	0.6139	12.13	0.00131
3	R, V_1^2	112.13	0.6139	12.13	0.00131
4	D, D^2, V_2^2	105.70	0.7710	16.71	0.00021
5	R, V_1^2	112.13	0.6139	12.13	0.00131

Table 10 Regressors for models developed based on AICc using the first-level iterations of glmulti for physical properties after the removal of insignificant regressors using ANOVA

Rank	Regressor	AICc	Adjusted R^2	Overall F	$p$value
1	R, V_2^2	105.94	0.7113	18.25	0.00023
2	R^2, V_2^2	108.82	0.7135	18.43	0.00022
3	D, D^2, V_2^2	108.82	0.7710	16.71	0.00021
4	R, V_2^2	108.94	0.7113	18.25	0.00023
5	R, V_2	106.73	0.6956	17.00	0.00032

1993; Wang et al. 2010; Bien et al. 2013). In this case, no R^2 main effect is included in the model. Disregarding this model, the only model with inclusion of interaction effects with their corresponding main effects for both criteria is the sixth-ranked model (not displayed here), which is the same model as the best-ranked model from the first-level iterations that we initialized the second-level iterations with. This means that including pairwise interaction into the model does not improve the strength of its correlation if we abide by the hierarchy of regression models in which main effect terms must be present or are added in tandem with interaction terms.

Reassessing the best-ranked model obtained from the first-level iterations as presented in Eq. (5), the same problem of substantially decreased adjusted R^2 value (to 0.6139) occurs (as seen using regsubsets when partial F-test is performed to remove insignificant regressors). To ensure that we select only models with statistically significant regressors, we eliminate insignificant regressors based on partial F-test using the ANOVA function on all the top-five models developed based on each of the three ICs. The results are summarized in Tables 8, 9 and 10.

The removal of insignificant terms from the initial models developed using glmulti results in several repeated models. For instance, based on the AIC, only two models are obtained from the initial top-five models. After the removal of insignificant terms, the first-ranked model developed via AIC with an adjusted R^2 of 0.6139 is similar to the model obtained from step and regsubsets for the second-order model without pairwise interaction (see Table 2). This outcome shows the consistency of results obtained from glmulti with the other functions available in R such as step and regsubsets when the same underlying principle (i.e., AIC) is used.

Comparing all models with insignificant terms removed, the model with the highest adjusted R^2 of 0.7710 is obtained from the fourth-ranked model based on BIC and third-ranked model based on AICc with a lowest p value of 0.00021. This model has the following correlation:

$$S = -2.807x10^3 + 6.738D - 3.993x10^{-3}D^2 - 1.517x10V_2^2. \quad (6)$$

Based on the coefficient magnitudes, the most influential regressor is determined to be the square of kinematic viscosity at 100 °C (i.e., V_2^2 term) for all significance levels greater than 0.00021 (which include typically reported levels such as 1%, 5% and 10%). This model developed via glmulti has the highest adjusted R^2 compared to all other regression techniques. Note that this model is justified in comparison with model (4) although the R term is not included since it was mentioned earlier that R and D exhibit a high degree of collinearity. (Hence, only one of them needs to be incorporated in a model.)

Regression modeling based on physical and chemical properties

Similar steps are repeated for the development of model using physical and chemical properties where two additional chemical properties of condensates and light crude oils, namely sulfur content, Su, and total acid number, T, are included. The preliminary step of plotting scatterplots as shown in Fig. 7 to represent the relationship between Saybolt color with all six physical and chemical properties of condensates and light crude oils again shows an absence of linear relationship between these properties, hence higher-order powers and interaction terms ought to be considered.

Considering $M = 2$, the full model without interaction terms is given by:

$$S = \beta_0 + \beta_1 R + \beta_2 D + \beta_3 V_1 + \beta_4 V_2 + \beta_5 Su + \beta_6 T + \beta_7 R^2 + \beta_8 D^2 + \beta_9 V_1^2 + \beta_{10} V_2^2 + \beta_{11} Su^2 + \beta_{12} T^2 + \varepsilon, \quad (7)$$

Table 11 Regressors for best models with and without pairwise interaction developed using stepwise regression for physical and chemical properties

Regression technique	Regressor	Adjusted R^2	Overall F	p value
Second-order model without pairwise interaction				
Forward selection	T^2, V_2^2, R^2	0.8342	24.48	0.00004
Backward elimination	R, V_1, V_2, T	0.8344	18.64	0.00013
Bidirectional elimination based on partial F-test	R, V_1, T	0.7990	19.55	0.00010
Bidirectional elimination based on overall F-test	R, V_1, T	0.7990	19.55	0.00010
Second-order model with pairwise interaction				
Forward selection	V_2T, R^2	0.8608	44.32	0.00000
Backward elimination	–	–	–	–
Bidirectional elimination based on partial F-test	R, V_1, T	0.7990	19.55	0.00010
Bidirectional elimination based on overall F-test	R, V_1, T	0.7990	19.55	0.00010

Table 12 Regressors for best models with and without pairwise interaction developed using *R* functions for physical and chemical properties

R function	Regressor	Adjusted R^2	Overall *F*	*p* value
Second-order model without pairwise interaction				
Step (forward selection)	T^2, V_2^2, R^2	0.8342	24.48	0.00004
Step (backward elimination)	R, V_1, T	0.7990	19.55	0.00010
Regsubsets	R, V_2, T	0.8490	27.25	0.00002
Second-order model with pairwise interaction				
Step (forward selection)	V_2T, R^2	0.8608	44.32	0.00000
Step (backward elimination)	–	–	–	–
Regsubsets	$V_2, Su, D^2, RV_1, RSu, DV_1, DSu, V_1T$	0.9968	549.60	0.0000

Table 13 Regressors for models developed based on AICc using first-level iteration of glmulti for physical and chemical properties

Rank	Regressor	AICc	Adjusted R^2	Overall *F*	*p* value
1	R^2, V_2^2, T	102.37	0.8510	27.66	0.00002
2	R, V_2^2, T	102.45	0.8502	27.49	0.00002
3	R^2, V_2, T	102.50	0.8498	27.40	0.00002
4	R, V_2, T	102.57	0.8490	27.25	0.00002
5	D^2, V_2, T	103.37	0.8408	25.64	0.00003

while the full model with all pairwise interaction terms has the following explicit form:

$$S = \beta_0 + \beta_1 R + \beta_2 D + \beta_3 V_1 + \beta_4 V_2 + \beta_5 Su + \beta_6 T + \beta_7 R^2 + \beta_8 D^2 + \beta_9 V_1^2 + \beta_{10} V_2^2 + \beta_{11} Su^2 + \beta_{12} T^2 + \beta_{13} RD + \beta_{14} RV_1 + \beta_{15} RV_2 + \beta_{16} RSu + \beta_{17} RT + \beta_{18} DV_1 + \beta_{19} DV_2 + \beta_{20} DSu + \beta_{21} DT + \beta_{22} V_1 V_2 + \beta_{23} V_1 Su + \beta_{24} V_1 T + \beta_{25} V_2 Su + \beta_{26} V_2 T + \beta_{27} SuT + \varepsilon. \quad (8)$$

Comparing full models (8) and (3), the former has 28 unknown parameters, a significant increase from the 15 parameters in the latter. Due to a less than zero residual degree of freedom, performing backward elimination is not possible, but other techniques are still available provided that the model developed has a maximum of 15 terms including the intercept. We first perform stepwise regression manually and utilize R functions to obtain the best model for both variants as summarized in Tables 11 and 12, respectively.

For models developed using stepwise regression, similar trends are observed where models fitted using forward selection have higher adjusted R^2 and lower *p* values than those developed using bidirectional elimination. When *R* functions are utilized, regsubsets performs better for models without and with pairwise interaction where the adjusted R^2 for the latter goes as high as 0.9968, significantly higher than all the other models developed using only physical properties.

Table 14 Regressors for models developed based on AIC using first-level iteration of glmulti for physical and chemical properties

Rank	Regressor	AIC	Adjusted R^2	Overall *F*	*p* value
1	$R, D^2, V_1, V_1^2, V_2^2, Su, Su^2, T, T^2$	84.76	0.9290	21.36	0.00179
2	$R^2, D^2, V_1, V_1^2, V_2^2, Su, Su^2, T, T^2$	84.99	0.9279	21.03	0.00186
3	$R, R^2, D^2, V_1, V_1^2, V_2^2, Su, Su^2, T, T^2$	85.47	0.9186	16.80	0.00762
4	$R, D, D^2, V_1, V_1^2, V_2^2, Su, Su^2, T, T^2$	85.50	0.9184	16.76	0.00765
5	$R, R^2, D, V_1, V_1^2, V_2^2, Su, Su^2, T, T^2$	85.50	0.9184	16.76	0.00766

Table 15 Regressors for models developed based on BIC using first-level iteration of glmulti for physical and chemical properties

Rank	Regressor	BIC	Adjusted R^2	Overall *F*	*p* value
1	$R, D^2, V_1, V_1^2, V_2^2, Su, Su^2, T, T^2$	92.55	0.9290	21.36	0.00179
2	$R, D^2, V_1^2, V_2, Su, Su^2, T, T^2$	92.67	0.9286	23.74	0.00054
3	$R^2, D^2, V_1^2, V_2, Su, Su^2, T, T^2$	92.76	0.9281	23.60	0.00054
4	$R^2, D^2, V_1, V_1^2, V_2^2, Su, Su^2, T, T^2$	92.78	0.9279	21.03	0.00186
5	$R, D, V_1^2, V_2, Su, Su^2, T, T^2$	93.10	0.9265	23.05	0.00058

This model with an F-statistic of 549.6 and corresponding p value of 0.0000 is explicitly expressed as:

$$lS = 4.766 \times 10^2 - 4.977 \times 10V_2 - 4.659Su - 6.175\times 10^{-4}D^2 - 9.009 \times 10^2 RV_1 + 5.978RSu + 1.612DV_1 - 4.898 \times 10^{-3}DSu - 2.557 \times 10^2 V_1 T. \quad (9)$$

However, one drawback from this model is that the interaction terms are fitted without their main effects. Hence, if the hierarchy of a model is to be abided, the best subset regression model would be the model developed by regsubsets based on full model (7) without pairwise interaction with an adjusted R^2 of 0.8490, overall F-statistic of 27.25 and corresponding p value of 0.00002 as given by the following:

$$S = 361.3 - 211.5R - 19.77V_2 - 278.2T. \quad (10)$$

Similarly, we reproduce results using the **glmulti** package with the inclusion of chemical properties and find that model (10) is returned as the top-fourth model via AICc as shown in Table 13, further supporting the suitability of AICc for small sample size. Regressors for the top-five models developed using AIC and BIC are shown in Tables 14 and 15, respectively.

Comparing all three criteria, the best model developed is found as the topmost model via the AIC and BIC with the highest adjusted R^2 of 0.9290, overall F-statistic of 21.36 and corresponding p value of 0.00179 expressed in the following correlation:

$$S = -5.844 \times 10^2 + 6.306 \times 10^2 R - 2.367 \times 10^{-4} D^2 - 1.396 \times 10^2 V_1 + 7.198\times 10V_1^2 - 1.136 \times 10^2 V_2^2 + 1.942 \times 10^{-1} Su - 1.156 \times 10^{-4} Su^2 - 3.140 \times 10^3 T + 4.420 \times 10^4 T^2. \quad (11)$$

We then proceed with the second-level glmulti iterations using the main effects obtained from the best-ranked model via AIC, BIC and AICc to observe for any improvement in models. Since the second-level iterations include pairwise interaction terms, they do not converge for the topmost first-level model obtained via AIC and BIC with nine main effects, namely R, D^2, V_1,V_1^2, V_2^2, Su, Su^2, T and T^2 (see Table 14 and Table 15), as there is no degree of freedom once interaction effects are included. The top-five models for second-level iterations based on AICc criterion are tabulated in Table 16 by considering the topmost first-level model obtained via AICc with three main effects, namely R^2, V_2^2 and T (see Table 13).

From Table 16, the topmost model has a highest adjusted R^2 of 0.8558 with main effect of R^2 as well as interaction effect ofTV_2^2. However, due to the absence of main effect T and V_2^2 which violates model hierarchy, we reassess the best model obtained from first-level iterations as presented in Eq. (11). Performing partial F-test to remove insignificant regressors from the model results in a significant drop in adjusted R^2 value from 0.9290 to 0.8354. Similar partial F-test is conducted on all models developed using first-level iterations to ensure that we select only models with statistically significant regressors. Note that all first-level models developed on AICc criterion have regressors that are already statistically significant; thus, they do not require subsequent F-tests. The results are summarized in Tables 17 and 18 for AIC and BIC, respectively.

Comparing all models with insignificant terms removed, the model with the highest adjusted R^2 of 0.8510 (see

Table 16 Regressors for models developed based on AICc using second-level iteration of glmulti for physical and chemical properties

Rank	Regressor	AICc	Adjusted R^2	Overall F	p value
1	R^2, TV_2^2	98.52	0.8558	0.00000	0.00000
2	R^2, T, TV_2^2	102.13	0.8535	0.00002	0.00002
3	R^2, TR^2, TV_2^2	102.16	0.8531	0.00002	0.00002
4	R^2, V_2^2, TR^2	102.25	0.8523	0.00002	0.00002
5	$R^2, V_2^2R^2, TV_2^2$	102.37	0.8511	0.00002	0.00002

Table 17 Regressors for models developed based on AIC using first-level iteration of glmulti for physical and chemical properties after the removal of insignificant regressors using ANOVA

Rank	Regressor	AIC	Adjusted R^2	Overall F	p value
1	R, V_1, V_2^2, T	97.77	0.8354	18.77	0.00012
2	R^2, V_1, V_2^2, T	97.69	0.8363	18.88	0.00012
3	R, V_1, V_2^2, T	97.77	0.8354	18.77	0.00012
4	R, V_1, V_2^2, T	97.77	0.8354	18.77	0.00012
5	R, V_1, V_2^2, T	97.77	0.8354	18.77	0.00012

Table 18 Regressors for models developed based on BIC using first-level iteration of glmulti for physical and chemical properties after the removal of insignificant regressors using ANOVA

Rank	Regressor	BIC	Adjusted R^2	Overall F	p value
1	R, V_1, V_2^2, T	102.02	0.8354	18.77	0.00012
2	R, V_1^2, V_2, T	102.15	0.8339	18.58	0.00013
3	R^2, V_1^2, V_2, T	102.08	0.8347	18.68	0.00012
4	R^2, V_1, V_2^2, T	101.94	0.8363	18.88	0.00012
5	R, V_1^2, V_2, T	102.15	0.8339	18.58	0.00013

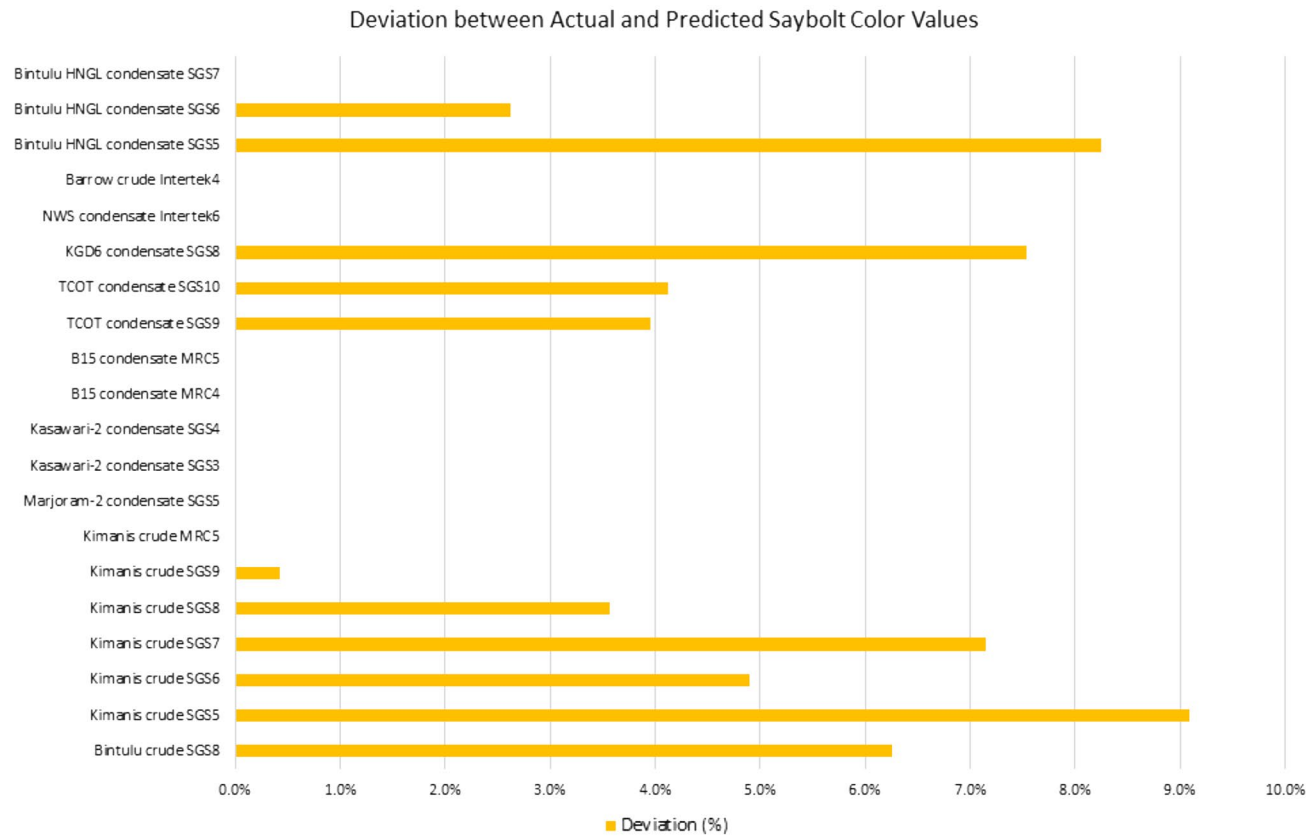

Fig. 8 Validation results as based on deviations between actual and predicted Saybolt color values

Table 19 Selection of best models obtained in this work

Model structure	Best model	Remark
Lower-order model with physical properties only	$S = -2.807x10^3 + 6.738D - 3.993 \times 10^{-3}D^2 - 1.517 \times 10V_2^2$ (Eq. 6)	Obtained using first-level iteration of glmulti based on the BIC, this model has the lowest BIC compared to all other models. Although the adjusted R^2 is the highest, the overall F-statistic still falls behind models developed using forward selection
Lower-order model with physical and chemical properties	$S = 199.8 - 73.89R^2 - 9.212V_2^2 - 268.8T$ (Eq. 10)	Obtained using first-level iteration of glmulti based on AICc, this model has the highest overall F-statistic and adjusted R^2 compared to all other first-level iteration models developed based the three ICs as well as models developed manually and using R functions

Table 13) is obtained as the topmost model based on AICc with the following correlation:

$$S = 199.8 - 73.89R^2 - 9.212V_2^2 - 268.8T. \quad (12)$$

Based on the coefficient magnitudes, the most influential regressor is determined to be the total acid number (i.e., T term) for all significance levels greater than 0.00002 (which include typically reported levels such as 1%, 5% and 10%). This model developed via **glmulti** has the highest adjusted R^2 compared to all other regression techniques and is slightly better than model (10) in which both the main effects of R and V_2 are raised to a second order. This implies that the inclusion of chemical properties into Saybolt color correlation improves the fit of the model with an increase in adjusted R^2 from 0.7710 to 0.8510.

Model validation

Model validation is performed on the developed correlation model by comparing its predicted Saybolt color values against those measured using conventional method (i.e., laboratory analysis). The latter (data from conventional measurements) are obtained from other assay reports for which the data are not used in developing (i.e., training) the model. We conduct the validation for 20 selected samples and present result on deviations between the actual (from assay reports) and predicted values by our proposed regression model given by Eq. (12). As shown in Fig. 8, the deviations are largely less than 5% except for five samples (but which yet do not exceed 10%) with a mean of 2.9%. The validation indicates a prediction error of less than 5% for 75% of the samples, which is acceptable (Simpson et al. 2004).

Concluding remarks

Based on the models developed using various regression approaches for this class of machine learning problems arising in prediction of petroleum properties, the best models representing different model structures are summarized in Table 19. As proposed by Draper and Smith (1998), no regression technique is the best, especially when there are constraints in terms of the patterns of the data as well as the practicality of the problem. Choosing the right regression approach would depend on the type of model that we aim to develop. If the significance of regressors in a model is an important factor, forward selection ensures that only significant regressors are added to the model, but this method may result in lower adjusted R^2 values. Hence, if we want to develop highly correlated models with high adjusted R^2 values, we can opt for the functions regsubsets (in **leaps** package) and glmulti (in **glmulti** package) which have been shown to give consistent results with **glmulti** offering the capability of providing multi-model inferencing. Since both functions are susceptible to overfitting in which redundant regressors may be present in the model, partial F-tests can be performed to remove insignificant terms. But it is important to note that the adjusted R^2 values typically drop with elimination of insignificant terms.

Acknowledgements We are grateful for the financial support provided from YUTP grant no. 015LCO-166 to perform this study. We also acknowledge data and technical advice given by engineers from Process Optimization Group of PETRONAS Group Technical Solutions (GTS), namely Shahrul Azman Zainal Abidin and Farah Syamim Anuar.

Authors' contributions Cheng Seong Khor designed the study; *Jia jia* Leam performed most of the study under the supervision of Cheng Seong Khor and Sarat C. Dass; Cheng Seong Khor and Sarat C. Dass reviewed the results and outcome of the study.

Funding This work is partially supported by YUTP (Yayasan Universiti Teknologi PETRONAS) grant no. 015LCO-166.

Code availability (software application or custom code) The computer code for this study is available upon request from the corresponding author.

Compliance with ethical standards

Conflict of interest/Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data and material (data transparency) The data and material used to perform this work is available upon request from the corresponding author.

References

ASTM International (2003) Standard test method for Saybolt color of petroleum products (Saybolt chromometer method). ASTM International. https://doi.org/10.1520/D0156-15

ASTM International (2008) Standard test method for ASTM color of petroleum products (ASTM color scale). ASTM International

Berk KN (1978) Comparing subset regression procedures. Technometrics 20(1):1–6. https://doi.org/10.1080/00401706.1978.10489609

Bien J, Taylor J, Tibshirani R (2013) A lasso for hierarchical interactions. Ann Stat 41(3):1111–1141

Burnham KP, Anderson DR (2002) Information and likelihood theory: a basis for model selection and inference. In: Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York, pp 66–70

Calcagno V, de Mazancourt C (2010) multi: an R package for easy automated model selection with (generalized) linear models. J Stat Softw 34(1):1–29. https://doi.org/10.18637/jss.v034.i12

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. Stat Probab Lett, 33(2), 201–208

Clarkson Laboratory and Supply Inc (2019) Available online: http://store.clarksonlab.com. Accessed on 11 July 2020

Coulton C, Chow J (1993) Interaction effects in multiple regression. J Soc Serv Res 16(1–2):179–199. https://doi.org/10.1300/J079v16n01_09

Dempster AP, Schatzoff M, Wermuth N (1977) A simulation study of alternatives to ordinary least squares. J Am Stat Assoc 72(357):77–91. https://doi.org/10.2307/2286909

Diller I, Dean J, DeGray R, Wilson J Jr (1943) Color index. Light-colored petroleum products. Ind Eng Chem Anal Ed 15(6):365–373. https://doi.org/10.1021/i560118a003

Douglas, R. K., Nawar, S., Alamar, M. C., Mouazen, A. M., & Coulon, F. (2018). Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. Sci Total Environ, 616–617, 147–155

Draper NR, Smith H (1998) Selecting the "best" regression equation. In: Applied regression analysis. Wiley, New York, pp 327–368

Hagemann HW, Hollerbach A (1986) The fluorescence behaviour of crude oils with respect to their thermal maturation and degradation. Org Geochem 10(1):473–480. https://doi.org/10.1016/0146-6380(86)90047-1

Hu S-P (2016) Generalized degrees of freedom. J Cost Anal Paramet 9(2):93–111. https://doi.org/10.1080/1941658X.2016.1191388

IHS Markit (2018) The shale gale turns 10: a powerful wind at America's back. IHS Markit

IndiaMart (2019) Color measuring instruments—Lovibond Model Fx Digital—New Exporter from Mumbai. https://www.indiamart.com/akumarlab/color-measuring-instruments.html. Accessed on 11 July 2020

Kassambara A (2018) Machine learning essentials: practical guide in R. CreateSpace Independent Publishing Platform

Khor CS et al (2020) Correlation model development for Saybolt color of condensates and light crude oils. ASM Sci J 13:7

Lumley T (2014) Package "leaps" in R (Version 2.9) [Fortran; R]

Montgomery DC, Peck EA, Vining GG (2012) Introduction to linear regression analysis. Wiley, Hoboken

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/

Rawlings JO, Pantula SG, Dickey DA (2006) Applied regression analysis: a research tool. Springer, New York

Rodriguez JD, Comstock MJ, Auz B, Olmstead T (2017) A spectroscopic method of determining color of petroleum products using CIELab color space with LED illumination. OPTO. https://doi.org/10.1117/12.2252399

Simpson R, Almonacid S, Mitchell M (2004) Mathematical model development, experimental validation and process optimization: retortable pouches packed with seafood in cone frustum shape. J Food Eng 63:153–162. https://doi.org/10.1016/S0260-8774(03)00294-2

Speight JG (2015) Chapter 1—occurrence and formation of crude oil and natural gas. In: Speight JG (ed) Subsea and Deepwater oil and gas science and technology. Gulf Professional Publishing, pp 1–43. https://doi.org/10.1016/B978-1-85617-558-6.00001-5

Steffens J, Landulfo E, Courrol LC, Guardani R (2011) Application of fluorescence to the study of crude petroleum. J Fluoresc 21(3):859–864. https://doi.org/10.1007/s10895-009-0586-4

Tomren AL, Barth T (2014) Comparison of partial least squares calibration models of viscosity, acid number and asphaltene content in petroleum, based on GC and IR data. Fuel 120:8–21. https://doi.org/10.1016/j.fuel.2013.11.065

Wang, X., Elston, R. C., & Zhu, X. (2010). The meaning of interaction. Hum Hered, 70(4), 269–277