Brazilian Journal
of Chemical
Engineering

ISSN 0104-6632
Printed in Brazil
www.abeq.org.br/bjche

Vol. 33, No. 04, pp. 1083 - 1090, October - December, 2016
dx.doi.org/10.1590/0104-6632.20160334s20150374

TOWARD PREDICTIVE MODELS FOR ESTIMATION OF BUBBLE-POINT PRESSURE AND FORMATION VOLUME FACTOR OF CRUDE OIL USING AN INTELLIGENT APPROACH

D. Abooali[1] and E. Khamehchi[2*]

[1]School of Chemical Engineering, Iran University of Science and Technology, (IUST), Postal Code, 16765-163, Tehran, Iran.
[2]Faculty of Petroleum Engineering, Amirkabir University of Technology, Hafez Avenue, 15914, Tehran, Iran.
Phone: + 98 (21) 64545154; Fax: + 98 (21) 64543528
E-mail: khamehchi@aut.ac.ir

(Submitted: June 14, 2015 ; Revised: September 1, 2015 ; Accepted: September 2, 2015)

Abstract - Accurate estimation of reservoirs fluid properties, as vital tools of reservoir behavior simulation and reservoir economic investigations, seems to be necessary. In this study, two important properties of crude oil, bubble point pressure (P_b) and formation volume factor (B_{ob}), were modelled on the basis of a number of basic oil properties: temperature, gas solubility, oil API gravity and gas specific gravity. Genetic programming, as a powerful method, was implemented on a set of 137 crude oil data and acceptable correlations were achieved. In order to evaluate models, two test datasets (17 data for P_b and 12 data for B_{ob}) were used. The squared correlation coefficient (R^2) and average absolute relative deviation (AARD %) over the total dataset (training + test) are 0.9675 and 8.22% for P_b and 0.9436 and 2.004% for B_{ob}, respectively. Simplicity and high accuracy are the advantages of the obtained models.

Keywords: Crude oil; Bubble point pressure; Formation volume factor; Genetic programming.

INTRODUCTION

Thermodynamic quantities of crude oil are a set of important features in order to determine technical specifications of oil production process equipment. Designing plenty of systems such as upstream and underground devices, surface operation equipment, etc., requires adequate and accurate information about oil parameters which are achieved, in many cases, from experimental tests along with mathematical correlations and formulas.

Laboratory tests are usually expensive and sometimes difficult and time-consuming. However, the application of correlations is economically advantageous and increases the speed of works. Furthermore, the other great use of the correlations is to determine oil future specifications and changes taken into great consideration in reservoir simulators.

Various pressure-volume-temperature (PVT) properties of crude oil can be estimated by means of equations of state or oil PVT analysis, if a complete set of variables of the oil including temperature, pressure and fluid composition are available. But in many cases, the composition of reservoir fluid is not predetermined, especially in the primary stages of recovery processes. Thus, some correlations are required to be functions of a number of readily available reservoir parameters in order to be used by engineers and scientists in this area.

In fact, the main aim of this project was to provide simple and accurate models for prediction of bubble point pressure (P_b) and bubble point formation

*To whom correspondence should be addressed volume factor (B_{ob}) solely as functions of simple and quickly accessible live crudeoil parameters. The parameters are temperature (T), gas solubility (R_s), oil API gravity and gas specific gravity (γ_g).

In a hydrocarbon system at constant temperature, whether single-component or mixture, the bubble point pressure is the maximum pressure at which the first gas bubbles appear (Ahmed, 2010). The state of the system in this condition is called "saturated liquid".

The oil formation volume factor (FVF) is the ratio of the specific volume of oil at its natural temperature and pressure to the specific volume of the oil at standard conditions (i.e. P = 1 atm and T = 60 ˚F). If B_o is measured in the bubble point condition, it will be the bubble point oil formation volume factor (Bob).

There are several correlations and methodologies developed and proposed so far for prediction of P_b and B_{ob}. Methods of Standing (1947), Vasquez and Beggs (1980), Glaso (1980), Marhoun (1988) and Petrosky and Farshad (1993), as famous correlations, have been introduced in the literature (Ahmed, 2010). Elsharkawy and Alikhan (1997) presented a set of correlations for gas solubility, oil compressibility (C_o) and B_{ob}. Their relation for B_{ob} is as follows:

$$B_{ob} = 1 + 40.428\left(10^{-5}\right) \times R_s + 63.802\left(10^{-5}\right) \times (T-60) + 0.78\left(10^{-5}\right) \times R_s (T-60)\left(\gamma_g / \gamma_o\right) \quad (1)$$

in which, γ_g is gas specific gravity. Khamehchi *et al.* (2009) also proposed three models for P_b, R_s and bubble point oil viscosity (μ_{ob}). They called the achieved models "AUT". Their P_b correlation is given below:

$$P_b = 107.93\, R_s^{0.9129} \times \gamma_g^{-0.666} \times T^{0.2122} \times API^{-1.08} \quad (2)$$

Some presented correlations or algorithms are based on consistencies of a number of oil components or assays, which should be predetermined (Elsharkawy, 2003; AlQuraishi, 2009; Bandyopadhyay and Sharma, 2011; Farasat *et al.*, 2013). However, the composition-based models have some limitations in their uses in preliminary reservoir investigations and simulations.

There are also several methods using the artificial neural network (ANN) technique to predict P_b and B_{ob} (Rasouli *et al.*, 2008; Asadisaghandi and Tahmasebi, 2011). Adaptive network-based fuzzy inference system (ANFIS) is another new approach that has been applied in this area (Shojaei *et al.*, 2014).

Different procedures and methodologies can be used for model development. Artificial neural network (ANN), generalized regression neural networks (GRN), imperialist competitive algorithm (ICA), particle swarm optimization (PSO), adaptive network-based fuzzy inference system (ANFIS), genetic programing (GP), etc. are applied as famous methods in various fields, especially for optimization and prediction purposes. In the present study, a genetic programming based multi-gene symbolic regression algorithm called "GPTIPS" (Searson, 2009) was applied. This is an approved method used by the authors in some projects (Abooali and Khamehchi, 2014).

The application of genetic programming for developing simple-to-use correlations for P_b and B_{ob} seems novel. Moreover, applying natural ranges of bubble point pressure, bubble point formation volume factor, temperature, gas solubility, oil API gravity and gas specific gravity has increased the applicability and accuracy of the new developed models.

MATERIALS AND METHODS

Data Set

The total dataset includes 137 training sets of data from 137 oil samples. Each set includes temperature, solution gas ratio, oil API gravity, gas specific gravity, oil bubble point pressure and formation volume factor. The data were collected from different geographical zones.

In order to determine the predictive capability of the models and also to implement a comparison between the new developed models and other correlations, two additional sets – one for P_b including 17 sets of data and the other for B_{ob} which has 12 sets of data – were applied. The data of the additional sets known as "test sets" were gathered from several papers and reference books (Ahmed, 2010; McCain, 1990; Shojaei *et al.*, 2014). The ranges of all parameters are presented in Table 1.

Table 1: The range of parameters in the dataset of present study.

Quantity	Range
Temperature (˚F)	85 – 306
Solution gas oil ratio (SCF/STB)	83 – 2217
Oil API gravity	21.143 – 124.054
Gas specific gravity (air = 1)	0.216 – 1.872
Bubble point pressure (psia)	350 – 5152
Bubble point formation volume factor (bbl/STB)	1.0955 – 2.027

Model Development Procedure

Genetic programing (GP) is a powerful methodology and its programing procedure was patterned from biological generation systems. Genetic programming was introduced in the early 1990s and has been developed mostly by its innovator John Koza (1992). As an efficient tool, it has a wide variety of applications in mathematical and also computational and modeling projects. Genetic algorithm (GA), as a famous algorithm based on genetic programing, is a trusted method in correlation development. Many new models have been produced by use of GA, so far, for different phenomena. In the present study, a kind of regression method was applied on the basis of GP. GP is an evolutionary methodology involving computer programs in order to perform tasks. It is a machine learning method that biologically generates the random population of mathematical functions under predetermined restrictions. The population is represented as chromosomes like syntactic tree structures. The primary population of functions includes mathematical operators operating on input data. Each tree structure is often known as a "gene". A simple gene is shown in Figure 1. Each gene mainly stands for a function. The number of genes, number of populations and complexity of genes can be determined by the user.

When the process is specification of mathematical models or functions, the GP is often known as "symbolic regression". In conventional regressions, at first, the form of the model should be determined by the user and then, model parameters will be fitted. But in a symbolic regression, the algorithm itself searches for both the model form describing the data behavior and also the model parameters.

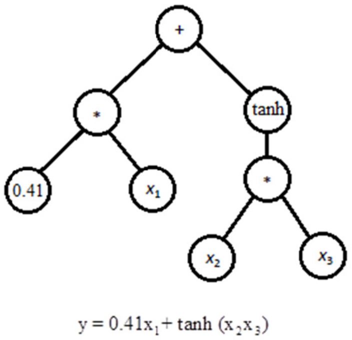

Figure 1: A simple gene (tree structure) with operators: ×, + and tanh.

In GA, after random generation of the first population (parents), the overall primary model is achieved by weighted summation of all functions represented as genes with a bias term. The weight of each tree and the bias term are calculated by an ordinary least squares technique. A simple schematic of a model including two gene structures is shown in Figure 2. As can be seen in Figure 2, in each gene there are some mathematical operators which are applied on the input variables. d_0 is a bias term and d_1 and d_2 are weights of genes. The optimal weights for the genes are automatically obtained by use of ordinary least squares. x_1, x_2 and x_3 are input variables.

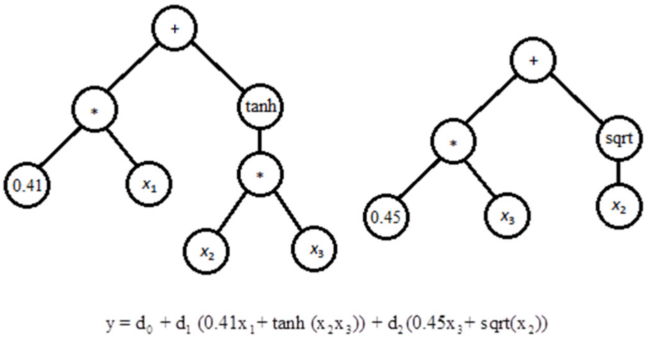

Figure 2: Overall model of two genes with a bias term. d_0, d_1 and d_2 are linear coefficients of the genes.

In the next step, crossing over the best performing trees and modifications of trees (cutting some parts of trees and exchanging cut parts between themselves) are implemented to make a new population (children), i.e., new tree structures. The weighted summation of all new genes is repeated and the new weights and also the new bias term are determined. This process is iterated and new populations are created until the last population contains new trees (functions) that are able to solve the problem successfully (Searson, 2010). Complementary explanations about GA are found elsewhere (Searson *et al.*, 2010; Koza, 1992).

If the algorithm creates a number of genes instead of one, in fact, a more accurate methodology will be applied for producing a population of mathematical relations (multi-gene symbolic regression). A multigene consists of one or more genes each one being an individually usual GP tree. Thus, multi-gene approaches often give simpler functions than other models consisting of one monolithic GP tree (Searson, 2010). The flowchart of genetic algorithm is shown in Figure 3.

A genetic algorithm toolbox called "GPTIPS" was prepared by Searson (2009) for use with MATLAB software. It was written mainly for multigene symbolic regression applications. So, all the previous operations (generating parent genes, crossing over the best trees, mutating, producing children, etc.) are carried out by GPTIPS to achieve the best correlation. More details about GPTIPS can be found in the references (Searson, 2009; Searson *et al.*, 2010).

In this study, GPTIPS was used to develop nonlinear correlations. Input data (training and test subsets), including experimental sets of temperature, gas solubility, oil API gravity and gas specific gravity along with bubble point pressure and bubble point formation volume factor, were fed to the GPTIPS. Then, tuning parameters were adjusted and controlled. Running the program, each correlation was obtained with acceptable values of statistical criteria and good accuracy.

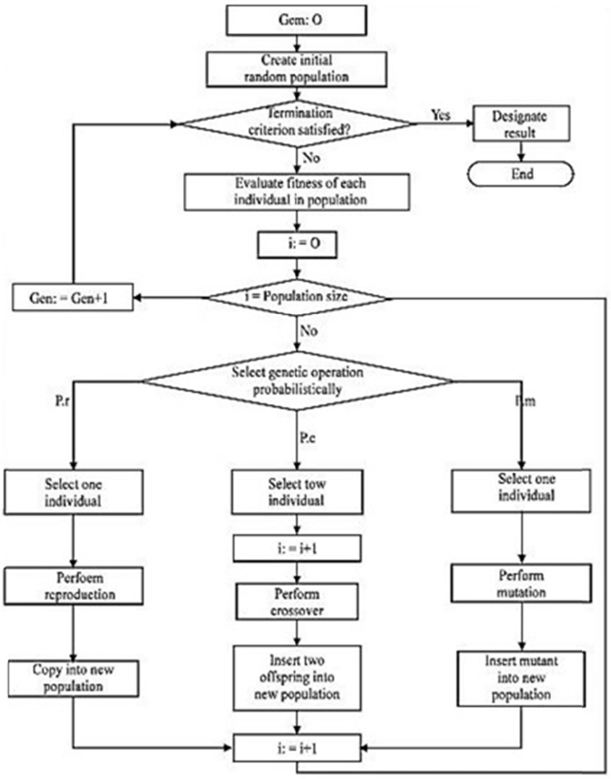

Figure 3: Genetic algorithm flowchart.

Evaluation of Model Validity

For evaluation of the developed models, three common statistical parameters were calculated: squared correlation coefficient (R^2), root-mean-square deviation (RMSD) and average absolute relative deviation percentage (AARD %). A low value of RMSD and AARD is preferred. R^2 should be near to one.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i^{exp.} - y_i^{cal.})^2}{\sum_{i=1}^{n} (y_i^{exp.} - \overline{y^{exp.}})^2} \tag{3}$$

$$RMSD = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i^{exp.} - y_i^{cal.})^2} \tag{4}$$

$$ARD(\%) = \left| \frac{y_i^{exp.} - y_i^{cal.}}{y_i^{exp.}} \right| \times 100 \tag{5}$$

$$AARD\ (\%) = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \left| \frac{y_i^{exp.} - y_i^{cal.}}{y_i^{exp.}} \right| \times 100 \tag{6}$$

where $y_i^{exp.}$, $y_i^{cal.}$, $\overline{y}^{exp.}$ and *n* are the experimental, predicted, and average of experimental dependent variables (B_{ob} and P_b) and number of samples in the dataset, respectively.

RESULTS AND DISCUSSIONS

Applying the genetic programming approach, two correlations for bubble point pressure and formation volume factor of crude oil were obtained. They are as follows:

$$\begin{aligned} P_b = {} & 169 \ln(R_s \gamma_g{}^3) - 2614 \ln\left(\gamma_g \ln\left(R_s\right)\right) \left(\gamma_g + \frac{1}{\gamma_g{}^2 + API} \right) + \frac{11.54 R_s}{\gamma_g{}^3 + 0.05948 API} \\ & + 1.934 \ln \left(\frac{R_s \gamma_g}{2 API} \right) \left(T + \frac{R_s \ln(API)}{T + API} \right) - \frac{0.004272 R_s{}^2}{2\gamma_g{}^2 + 0.05948 API} + 2746 \gamma_g{}^2 + 472.5 \end{aligned} \tag{7}$$

$$\begin{aligned} B_{ob} = {} & 0.0007004\, T + 0.003542\, R_s + 0.0003534\, API + 0.0004275 \left(\gamma_g{}^2\right) \left(API - \ln\left(R_s\right)\right) \\ & -0.0003518 \left(\exp\left(\gamma_g\right) \ln\left(R_s\right) + R_s \ln\left(2\, T^2\right) - \gamma_g API\right) - 0.003894 \left(\gamma_g{}^2\right) \exp\left(\gamma_g\right) + \\ & 1.622 \left(10^{-6}\right) \left(\left(R_s - \exp\left(\gamma_g\right) \ln(T)\right) \left(3\, T - API\right) + R_s API \ln\left(\gamma_g\right)\right) + 0.945 \end{aligned} \tag{8}$$

The statistical parameters of the new developed models are given in Table 2 and Table 3 for P_b and B_{ob}, respectively. Figure 4 and Figure 5 show the values predicted by Equation (7) and Equation (8) versus experimental data. According to Table 2, Table 3, Figure 4 and Figure 5, the new developed models are capable of prediction and estimation of P_b and B_{ob}. In Figure 6 and Figure 7, cumulative frequency percent versus maximum absolute relative deviation percent are shown over the entire data set for P_b and B_{ob}, respectively. As shown in these figures, the absolute relative deviation percent for 90.26% of all data estimated by Equation (7) is less than 20%. For B_{ob}, the absolute relative deviation percent for 97.315% of all the dataset is lower than 10%.

Table 2: Statistical parameters of the new developed model for P_b.

n_{total} = 154 $RMSD_{total}$ = 190.8408 psia R^2_{total} = 0.9675 $AARD_{total}$ = 8.2206%	n_{train} = 137 $RMSD_{train}$ = 169.1126 psia R^2_{train} = 0.9695 $AARD_{train}$ = 8.0395%	n_{test} = 17 $RMSD_{test}$ = 315.3562 psia R^2_{test} = 0.9131 $AARD_{test}$ = 9.6795%

Subscripts “total”, “train”, and “test” are used to distinguish the results related to total data set, training data set, and test data set, respectively.

Table 3: Statistical parameters of the new developed model for B_{ob}.

n_{total} = 149 $RMSD_{total}$ = 0.0449 bbl/STB R^2_{total} = 0.9436 $AARD_{total}$ = 2.0040%	n_{train} = 137 $RMSD_{train}$ = 0.0456 bbl/STB R^2_{train} = 0.9419 $AARD_{train}$ = 2.0069%	n_{test} = 12 $RMSD_{test}$ = 0.0360 bbl/STB R^2_{test} = 0.9322 $AARD_{test}$ = 1.9700%

Subscripts “total”, “train”, and “test” are used to distinguish the results related to total data set, training data set, and test data set, respectively.

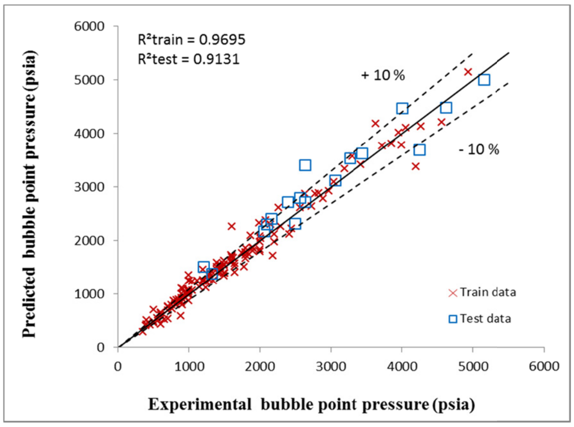

Figure 4: Predicted versus experimental bubble point pressure. R^2_{train} and R^2_{test} are correlation coefficients of training and test data, respectively.

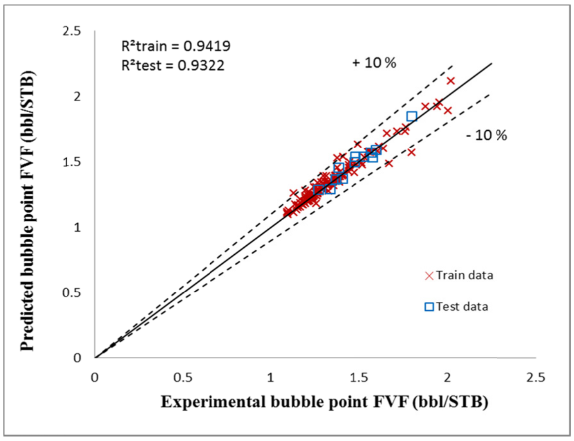

Figure 5: Predicted versus experimental bubble point oil formation volume factor. R^2_{train} and R^2_{test} are correlation coefficients of training and test data, respectively.

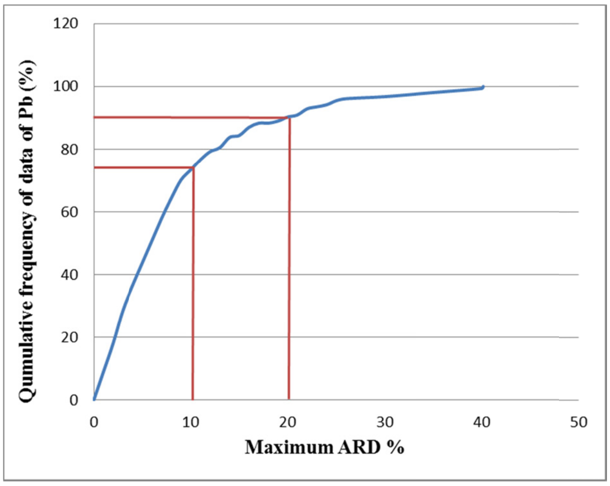

Figure 6: Cumulative frequency percent versus maximum absolute relative deviation of the new developed model for bubble point pressure over the whole data set (154 data). As can be seen, the absolute relative deviations for 73.377% of all data are less than 10% and absolute relative deviations for 90.26% of all data are less than 20%.

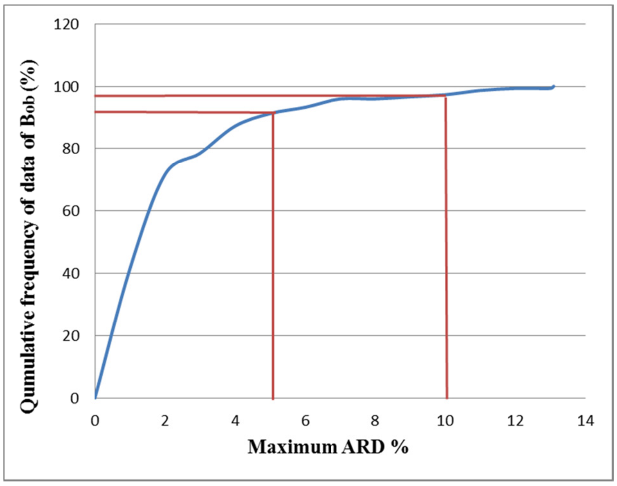

Figure 7: Cumulative frequency percent versus maximum absolute relative deviation of the new developed model for bubble point formation volume factor over all the data set (149 data). As can be seen, absolute relative deviations for 91.275% of all the data are less than 5% and absolute relative deviations for 97.315% of all the data are less than 10%.

In order to evaluate the new correlations along with other models, a comparison has been implemented over test datasets and the results are presented in Table 4, Figure 8 and Figure 9. As a result, the prediction capability of the new developed models is higher than that of previous relations.

Table 4: Comparison between empirical correlations and the new developed models over test data set.

Method	P_b (Number of data=17)				B_{ob} (Number of data=12)			
	AARD %	R^2	RMSD (psia)	Maximum AARD%	AARD %	R^2	RMSD (bbl/STB)	Maximum AARD%
Standing	16.951	0.431	807.127	59.054	2.621	0.888	0.0462	5.773
Glaso	21.740	0.461	785.440	50.763	2.171	0.901	0.0435	6.484
Marhoun	11.190	0.676	608.592	43.582	2.292	0.902	0.0434	6.774
Petrosky and Farshad	-	-	-	-	2.109	0.916	0.0402	7.074
AUT	35.199	-0.944	1491.11	94.462	-	-	-	-
Present study	9.680	0.913	315.356	29.317	1.970	0.932	0.0360	4.788

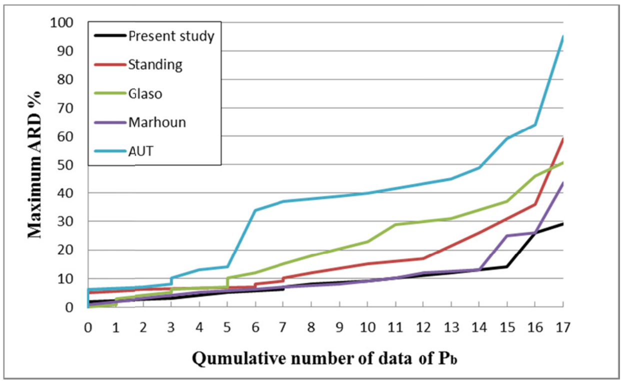

Figure 8: Maximum absolute relative deviation versus cumulative number of data for bubble point pressure of the test set (17 data). As can be seen, the absolute relative deviation curve of the new developed model is lower than that of other empirical correlations.

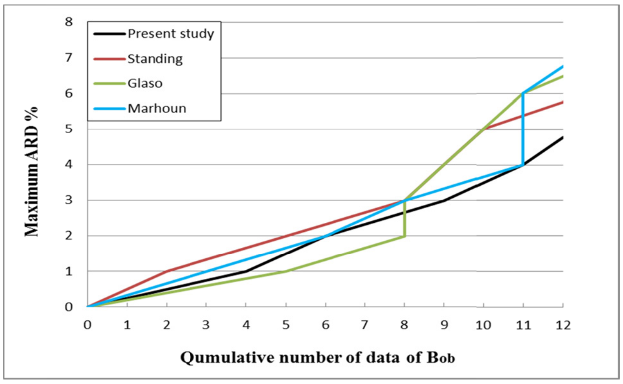

Figure 9: Maximum absolute relative deviation versus cumulative number of data for the bubble point formation volume factor for the test set (12 data). The average of the absolute relative deviation of the new developed model is lower than that of other empirical correlations.

These comparisons demonstrate the superiority of the correlations developed in the present project among proposed models.

The experimental values of all dataset along with predicted data have been provided in the supporting materials and information.

CONCLUSIONS

By application of genetic programing methodology, two new models have been achieved for estimation and prediction of bubble point pressure and bubble point formation volume factor, as functions of a number of rapidly measurable oil parameters. One of the useful applications of this kind of model is prediction of oil properties in the future during the reservoir lifetime that is very important, especially for economic studies as well as effective uses in reservoir simulators. A comparison between the new proposed models and some other correlations shows the greater accuracy of the proposed models over previous works.

NOMENCLATURE

AARD%	average absolute relative deviation percentage
ANN	artificial neural network
API	oil API gravity
ARD%	absolute relative deviation percent
bbl STB^{-1}	barrel(s) per standard barrels
B_o	oil formation volume factor
B_{ob}	bubble point oil formation volume factor
C_o	oil compressibility at constant temperature
FVF	formation volume factor
GP	genetic programing
GRN	generalized regression neural networks
ICA	imperialist competitive algorithm
n	number of samples in the dataset
P_b	bubble point pressure
PSO	particle swarm optimization
PVT	pressure – volume – temperature
R^2	squared correlation coefficient
RMSD	root-mean-square deviation
R_s	gas solubility
SCF STB^{-1}	standard cubic feet of solution gas per standard barrels of oil
T	temperature
$y_i^{cal.}$	predicted dependent variable of component i
$y_i^{exp.}$	experimental dependent variable of component i
$\overline{y}^{exp.}$	average of experimental dependent variables
γ_g	gas specific gravity
γ_o	oil specific gravity
μ_{ob}	oil bubble point viscosity

REFERENCES

Abooali, D. and Khamehchi, E., Estimation of dynamic viscosity of natural gas based on genetic programming methodology. J. Nat. Gas. Sci. Eng., 21, 1025 (2014).

Ahmed, T., Reservoir Engineering Handbook. Gulf Professional Publishing (2010).

AlQuraishi, A. A., Determination of crude oil saturation pressure using linear genetic programming. Energy Fuels, 23, 884 (2009).

Asadisaghandi, J. and Tahmasebi, P., Comparative evaluation of back-propagation neural network learning algorithms and empirical correlations for prediction of oil PVT properties in Iran oilfields. J. Pet. Sci. Eng., 78, 464 (2011).

Bandyopadhyay, P. and Sharma, A., Development of a new semi analytical model for prediction of bubble point pressure of crude oils. J. Pet. Sci. Eng., 78, 719 (2011).

Elsharkawy, A. M., An empirical model for estimating the saturation pressures of crude oils. J. Pet. Sci. Eng., 38, 57 (2003).

Elsharkawy, A. M. and Alikhan, A. A., Correlations for predicting solution gas/oil ratio, OFVF and undersaturated oil compressibility. J. Pet. Sci. Eng., 17, 291 (1997).

Farasat, A., Shokrollahi, A., Arabloo, M., Gharagheizi, F. and Mohammadi, A. H., Toward an intelligent approach for determination of saturation pressure of crude oil. Fuel Process Technol., 115, 201 (2013).

Glaso, O., Generalized pressure-volume-temperature correlations. J. Pet. Technol., 32, 785 (1980).

Khamehchi, E., Rashidi, F., Rasouli, H. and Ebrahimian, A., Novel empirical correlations for estimation of bubble point pressure, saturated viscosity and gas solubility of crude oils. J. Petrol Sci., 6, 86 (2009).

Koza, J., Genetic Programming. Massachusetts Institute of Technology. New York (1992).

Marhoun, M. A., PVT Correlations for Middle East Crude Oils. J. Pet. Technol., 40, 650 (1988).

McCain, W. D., The Properties of Petroleum Fluids. PennWell Publishing Co (1990).

Petrosky, Jr. G. E. and Farshad, F., Pressure-volume-temperature correlations for Gulf of Mexico crude oils. SPE Res. Eval. Eng., 416-420 (1998).

Rasouli, H., Rashidi, F. and Ebrahimian, A., Estimating the bubble point pressure and formation volume factor of oil using artificial neural networks. Chem. Eng. Technol., 31, 493 (2008).

Searson, D. P., GPTIPS: Genetic Programming & Symbolic Regression for MATLAB. http://gptips.sourceforge.net (2009).

Searson, D. P., Leahy, D. E. and Willis, M. J., GPTIPS: An open source genetic programming toolbox for multigene symbolic regression. IMECS, March 17-19, Hong Kong (2010).

Shojaei, M-J., Bahrami, E., Barati, P. and Riahi, S., Adaptive neuro-fuzzy approach for reservoir oil bubble point pressure estimation. J. Nat. Gas. Sci. Eng., 20, 214 (2014).

Standing, M. B., A Pressure-volume-temperature correlation for mixtures of California oils and gases. API Drill Prod. Pract., 275-287 (1947).

Vazquez, M. and Beggs, H. D., Correlations for fluid physical property prediction. J. Pet. Technol., 32, 968 (1980).