Computers & Geosciences 44 (2012) 109–119

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Hybrid soft computing systems for reservoir PVT properties prediction

Amar Khoukhi *

Systems Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia

ARTICLE INFO

Article history:
Received 25 April 2011
Received in revised form
15 March 2012
Accepted 18 March 2012
Available online 1 April 2012

Keywords:
Pressure–Volume–Temperature
Oil formation volume factor
Bubble point pressure
Correlation
Genetically-optimised neural networks
Genetic-adaptive neuro-fuzzy inference systems

ABSTRACT

In reservoir engineering, the knowledge of Pressure–Volume–Temperature (PVT) properties is of great importance for many uses, such as well test analyses, reserve estimation, material balance calculations, inflow performance calculations, fluid flow in porous media and the evaluation of new formations for the potential development and enhancement oil recovery projects. The determination of these properties is a complex problem because laboratory-measured properties of rock samples ("cores") are only available from limited and isolated well locations and/or intervals. Several correlation models have been developed to relate these properties to other measures which are relatively abundant. These models include empirical correlations, statistical regression and artificial neural networks (ANNs). In this paper, a comprehensive study is conducted on the prediction of the bubble point pressure and oil formation volume factor using two hybrid of soft computing techniques; a genetically optimised neural network and a genetically enhanced subtractive clustering for the parameter identification of an adaptive neuro-fuzzy inference system. Simulation experiments are provided, showing the performance of the proposed techniques as compared with commonly used regression correlations, including standard artificial neural networks.

1. Introduction

1.1. Motivation and background

A reservoir is a volume of porous sedimentary rock, which has been filled with a substantial amount of hydrocarbons, such as crude oil and natural gas. High-quality rock and fluid data are critical for reliable modelling, reservoir-engineering calculations, and performance predictions by use of reservoir simulators and for subsequent economic analysis (Al-Hussainy and Humphreys, 1996).

Typically, reservoir properties consist of a set of parameters which are used to characterise the spatially varied geological information. Fluid characterisation quantifies the reservoir phase behaviour, fluid compositional changes throughout the reservoir, and changes in fluid properties as a result of production and injection processes.

The determination of reservoir properties is however a complex problem, because laboratory-measured properties of rock samples ("cores") are only available from limited and isolated well locations and/or intervals. There is a great demand for the development of correlation models to relate these properties to other, more relatively abundant, measures. One example of such a kind of measures is "well logs", which are a series of multi-type digital measurements along the vertical depth of drilled wells. These models are used to transform the well log data into reservoir properties at locations where no cores are available.

Pressure–Volume–Temperature (PVT) properties are key parameters associated with the characterisation of any hydrocarbon reservoir. In fact, it is not possible to have accurate solutions to many petroleum engineering problems without having accurate estimates of these properties. Furthermore, the control of the relationship between the surface and reservoir hydrocarbon volumes and the underground withdrawal is made possible by knowing the oil PVT parameters.

To overcome the abovementioned problems of expensive laboratory-based tests and the possible non-availability of the right samples from the well-bore or well surface, equations of state models and several empirically derived correlations between those properties and well data, and eventually artificial neural networks, have been developed based on available data for different regions of the world.

In this paper, the capabilities of soft computing techniques are explored and investigated by gathering their respective powerfulness in hybrid optimised structures, so as to improve their performance compared to the use of each technique alone. A case study of two important PVT properties prediction problem is considered. These two properties are bubble point pressure (P_b), and oil formation volume factor (B_{ob}). The former defines the pressure at which the first gas comes out of solution in oil. It is a function of temperature, oil gravity, gas gravity and gas/oil ratio,

* Tel.: +966 3 860 7614; fax: +966 3 860 2965.
E-mail address: amar@kfupm.edu.sa

0098-3004/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2012.03.016

Nomenclature

EOS	Equation-of-state
FVF	Formation volume factor
PVT	Pressure/Volume/Temperature
API	American Petroleum Institute
ANN	Artificial neural networks
RBFNN	Radial basis feed-forward neural networks
GA	Genetic Algorithm
ANFIS	Adaptive neuro-fuzzy inference system
GANFIS	Genetic adaptive neuro-fuzzy inference system
GONN	Genetically optimised neural networks
Er	Average percent relative error
Ea	Average absolute percent relative error
$E_{\min}$	Minimum root mean square error
$E_{\max}$	Maximum root mean square error
RMSE	Root mean square error
SD	Standard deviation
R^2	Correlation coefficient
Res_Temp	Reservoir temperature (°F)
Mol_N_2	Mole fraction of N_2 (mol%)
Mol_CO_2	Mole fraction of CO_2 (mol%)
Mol_H_2S	Mole fraction of H_2S (mol%)
p	Pressure (psi)
P_b	Bubble point pressure (psi)
P_{od}	Pressure at dead oil viscosity (psi)
B_{ob}	Oil formation volume factor (RB/STB)
R_s	Solution gas/oil ratio, SCF/STB (m^3/m^3)
R_{sb}	Bubble point solution gas/oil ratio, SCF/STB (m^3/m^3)
T	Temperature (°F)
V	Volume (m^3)
μ_a	Viscosity above bubble point (cP)
μ_b	Viscosity below bubble point (cP)
μ_o	Oil viscosity (cP)
μ_{ob}	Bubble point/gas-saturated oil viscosity (cP)
μ_{od}	Dead oil viscosity (cP)
γg	Gas relative density (API)

and it is needed to know when the gas is dissolved from oil. B_{ob} is a function of pressure and temperature, which represents the ratio of oil and dissolved gas volume under reservoir conditions to the oil volume at standard conditions.

1.2. Correlation-based prediction

Since the pioneering work of Katz (1942), who developed five methods for predicting crude oil shrinkage in 1942, a considerable volume of literature has accumulated on the study of PVT correlations for crude oil. Various correlations were proposed by different researchers from all over the world, with varying degrees of accuracy in terms of average error, and based on crude samples from different oil fields (Katz, 1942; Glaso, 1980). Most of these correlations were developed empirically using graphical or regression methods. Several efforts were also made to develop a universal correlation.

In Al-Marhoun (1988), a correlation was published for B_{ob} using 11,728 experimental values representing samples taken from 700 reservoirs, mostly from the Middle East and North America. Other correlations were also developed for the Gulf of Suez and Malaysian Crude oils by Al-Marhoun (1988), Macary and El-Batanoney (1992), respectively. In Omar and Todd (1993), a reliability analysis was conducted for PVT correlations. In De Ghetto et al. (1995), correlations were published for properties of UAE crude oils using 62 data sets from UAE reservoirs. A comprehensive review of various empirical correlations is reported in Almehaideb (1997) and Sutton (2008).

1.3. Artificial neural networks based prediction

Several artificial neural network (ANN) correlations have also been proposed. The prediction performance was assessed through four main criteria. The first is the average percent relative error (Er), which measures the relative deviation of the results of the prediction from the experimental data. The second is the average absolute percent relative error (Ea), which measures the relative absolute deviation of the prediction results from the experimental values. The standard deviation (SD) measures the variability of the obtained results with respect to the average value. The forth criterion consists of the correlation coefficient R^2, which indicates the degree of success in reducing the standard deviation by regression analysis.

In Elsharkawy (1998) a two hidden layers radial basis function neural network model (RBFNM) was proposed to predict PVT properties for crude oil and gas. The model predicts oil formation volume factor, solution gas oil ratio, oil viscosity, saturated oil density, under-saturated oil compressibility, and evolved gas gravity. The input data to the RBFNM were reservoir pressure, temperature, stock tank oil gravity and separator gas gravity. For the oil formation volume factor, the RBFNM resulted in a better accuracy than as-to-date available PVT correlations. In this study, the average percent relative error (Er) was estimated for the training to -0.06% and 0.08% for testing. The average absolute percent relative error (Ea) was obtained as 0.87% for training and 0.53% for testing. The standard deviation (SD) was found to be 1.28% for training and 0.57% for testing. The correlation coefficient R^2 was determined as 99.46% for the training samples and 98.24% for testing.

In Varotsis et al. (1999), a novel approach for predicting the complete PVT behaviour of reservoir oil and gas condensates was introduced using an ANN. The method used key measurements that can be performed rapidly, either in a lab or at the well site, as input to the ANN. The ANN was trained by a PVT database of over 650 reservoir fluids originating from all parts of the world. The testing of the trained ANN indicated that, for all fluid types, most of the PVT properties estimates can be obtained with a very low mean relative error of 0.5–2.5%, with no data set producing a relative error of more than 5%. This level of error is considered as better than that provided by tuned equation-of-state (EOS) models, which are still in common use for the estimation of reservoir fluid properties.

In Al-Marhoun and Osman (2002), two new models were presented to predict P_b and B_{ob} at the bubble-point pressure using a Multi-layer Preceptron (MLP), trained by backpropagation with early stopping, for Saudi Arabian crudes. Both models were based on ANNs, and developed using 283 unpublished data sets collected from different Saudi Arabian fields. Off the 283 data sets, 142 were used to train the B_{ob} and P_b ANNs, 71 to cross-validate the relationships established during the training process and adjust the calculated weights, and the remaining 70 were used to test the model and evaluate its accuracy. The results showed that the developed B_{ob} model provides better predictions and higher accuracy than previously published empirical correlations.

In Goda et al. (2003) another ANN correlation was developed to predict both P_b and B_{ob} with the aid of two separate networks.

The data used was a set of 160 measured points collected from the Middle East, where 120 points were dedicated to training, and 20 for testing. The Bubble point network was constructed of two hidden layers, with ten neurons for each layer. All hidden neurons were activated by a log sigmoid function. The four input data were temperature, API gravity, gas oil ratio and gas relative density. The output neuron was designed to be activated with pure linear functions. The results showed that the network gives higher accuracy in the prediction task than other published empirical correlations. The network has an average percent relative error of 0.030704 and correlation coefficient of 0.9981. Other correlations were introduced in Labedi (1990) and Suttan and Farshad (1990) for Libyan and Gulf of Mexico crude oils.

In Osman and Al-Marhoun (2005), two new models were developed to predict different brine properties. The first model predicted brine density, oil formation volume factor (B_{ob}), isothermal compressibility as a function of pressure and temperature and salinity. The second model was developed to predict brine viscosity as a function of temperature and salinity alone. The models were developed using 1040 published data sets. These data were divided into three groups: training, cross-validation and testing. Radial Basis Functions (RBF) and Multi-layer Preceptron (MLP) neural networks were utilised in the study. Trend tests were performed to ensure that the developed model would follow physical laws. The results showed that the developed models outperform the published correlations in terms of absolute average percent relative error, correlation coefficient and standard deviation.

1.4. Hybrid soft computing based techniques

Soft and mimetic computing techniques other than ANN were also used (Ouenes, 2000; Jang et al., 1996). In Ouenes (2000), fuzzy neural networks were used to evaluate the hierarchical effect of each geologic driver on fractures, in an effort to assist a geologist or reservoir engineer in being able to identify, locally and globally, the key geologic drivers affecting fractures. Aulia et al. (2010) used data mining for smart oilfield reservoir analysis. In El-Sebakhy (2009), a support vector regression (SVR) was used for generating correlations for forecasting bubble point pressure using three different published PVT databases, which was claimed to outperform empirical correlation and standard neural network models.

All these studies prove that correlations based on data mining techniques are more accurate than empirical correlations. However, most of these correlations were found to be appropriate for the specific region where the parameters were measured, but not for other regions. Furthermore, the neural network correlations developed are often limited, and global correlations are usually less accurate compared to local correlations. Nevertheless, the achievements of neural networks opened the door to soft computing techniques to play a major role in the oil and gas industry. Adaptive neuro-fuzzy inference systems have been proposed as a new intelligence framework for both prediction and classification based on fuzzy clustering optimisation criterion and ranking (Jang et al., 1996).

This paper is an upgraded and extended version of two other papers (Khoukhi and Albukhitan, 2010, 2011; Khoukhi et al., 2011) that are concerned with estimating two PVT properties of crude oil systems; namely bubble point (P_b) and oil formation volume factor (B_{ob}). The main extension is the performance of further experiments, including a genetically optimised neural network (GONN), and a genetic adaptive neuro-fuzzy inference system (GANFIS), along with comparisons with standard ANN, ANFIS and state-of-the-art regression based correlations.

In Section II, the problem is formulated and the datasets used are introduced. Section III presents the proposed approach to the problem and the implemented techniques. Section IV reports on the simulation experiments and a comparative study. Section V concludes this work and offers perspectives and future trends.

2. Problem statement

2.1. Data acquisition and pre-processing

Three distinct databases were used in this study to implement the proposed techniques, to forecast both bubble point pressure (P_b) and oil formation volume factor (B_{ob}) based on the same four input parameters; namely solution gas–oil ratio (Rs), reservoir temperature (T), oil gravity (API), and gas relative density (γg). The properties of the three datasets acquired are explained below; a statistical description of the datasets appears in Appendix I:

– The first dataset refers to work conducted by Al-Marhoun, who published correlation for estimating P_b and B_{ob} for Middle Eastern oils (Al-Marhoun, 1988). The dataset consists of 160 observations collected from 69 Middle Eastern reservoirs.
– The second dataset were retrieved from works conducted by Al-Marhoun and Osman (2002). The dataset consists of 283 observations collected from different Saudi Arabian oil fields for the prediction of P_b and B_{ob} at P_b for Saudi crude oils.
– The third dataset was obtained from the work conducted by Osman and Al-Marhoun (2005), this dataset contains 782 observations collected from oil fields in Malaysia, the Middle East, Gulf of Mexico and Columbia.

2.2. Statistical quality measures

To compare the performance and accuracy of the proposed framework to other empirical correlations, statistical error analysis and quality measures are performed. Three main standards are used in assessing the suggested techniques and their comparisons with regression-based models.

The first standard is an assessment through six error criteria; including the average percent relative error (Er), average absolute percent relative error (Ea), minimum and maximum absolute percent error ($E_{\min}$ and $E_{\max}$), root mean square errors (RMSE), standard deviation (SD), and correlation coefficient (R^2). These performance indices are detailed in Appendix II.

The second comparison is through graphical representation of the correlations between the actual and predicted values of P_b and B_{ob}.

The third approach is through graphical representation of errors as a function of correlations (Al-Marhoun, 1988, 1992; Al-Marhoun and Osman, 2002). Scatter plots are shown for the absolute percent relative error (Ea) versus the correlation coefficient for all computational intelligence forecasting schemes and the most common empirical correlations. Each modeling scheme is represented by a symbol; the good forecasting scheme should appear in the upper left corner of the graph.

3. Proposed approach

3.1. Adaptive neuro-fuzzy inference systems (ANFIS)

Neuro-fuzzy inference systems are hybrid classification/forecasting frameworks, which learn the rules and membership functions from data. It is a network of nodes and directional

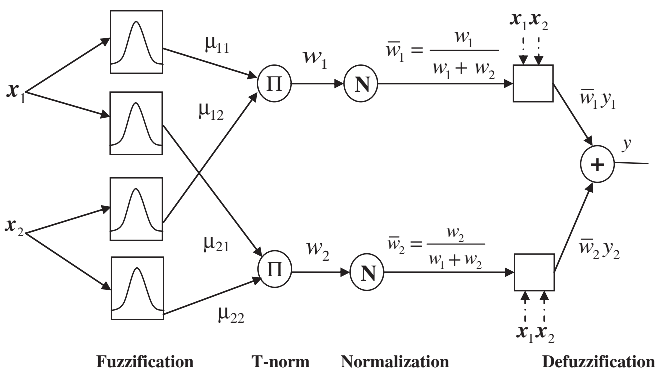

Fig. 1. ANFIS system with two inputs two-rule one-output.

links. Associated with the network is a learning rule, for instance, back-propagation. These networks are learning a relationship between inputs and outputs. This type of network covers a number of different approaches, namely Mamdani type and Takagi–Sugeno–Kang (TSK) type (see Jang et al., 1996) for more detail. Unlike the Mamdani method, also referred to as subjective fuzzy modelling as it builds the fuzzy if-then rules through expert statements which might involve vagueness and subjectivity, the TSK fuzzy objective modelling method is a framework for generating fuzzy if-then rules from input/output numerical data. A way to construct a TSK fuzzy model from numerical data proceeds in three steps: fuzzy clustering, setting of the membership functions, and parameter estimation (Nikravesh et al., 2003; Jang et al., 1996). Fig. 1 shows an ANFIS System with a two-inputs two-rules one-output arrangement.

The implemented ANFIS in the study at hand is made up of six layers. The first layer is the input layer, characterising the crisp inputs. The second layer performs the fuzzification of the crisp inputs into linguistic variables, through Gaussian transfer functions. The third is the rule layer, which applies the product *t*-norm to produce the firing strengths of each rule. This is followed by a normalisation layer, at which each node calculates the ratio of a rule's firing strength to the sum of the firing strengths of all rules. The fifth layer performs the defuzzification. The last layer conducts the aggregation, where an output is obtained as the summation of all incoming signals. The training rule option used is the Levenberg–Marquard version of the gradient back-propagation algorithm.

3.2. *Genetic adaptive neuro-fuzzy inference system (GANFIS)*

Genetic Algorithms (GAs) are intelligent search mechanisms based on the principle of natural selection and population genetics that are transformed by three genetic operators: selection, crossover and mutation. Each string (chromosome) is a possible solution to the problem being optimised, and each bit (or group of bits) represents a value or of some variable (gene) of the problem. These solutions are classified by an evaluation function, giving better values, or fitness, for better solutions individuals. Each solution must be evaluated by the fitness function to produce a value. Different crossover and mutation rates are used for the optimisation using genetic algorithms. The ability of genetic programming and adaptive neuro-fuzzy inference system (ANFIS) techniques were considered for ground water depth forecasting (Shiri and Kişi, 2010).

To reduce the number of rules that are generated during ANFIS implementation, different clustering methods have been proposed, e.g., subtractive clustering and fuzzy-C means (Jang et al., 1996). Subtractive clustering is used in the problem at hand as it provides a fast one-pass algorithm to take input-output training data to generate an adaptive fuzzy inference system that is better tailored to the dataset. The subtractive clustering method assumes each data point is a potential cluster centre, and calculates a measure of likelihood that each data point will define a cluster centre, based on the density of the surrounding data points. It starts with the normalisation of all values in the dataset to fit in a hypercube unit (Jang et al., 1996). Each cluster centre may be translated into a fuzzy rule for identifying a class. The cluster radius indicates the range of its influence. Specifying a small cluster radius yields many small clusters in the data, thus resulting in many rules. Conversely, choosing a large radius will lead to few rules, misrepresenting therefore the data. Many trials considering different radii values have led to different prediction accuracies. A genetic algorithm (GA) is used to fine-tune the clustering parameter radii (ra), which then leads to the genetic adaptive neuro-fuzzy inference system (GANFIS). An important literature had been published recently on genetic fuzzy systems (GFS). Zanganeh et al., 2006 used a hybrid genetic fuzzy inference system for wave parameters prediction.

The step-by-step implementation of GANFIS is clearly defined and given as follows (Fig. 2). For the training data set, each radius has a value between a lower bound of 0.10 and an upper bound of 0.90. The fitness function used for the genetic search is the least squared error between the predicted and target value for the output variable.

A population initialisation of size N and a termination criterion are first defined. The termination is either the maximum number of generations or errors. Then the number n of radii are generated randomly radii$=[r_1, r_2, \ldots, r_n]$ in the solution space. Note here radii's are synonymous with chromosomes, also called individuals. Therefore each population now has $(N \times n)$ chromosomes.

The whole population in the generation is evaluated through ANFIS, using the root mean squared error of the testing data as a criterion of best fit. Candidate chromosomes for the new generation are selected from the population, using a tournament selection procedure. Then, crossover and mutation operations are performed

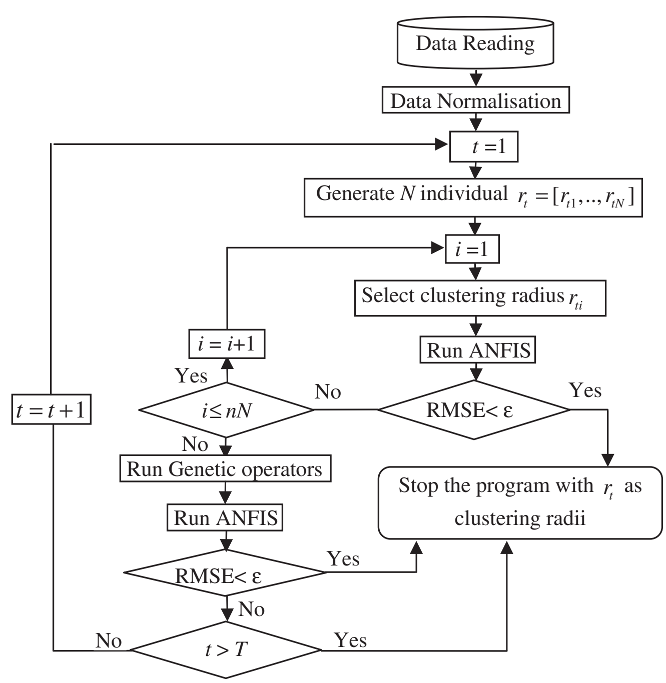

Fig. 2. Flowchart of the proposed GANFIS approach.

on the selected chromosomes. The $N \times n$ best individuals are selected from the parent and child populations (elitism). These new individuals will form the parent population for the next generation. The process is repeated until the termination criterion is met.

Other GA algorithm settings consist of the number of generations (20), and population size (10). Population type (double), Selection function: (Stochastic uniform), Crossover function: (Scattered with crossover rate: 0.80), Mutation function: (Gaussian with mutation rate: 0.20). The other "options" of ANFIS are set to Matlab default. The results for GANFIS show a significant improvement in both the bubble point pressure (P_b) and oil formation volume factor (B_{ob}) predictions than when the trial-and-error selection method was used for the radii.

3.3. Genetically optimized neural networks (GONN)

At this stage we consider a hybrid genetic neural network for the P_b and B_{ob} (at P_b) prediction problem.

Recently, several hybrids of genetic algorithm with ANNs were proposed [Balan et al., 1996]. In Oloso et al. (2009) a differential evolutionary artificial neural network was introduced for predicting viscosity and gas/oil ratio curves.

For the problem at hand, a Levenberg Marquardt Back Propagation (LMBP) ANN is first considered to predict B_{ob} and P_b with a two hidden layers feed forward neural network and both linear and sigmoid activation functions were used. The best network outputs of 1000 runs were taken in each case. Fig. 3 shows how the fitness evaluation is performed in the case of neural network optimisation. A bit-string from the population is decoded into the description of the network weights. The fitness evaluation of this neural network is set to measure the prediction performance. This performance determines the selection probability for the respective chromosome. Then this chromosome is manipulated by genetic operators to form the next generation of potential solutions. The genetic operators include mutation, which randomly changes the bit-positions of an individual, and crossover, which exchanges sub-strings between the same positions of different individuals. The GA search stops if performance is not improved any further after a number of generations, or if the population of bit-strings converges to (nearly) identical patterns. Matlab functions were used for building the GONN model.

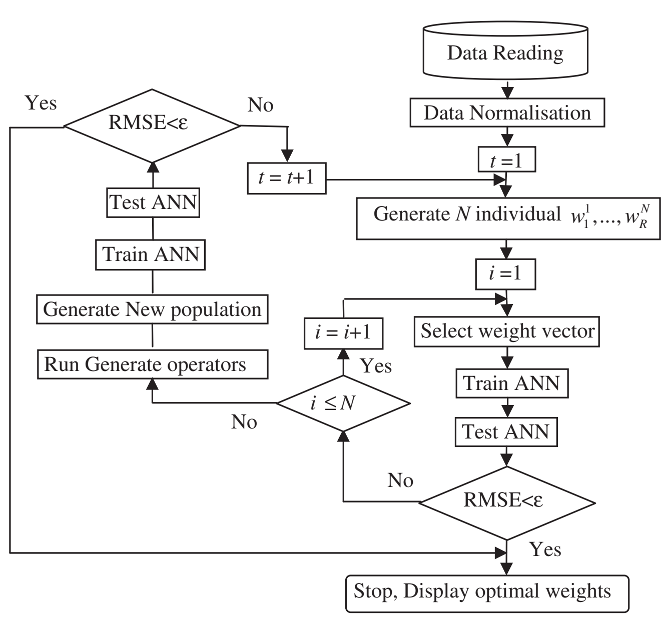

Fig. 3. Flowchart of the proposed GONN approach.

4. Simulation experiments and comparative studies

4.1. Introduction and experimental set up

The capabilities of GONN and GANFIS for bubble point pressure and oil formation volume factor prediction are investigated using the above provided datasets in Section 2. The study includes a comparison with a standalone ANN and ANFIS, along with three common published empirical correlations, namely Standing (1947), Glaso (1980), and Al-Marhoun (1992). These correlations are outlined for quick reference in Appendix III.

To evaluate the performance of each scheme, the entire database is divided using random selection. Both internal and external validation processes are repeated 100 times. Therefore, out of the 782 data points, 382 are used to train the hybrid neural network models, 200 are used to cross-validate the relationships established during the training process, and 200 to test the model by evaluating its accuracy and trend stability. For testing data, a statistical summary to investigate the different quality measures corresponding to the genetic neuro-fuzzy system and genetic neural network are provided.

For the implemented neural network we follow the same procedures as in Al-Marhoun and Osman (2002), Osman and Al-Marhoun, (2005), and Khoukhi and Albukhitan (2010, 2011). The initial weights were generated randomly, and the learning technique is achieved based on 1000 epochs or 0.001 goal error and 0.01 learning rate. For both models, the first layer consists of four neurons representing the input values of the reservoir temperature, solution gas–oil ratio, gas specific gravity and API oil gravity. The second (hidden) layer consists of seven neurons for the P_b model, and eight neurons for the B_{ob} model. The third layer contains one neuron representing the output values of either P_b or B_{ob}. Each layer contains neurons that are connected to all neurons in the neighbouring layers. The connections have numerical values (weights) associated with them, which will be adjusted during the training phase. Training is completed when the network is able to predict the given output. The tangent sigmoid was chosen as the transfer function to propagate the signal through the different layers of the network. This gives the ability to monitor the generalisation performance of the network and prevents the network to over fit the training data based on repeating the computations 1000 times and taking the average of all runs. The architecture of the abovementioned ANN was retained during the implementation of GONN. The fitness function used for the genetic search is the least squared error between the predicted and target values for the output variable.

For GANFIS, the implementation process starts by optimising the subtractive clustering radii using a genetic algorithm. Then, the clustering results are used to build the initial neuro-fuzzy inference system from the available input data sets using a Sugeno-type FIS system. The ANFIS parameters are then tuned through several passes in the available training data.

4.2. Simulation results and comparative studies

For the purpose of comparison, we use the same feature selection criterion and stratified sampling cross-validation schemes. Based on the results obtained within the external validation checks (testing and validation), the simulation results show that the genetic neuro-fuzzy scheme and genetic neural network, are very competitive, while both outperform standalone neural network and ANFIS and many common empirical correlations methods. In addition, these schemes have a high accuracy in predicting P_b and B_{ob} values with stable performances and achieved the lowest absolute percent relative errors, lowest minimum errors, lowest maximum errors, lowest root mean square errors, and highest correlation coefficients among the other correlations for the three distinct databases used. Nevertheless, because the three datasets used are from different locations and reservoirs, the impact produced on the prediction performance of each dataset is independent and not significantly correlated to the other. Henceforth, the results of the present study are discussed for each data set. Moreover, as noted in the introduction, most of the empirical correlations developed, or those developed using graphical or regression methods, are only applicable to a specific region and data set, effort is still being made to come up with a universal correlation.

Figs. (4)–(8) illustrate a sample of the scatter plots of the predicted results versus the experimental data for P_b and B_{ob} values using the distinct data sets provided at training and testing stages. These cross-plots indicate the degree of agreement between the actual (experimental) and the predicted values. For an ideal model, all points on the plot should appear on a line of angle of 45° with respect to the x-axis, with equidistant points from both axes and with a correlation coefficient equal to 1.

Also as it is clear from the Figures, that the regression models have not good generalization capabilities as compared to soft computing based models. This fact was also reported in a previous study by Al-Marhoun (1988, 1992), Al-Marhoun and Osman (2002). They noticed especially that the developed regression models using Middle East crude oils data, when tested on new data from different crudes, they fail to provide with acceptable performance.

Fig. 9 shows a scatter plot of Ea versus R^2 at the training stage for all modeling schemes that are used to determine B_o based on the data set used in Osman and Al-Marhoun (2005). We observe

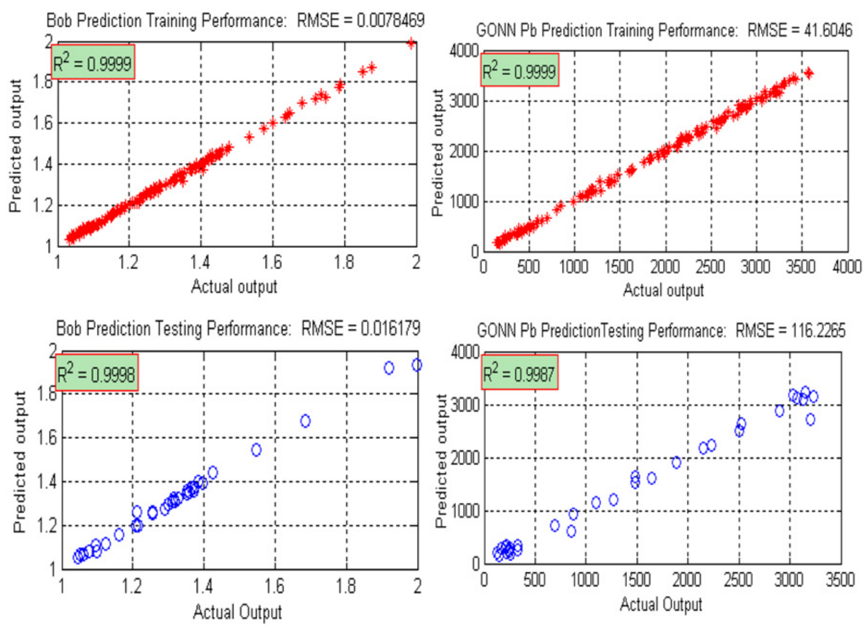

Fig. 4. Correlation coefficient and cross-plot for GONN B_{ob} (left) and P_b (right) prediction (1st dataset).

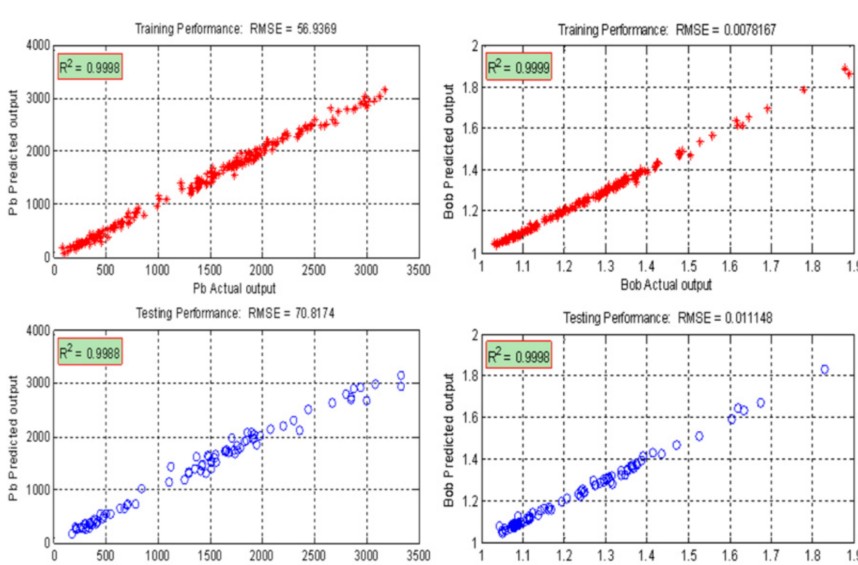

Fig. 5. Correlation coefficient and cross-plot of GANFIS for B_{ob} and P_b (1st dataset).

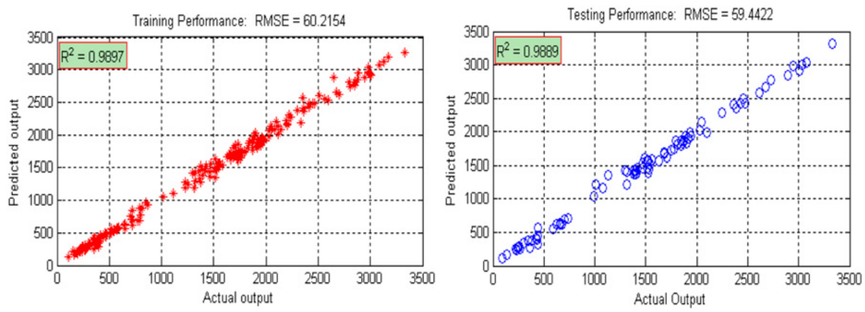

Fig. 6. P_b Correlation coefficient cross-plot for GANFIS (2nd dataset).

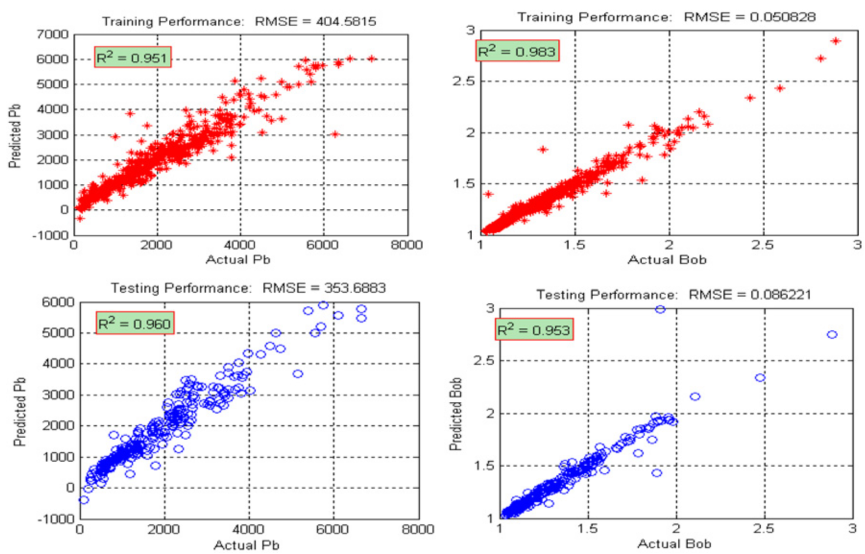

Fig. 7. Correlation coefficient and cross-plots of ANN for B_{ob} and P_b (1st dataset).

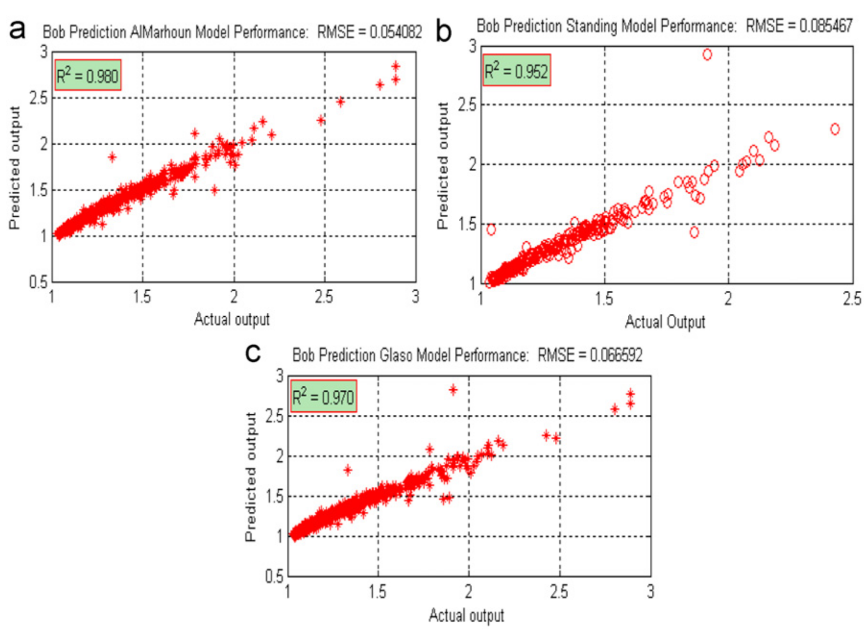

Fig. 8. B_{ob} Correlation coefficient and cross-plot of regression models. (top to bottom: (a) Al-Marhoun (b) Standing, (c) Glaso) (1st dataset).

that the symbol corresponding to the GONN scheme falls in the upper left corner with Ea=0.10015% and R^2=0.9999, while GANFIS very close to GONN with Ea=0.1013% and R^2=0.9998; whereas standalone ANN got Ea=1.7886% and R^2=0.9878. Empirical correlations indicate higher error values with lower correlation coefficients, such as, Standing with Ea=2.1923% and R^2=0.9874; Al-Marhoun with Ea=2.1037% and R^2=0.9846; and Glaso Correlation with Ea=2.8173% and R^2=0.97813.

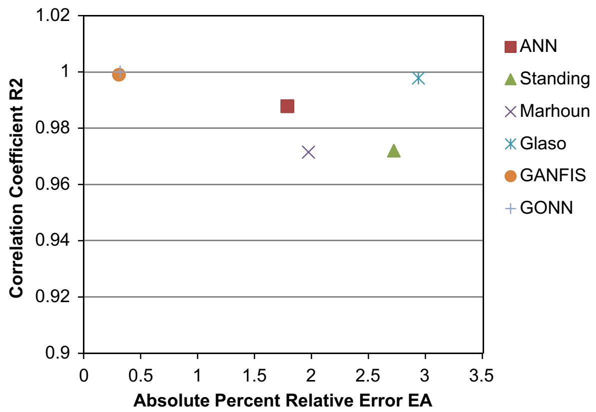

Fig. 9. Performance of different techniques (3rd dataset Correlation coefficient vs. absolute percent relative error).

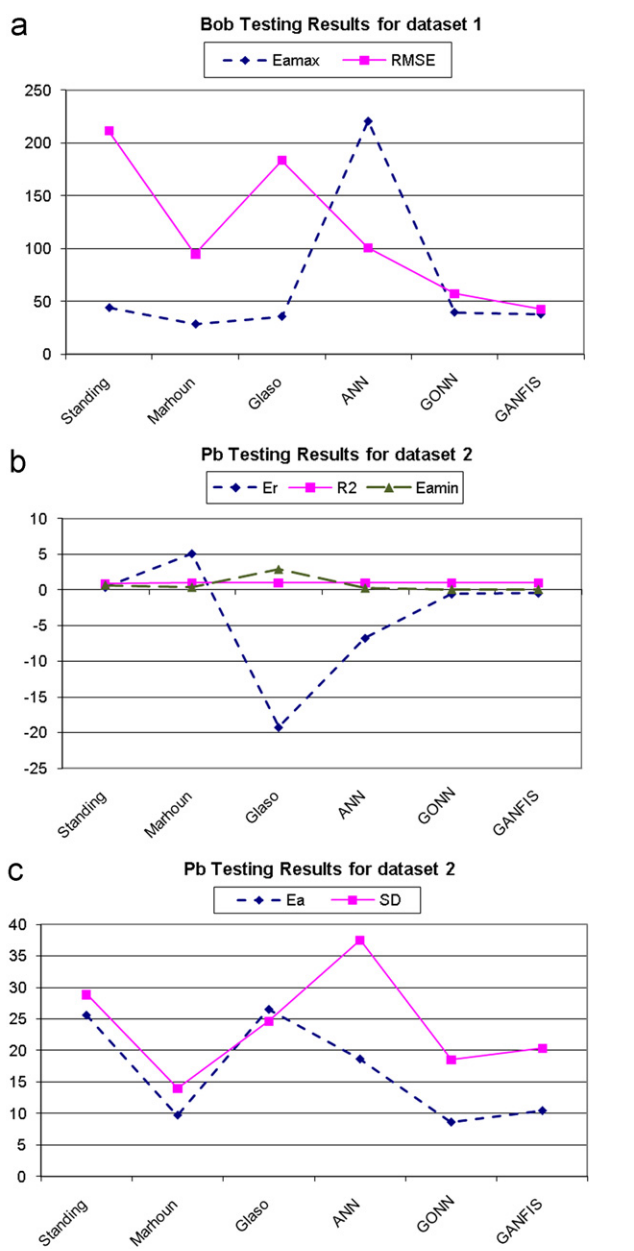

Fig. 10. Performance criteria obtained using dataset 2 for the eight methods for bubble point pressure (P_b).

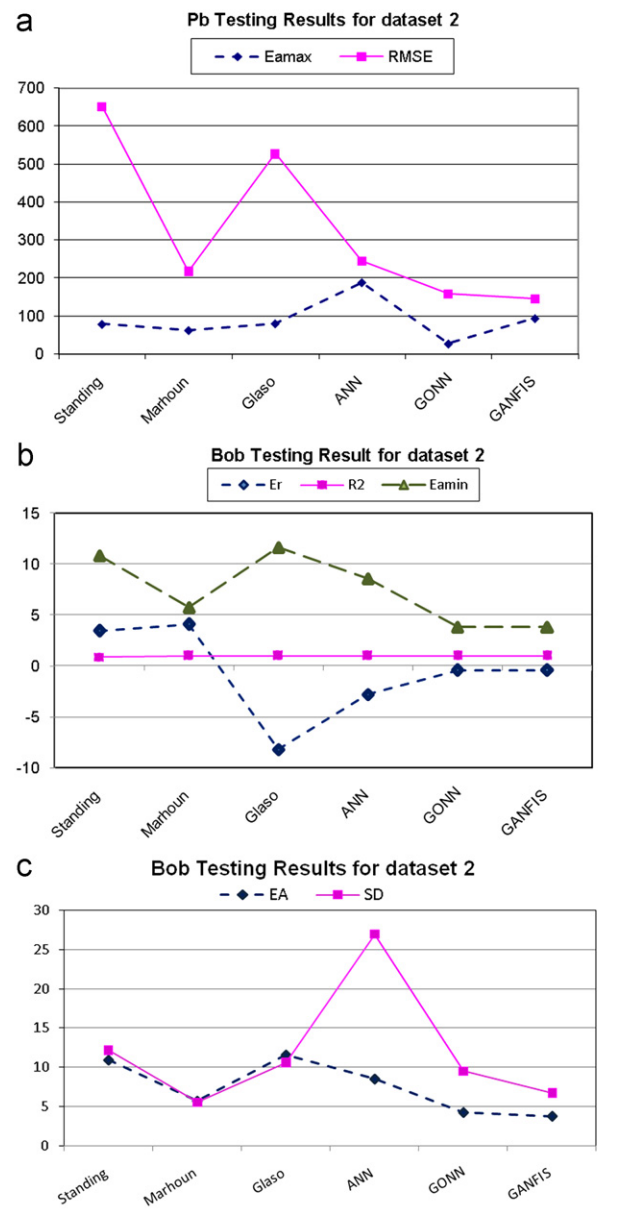

Fig. 11. Performance criteria obtained using dataset 2 for the eight methods for oil formation volume factor (B_{ob}).

Table 1
Time complexity of all models for P_b Prediction (1st dataset).

Model	Training	Testing
Cputime (s)		
Standing	34.819	0.0682
Marhoun	15.254	0.0782
Glaso	19.025	0.072
ANN	724.54	0.088
GONN	~57600	0.0182
GANFIS	~69200	0.0096

Figs. 10 and 11 show different error criteria obtained using dataset 2 for the six methods for bubble point pressure (P_b) and oil formation volume factor (B_{ob}), respectively. The figures show clearly that the GANFIS and GONN are very competitive and produce the best performance, as compared to standalone ANN and common empirical correlations including the correlations by Al-Marhoun (1988), Osman and Al-Marhoun (2005), and Al-Marhoun and Osman (2002)) and multi-layer back-propagation neural networks. The six considered error criteria, include the average percent relative error (Er), average absolute percent relative error (Ea), minimum and maximum absolute percent error (E_{min} and E_{max}), root mean square errors (RMSE), standard deviation (SD), and correlation coefficient (R^2).

Finally, regarding time complexity, Table 1 shows the computational processing time for all models for the P_b prediction. It is clear that GANFIS and GONN are very demanding in computational time, as compared to stand alone ANN and regression based models. However this does not jeopardize usage as these computations are performed offline in the development stage, and once the parameters of GANFIS and GONN are optimized these structures are then used for online calculations with significantly lower computational time as compared to ANN, but with better prediction accuracy.

5. Conclusion and future work

In this paper, a comprehensive study is conducted on the prediction problem of Pressure–Volume–Temperature (PVT) properties, namely bubble point pressure and oil formation volume factor. Several soft computing techniques are explored; including artificial neural networks, and adaptive neuro-fuzzy inference systems. The paper also presented two hybrids; a genetically optimised neural network and a genetically enhanced subtractive clustering technique for parameter identification of the adaptive neuro-fuzzy inference system.

Three distinct published databases were utilised to investigate the capabilities of the soft computing techniques. Based on simulations, and by comparing the results obtained, one can conclude that, the genetic neural network and the genetic neuro-fuzzy inference system schemes showed better performance in predicting P_b and B_{ob} values with stable performance, and achieved the lowest absolute percent relative error, lowest minimum error, lowest maximum error, lowest RMSE and the highest correlation coefficient R^2 compared to the many existing empirical correlations used for the three distinct data sets. The plan is to push this hybrid computing further by incorporating some recent advances in soft computing, like optimized support vector regression and extreme learning machines and granular computing (Khoukhi et al., 2011; Zhu et al., 2005; Xu and Shu, 2006; Schölkopf and Smola, 2002; Yu and Pedrycz, 2009) and other intelligent search techniques. Some these hybrids are being tested in ongoing works. The resulting massive computational time required to implement these hybrids will be tackled using the High Performance Computing (HPC) facilities available to achieve higher performance with a reasonable time complexity. Furthermore, the proposed hybrid evolutionary soft computing techniques are flexible, reliable and can be implemented for other related oil and gas industry problems, especially in the prediction of permeability and porosity, history matching, data management and rock mechanics properties. Some of these tasks are being undertaken in ongoing works.

Acknowledgments

This work is supported by King Fahd University of Petroleum and Minerals under Grant no. SB100014.

Appendix I. Statistical description of datasets used

See Tables (A.1)–(A.3).

Table A.1
Statistical description of dataset 1 used for PVT models (160 records).

Parameter	Min	Max	Average	SD
Input variables				
Temperature (Tf), (°*F*)	74	240	144.43	39.036
Gas–oil ratio (*Rs*), (SCF/STB)	26	1602	557.66	403.12
Gas relative density (γg)	0.69	1.367	0.96417	0.17103
Api oil gravity, (degrees api)	19.4	44.6	32.388	5.7444
Output variables				
Bubble point pressure (P_b), (psi)	130	3573	1731.1	1084.6
Oil FVF at P_b, (RB/STB)	1.032	1.997	1.3036	0.20564

Table A.2
Statistical description of dataset 2 used for PVT models (283 Records).

Parameter	Min	Max	Average	SD
Input variables				
Temperature (Tf), (°F)	75	240	147.35	47.772
Gas–oil ratio (*Rs*) (SCF/STB)	24	1453	432.46	303.58
Gas relative density (γg)	0.7527	1.8195	1.008	0.14826
API oil gravity (degrees API)	17.5	44.6	31.622	5.2518
Output variables				
Bubble point pressure (P_b, psi)	90	3331	1390.2	860.66
Oil FVF at P_b, (RB/STB)	1.0308	1.889	1.2504	0.15824

Table A.3
Statistical description of dataset 3 used for PVT models (782 Records).

Parameter	Min	Max	Average	SD
Input variables				
Temperature (Tf), (°F)	58	341.6	181.9	51.984
Gas–oil ratio (*Rs*) (SCF/STB)	8.61	3617.3	541.75	483.68
Gas relative density (γg)	0.511	1.789	0.88825	0.18556
API oil gravity (degrees API)	11.4	63.7	34.588	8.7286
Output variables				
Bubble point pressure (P_b), (psi)	107.33	7127	2006.1	1291.2
Oil FVF at P_b, (RB/STB)	1.028	2.887	1.3362	0.27201

Appendix II. Performance indices

Average percent relative error: Measures the relative deviation from the experimental data, given as:

$$E_r = \frac{1}{n}\sum_{i=1}^{n} E_i \tag{II.1}$$

where E_i is a relative deviation of an estimated value from an experimental value.

$$E_i = \left[\frac{(B_{ob})_{\exp} - (B_{ob})^{est}}{(B_{ob})_{\exp}}\right]_i \times 100 \quad i = 1,2,\ldots n \tag{II.2}$$

Average absolute percent relative error: Measures the relative absolute deviation from the experimental values, defined as:

$$E_a = \frac{1}{n}\sum_{i=1}^{n} |E_i| \tag{II.3}$$

Minimum absolute percent relative error: To define the range of error for each correlation, the calculated absolute percent relative error values are scanned to determine the minimum values. They are defined by:

$$E_{\min} = \min_{i=1}^{n} |E_i| \tag{II.4}$$

Maximum absolute percent relative error: Similarly, the maximum absolute percent relative error is

$$\text{defined as}: E_{\max} = \max_{i=1}^{n} |E_i| \tag{II.5}$$

Root mean squares error: Measures the data dispersion around zero deviation, defined by:

$$\text{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n} E_i^2\right]^{1/2} \tag{II.6}$$

The correlation coefficient: Represents the degree of success in reducing the standard deviation by regression analysis. It is defined by:

$$R = \sqrt{1 - \sum_{i=1}^{n}[(B_{ob})_{\exp} - (B_{ob})_{\text{est}}]_i^2 \Big/ \sum_{i=1}^{n}[(B_{ob})_{\exp} - \overline{B}_{ob}]_i^2} \tag{II.7}$$

where

$$\overline{B}_{ob} = \frac{1}{n}\sum_{i=1}^{n}[(B_{ob})_{\exp}]_i \tag{II.8}$$

Appendix III. Regression models used (Standing, 1947; Glaso, 1980; Al-Marhoun, 1992)

Standing (1947) proposed the following empirical expressions for the formation volume factor and bubble point pressure:

$$B_{bob} = 0.9759 + 0.00012\left[R_s\left(\frac{\gamma_g}{\gamma_o}\right)^{0.5} + 1.25t\right]^{1.2} \tag{III.1}$$

$$P_b = 18.2[(R_s/\gamma_g)^{0.83}(10)^a - 1.4] \tag{III.2}$$

$$a = 0.00091(T({}^\circ{}_R) - 460) - 0.0125(\text{API})$$

Glaso (1980) suggested the following expression:

$$B_{bob} = 1 + 10^A \tag{III.3}$$

where

$$A = -6.58511 + 2.91329\log(B_{ob}^*) - 0.27683(\log(B_{ob}^*))^2 \tag{III.4}$$

With B_{ob}^* given as:

$$B_{ob}^* = Rs\left[\frac{\gamma_g}{\gamma_0}\right]^{0.526} + 0.968(T - 460) \tag{III.5}$$

With respect to the bubble point pressure, he found:

$$\log(P_b) = 1.7669 + 1.7447\log(P_b^*) - 0.30218[\log(B_{ob}^*)]^2 \tag{III.6}$$

$$P_b^* = (R_s/\gamma_g)^a (T)^b (API)^c \tag{III.7}$$

$$a = 0.816, \quad b = 0.172, \quad c = -0.989$$

Al-Marhoun, 1992, determined the following empirical correlation of the formation volume factor with respect to the gas/oil ratio, gas gravity, oil gravity and temperature:

$$B_0 = 0.497069 + 0.862963.10^{-3}T + 0.182594.10^{-2}F + 0.318099.10^{-5}F^2 \tag{III.8}$$

with

$$F = R_s^2\gamma_g^b\gamma_0^c \quad \text{and} \quad a = 0.74239, \quad b = 0.323294, \quad c = -120204 \tag{III.9}$$

He proposed the following for the bubble point pressure:

$$P_b = aR_s^b\gamma_g^c\gamma_0^d T^c \tag{III.10}$$

with $a = 5.38088 \times 10^{-3}$, $b = 0.715082$, $c = -1.87784$,

$d = 3.1437 \quad e = 1.32657$

References

Al-Hussainy, R., Humphreys, N., 1996. Reservoir management principles and practices. Journal of Petroleum Technology 48 (12), 1129–1135.
Al-Marhoun, M.A., 1988. PVT correlations for Middle East crude oils. Journal of Petroleum Technology 40 (5), 650–666.
Al-Marhoun, M.A., 1992. New correlations for formation volume factor of oil and gas mixture. Journal of Canadian Petroleum Technology 31 (3), 22–26.
Al-Marhoun, M.A., Osman, E.A., 2002, Using Artificial neural networks to develop New PVT correlations for saudi crude oils, in 10th Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC), Abu Dhabi, UAE, paper SPE 78592.
Almehaideb, R.A., 1997. Improved PVT correlations for UAE crude oils: SPE Middle East Oil Show and Conference, Manama, Bahrain, V1, SPE 37691, pp. 109–120.
Aulia, A., Keat, T.B., Bin Maulut, M.S., El-Khatib, N., Jasamai, M., 2010. Smart oilfield data mining for reservoir analysis. International Journal of Engineering & Technology 10, 06.
Balan, B., Ameri, S., Mohaghegh, S., 1996. A hybrid neuro-genetic approach to hydraulic fracture treatment design and optimization. SPE 36602.
De Ghetto, G., Paone, F., Villa, M., 1995. Pressure–Volume–temperature correlations for heavy and extra heavy oils. In: SPE International Heavy Oil Symposium, Calgary, Alberta, Canada.
El-Sebakhy, E., 2009. Data mining in forecasting PVT correlations of crude oil systems based on type 1 fuzzy logic inference systems. Computers & Geosciences 35 (9), 1826–1917.
Elsharkawy, A.M., 1998. Modelling the properties of crude oil and gas systems using RBF network. In: SPE Asia Pacific Oil & Gas Conference, Perth, Australia, paper SPE 49961.
Glaso, O., 1980. Generalized Pressure–Volume–Temperature correlations. Journal of Petroleum Technology 32 (5), 785–795.
Goda H.M., Fattah K.A., Shokir E.M., Sayyouh M.H., 2003.Prediction of the PVT data using neural network computing theory. In: 27th Annual SPE International Technical Conference and Exhibition, Abuja, Nigeria, paper SPE 85650.
Jang, S.R., Sun, C.T., Mizutani, E., 1996. Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Prentice-Hall.
Katz, D.L., 1942. Prediction of the shrinkage of crude oils. Drilling and Production Practice, API, 137–147.
Khoukhi, A., Albukhitan, S., 2010.A data driven genetic neuro-fuzzy system to PVT properties prediction. In: Proceedings of the North American Fuzzy Information Processing Society (NAFIPS'10), Toronto, Canada.
Khoukhi, A., Albukhitan, S., 2011. PVT properties prediction using hybrid genetic neuro-fuzzy systems. International Journal of Oil Gas and Coal Technology 4 (1), 47–63.
Khoukhi, A., Oloso, M., Abdulraheem, A., El-Shafei, M., Al-Majed, A., 2011. Support vector machines and functional networks for viscosity and gas/oil ratio curves prediction. International Journal of Computational Intelligence and Applications 10 (3), 269–293.

Labedi, R., 1990. Use of production data to estimate volume factor density and compressibility of reservoir fluids. Journal of Petroleum Science & Engineering 4, 357–390.

Macary, S.M., El-Batanoney, M.H., 1992. Derivation of PVT correlations for the Gulf of Suez crude oils. In: EGPC 11th Petroleum Exploration & Production Conference.

Nikravesh, M., Zadeh, L.A., Aminzadeh, F., 2003. Soft Computing and Intelligent Data Analysis in Oil Exploration, 51. (Developments in Petroleum Science), Elsevier Science.

Oloso, M., Khoukhi, A., Abdulraheem, A., El-Shafei, M., 2009. A genetic-optimized artificial neural networks for predicting viscosity and gas/oil ratio curves.SPE/EAGE Reservoir Characterization and Simulation Conference & Exhibition, Abu-Dhabi, UAE.

Omar, M.I., Todd, A.C., 1993, Development of new modified black oil correlation for Malaysian crudes. In: SPE Asia Pacific Oil and Gas Conference, Singapore, SPE 25338.

Osman, E.A., Al-Marhoun, M.A., 2005. Artificial neural networks models for predicting PVT properties of oil field brines. In: 14th SPE Middle East Oil & Gas Show and Conference, Bahrain, paper SPE 93765.

Ouenes, A., 2000. Practical application of fuzzy logic and neural networks to fractured reservoir characterization. Computers & Geosciences 26 (8), 953–962.

Shiri, J., Kişi, Ö., 2010. Comparison of genetic programming with neuro fuzzy systems for predicting short-term water table depth fluctuations. Computers & Geosciences 37 (10), 1692–1701.

Schölkopf, P., Smola, A.J., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge.

Suttan, R.P., Farshad, F.F., 1990. Evaluation of empirically derived PVT properties for Gulf of Mexico crude oils. SPE Reservoir Engineering 5 (1), 79–86.

Sutton, R.P., 2008. An accurate method for determining oil PVT properties using the Standing–Katz gas *Z*-factor chart. SPE Reservoir Evaluation & Engineering 11 (2), 246–266.

Standing, M.B., 1947A Pressure–Volume–Temperature correlation for mixture of California oils and gases. In: API Drilling and Production Practice, Dallas, pp. 275–287.

Varotsis, N., Gaganis, V., Nighswander, J., Guieze, P., 1999. A novel non-iterative method for the prediction of the PVT behavior of reservoir fluids. In: SPE Annual Technical Conference and Exhibition, Houston, Texas, paper SPE 56745.

Yu, F., Pedrycz, W., 2009. The design of fuzzy information granules: tradeoffs between specificity and experimental evidence. Journal of Applied Soft Computing 9 (1), 264–273.

Xu, Y., Shu, Y., 2006. Evolutionary extreme learning machine based on particle swarm optimization. In: Wang, J., et al. (Eds.), ISNN, LNCS 3971. Springer-Verlag, pp. 644–652.

Zanganeh, M., Mousavi, S., Etemad-Shahidi, A., 2006. A genetic algorithm-based fuzzy inference system in prediction of wave parameters, computational intelligence. Theory and Applications (Part 25), 741–750, http://dx.doi.org/10.1007/3-540-34783-6_72.

Zhu, Q., Qin, A.K., Suganthan, P.N., Huang, G.B., 2005. Evolutionary extreme learning machine. Pattern Recognition 38, 1759–1763.