Advances in Geo-Energy Research

Vol. 3, No. 3, p. 225-241, 2019
www.astp-agr.com

Original article

Reliable predictions of oil formation volume factor based on transparent and auditable machine learning approaches

David A.Wood[1] *, *Abouzar Choubineh*[2]

[1]*DWA Energy Limited, Lincoln, United Kingdom*

[2]*Petroleum University of Technology, Ahwaz, Iran*

(Received March 10, 2019; revised April 10, 2019; accepted April 13, 2019; available online May 2, 2019)

Citation:
Wood, D.A., Choubineh, A. Reliable predictions of oil formation volume factor based on transparent and auditable machine learning approaches. *Advances in Geo-Energy Research*, 2019, 3(3): 225-241, doi: 10.26804/ager.2019.03.01.

Corresponding author:
*E-mail: dw@dwasolutions.com

Keywords:
Machine learning transparency
non-correlation-based machine learning
oil formation volume factor prediction
sparse data impacts
avoidance of overfitting

Abstract:
Neural-network, machine-learning algorithms are effective prediction tools but can behave as black boxes in many applications by not easily providing the exact calculations and relationships among the underlying input variables (which may or may not be independent of each other) involved each of their predictions. The transparent open box (TOB) learning network algorithm overcomes this limitation by providing the exact calculations involved in all its predictions and achieving acceptable and auditable levels of prediction accuracy. The TOB network, based on an optimized data-matching algorithm, can be applied in spreadsheet or fully-coded configurations. This algorithm offers significant benefits to analysis and prediction of many complex and difficult to measure non-linear systems. To demonstrate its prediction performance, the algorithm is applied to the prediction of crude oil formation volume factor at bubble point (B_{ob}) using published datasets of 166, 203 and 237 data records involving 4 variables (reservoir temperature, gas-oil ratio, oil gravity and gas specific gravity). Two of these datasets display uneven and irregular data coverage. The TOB network demonstrates high prediction accuracy for B_{ob} (Root Mean Square Error (RMSE)$\sim$0.03; $R^2 > 0.95$) for the more evenly distributed dataset. The performance of the TOB readily reveals the risk of overfitting such datasets. With its high levels of transparency and inhibitions to being overfitted, the TOB learning network offers an insightful approach to machine learning applied to predicting complex non-linear systems. Its results complement and benchmark the prediction contributions of neural networks and empirical correlations. In doing so it provides further insight to the underlying data.

1. Introduction

Neural networks and other machine learning now represent the mainstay of tools applied to generate reliable predictions from systems dependent upon multiple variables. This is particularly so when some or all of the underlying variables are either difficult and/or expensive and/or time consuming to measure experimentally. Where complex, non-linear relationships between the input variables cannot be defined in terms of simple universally applicable formulas are systems that also lend themselves to machine learning approaches to prediction. The most commonly applied machine learning tools to complex non-linear systems are artificial neural networks (ANN) (Bishop, 1995; Haykin, 1999). Two distinct ANN architectures commonly applied are multi-layer perceptron (MLP) and radial basis function networks (RBFN) (Broomhead and Lowe, 1988). ANNs are often combined with various optimization, back-propagation and forward-feeding training algorithms to improve their predictions. Other commonly applied machine -learning algorithms are adaptive neuro-fuzzy inference systems (ANFIS) (Jang, 1993; Jang et al., 1997), support vector machines (SVM) (Vapnik, 1998) and least-squares support vector machine (LSSVM) (Espinoza et al., 2003). The application of such machine-learning tools is growing rapidly (Schmidhuber, 2015). However, their applications often pose problems and frustration for their users. Their lack of transparency also poses difficulties in assessing their reliability when applied to datasets with subtle differences to the one used to calibrate them.

The lack of transparency provided by many machine learning algorithms makes many researchers sceptical about the benefits of their widespread uptake. Failure to provide a clear explanation of how individual predictions are calculated is a barrier to their universal acceptance. Also, the lack of information on the relative weights and significance of each input variable applied to each data record in the dataset

https://doi.org/10.26804/ager.2019.03.01.

contributing to a set of predictions associated with specific data records does not help their cause. Some view them sceptically and suspiciously as obscure black-box tools (Heinert, 2008) of secondary value to more rigorous experimental and analytical methods. It is possible to provide some insight to the inner workings of the correlation-based machine-learning tools mentioned, but that typically requires detailed simulation studies to measure the significance of each input variable in generating the prediction for a specified set of test data. That approach can be informative, but only provides partial and indirect insight to these complex machine-learning algorithms (e.g., Elkatatny and Mohamed, 2017). Variable importance algorithms and other data mining algorithms (e.g., random forest algorithm) can also be used to provide the covariances between the influencing variables of machine-learning methods (Auret and Aldrich, 2012). Such approaches reveal the relative importance of the input variables in influencing the predictions, but again do not provide full transparency to the exact correlations involved in the predictions of many machine-learning methods.

A recently-proposed learning-network algorithm, based on simple matching and optimization heuristics (i.e., the transparent open-box (TOB); Wood, 2018a) provides an alternative approach and direction for machine learning. One of its primary objectives is to provide reliable predictions of high accuracy. Another is to provide complete access to the underlying contributions to each variable from specific records in the dataset that underpin the predictions generated by the learning network. Moreover, the TOB methodology reduces the risk of over-fitting data sets, a problem impacts many other machine-learning algorithms. This is because their opaque and complex correlations are only viable for the specific dataset studied (Lever et al., 2016), yet those correlations are not readily revealed for many machine-learning networks. This tends to be a more significant issue for datasets with irregular covering of records over the dependent -variable range to be predicted.

The TOB algorithm is quite distinct from other locally-weighted learning methods (Atkeson et al. 1997) but some of its attributes are consistent with lazy-learning principles (Birattari et al., 1999). Locally-weighted learning methods originate from the various earlier algorithms commonly distinguished as nearest-neighbour prediction methods (Fix and Hodges, 1951; Cover and Hart, 1967). Such algorithms can also be readily configured to provide transparency in the predictions that they generate (Shakhnarovich et al., 2006). Currently, such approaches constitute part of some pattern recognition algorithms (Garcia et al., 2012; Chen and Shah, 2018), but are not so widely applied to improve the prediction of dependent variables related to multiple independent variables by a suite of highly non-linear distributions. The more-opaque neural network, correlation-based, machine-learning algorithms mentioned now tend to be favoured for such predictions, despite their general lack of transparency. The majority of nearest-neighbour-prediction algorithms seek to linearize non-linear and irregularly distributed systems to provide approximations at the local level (Bontempi et al., 1999). The TOB algorithm goes beyond such simple approximations, because its approach relates the underlying independent variables in optimized non-linear relationships to generate each of its predictions.

Here, we describe how the TOB approach can be successfully applied to predict formation volume factor from published crude oil datasets and compare its predictions to those generated from a trained ANN algorithm and well-established empirical correlations. Section 2 summarizes the TOB methodology and how it is applied to generate individual predictions. Section 3 describes the crude oil formation volume factor datasets used to demonstrate TOB's prediction capabilities. Section 4 defines the statistical prediction measures used to assess prediction performance. Sections 5, 6 and 7 compare the prediction performances of TOB and ANN applied to the datasets analysed. Section 8 compares the machine learning predictions with those achieved by published empirical correlations for crude oil formation volume factor.

2. The TOB prediction approach to non-linear systems

The transparent open-box algorithm involves 14 steps (Wood, 2018a; Wood et al., 2019) which are tailored to provide reliable and auditable predictions for non-linear systems with comparable accuracy to the more-opaque machine-learning algorithms (i.e., ANFIS, ANN, LSSVM and SVM). TOB Stage 1 builds broadly upon lazy learning principles (Birattari et al., 1999), which commonly underpin nearest-neighbour prediction methods (Chen and Shah, 2018). However, the TOB algorithm applies quite specific variable-error metrics to derive its initial predictions. TOB Stage 2 goes far beyond k-learning methods by applying standard optimizers to optimize the weights applied to its input variables in high-ranking nearest-neighbour data records. This approach leads to a more flexible, effective and versatile weighting regime than typically applied in k-nearest neighbour prediction methods (Samworth, 2012). The mathematical basis for the TOB methodology and processes involved in the calculation steps are explained in detail in Appendix 1 and are summarized in Fig. 1. Tables 1 to 3 display the intermediate TOB calculation results associated with a specific data record from one of the datasets analysed in this study. These displays highlight the level of calculation and transparency detail the TOB method routinely provides.

TOB-Stage-one predictions (steps 1 to 10) are found to provide credible but sub-optimal prediction accuracy (e.g., comparable to those provided by simple data-matching algorithms). However, the TOB-stage-one predictions are almost always significantly improved upon by applying TOB-stage-two optimization, provided that the tuning subset is of sufficient size.

It is straightforward to fully code the TOB algorithm using Eqs. (A-1) to (A-9) (Appendix 1) in various mathematically-focused software packages (Octave, R, Python, MatLab etc.). For large data sets (i.e., with many thousands of data records and large numbers of input variables) it is more efficient to apply the TOB in that way.

For small and medium numbers of data records (e.g., up to several thousand records and ten or so variables) hybrid-VBA-spreadsheet can be a convenient way of maximizing the

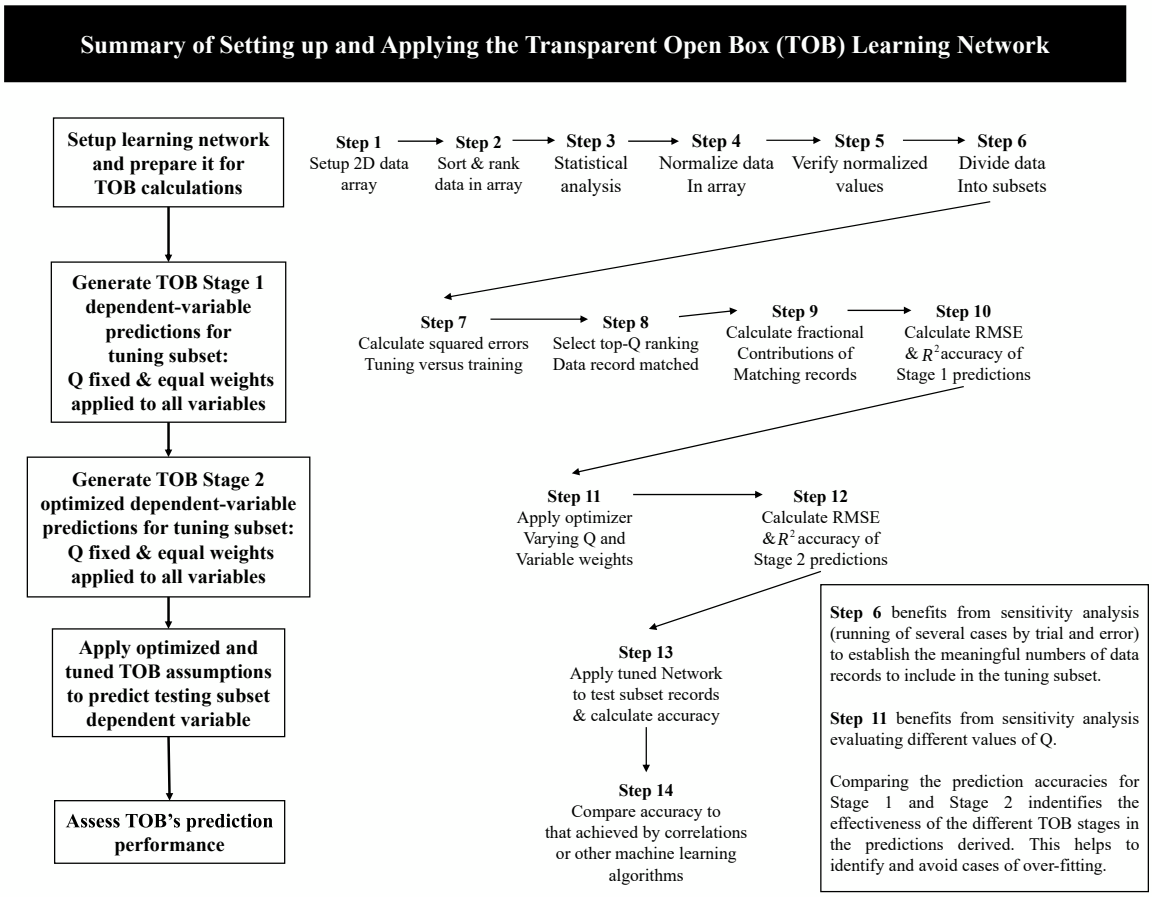

Fig. 1. Diagrammatic illustration of the stages and steps constituting the TOB non-linear system prediction method (Wood, 2018a; Wood et al., 2019). Text section 2 and Appendix 1 provide the calculation details of the algorithm.

benefits of the TOB algorithm's transparency, by systematically recording the intermediate calculations, and exploiting the standard optimizers available (i.e., the generalized reduced gradient and/or evolutionary optimizers contained within Excel's Solver package).

The crude oil formation volume factor at bubble point (B_{ob}) is predicted in this study using small published datasets of 166 to 237 data records involving four independent variables. A VBA-driven Excel-Solver-based algorithm is involved, exploiting spreadsheet attributes to transparently record the intermediate calculations involved in the predictions generated and display the results backed up by Excel cell formulas.

The TOB learning network provides easy access to detailed calculations for the contribution of each of the top-matching data records in the training subset to each dependent-variable prediction value for records in the tuning and testing subsets. There are no hidden correlations or inaccessible calculations involved in the TOB algorithm. This key attribute of transparency, is the detail it provides of the calculations involved in the predictions generated for each specific data record from the datasets studied (for such details see Wood, 2018b). An example shown in the supplementary file identifies exactly which top-ten matching data records are selected by the TOB algorithm for a specific data record prediction. This provides forensic-like information from the underlying dataset, i.e., close matches between new samples and existing samples in a dataset can be revealed that would otherwise not be obvious. For some datasets this TOB attribute can be used to identify the likely provenance (e.g., a specific location origin) of a sample from an unspecified source. If the TOB stage 2 optimizer is setup to use Excel's Solver, which is convenient for small and mid-sized data sets, the intermediate calculations for each data record are displayed in Excel cell formulas, making the calculations even more visible and easy interrogate.

ANN and other neural networks do not provide easy transparent access to their intermediate calculations in detail and certainly not routinely on a record by record basis as disclosed by the TOB. With effort, it is possible to extract some parametric information from ANN models, but it is very difficult and time consuming for a user to get to the exact step-by-step calculations made in the prediction of each individual data record. TOB makes forensic assess readily available for the predictions it makes.

3. Predicting crude oil formation volume factor at bubble point

Formation volume factor (FVF) is a key property of crude

Table 1. Statistical summary of the key metrics measured in the PVT dataset of 166 Pakistani crude oils (Al-Marhoun, 1998). The TOB learning network, the MLP-ANN network and several published B_{ob} correlations are applied to predict B_{ob} for this dataset.

Dataset #1 Used for Formation Volume Factor (B_{ob}) Prediction (B_{ob} Range Covered: 1.20 to 2.92)			
Dataset: 166 data records	Min	Max	Mean
Reservoir temperature T (°F)	182	296	242
Solution gas to oil ratio R_s (scf/stb)	92	2,496	500
Specific gravity of gas γ_g (Air = 1)	0.8253	3.4445	1.760
Oil Gravity (degrees API)	29	56.5	39.1
Formation Volume Factor at bubble point (B_{ob})	1.200	2.916	1.479

oil derived from sub-surface reservoirs. The FVF of crude oil establishes the ratio of the volume of crude oil in a sub-surface reservoir at temperature and pressure higher than at the earth's surface to the volume of that same crude oil composition in stock tank conditions at the earth's surface. It is measured as a fundamental crude-oil attribute during pressure-volume-temperature (PVT) analysis. PVT data provide essential information with which to characterize crude oils. PVT data are also used in the calculation of the recoverable resources that could potentially be recovered to the surface from specific oil reservoirs. However, measuring the input variables required to determine FVF is time-consuming and expensive.

Estimating FVF at bubble point (B_{ob}) from PVT (pressure-volume-temperature) data using formulaic relationships between the input metrics is also fraught with inaccuracies. This is because the input-metric assumptions required for highly non-linear and complex sensitivities for different ranges of pressure and temperature conditions and crude-oil composition types are difficult to define consistently. This is particularly so for the compressibility factor (Z) of the entrained and associated natural gas (solution gas) present in variable quantities in all crude oils. Over past decades many correlations were proposed and widely applied to relate bubble point pressure and formation volume factor of crude oils of different compositions (Katz, 1942; Standing, 1947; Al-Marhoun, 1992; Karimnezhad et al., 2014; Jarrahian et al., 2015). However, such correlations are difficult to apply with confidence to crude oils from outside the dataset on which the correlations are based. Various non-linear optimization methods have been applied to predict PVT properties, including B_{ob} (Arabloo et al., 2014; Oloso et al., 2017; El-Hoshoudy and Desouky, 2018; Fattah and Lashin, 2018). Also, a number of neural network models have been developed to estimate B_{ob} from PVT data (Gharbi and Elsharkawy, 1997; Vartosis et al., 1999; Dutta and Gupta, 2010; Moghadam et al., 2011; Irene and Sunday, 2013; Elkatatny and Mahmoud, 2017). These models attempt to predict B_{ob} and FVF from the more easily measured and readily available input metrics, avoiding the need to derive Z-factors.

Here, we apply the TOB learning-network to a published PVT dataset (Mahmood and Al-Marhoun, 1998) for 22 bottom-hole samples of crude oils from Pakistan (Table 1). An ANN model was recently applied (Rammay and Abdulraheem, 2017) to this dataset to predict FVF at bubble point pressure (B_{ob}). Here, we apply the TOB model to predict B_{ob} from the 166 data records of that dataset using only four easy-to-measure variables:

- Temperature – T (°F);
- Solution gas-to-oil ratio – R_s measured in standard cubic feet per stock tank barrel (scf/stb);
- Specific gravity of the solution gas – γ_g, and,
- *API* gravity of the crude oil.

We compare the B_{ob} predictions of the TOB algorithm for three sets of data with a standard multi-layer perceptron (MLP) artificial neural network ANN model. The two models were specifically set up to evaluate a training subset, a tuning subset and an independent testing subset, each containing specified data records.

3.1 Dataset #1 including a range with B_{ob} sparse data representation

The 166 data records of the entire dataset of crude oils from Pakistan (Al-Marhoun, 1998) are divided into: a training subset consisting of 115 data records, a tuning subset consisting of 18 data records; and, a testing subset (33 records). The data ranges sampled by dataset #1 are listed in Table 1.

A challenge to prediction using dataset #1 is that the density of samples is skewed towards the lower end of the B_{ob} scale: 134 of the samples have B_{ob} values of < 1.6; only 32 samples cover the B_{ob} range 1.6 to 2.9, with only 7 of those samples having B_{ob} values of > 2.0. For this reason, two other modified datasets are also evaluated to reveal the capabilities and limitations of the TOB method.

3.2 Dataset #2 introducing data from other sources

The 237 data records of the entire dataset #2 includes the Pakistan crude oils from dataset #1 expanded with additional crude oils from the United Arab Emirates (Dokla and Osman, 1992), Malaysia (Omar and Todd, 1993) and Iran (Moghadam et al., 2011) in the B_{ob} data range 1.1 to 2.1. The data records are divided into: a training subset consisting of 172 data records, a tuning subset consisting of 30 data records; and, a testing subset (35 records). The data ranges sampled by dataset #2 are listed in Table 2.

Although the density of data records for the dependent variable range considered is increased in dataset #2 compared

Table 2. Statistical summary of the key metrics measured in the PVT dataset #2 made up of 237 crude oils with published data from four countries (Iran, Malaysia, Pakistan, and UAE). The TOB learning network, the MLP-ANN network and several published B_{ob} correlations are applied to predict B_{ob} for this dataset in the range 1.10 to 2.1.

Dataset #2 Used for Formation Volume Factor (B_{ob}) Prediction (B_{ob} Range Covered: 1.10 to 2.10)			
Dataset: 237 data records	Min	Max	Mean
Reservoir temperature T (°F)	120	296	223
Solution gas to oil ratio R_s (scf/stb)	92	1,784	552
Specific gravity of gas γ_g (Air = 1)	0.5210	3.4445	1.3204
Oil Gravity (degrees *API*)	21.1	53.2	37.6
Formation Volume Factor at bubble point (B_{ob})	1.102	2.055	1.429

Table 3. Statistical summary of the key metrics measured in the PVT dataset #3 made up of 206 crude oils with published data from four countries (Iran, Malaysia, Pakistan, and UAE). The TOB learning network, the MLP-ANN network and several published B_{ob} correlations are applied to predict B_{ob} for this dataset in a range 1.10 to 1.62.

Dataset #3 Used for Formation Volume Factor (B_{ob}) Prediction (B_{ob} Range Covered: 1.10 to 1.62)			
Dataset: 206 data records	Min	Max	Mean
Reservoir temperature T (°F)	120	296	220
Solution gas to oil ratio R_s (scf/stb)	92	1,376	467
Specific gravity of gas γ_g (Air = 1)	0.5210	3.4445	1.3419
Oil Gravity (degrees *API*)	21.1	53.2	37.6
Formation Volume Factor at bubble point (B_{ob})	1.102	1.619	1.374

to dataset #1, particularly in the B_{ob} range 1.1 to 1.6 (200 data records), the density of data records remains relatively sparse for B_{ob} range 1.6 to 2.1 (only 37 samples). For this reason, dataset #3 is also evaluated.

3.3 Dataset #3 focusing on a B_{ob} range covered more densely by the data records

The 206 data records of the entire dataset #3 consists of those crude oils (from Pakistan, United Arab Emirates, Malaysia and Iran) from dataset #2 in the B_{ob} data range 1.10 to 1.62. The data records are divided into: a training subset consisting of 145 data records, a tuning subset consisting of 31 data records; and, a testing subset (30 records). The data ranges sampled by dataset #3 are listed in Table 3.

The density of data records for the dependent variable range covered by dataset #3 is spread more evenly and densely compared to that in datasets #1 and #2, for the B_{ob} range 1.10 to 1.62. The 206 data records in dataset #3 are split as follows: B_{ob} range 1.1 to < 1.2 contains 11 records; B_{ob} range 1.2 to < 1.3 contains 46 records; B_{ob} range 1.3 to < 1.4 contains 64 records; B_{ob} range 1.4 to < 1.5 contains 47 records; B_{ob} range 1.5 to 1.62 contains 38 records.

4. Statistical measures of prediction accuracy

Several widely-used statistical error measures are calculated to determine the accuracy, precision and correlation of measured versus predicted B_{ob}. These measures, and components used in their calculation, are expressed in Eq. (1) to Eq. (8) where X_i refers to the measured value and Y_i the predicted value of data record i in a dataset.

Mean Squared Error (MSE) (Mood et al., 1974; Lehmann and Casella, 1998):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - Y_i\right)^2 \tag{1}$$

Also used for the GA fitness function f (Eqs. (4) and (5)).

Root Mean Square Error (RMSE) (Carbone and Armstrong, 1982; Hyndman and Koehler, 2006):

$$RMSE = \sqrt{MSE} \tag{2}$$

RMSE calculated with Eqs. (1) and (2) is used as the objective function of the TOB algorithm.

Percent Deviation between measured and predicted values for data set record i (PDi) (Makridakis, 1993):

$$PD_i = \frac{X_i - Y_i}{Y_i} \times 100 \tag{3}$$

Average Percent Deviation (APD) (Makridakis, 1993):

$$APD = \frac{\sum_{i=1}^{n} PD_i}{n} \tag{4}$$

APD combines both positive and negative percent deviations Eq. (3) and is expressed in percentage terms.

Absolute Average Percent Deviation (AAPD) (Makridakis, 1993):

$$AAPD = \frac{\sum_{i=1}^{n} |PD_i|}{n} \tag{5}$$

Table 4. Formation volume factor at bubble point (B_{ob}) prediction accuracy achieved for the TOB-tuning and TOB-testing data subsets for the 166-record B_{ob}-dataset #1.

Transparent Open Box (TOB) Learning Network Results and Variable Weightings in the Prediction of Formation Volume Factor at Bubble Point (Dataset #1 with 166 records)													
Variable Description	Variable Number	Pre-optimization Equal Weightings	Best Solution Solver GRG Multi-start	Best Solution Solver Evolutionary Algorithm	Sensitivity Analysis with Q constrained to integers progressively from 10 to 2 (All cases runs with for the 18 records of the tuning subset with the Solver GRG optimizer configured in the same way)								
Q Constained to	Integer Constraints		2 to 10	2 to 10	10	9	8	7	6	5	4	3	2
Q selected for solution	Integer #		2	2	10	9	8	7	6	5	4	3	2
Prediction Performance of Optimum and Constrained Optimum Solutions Applied to the Tuning Subset (18 records: $\sim$ 11.0% of total dataset)													
RMSE	rsb/stb	0.1400	0.0738	0.0738	0.1174	0.1121	0.1085	0.1044	0.0993	0.0949	0.9058	0.0774	0.0738
R^2	fraction	0.9388	0.9843	0.9843	0.9514	0.9541	0.9558	0.9549	0.9572	0.9569	0.9629	0.973	0.9843
Weightings ($0 \leq w \leq 1$) Applied to constrained optimum solutions for the tuning subset													
Temperature	#1	0.5	0.00000	0.00000	0.71357	1.00000	0.95896	0.78098	0.19709	0.97291	0.37198	0.03670	0.00000
Gas Oil Ratio	#2	0.5	0.03268	1.00000	0.76145	0.83300	0.67069	0.63378	0.18953	0.73867	0.23491	0.29306	0.03268
Gas Gravity	#3	0.5	0.00000	0.00000	0.00947	0.01160	0.00000	0.00000	0.00281	0.00000	0.00000	0.00000	0.00000
Oil API Gravity	#4		0.00077	0.02359	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.07983	0.00077
Ratio of weight for Rs to weight for API			42.44156	42.39084	N/A	N/A	N/A	N/A	N/A	N/A	N/A	3.67105	42.44156
Prediction Performance of Optimum Solution Variable Weightings and Q Value Applied to the Testing Subset (33 records: $\sim$ 20% of total dataset)													
RMSE	rsb/stb		0.0509										
R^2	fracture		0.9599										

AAPD combines both positive and negative percent deviations Eq. (3) and is expressed in percentage terms.

Standard Deviation (SD) (Pearson, 1894):

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (D_i - D_{imean})^2}{n-1}} \tag{6}$$

Where, D_i is $(X_i - Y_i)$ for each (i^{th}) data record of a dataset; and, D_{imean} is the mean of the D_i values of all the data records in a dataset:

$$D_{imean} = \frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i) \tag{7}$$

Correlation Coefficient (R) between variables X_i and Y_i (on a scale between -1 and +1) (Pearson, 1894):

$$R = \frac{\sum_{i=1}^{n} (X_i - X_{mean})(Y_i - Y_{mean})}{\sqrt{\sum_{i=1}^{n} (X_i - X_{mean})^2 \sum_{i=1}^{n} (Y_i - Y_{mean})^2}} \tag{8}$$

Coefficient of Determination = R^2 (on scale between 0 and 1) (Wright, 1921).

5. B_{ob} prediction performances of TOB and ANN compared for dataset #1 (166-samples extending over a B_{ob} range of 1.20 to 2.90)

The TOB model applied to dataset #1 achieves its optimum B_{ob} prediction performance (after TOB stage 2; based on its objective function RMSE value) with Q = 2 and weights wT = 0, wRs = 0.03268, wγ_g = 0; wAPI = 0.00077. Sensitivity analysis (not shown) applies several Q values and a number of distinct allocations (selected on a trial and error basis) of data records between the three data subsets into which the dataset is divided. This leads to slightly different but broadly acceptable levels of B_{ob} prediction performance. The optimum solution with a Q value of 2 suggests that the dataset #1 is prone to overfitting and this is further indicated by sensitivity analysis (Table 4).

A multi-layer perceptron artificial neural network (MLP-ANN) was also developed and tuned using standard Matlab codes and functions to predict B_{ob} for this dataset. The ANN methodology is well documented (Bishop, 1995; Haykin, 1999) and widely applied. Conceptually, the ANN method can be described in simple terms by the simple learning law expressed as Eq. (9) indicative of underlying formulaic correlations:

$$f : X \rightarrow Y \tag{9}$$

That function varies in the way it is applied (Bose and Liang, 1995; Choubineh et al., 2017). MLP-ANN functions, $f(x)$, are comprised of I underlying contributing functions. These contributing functions, $g_i(x)$, are formulated single vectors, e.g., $g = (g_1, g_2, ..., g_I)$. Each component of vector g is then related to the overall function $f(x)$, by applying a set of non-linear weights in a summed relationship constituting Eq. (10):

$$f(x) = K\left(\sum_{i=1}^{n} w_i g_i(x)\right) \tag{10}$$

where, K = non-linear activation functions that determine the MLP-ANN output; w_i = weights applied to each function of vector g.

A supervised-learning method further develops the MLP-ANN model. Sets (x, y) are selected for which $x \in X$, $y \in Y$ to derive the function f: $X \rightarrow Y$ applying a specified cost function. Mean squared error (MSE, Eq. (1)) between the predicted and measured B_{ob} values in the dataset is the cost factor that is minimized as the objective function for the MLP-ANN model developed. A gradient-descent algorithm is applied to minimize MSE as the backpropagation algorithm that is able to train the MLP-ANN rapidly and effectively to optimize its predictions.

The MLP-ANN model applied to the B_{ob} dataset #1, involves two hidden layers in its network architecture. Hidden layer 1 has 4 neurons, whereas hidden layer 2 has 3 neurons. The developed MLP-ANN model, selected using a trial-and-error sensitivity analysis to determine the optimum number of hidden layers, neurons and activation functions, applies the following activation functions:

- purelin between the input layer and hidden layer 1;
- logsig between hidden layer 1 and hidden layer 2; and,
- purelin between hidden layer 2 and the network's output layer.

1000 iterations were evaluated to tune the MLP-ANN applying a back-propagation algorithm that optimizes the mean-squared error (MSE) between the measured and predicted B_{ob} values.

The value ranges and non-linear relationships between the input variables, T, R_s, γ_g, API and the dependent variable B_{ob} are illustrated in Figs. 2A to 2D. These relationships are primarily non-linear and, as revealed by Fig. 2, are distributed irregularly for the wide B_{ob} range covered by this dataset. The most-dense coverage of samples is for B_{ob} less than about 0.6 with sparse coverage of B_{ob} at greater than 0.6. Gas-to-oil ratio (R_s) is the input metric best correlated with B_{ob} and therefore has the greatest discriminatory impact in the record -match selections for stage 1 of the TOB network.

Figs. 3 and 4 compare the B_{ob} prediction results obtained by the TOB algorithm with those derived from a multi-layer perceptron ANN algorithm applied to the same dataset. Both algorithms achieve a high degree of accuracy in the predictions they generate for this data set. For the tuning and testing subsets (Figs. 3 and 4) the ANN yields a higher correlation coefficient and a lower RMSE value compared to the TOB model. The ANN architecture (number of hidden layers, number of neurons, type of activation function) selected by the sensitivity analysis performs very well, so we believe that the developed ANN structure should be considered fit for purpose.

The TOB model is clearly unable to match the apparent prediction accuracy of the ANN model. This is due to the

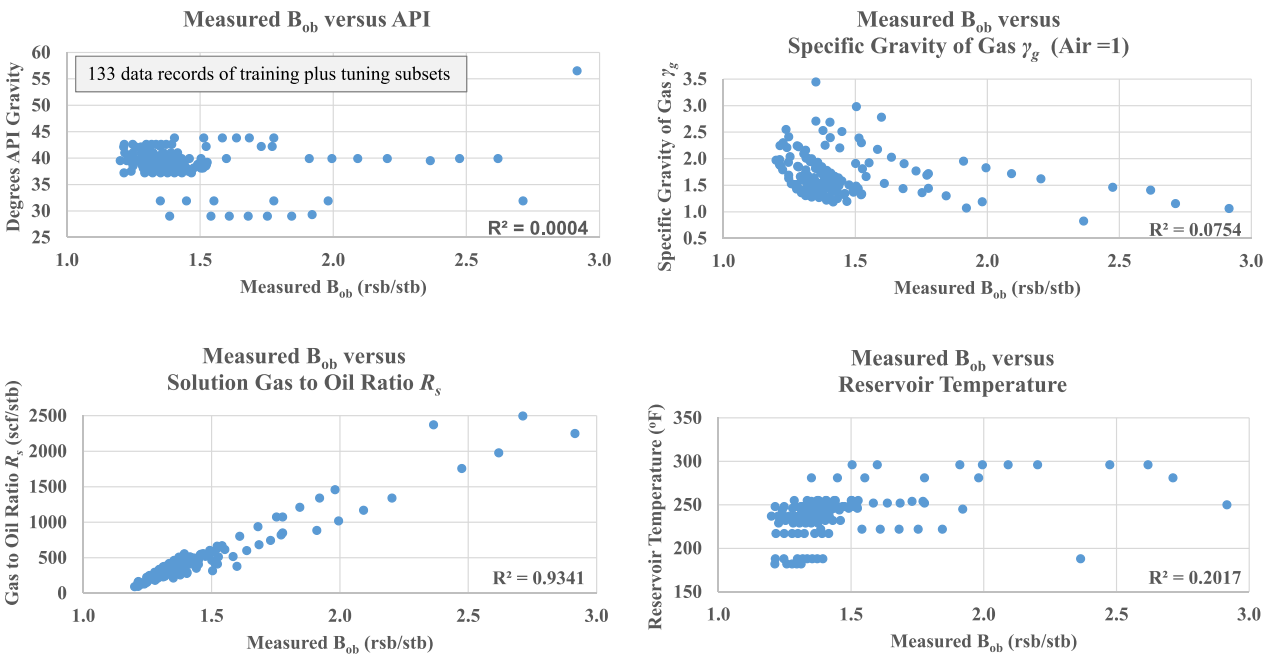

Fig. 2. T, R_s, γ_g, and API relationships for data records in the TOB-training subset plus those in the TOB-tuning subsets used to evaluate dataset #1 to which the TOB-algorithm is applied to predict formation volume factor at bubble point (B_{ob}).

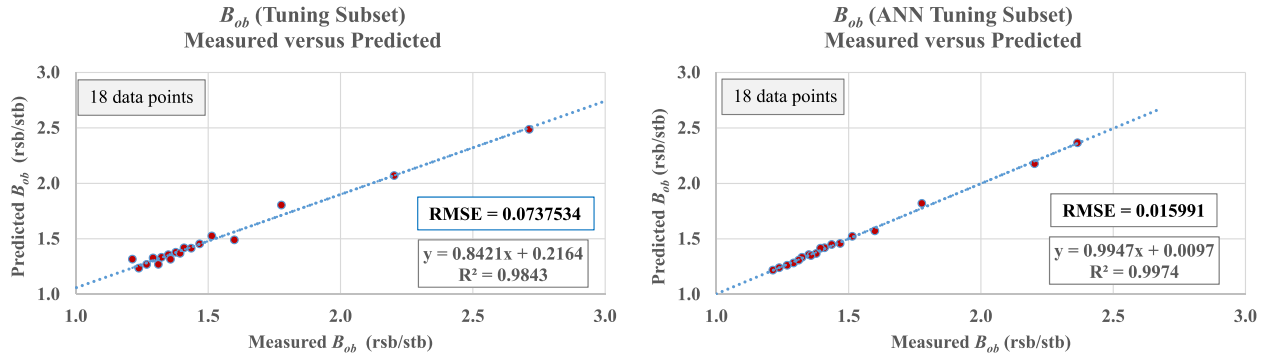

Fig. 3. Predicted versus measured formation volume factor at bubble point (B_{ob}) for the tuning subset of dataset #1 (18 records).

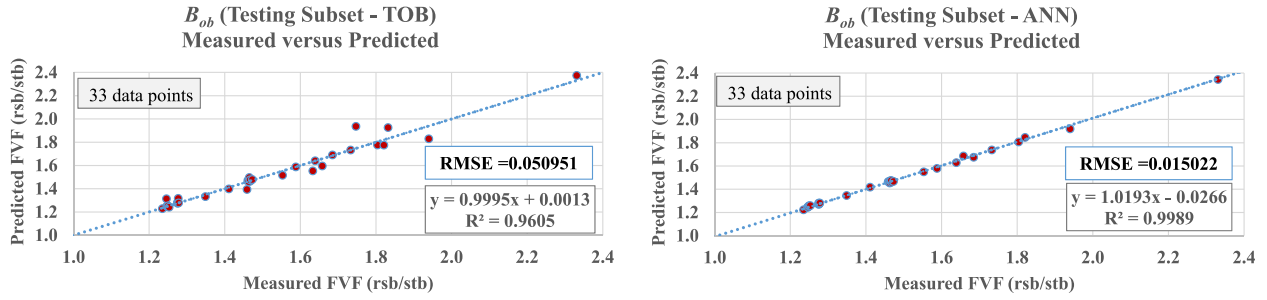

Fig. 4. Predicted versus measured formation volume factor at bubble point (B_{ob}) for the testing subset of dataset #1 (33 records).

sparsity of data points in the full dataset for B_{ob} values above 1.6 rsb/stb and especially above 2 rsb/stb.

The ANN model, assessed in terms of the statistical accuracy measures described by Eq. (1) to Eq. (8), performs particularly well with sparse and clustered datasets, such as dataset #1 (Table 5). On the other hand, the TOB model, which does not in any way involve the construction of a correlation, is less prone to overfitting. The TOB relies on surrounding data points with close matches to the data record being predicted rather than on complex correlations established by neural networks or those involved in published empirical B_{ob}/PVT correlations.

The more-widely spaced the surrounding data points and the more non-linear and clustered the relationship between the input variables, the less precise the TOB's predictions become, and rightly so. In practice, for datasets such as dataset #1, a lower R^2 value obtained for TOB versus that obtained for ANN may not be such a bad outcome, as it highlights some of the limitations of the dataset (e.g., sparse data over half of the dependent variable's value range). The R^2 values of > 0.994 achieved by the ANN may be overstating the reliability of its correlation to predict new unknown data points for that sparsely sampled upper range of the B_{ob} distribution constituting dataset #1. As its correlations depend on few data points in that region it may be overfitting the available data. It suggests that a more reliable TOB prediction model could be constructed for B_{ob} values less than 1.6 rsb/stb, because of the greater density of data records. On the other hand, for values greater than 1.6 rsb/stb the ANN algorithm provides a more credible prediction model, albeit with a correlation that is probably somewhat overfitted for the limited data points available. Evaluations of datasets #2 and #3 further explore these possibilities.

6. B_{ob} prediction performances of TOB and ANN compared for dataset #2 (237-samples extending over a B_{ob} range of 1.10 to 2.10)

One TOB model applied to dataset #2 achieves its optimum B_{ob} prediction performance with Q = 7 and weights wT = 0.0059, wRs = 0.6214, $w\gamma_g$ = 0; wAPI = 1. Note that although the weight applied to the gas gravity data is zero in TOB stage 2, gas gravity input data is still used (with equal weight to the other input variables) to select the highest- ranking matching records in TOB stage 1. A second TOB model applied to dataset #2 ignored the gas gravity input data in TOB stages 1 and 2. It achieved its optimum B_{ob} prediction performance (after TOB stage 2; based on its objective function RMSE value) with Q = 7 and weights wT = 0.0052, wRs = 0.418, $w\gamma_g$ = 0; wAPI = 1.

The MLP-ANN model applied to the B_{ob} dataset #2, involves two hidden layers in its network architecture. Hidden layer 1 has 5 neurons and hidden layer 2 has 6 neurons. The developed MLP-ANN model, selected using sensitivity analysis, applies the following activation functions:

- purelin between the input layer and hidden layer 1;
- tansig between hidden layer 1 and hidden layer 2; and,
- purelin between hidden layer 2 and the network's output layer.

The MLP-ANN was evaluated in the same way as for dataset #1.

Table 5 displays the prediction accuracies achieved by the TOB and ANN models applied to dataset #2. Although dataset #2 has a greater sample density in the B_{ob} range 1.1 to 1.6, the B_{ob} range > 1.6 is only sparsely sampled (37 samples only for that region). For the full testing subset, the TOB achieves RMSE = 0.06698 and AAPD = 3.000%. For the 29 samples of the testing subset within the B_{ob} range 1.1 to 1.6 the TOB achieves RMSE = 0.04367 and AAPD = 2.600%. For the full testing subset, the ANN achieves RMSE = 0.03870 and AAPD = 1.678%. For the 29 samples of the testing subset within the B_{ob} range 1.1 to 1.6 the ANN achieves RMSE = 0.02947 and AAPD = 1.462%. The accuracy metrics show slightly less accuracy for TOB and ANN for dataset #2 compared to dataset #1, particularly for the ANN for which the RMSE has more than doubled (Table 5). Although, the ANN clearly outperforms the TOB in terms of the accuracy achieved for dataset #2, both methods are adversely affected in terms of the accuracy they achieve when an even greater sampling density contrast exists between the upper and lower ends of the B_{ob} range.

Interestingly, the TOB model that completely disregards gas gravity as an input variable in stages 1 and 2 of its analysis, yields slightly more accurate predictions (RMSE = 0.06282 and AAPD = 2.780%) than the TOB model that involves gas gravity in the stage 1 data record matching process (RMSE = 0.06698 and AAPD = 3.000%). This implies that gas gravity as an input variable is actually a hinderance to the TOB model in achieving accurate B_{ob} predictions for dataset #2. Table 5 reveals that this is also the case for datasets #1 and #3. On the other hand, limiting the ANN to ignore gas gravity input data reduces its accuracy; significantly so for dataset #3.

7. B_{ob} prediction performances of TOB and ANN compared for dataset #3 (206-samples extending over a B_{ob} range of 1.10 to 1.62)

One TOB model applied to dataset #3 achieves its optimum B_{ob} prediction performance with Q = 7 and weights wT = 0.0014, wRs = 1, $w\gamma_g$ = 0; wAPI = 0.0981. A second TOB model applied to dataset #3 ignored the gas gravity input data in TOB stages 1 and 2. It achieved its optimum B_{ob} prediction performance with Q = 6 and weights wT = 0.0006, wRs = 1, $w\gamma_g$ = 0; wAPI = 0.0586.

The MLP-ANN model applied to the B_{ob} dataset #3, involves three hidden layers in its network architecture. Hidden layer 1 has 6 neurons, hidden layer 2 has 5 neurons and hidden layer 3 has 3 neurons. The developed MLP-ANN model, selected using sensitivity analysis, applies the following activation functions:

- purelin between the input layer and hidden layer 1;
- logsig between hidden layer 1 and hidden layer 2;
- logsig between hidden layer 2 and hidden layer 3; and,
- purelin between hidden layer 3 and the network's output

Table 5. Formation volume factor at bubble point (B_{ob}) prediction accuracy achieved for TOB-testing subsets for datasets #1, #2 and #3 displaying values for a range of statistical accuracy metrics and comparing the results with accuracies achieved by several published correlations applied to the same testing subset data records.

Accuracy of Various Prediction Models for Formation Volume Factor of Oil at Bubble Point Pressure Appled to Datasets #1, #2 and #3							
		RMSE(*)	APD%	AAPD%	SD	R	R^2
Dataset (1) sparsely sampled $B_{ob} > 1.6$							
# Total Samples	166						
B_{ob} Range	1.20 to 2.90						
# Samples in Testing Subset	33						
TOB		0.05095	-0.056%	1.8515%	0.0517	0.9800	0.9605
TOB (ignoring gas gravity)		0.04861	0.245%	1.949%	0.0492	0.9817	0.9638
ANN		0.01502	-0.168%	0.582%	0.0148	0.9995	0.9989
ANN (ignoring gas gravity)		0.02112	-0.005%	0.796%	0.0214	0.9986	0.9971
Standing (1947) Correlation		0.03804	1.073%	2.112%	0.0363	0.9939	0.9879
Vazquez & Beggs (1980) Correlation		0.25860	15.064%	15.064%	0.1066	0.9676	0.9362
Glaso (1980) Correlation		0.06301	3.843%	4.066%	0.0306	0.9937	0.9875
Al-Marhoun (1988) Correlation		0.07271	3.036%	3.079%	0.0521	0.9950	0.9901
Petrosky & Farshad (1993) Correlation		0.10892	-5.423%	5.507%	0.0654	0.9917	0.9835
Arabloo et al. (2014) Correlation		0.25992	-13.898%	13.898%	0.1342	0.9960	0.9919
Dataset (2) sparsely sampled $B_{ob} > 1.6$							
# Total Samples	237						
B_{ob} Range	1.10 to 2.10						
# Samples in Testing Subset	35						
TOB		0.06693	0.218%	3.000%	0.0674	0.9491	0.9007
TOB (ignoring gas gravity)		0.06282	0.463%	2.780%	0.0627	0.9550	0.9119
ANN		0.03870	0.541%	1.678%	0.0384	0.9806	0.9616
ANN (ignoring gas gravity)		0.04074	0.252%	1.866%	0.0411	0.9778	0.9560
Standing (1947) Correlation		0.03755	0.690%	1.724%	0.0367	0.9826	0.9654
Vazquez & Beggs (1980) Correlation		0.14178	7.689%	7.722%	0.0859	0.8994	0.8089
Glaso (1980) Correlation		0.06173	3.412%	3.636%	0.0379	0.9810	0.9624
Al-Marhoun (1988) Correlation		0.05241	1.964%	2.110%	0.0422	0.9906	0.9813
Petrosky & Farshad (1993) Correlation		0.06468	-1.496%	2.845%	0.0617	0.9570	0.9158
Arabloo et al. (2014) Correlation		0.19860	-11.596%	11.596%	0.0982	0.9801	0.9606
Dataset (3) evenly sampled over B_{ob} range							
# Total Samples	206						
B_{ob} Range	1.10 to 1.62						
# Samples in Testing Subset	30						
TOB		0.02969	0.221%	1.729%	0.0299	0.9719	0.9446
TOB (ignoring gas gravity)		0.02664	0.542%	1.494%	0.0258	0.9767	0.9539
ANN		0.02263	0.488%	1.114%	0.0220	0.9833	0.9669
ANN (ignoring gas gravity)		0.03937	1.383%	2.131%	0.0350	0.9579	0.9176
Standing (1947) Correlation		0.02815	0.620%	1.474%	0.0274	0.9760	0.9526
Vazquez & Beggs (1980) Correlation		0.12005	7.421%	7.460%	0.0620	0.8582	0.7366
Glaso (1980) Correlation		0.05335	3.298%	3.478%	0.0292	0.9717	0.9442
Al-Marhoun (1988) Correlation		0.02971	1.274%	1.444%	0.0236	0.9843	0.9688
Petrosky & Farshad (1993) Correlation		0.05422	-1.774%	2.582%	0.0488	0.9418	0.8870
Arabloo et al. (2014) Correlation		0.16838	-10.710%	10.710%	0.0762	0.9610	0.9234

(*) RMSE is used as the objective function of the TOB algorithm

layer.

The MLP-ANN was evaluated in the same way as that used to evaluate dataset #1.

Table 5 displays the prediction accuracies achieved by the TOB and ANN models applied to dataset #3. Dataset #3 is characterized by a much more evenly distributed sampling by the data records for the B_{ob} range it covers (1.10 to 1.62). For the full testing subset, the TOB achieves RMSE = 0.02969 and AAPD = 1.729%, whereas the full testing subset the ANN achieves RMSE = 0.02263 and AAPD = 1.114%. The accuracy metrics show much improved accuracy for dataset #3 compared to datasets #1 and #2 (Table 5). On the other hand, although the ANN prediction performance for dataset #3 is improved in comparison with dataset #2, it is significantly worse than for the ANN applied to dataset #1 (RMSE = 0.01502 and AAPD = 0.582%) (Table 5). This finding suggests that the accuracy recorded by ANN for dataset #1 is likely to have included an element of overfitting, as when applied to a much more densely sampled dataset it achieves inferior accuracy. Although, the ANN does slightly outperform the TOB in terms of the accuracy achieved for dataset #3, their prediction performances are much more closely matched (Table 5). Moreover, TOB is behaving as it should, i.e., improving its performance as sampling density increases and sparsely sampled areas of the distribution are excluded. What is encouraging about this finding is that as databases are expanded with more and more samples added from around the world the prediction performance of the TOB method should continue to improve. Data augmentation does often also improve the prediction performance of ANN methods, but the way it achieves this is more complex as it depends upon the multiple correlations it establishes between the variables.

8. B_{ob} prediction performances compared to published correlations

As mentioned in the section 3 there are many published correlations developed using different crude oil datasets that provide predictions of B_{ob}. Here, we evaluate the performance of some selected correlations applied to the testing subsets used to evaluate datasets #1, #2 and #3 in order to compare their performances with the TOB and ANN models (Table 5). The selected correlations are expressed in Eq. (11) to Eq. (18).

Standing (1947) correlation:

$$B_{ob} = k_1 + k_2\left[R_s\left(\frac{\gamma_g}{\gamma_o}\right)^{k_3} + k_4 T\right]^{k_5} \tag{11}$$

where, k_1 = 0.972; k_2 = 0.000147; k_3 = 0.5; k_4 = 1.25; k_5 = 1.175.

Vazquez-Beggs (1980) correlation:

$$B_{ob} = 1 + k_1 R_s + (T-60)\left(\frac{\gamma_{API}}{\gamma_g}\right)(k_2 + k_3 R_s) \tag{12}$$

where, If API $\leq$ 30: k_1 = 0.0004677; k_2 = 0.00001751; k_3 = -0.000000018106; If API >30: k_1 = 0.000467; k_2 = 0.000011; k_3 = 0.000000001337.

Glaso (1980) correlation:

$$X = R_S\left(\frac{\gamma_g}{\gamma_o}^{k_1}\right) + k_2 T \tag{13}$$

$$B_{ob} = 1 + 10^{k_3 + k_4 logX + k_5 (\log X)^2} \tag{14}$$

where, k_1 = 0.526; k_2 = 0.968; k_3 = -6.58511; k_4 = 2.91329; k_5 = -0.27683.

Al-Marhoun (1988) correlation:

$$X = R_S^{k_1}\gamma_g^{k_2}\gamma_o^{k_3} \tag{15}$$

$$B_{ob} = k_4 + k_5(T + 459.67) + k_6 X + k_7 X^2 \tag{16}$$

where, k_1 = 0.74239; k_2 = 0.323294; k_3 = -1.20204; k_4 = 0.497069; k_5 = 0.000862963; k_6 = 0.00182594; k_7 = 0.00000318099.

Petrosky-Farshad (1998) correlation:

$$B_{ob} = k_1 + k_2\left[\frac{R_s^{k_3}\gamma_g^{k_4}}{\gamma_o^{k_5}} + k_6 T^{k_7}\right]^{k_8} \tag{17}$$

where, k_1 = 1.0113; k_2 = 0.000072046; k_3 = 0.3738; k_4 = 0.2914; k_5 = 0.6265; k_6 = 0.24626; k_7 = 0.5371; k_8 = 3.0936.

Arabloo et al. (2014) correlation:

$$B_o = 1 + a_1\left[(R_s + 2a_2)(\gamma_g + 1)(\log(API)) + a_2\right]^{\left(a_3 + T_R^{a_4}\right)} \tag{18}$$

where, a_1 = 0.0003348062; a_2 = 25; a_3 = -0.2856905; a_4 = 0.03640287.

For all these correlations the temperature T is in degrees Fahrenheit (F), solution gas to oil ratio R_s is expressed in scf/stb, and formation volume factor at bubble point pressure B_{ob} is expressed in rsb/stb.

For dataset #1 the ANN model outperforms all these selected correlation predictions in all accuracy measures (Table 5). TOB outperforms all the correlations in terms of AAPD% for dataset #1 but is outperformed in terms of RMSE only by the Standing correlation. All of the correlations achieve a higher R and R^2 values for dataset #1 than the TOB model even though they show (except for the standing Correlation) much poorer RMSE and AAPD values. this is a very credible prediction performance for the TOB method applied to a dataset that is in part sparsely sampled.

For dataset #2 the ANN model outperforms all the selected correlation predictions in most accuracy measures except the Standing correlation (Table 5), which achieves slightly lower RMSE, SD, and slightly higher R and R^2 values. TOB is outperformed by all but the Vazquez and Beggs and Arabloo correlations for dataset #2, struggling to achieve high degrees of accuracy in the sparsely populated B_{ob} range of > 1.6. The superior performance of Standing's correlation compared to other correlations is clear for dataset #2.

For dataset #3 the ANN model only just outperforms the Standing's correlation, the TOB model and the Al-Marhoun's model in all accuracy measures (Table 5). In fact, the performance of the ANN, Standing's correlation, Al-Marhoun's

correlation and TOB models achieve similar levels of accuracy for this densely and evenly sampled dataset #3. On the other hand, the Vazquez and Beggs' and Arabloo et al's correlations achieve much lower levels of accuracy than the other models for dataset #3. Interestingly the TOB model based on just three input variables (i.e., excluding gas gravity completely) performs almost exactly as the Standing's and Al-Marhoun's correlations for dataset #3. This is an impressive performance for the TOB method considering no correlations only record matching and input variable weighting are involved in this learning algorithm. Clearly TOB has the potential to improve its prediction performance as more data records are added to a dataset to sparsely populated regions are infilled.

Although the ANN clearly provides superior prediction performance to TOB for two out of three of the B_{ob} datasets studied, the performance of the TOB is both credible and comparable with ANN, and the higher-performing published B_{ob} correlations, when applied to dataset #3. This makes the TOB a useful tool for predicting B_{ob} because it reveals details of its underlying prediction calculations and can potentially simplify those calculations by identifying non-contributing variables (e.g., gas gravity in the case of TOB B_{ob} predictions).

At first sight the relative prediction performances indicated in Table 5 for the data set studied imply that it easier to simply apply a published empirical correlation (in this case Standing's 1947 equation) than to bother with TOB or ANN machine learning networks. However, the value of the information empirical correlations provide is quite different from that provided by TOB. In the dataset presented, Standing's (1947) equation performs the best (compared to other published correlations), but in other areas it is not unusual for some of the other published empirical equations to outperform Standing's equation. Indeed, empirical correlations raise several issues relating to the predictions they provide. They were developed based on a specific set of data usually with some biases to specific geographic regions and oil-producing provinces. These correlations work well with data from some regions but provide lower prediction accuracy when applied to other regions. When applied to new areas it is not clear which one of the published correlations is the best one to use to provide the greatest accuracy. As TOB involves no correlations its method works in the same way when applied to different data sets. Moreover, as more data becomes available over time, which occurs in most developing oil-provinces, the TOB training subset becomes larger and the accuracy of the TOB predictions typically improves. For empirical calculations that progression does not occur as their parametric constant and exponent values typically remain fixed.

Considering the pros and cons of the two approaches evaluated it is apparent that the TOB and ANN algorithms have the ability to complement each other's performance. Certainly, the TOB provides a useful method for evaluating to credible levels of accuracy. Moreover, for more densely-populated data sets the accuracy it achieves rivals ANN and the best-performing prediction correlations.

9. Conclusions

The learning network applied in this study (i.e., transparent open-box (TOB)) generates its predictions in quite a different way to most neural-networks and other correlation-based machine learning tools. Its predictions depend on closely matching the record for prediction with a number of specific records existing in the underlying training data subset. It does so in a manner that is quite distinct from other nearest-neighbour matching algorithms.

When applied to three data sets to predict crude oil formation volume factor, and the prediction results are compared to those produced by an artificial neural network (ANN) and empirical correlations, the TOB's prediction performance is impressive. The prediction analysis with respect to crude oil formation volume factor identifies the following key advantages of the TOB approach:

- individual predictions are auditable revealing exactly how they are calculated;
- when data sets are evenly distributed across an objective function range TOB can generate prediction with accuracies that rival neural networks (e.g., for data sets (3) analysed prediction accuracy achieved by TOB is: RMSE$\sim$0.03 and $R^2\sim$0.95 compared to ANN prediction accuracy of RMSE$\sim$0.2 and $R^2\sim$0.97);
- optimized solutions are linked to specific weightings applied to input variables;
- standard optimizers (e.g., Excel's Solver options) or customized optimizers can achieve the necessary optimization;
- sensitivity to its Q-factor helps prevent it overfitting sparse datasets;
- TOB can act as a useful performance benchmark for more complex neural and fuzzy network algorithms, and for densely populated dataset rival their level accuracy.

The disadvantages of the TOB are:

- Its prediction capabilities are reasonable, but constrained, when applied to sparse and/or clustered data sets (e.g., for data sets (1) and (2) analysed prediction accuracy achieved by TOB is: RMSE ranges from 0.05 to 0.07 and R^2 ranges from 0.90 to 0.96). It does not attempt to overfit such data sets;
- As it involves no correlations, predictions cannot be extended beyond the range of dependent variable values in its training set.

By combining the application of a TOB algorithm with the ANN algorithm (i.e., running them both in parallel), it is possible to provide useful insight to sparsely populated datasets and to better assess the issues of overfitting that they pose. The TOB algorithm has the potential to improve the transparency of predictions, enabling them to be audited, and highlight overfitting risks in many complex oil and gas datasets, as well as achieving credible levels of accuracy.

References

Al-Marhoun, M.A. New correlation for formation volume factor of oil and gas mixtures. J. Can. Pet. Technol. 1992, 31(3): 22-26.

Al-Marhoun, M.A. PVT correlations for Middle East crude oils. J. Pet. Technol. 1998, 40(5): 650-666.

Arabloo, M., Amooie, M.A., Hemmati-Sarapardeh, A., et al. Application of constrained mult i-variable search methods for prediction of PVT properties of crude oil systems. Fluid Phase Equilibr. 2014, 363: 121-130.

Atkeson, C.G., Moore, A.W., Schaal. S. Locally Weighted Learning for Control. Dordrecht, Netherlands, Springer, 1997.

Auret, L., Aldrich, C. Interpretation of nonlinear relationships between process variables by use of random forests. Miner. Eng. 2012, 35: 27-42.

Birattari, M., Bontempi, G., Bersini, H. Lazy learning meets the recursive least squares algorithm. Adv. Neural Inf. Process. Syst. 1999, 12: 375-381.

Bishop, C.M. Neural networks for pattern recognition, $2^n d$ ed. UK, Oxford University Press, 1995.

Bontempi, G., Birattari, M., Bersini, H. Lazy learning for local modelling and control design. Int. J. Control 1999, 72(7-8): 643-658.

Broomhead, D.H., Lowe, D. Radial basis functions, multi-variable functional interpolation and adaptive networks. Royal Signals and Radar Establishment Malvern (United Kingdom) 1988, 25(3): 1-8.

Carbone, R., Armstrong, J.S. Evaluation of extrapolative forecasting methods: results of a survey of academicians and practitioners. J. Forecast. 2010, 1(2): 215-217.

Chen, G.H., Shah, D. Explaining the success of nearest neighbor methods in prediction. Found. Tren. Mach. Learn. 2018, 10(5-6): 337-588.

Choubineh, A., Ghorbani, H., Wood, D.A., et al. Improved predictions of wellhead choke liquid critical-flow rates: modelling based on hybrid neural network training learning based optimization. Fuel 2017, 207: 547-560.

Dokla, M., Osman, M. Correlation of PVT properties for UAE crudes (includes associated papers 26135 and 26316). SPE Form. Eval. 1992, 7(1): 41-46.

Dutta, S., Gupta, J.P. PVT correlations for Indian crude using artificial neural networks. J. Pet. Sci. Eng. 2010, 72(1-2): 93-109.

El-Hoshoudy, A.N., Desouky, S.M. Numerical prediction of oil formation volume factor at bubble point for black and volatile oil reservoirs using non-linear regression models. Pet. Petrochem. Eng. J. 2018, 2(2): 000145.

Elkatatny, S., Mahmoud, M. Development of new correlations for the oil formation volume factor in oil reservoirs using artificial intelligent white box technique. Petroleum 2018, 4(2): 178-186.

Espinoza, M., Suykens, J.A.K., Moor, B.D. Least squares support vector machines and primal space estimation. IEEE Cat. No. 03CH37475 in IEEE 42nd Conference on Decision and Control, Maui, USA, 9-12 December, 2003.

Fattah, K.A., Lashin, A. Improved oil formation volume factor (Bo) correlation for volatile oil reservoirs: An integrated non-linear regression and genetic programming approach. J. King Saud. Uni. Eng. Sci. 2018, 30(4): 398-404.

Frontline Solvers. Standard excel solver-limitations of nonlinear optimization. 2018.

Garcia, S., Derrac, J., Cano, J., et al. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE. T. Pattern Anal. 2012, 34(3): 417-435.

Gharbi, R.B., Elsharkawy, A.M. Neural network model for estimating the PVT properties of Middle East crude oils. SPE Reserv. Eval. Eng. 1999, 2(3): 255-265.

Glaso, O. Generalized pressure-volume-temperature correlations. J. Pet. Technol. 1980, 32(5): 785-795.

Haykin, S. Neural Networks: A Comprehensive Introduction, 3rd Edition. New York, USA, Pearson/Prentice Hall, 1999.

Heinert, M. Artificial neural networks-how to open the black boxes. App. Art. Intell. Eng. Geo. 2008, 5: 42-62.

Hyndman, R.J., Koehler, A.B. Another look at measures of forecast accuracy. Int. J. Forecast. 2006, 22(4): 679-688.

Irene, A.I., Sunday, I.S. Forecasting oil formation volume factor for API gravity ranges using artificial neural network. Adv. Pet. Explor. Dev. 2013, 5(1): 14-21.

Jang, J.S.R. ANFIS: Adaptive-network-based fuzzy inference system. Syst. man and Cybern. 1993, 23(3): 665-685.

Jang, J.S.R., Sun, C.T., Mizutani, E. Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence. IEEE Trans. Autom. Control 1997, 42(10): 1482-1484.

Jarrahian, A., Moghadasi, J., Heidaryan, E. Empirical estimating of black oils bubblepoint (saturation) pressure. J. Pet. Sci. Eng. 2015, 126: 69-77.

Karimnezhad, M., Heidarian, M., Kamari, M., et al. A new empirical correlation for estimating bubble point oil formation volume factor. J. Pet. Sci. Eng. 2014, 18: 329-335.

Katz, D.L. Prediction of shrinkage of crude oils. Paper API-42-137 Presented at the American Petroleum Institute, New York, USA, 1 January, 1942.

Lehmann, E.L., Casella, G. Theory of Point Estimation (2nd Ed.). New York, USA, Springer, 1998.

Lever, J., Krywinski, M., Altman, N. Points of significance: Model selection and overfitting. Nat. Methods 2016, 13: 703-704.

Liang, P., Bose, N.K. Neural Network Fundamentals with Graphs, Algorithms, and Applications. New York, USA, McGraw-Hill, 1996.

Mahmood, M.A., Al-Marhoun, M.A. Evaluation of empirically derived PVT properties for Pakistani crude oils. J. Pet. Sci. Eng. 1996, 16(4): 275-290.

Makridakis, S. Accuracy measures: Theoretical and practical concerns. Int. J. Forecast. 1993, 9(4): 527-529.

Moghadam, J.N., Salahshoor, K., Kharrat, R. Introducing a new method for predicting PVT properties of Iranian crude oils by applying artificial neural networks. Pet. Sci. Technol. 2011, 29(10): 1066-1079.

Mood, A., Graybill, F., Boes, D. Introduction to The Theory of Statistics (3rd Ed.). New York, USA, McGraw-Hill, 1974.

Oloso, M.A., Hassan, M.G., Bader-El-Den, M.B., et al. Hybrid functional networks for oil reservoir PVT characterisation. Expert Syst. Appl. 2017, 87: 363-369.

Omar, M.I., Todd, A.C. Development of new modified black oil correlations for Malaysian crudes. Paper SPE25338MS Presented at the SPE Asia Pacific Oil and Gas Conference, Singapore, 8-10 February, 1993.

Pearson, K. On the dissection of asymmetrical frequency curves. Phil. Trans. Roy. Soc. A 1894, 185: 71-110.

Petrosky Jr, G.E., Farshad, F. Pressure-volume-temperature correlations for Gulf of Mexico crude oils. Paper SPE26644MS Presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, 3-6 October, 1993.

Rammay, M.H., Abdulraheem, A. PVT correlations for Pakistani crude oils using artificial neural network. J. Pet. Explor. Prod. Technol. 2017, 7(1): 217-233.

Rao, R.V., Savsani, V.J., Vakharia, D.P. Teaching-learning-based optimization: An optimization method for continuous non-linear large-scale problems. Inform. Sciences 2012, 183(1): 1-15.

Samworth, R.J. Optimal weighted nearest neighbour classifiers. Ann. Stat. 2012, 40(5): 2733-2763.

Schmidhuber, J. Deep learning in neural networks: An overview. Neural Netw. 2015, 61: 85-117.

Shakhnarovich, G., Darrell, T., Indyk, P. Nearest-neighbor Methods in Learning and Vision: Theory and Practice (neural information processing). London, England, The MIT Press, 2006.

Standing, M.B. A pressure-volume-temperature correlation for mixtures of California oils and gases. Paper API47275 Presented at the API Drilling and Production Practice, New York, USA, 1 January, 1947.

Vapnik, V. Statistical Learning Theory. New York, USA, Wiley, 1998.

Varotsis, N., Gaganis, V., Nighswander, J., et al. A novel non-iterative method for the prediction of the PVT behavior of reservoir fluids. Paper SPE56745MS Presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, 3-6 October, 1999.

Vazquez, M., Beggs, H.D. Correlations for fluid physical property prediction. Paper SPE6719MS Presented at the SPE Annual Fall Technical Conference and Exhibition, Denver, Colorado, 9-12 October, 1977.

Wood, D.A. A transparent open-box learning network provides insight to complex systems and a performance benchmark for more-opaque machine learning algorithms. Adv. Geo-Energy Res. 2018a, 2(2): 148-162.

Wood, D.A. Transparent open-box learning network provides auditable predictions for coal gross calorific value. Model Earth Syst. Environ. 2018b, 1-25.

Wood, D.A., Choubineh, A., Vaferi, B. Transparent open-box learning network provides auditable predictions: Pool boiling heat transfer coefficient for alumina-water-based nanofluids. J. Therm. Anal. Calorim. 2019, 136(3): 1395-1414.

Wright, S. Correlation and causation. J. Agri. Res. 1921, 20: 557-585.

Appendix 1. Calculations involved in the TOB method

The fourteen calculation steps of the recently introduced transparent open-box (TOB) is described in detail elsewhere (Wood, 2018a; Wood et al., 2019). The following summary information is provided to enable readers to reproduce the calculation steps involved.

TOB Stage 1 (data matching to provide an initial prediction)

Step 1: Construct a 2-D array of number of variables defining the system and number of data records (M) involved. The variables are distinguished as N independent or influencing variables and a dependent variable which constitutes the prediction target.

Step 2: Configure the data into a sorted order of the prediction variable's values. This order can be ascending or descending.

Step 3: Compute the maximum and minimum values for the data range sampled by each data record in the data set. Other standard statistical measures for each variable (such as mean range, variance and standard deviation) also provide useful, but optional, metrics to characterize the dataset of interest. Summary statistical information for the B_{ob} dataset evaluated here is listed in Table 1.

Step 4: Maximum and minimum values for each variable are used to normalize all the variables for each data record in the data set to a range varying from minus one to plus employing Eq. (A-1).

$$X_i^* = 2 * \left[\frac{X_i - X_{min}}{X_{max} - X_{min}} \right] - 1 \tag{A-1}$$

where, X_i = the i^{th} data record for X of $N+1$ variables, X_{min} = minimum of variable X for the entire dataset, X_{max} = maximum of variable X for all data records, X_i^* = normalized value for the i^{th} record for X of $N+1$ variables.

Step 5. Review summary statistics for the adjusted / normalized values of each variable to verify that each variable has been normalized correctly, i.e., $-1 \leq X_i^* \leq +1$. It is good practice to make this important verification at this point in the process.

Step 6. Allocate all the data records in the dataset to either a training subset, a tuning subset or a testing subset. Sensitivity analysis identifies the optimum allocations to provide the most accurate predictions for a specific data set. As a rule, for most datasets the training subset is likely to constitute more than seventy percent of the dataset records. The two-stage process of the TOB is exploited in to perform the sensitivity analysis required. A series of TOB cases are run with different sizes of tuning subsets. Comparing the prediction performances of TOB stages 1 and 2 identifies the minimum number of data records needed in the tuning subset for the Stage 2 TOB predictions to consistently and reliably outperform the stage 1 predictions.

The requirement for separate tuning and training subsets is to enable the optimizer to tune the weights applied to the independent variables in the records of the training subset to provide better predictions for a representative, but relatively small, tuning subset records. For quite small data sets it is possible to conduct TOB stage 1 for all records in the data set. However, for datasets of more than 100 or so data records this adds to the computational effort without providing any benefits to the optimization process of TOB stage 2. By focusing on relatively small tuning subsets computational effort is reduced.

Step 7. Compute the variable-squared error (VSE) for each of J tuning-subset records versus the K training-subset records using Eq. (A-2):

$$VSE(X)_{jk} = [X_k(tr) - X_j(tu)]^2 \tag{A-2}$$

where, $X_k(tr)$ = variable X value for the k^{th} training-subset record, $X_j(tu)$ = variable X for the j^{th}tuning-subset record, $VSE(X)_{jk}$ = variable-squared error (VSE) for variable X for the j^{th} tuning-subset record versus the k^{th} training-subset record.

$\sum VSE_{jk}$ computes the weighted sum of the computed VSE values applying Eq. (A-3):

$$\sum VSE_{jk} = \sum_{n=1}^{n=N+1} VSE(Xn)_{jk} * (Wn) \tag{A-3}$$

where, $VSE(Xn)_{jk}$ = the variable-squared error (VSE) for variable Xn for the j^{th} tuning-subset record versus the kth training-subset record, $\sum VSE_{jk}$ = sum of variable-squared errors (VSE) for the $N+1$ variables (including the dependent variable) for the j^{th} tuning-subset record versus the k^{th} data training-subset record, Wn = weights ($0<\text{Wn} \leq 1$) applied to the calculated VSE for all variables involved in the prediction (i.e., $N+1$). Each weight is set to a constant value (e.g., 0.5 or 1.0) in TOB stage 1. This avoids any bias being introduced into the ranked initial matches derived for the tuning versus training subsets.

Step 8. Rank the matching data records in the training subset versus each tuning-subset record. The training-subset record that possesses the smallest calculated $\sum VSE$ value is identified as the best matching record for a specific tuning-subset record. The top-Q-matching training-subset records, established for each tuning-subset record, are then selected for the initial TOB-stage-one prediction. Q = 10 has empirically been found to be sufficient to provide reasonably accurate TOB-stage-one predictions from a number of distinct small to large non-linear dataset.

Step 9. The best identified matching records in the training subset (up to the limit set by the value of Q) for the j^{th} tuning-subset record each contribute fractionally to the TOB-stage-one prediction for that record. The exact contribution fraction applied to the Q top-ranking-matching records is established with Eqs. (A-4) , (A-5) and (A-6). The computed contribution fraction depends upon the $\sum VSE$ values for each training-subset record versus the j^{th} tuning-subset record.

$$f_{jq} = \frac{\sum VSE_{jq}}{\sum_{r=1}^{r=Q} \sum VSE_{jr}} \tag{A-4}$$

where, q = the q^{th} of Q top-ranking training-subset records from the training subset for the j^{th} tuning subset record, r = the r^{th} of Q top-ranking training-subset records from the training subset for the j^{th} tuning subset record, f_q = the contribution fraction calculated for the q^{th} of Q top-ranking records for the j^{th} tuning subset record.

Eq. (A-5) imposes a key constraint that normalizes the f_q values to sum to 1.

$$\sum_{q=1}^{q=Q} f_q = 1 \tag{A-5}$$

The best-matching training-subset record (i.e., the one with the lowest $\sum VSE_{jk}$ value) must make the greatest contribution to the prediction of the dependent variable associated with the j^{th} tuning-subset record. To facilitate this outcome the contribution fractions are applied as $(1 - f_q)$ multipliers in Eq. (6).

$$(X_{N+1})_j^{predicted} = \sum_{q=1}^{q=Q} \left[(X_{N+1})_q * (1 - f_q) \right] \tag{A-6}$$

where, $(X_{N+1})_q$ = dependent variable for the q^{th} training-subset record (i.e., one of Q best-matching records), $(X_{N+1})_j^{predicted}$ = TOB-stage-one predicted-dependent-variable value for the j^{th} tuning-subset record.

This TOB-stage-one prediction is provisional because for this prediction equal weights (Wn) are applied to the variables in TOB stage 1. This prediction is further refined in TOB stage 2.

Step 10. Three statistical metrics that establish accuracy are computed for the TOB-stage-one predictions, although a number of other accuracy metrics could also be used for this purpose. The ones applied in this study to the measured versus predicted values of dependent variable for all J tuning-subset records are:

Coefficient of determination (R^2) defined by Eq. (A-7):

$$R^2 = 1 - \frac{\sum_{j=1}^{j=J} \left(X_j^{actual} - X_j^{predicted} \right)^2}{\sum_{j=1}^{j=J} \left(X_{ave}^{actual} - X_j^{predicted} \right)^2} \tag{A-7}$$

Mean square error (MSE) defined by Eq. (A-8):

$$MSE = \frac{1}{J} \sum_{j=1}^{j=J} \left(X_j^{actual} - X_j^{predicted} \right)^2 \tag{A-8}$$

Root mean square error (RMSE) defined by Eq. (A-9):

$$RMSE = \sqrt{MSE} \tag{A-9}$$

where, X_j = dependent variable (previously referred to as $(X_{N+1})_j$ in Eq. (A-6)) for the j^{th} tuning-subset record, X_j^{actual} = the directly measured value of the dependent variable for the j^{th} tuning-subset record, $X_j^{predicted}$ = predicted value of the dependent variable for the j^{th} tuning-subset record, X_{ave}^{actual} = average measured value of the dependent variable for all J tuning-subset records.

TOB Stage 2 (optimizing the weights and number of matching records)

Step 11. Optimization is the key focus of TOB stage 2. This is achieved by minimizing the RMSE metric (Eq. A-9) measured across the entire set of J tuning-subset records. Two metrics applied as constants in TOB stage 1 are applied as optimization control metrics with specified constraints imposed.

The two TOB-stage-two optimization control metrics are Q and Wn:

1. The N input-variable weights (Wn) are allowed to vary across the full constrained range ($0 < Wn \leq 1$) leaving the optimizer free to select the best values for them to minimize RMSE. It is not unusual for quite small non-zero values to be assigned to certain Wn by the optimizer. However, these low weights often have significant impacts that improve prediction accuracy of TOB stage 2.

2. The optimizer is allowed to vary Q defining how many of the best-matching records should be used for TOB-stage-two prediction calculations computed by Eqs. (A-4), (A-5) and (A-6). For most datasets the optimizer is free to apply values of Q in the integer range $2 \leq Q \leq 10$ in its quest to minimize RMSE.

Here, the Generalized Reduced Gradient (GRG) non-linear optimization algorithm option of the standard "Solver" optimizer available in Microsoft Excel spreadsheets (Frontline Solvers, 2018) is employed for TOB-stage-two optimization. This is applied in a coded algorithm with the visual basic for application (VBA) language available as a standard feature of the Excel software. Other customized evolutionary optimizers (including one available in Excel's Solver package) could be utilized for this TOB step. For mid-sized datasets, calculating the TOB-stage-one and stage-two predictions in Excel, aided by VBA code, enables the TOB algorithm to routinely display all its intermediate calculations, which is ideal for transparency purposes.

The advantage of Excel-based configurations is that all of the TOB Stage 2 optimization calculations are then displayed on spreadsheets linked with Excel's cell-by-cell calculation formulas. This means that the intermediate calculation results are not only visible, but a user can interrogate all the numbers in Tables 1 to 3 for example via cell formulas involved in the calculation of those results. TOB configured in a fully coded manner in any programming language could display the intermediate calculation results, but they cannot easily display the formulas relating each number in the calculation to the next calculation step in the same display. Excel (or other spreadsheets) can do that.

The top-matching training-subset records (typically applying Q = 10 in TOB stage 1) established for the J tuning-subset records are carried forward from TOB stage 1 for selection by TOB stage 2. Eq. (A-3) is re-evaluated by the optimizer algorithm by varying Wn across its constrained range. TOB stage-2 $\sum VSE_{jq}$ values are also recomputed with Eq. (A-4) for different Q values ($2 < Q \leq 10$) in each iteration of the optimizer. This contrasts with the fixed value of Q applied in TOB stage 1 and leads to one value of Q in that range being identified as yielding the most accurate predictions.

Step 12. Compute the RMSE and R^2 accuracy metrics for the TOB-stage-two predictions. Compare the TOB-stage-2 predictions with the TOB-stage-1 predictions in terms of their accuracy to quantify the prediction improvements achieved by TOB stage 2, if any. Performing sensitivity analysis by optimizing with different fixed values of Q (i.e., Q = 2 to 10) also generates a set of useful sub-optimal solutions. This set of solutions (all but one being sub-optimal) provide insight to potential underfitting or overfitting issues with the data set being evaluated.

Step 13. Compute TOB-stage-one and TOB-stage-two predictions for the independent testing-subset records applying the optimum values established for Wn and Q in step 11 with the tuning subset. RMSE and R^2 accuracy metrics calculated for the testing and tuning subset predictions facilitate a direct comparison of their prediction accuracies.

As part of this step it is often appropriate to audit the intermediate calculation steps to reveal which variables are having the greatest impact on the TOB predictions. Reviewing the intermediate calculations can also facilitate comprehensive outlier analysis (i.e., understanding why some data records lead to less-accurate predictions than the main trend of predictions) and identify regions of the dependent-variable range that could be under-fitted by the TOB method.

Step 14. Compare the prediction accuracy provided by the TOB algorithm with empirical calculations and/or other machine learning algorithms and empirical correlations. Such comparisons can be used to complement and benchmark the prediction performance achieved by the less-transparent machine-learning methods for the data sets of interest. Also, they typically provide further insight to the relationships between the variables in the underlying dataset. It is best to use the results of these methods collectively to aid interpretation regarding specific data records or dependent-variable-value segments of the underlying dataset.

TOBScalability. The dataset evaluated in this study are small in terms of the number of data records and the number of independent variables involved in the predictions. However, the TOB learning network is scalable. It is now routinely being applied to datasets of up to 10,000 data records and 10 to 15 independent variables configured to use Excel Solver optimizers and/or fully coded customized optimizers. As the datasets get larger than about 10,000 records and 10 to 15 variables it becomes more efficient and flexibly to use a fully coded configuration for TOB Stages 1 and 2. For small-sized and medium-sized dataset there are advantages to use Excel Solver optimizers in parallel with fully coded optimizers. Development plans for the TOB learning network are to test whether it can be usefully adapted for application to process "big data" (e.g., 50,000 data records and more independent variables). This work is ongoing.

Appendix 2. Supplementary file of the dataset evaluated

An Excel file is available for readers to download that includes all the data records evaluated in this study.