



Total Organic Carbon Predictions from Lower Barnett Shale Well-log Data Applying an Optimized Data Matching Algorithm at Various Sampling Densities

DAVID A. WOOD¹

Abstract—Accurately estimating total organic carbon (TOC) from suites of well logs is essential as it is too costly and time consuming to take direct measurements from core samples in many wells. Unfortunately, the several methods developed over recent decades, based on various correlations and correlation-based machine learning methods, do not provide universally reliable, accurate or easily auditable TOC predictions. A method is developed and its viability evaluated exploiting a promising correlation-free, data-matching routine. This is applied to published well-log curves, with supporting mineralogical data and measured TOC, for two wells penetrating the Lower Barnett Shale formation at distinct settings within the Fort Worth Basin (Texas, U.S.). The method combines between 5 and 10 well log features and evaluates, on a supervised learning basis, multiple cases for nine distinct models at data-record-sampling densities ranging from one record for every 0.5 ft to one record for every 0.04 ft. At zoomed-in sampling densities the model achieves TOC prediction accuracies for the models combining data from both wells of ($RMSE \leq 0.3\%$ and $R2 \geq 0.955$) for models involving 6 and 10 input variables. It is the models involving six input variables that have the potential to be applied in unsupervised circumstances to predict TOC in surrounding wells lacking measured TOC, but that potential requires confirmation in future multi-well studies.

Keywords: Well log TOC estimates, data-matching machine learning, zoomed-data interpolation, correlation-free feature selections, model transparency, data record sample densities.

1. Introduction

Since the recognition in the 1940s that oil yield from shales was linked to their Uranium contents which was recorded by gamma-ray (Gr) well logs (Beers 1945; Swanson 1960) there has been intense interest in developing accurate methods to predict organic matter (OM) content and source rock properties from single or multiple well log curves. OM

has quite distinctive petrophysical properties from the inorganic minerals that dominate sedimentary rocks, as well as its distinctive chemistry dominated by hydrogen (H) and carbon (C). Its lower density and acoustic velocity (higher travel time) and higher resistivity and H content mean that the basic wireline logs, Gr, bulk density (Pb), resistivity (Rs), neutron (Np) and acoustic travel time (DT recorded from a P-sonic log), can all be used to an extent as indicators of OM, total organic carbon (TOC) and other source rock properties (Meyer and Nederlof 1984; Mann 1986; Fertl and Chilingar 1988).

The most accurate way to obtain TOC values from profiles through rock formations is through laboratory measurements on samples from well-bore cores, side-wall lateral-cores and drilling cuttings. To obtain such measurements is time consuming and costly and only provides intermittent sample analysis. The attraction of being able to somehow convert continuous wireline or measurement-while drilling (MWD) log curves into reliable TOC and other source rock parameters is, therefore, clear to see. Many methods have been proposed, tested and are applied with respect to TOC estimation. All have limitations and many fail to provide consistent accuracy.

Schmoker (1979, 1981, 1993) and Schmoker and Hester (1983) with his colleagues (Schmoker and Hester 1983; Hester and Schmoker 1987) provided an early and relatively simple but effective method based on bulk density (Pb) linear regressions. El Sharawy and Gaafar (2012) summarize the various well-log relationships in Schmoker's equations some of which involve Gr to indicate the presence of OM. Pb is a viable log to select because of the high contrast between shale and OM densities. Non-organic shale

¹ DWA Energy Limited, Lincoln, UK. E-mail: dwa@dwasolutions.com

typically has a matrix density of close to 2.7 g cm^3 and that contrasts significantly with OM density of $\sim 1.3 \text{ g cm}^3$. These methods provide reasonable results in homogeneous formations and are useful quick-look tools to apply, but in heterogeneous or highly laminated formations the results become unreliable.

Multi-variate regression analysis, first applied to predict TOC from several wireline log curves by Mendelson and Toksoz (1985), continues to be widely applied with local success (e.g., Verma et al. 2016; Alshakhs and Rezaee 2017). However, the established correlations typically work well for one location, but cannot be reliably applied universally.

Passey et al. (1990) proposed the popular delta-log-resistivity method (ΔLogR) for estimating TOC from well logs. Log-scale resistivity (in the range 0.2–2000 mV) and linear-scale sonic logs (in the range 40–140 $\mu\text{s/ft}$) are overlain following the establishment of the baseline Rs to DT relationship and thermal maturity. Separation of the log overlays can then be calibrated to TOC for certain ranges of thermal maturity. As with Pb regressions the method works well in some formations but not others. Passey et al. (2010) extended these relationships to more thermally mature zones appropriate shale gas resources evaluations. As the relationships apply to a limited range specific conditions with some linear assumptions, the single formula typically used to calculate ΔLogR is restricted to limited conditions (e.g. Alizadeh et al. 2018), although approximations can be made. Another key over-simplification of the method is that it assumes that the matrix of the kerogen-rich and kerogen-poor shale fractions of a formation have the same matrix log responses. This is clearly not always the case.

Various attempts have been made to improve or modify the ΔLogR method with limited success, e.g. by incorporating Pb values (Huang et al. 2015). Cheng et al. (2016) combined the ΔLogR method with a shale volume (Vsh) calculation to identify organic-rich shale lithofacies from gamma ray corrected for Vsh continuously from well logs. Although, Pb-regression and ΔlogR methods have a tendency to overestimate rather than underestimate TOC compared to measured data in some formations, that is not always the case. Smectite-rich shales tend

to display low resistivity due to the influence of high bound water contents. Mallick and Raju (1995) showed that the ΔlogR method did not work in areas experiencing conditions beyond its correlations. Zhao et al. (2016) provided examples showing the ΔlogR method providing inconsistent results in marine shales of the Sichuan Basin (China).

Others, particularly well-logging companies, have proposed full-scale petrophysical formation evaluation attempting to derive accurate porosity values to accurately estimate the amount of TOC and/or kerogen present (Herron 1988; Boyer et al. 2006; Herron et al. 2011). These methods are time consuming and have to overcome challenges of clay and Vsh quantification in shales, and the inability of Np and Pb to distinguish between porosity and kerogen. In practice, this means that additional data from specialist well logs (e.g., NMR and SpectroLith) which are only available on a few wells. In wells where natural gamma ray spectral log data is available the U and U/Th ratio can be used to estimate OM and TOC, as higher U/Th corresponds to higher TOC (Luffel et al. 1992; Wang 2013). However, the limited availability and high-cost of such logs means that such an approach is not suitable to the bulk of wells with only basic well-log data available.

Since the 1990s (Huang and Williamson 1996) there have been many attempts to demonstrate that various machine learning methods represent the way forward to overcome the challenges with linear and multi-variate correlations and direct TOC measurements. These methods are able to use all the well logs available (typically, Gr, Pb, Rs, Np, DT), but require extensive training and testing with multiple samples to verify the reliability of their predictions. Artificial neural network (ANN) have been most widely applied (Arbogast and Utley 2003; Alizadeh et al. 2011, 2018; Kadkhodaie et al. 2009; Khoshnoodkia et al. 2011; Bolandi et al. 2015; Tabatabaei et al. 2015, 2017). Some have applied statistical clustering to distinguish lithofacies prior to applying ANN to predict TOC for specific lithofacies (e.g. Sfidari et al. 2012). Others have applied combined neuro-fuzzy methods to calculate OM (Kamali and Mirshady 2004). Some have also applied ANN to well log data to predict a broader range of pyrolysis source rock metrics, e.g. S2 peak value (Alizadeh et al. 2018), or

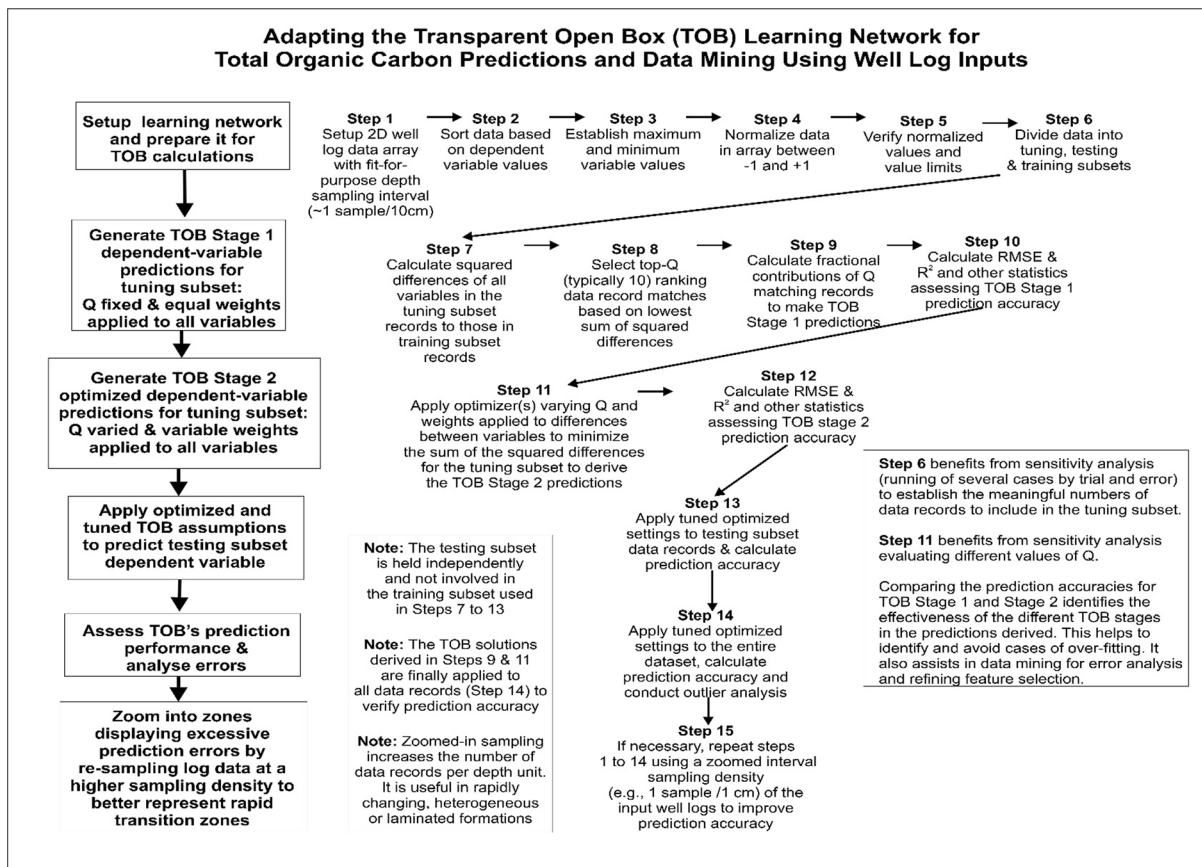


Figure 1

Workflow for applying TOB for total organic carbon prediction from well-log data

thermal maturity measures, e.g., vitrinite reflectance (Kadkhodaie and Rezaee 2017).

Support vector regression (SVR) has also been applied to predict TOC from multiple well logs (Liu et al. 2013; Tan et al. 2015) and shown to outperform ANN predictions applied to the same dataset but with coefficient of determination (R^2) less than 80%. Shi et al. (2016) demonstrated that an extreme learning machine (ELM) method and an ANN applied to the same dataset achieved broadly similar level of TOC prediction accuracy (R^2 from 0.88 to 0.93), but the ELM model executed much more rapidly.

There are some drawback with these regression-based machine-learning techniques. One is that their correlations are not transparently developed involving hidden layers or intricate and difficult-to-decipher

intermediate steps. Moreover, they are prone to over-fitting data in ways that are not easy to identify. Applied to laminated and heterogeneous formations their training and testing subsets may not adequately sample all sub-layers of a formation.

In this study, an optimized data-matching, machine-learning technique, the transparent-open-box algorithm, (TOB) is proposed (Wood 2018) to predict TOC from a suite of well logs and core data for two published Barnett Shale (Texas) well log sections. Significantly, this method does not involve the use of any correlations, but rather exploits, in a highly transparent manner, the local relationships between poorly-correlated and well-correlated variables. This study primarily seeks to validate this novel method for TOC prediction from well logs.

Table 1

Range, mean and units for the Barnett Shale variables measured in Wells (A and B, Verma et al. 2016)

Data sources	Variable	Symbol	Units	Min	Max	Mean
Standard logs recorded	Gamma ray	Gr	API	22.5	241.0	142.3
	Bulk density (RHOB)	Pb	g/cm ³	2.39	2.73	2.54
	Resistivity (deep—RTD)	Rs	ohm m	4.9	2074.9	182.4
	Neutron porosity (NPHI)	Np	fraction (v/v)	0.02	0.36	0.19
	Acoustic /P-sonic travel time	Dt	μs/ft	50.4	98.1	77.5
Core measurements	Quartz	Qz	volume %	0.28	58.93	25.23
	Limestone	Ls	volume %	0.00	59.14	11.26
	Clay	Cy	volume %	0.37	59.98	36.80
	Dolomite	Dl	volume %	0.00	49.41	3.53
Calculated petrophysical metric	Brittleness index	BI	fraction (v/v)	0.01	0.90	0.39
Core measurements	Total organic carbon	TOC	volume %	0.75	6.95	3.79

Table 2

Correlation matrix for Barnett shale variables recorded in wells A and B (Verma et al. 2016)

R	Gr	Pb	Rs	Np	Dt	Qz	Ls	Cy	Dl	BI	TOC
Gr	1	-0.1422	0.0906	0.4285	0.0368	0.1597	-0.2258	0.2512	0.0759	-0.0451	0.2740
Pb		1	-0.1060	-0.3552	-0.7044	-0.4032	0.3107	-0.1286	0.2447	-0.119	-0.38
Rs			1	-0.2209	-0.0919	0.2747	-0.0833	-0.0904	0.0935	0.2804	0.2299
Np				1	0.5902	-0.0634	-0.1474	0.3324	-0.1728	-0.355	0.1187
Dt					1	0.1742	-0.2417	0.1751	-0.2113	-0.042	0.1385
Qz						1	-0.5035	0.0303	-0.1997	0.6146	0.5097
Ls							1	-0.6669	0.0924	-0.139	-0.661
Cy								1	-0.319	-0.434	0.4947
Dl									1	0.3593	-0.221
BI										1	0.1091
TOC											1

For input abbreviations see Table 1

Future work is planned to apply the algorithm and its application methodology to predict TOC in multiple wells across a basin.

2. Methods

2.1. Barnett Shale Data for Wells A and B

The Barnett Shale is a TOC-rich prolific shale-gas producing formation of Mississippian age distributed across the Bend Arch-Fort Worth Basin with its well documented tectonic history (Walper 1982), lithofacies distribution and stratigraphic framework (Abouelresh and Slatt 2012). It is the basin that launched today's prolific shale gas industry (Jarvie

et al. 2007). The Barnett shale is deepest towards the northeast reaching more than 1000 ft (305 m) in thickness with substantial limestone interbeds but thins on to the Chappel shelf to less than 50 feet in places (Pollastro et al. 2007). The formation made up of organic-rich shales that alternate with thin shallow-marine limestones (Singh et al. 2008). Wells A and B (Verma et al. 2016) from which the data evaluated in this study is derived are situated in the northeast of the basin where most gas production comes from the lower part of the formation. Well B is some 30 miles (48 km) west-south-west of Well A. Both wells, for the Lower Barnett Shale (LBS) have extensive TOC measurements taken on rock samples by Rock-Eval, have a basic suit of well log data (Gr,

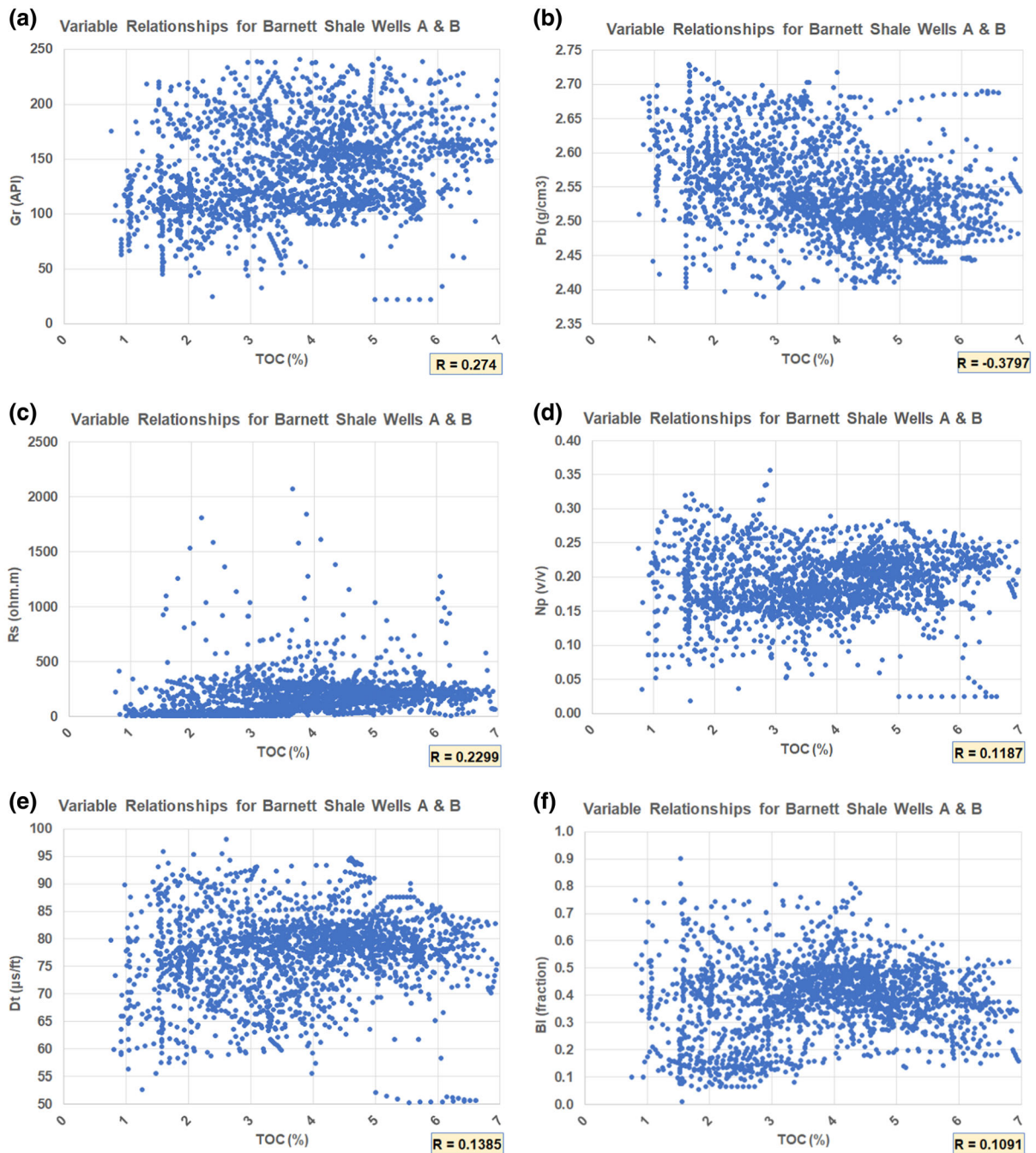


Figure 2.

a Gamma ray (Gr), **b** bulk density (Pb), **c** resistivity (Rs), **d** neutron porosity (Np), **e** P-sonic travel time (Dt), and **f** brittleness index (BI) values for Barnett Shale wells A and B (data for 2000 depth intervals derived from Fig. 4 of Verma et al. 2016).

Pb, Rs, Np, DT) and Fourier-transform infra-red (FTIR) mineralogical measurements (volume percent) for quartz, limestone, clay and dolomite (Qz, Ls, Cl, DI) all displayed as log versus depth curves

(Verma et al. 2016). They calculate ΔLogR and use the mineralogical data to calculate the Brittleness index (BI) using the Wang and Gale (2009) ratio, which includes measured TOC along with the four

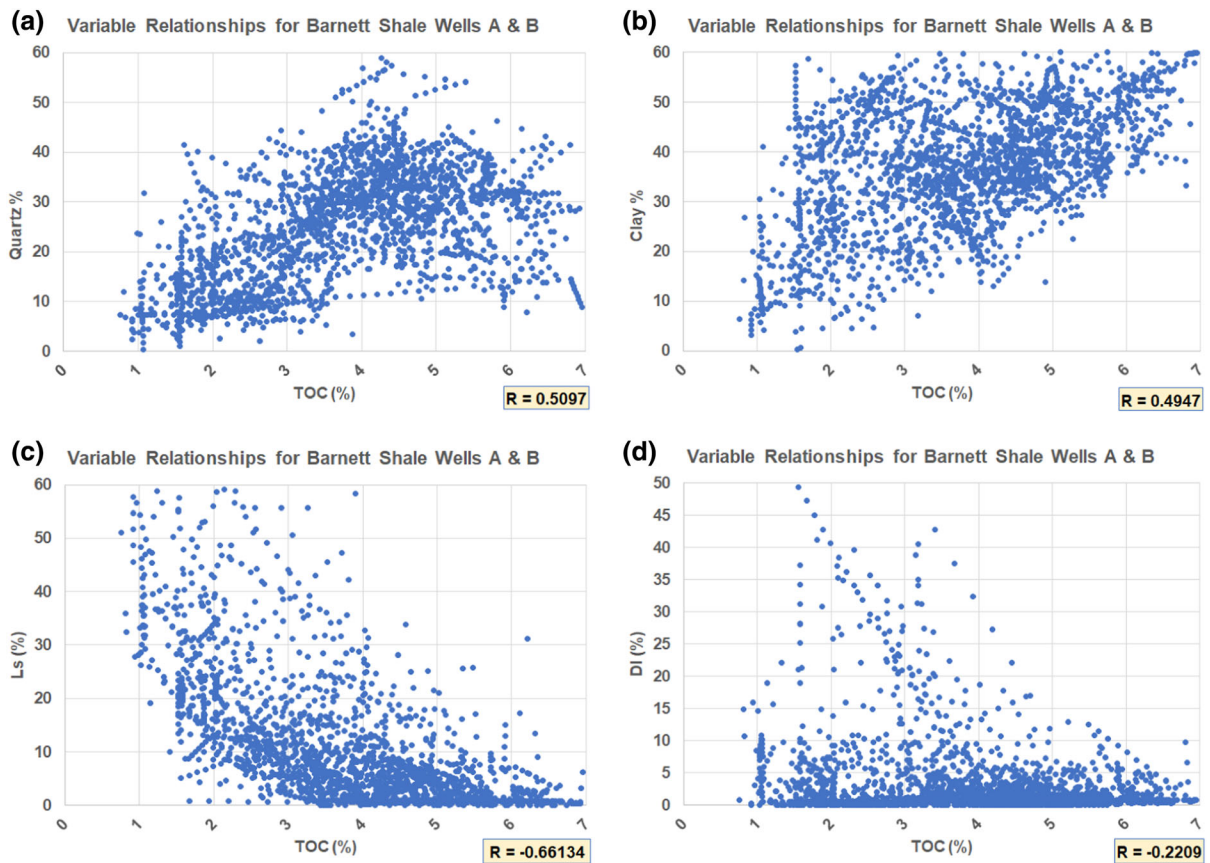


Figure 3

Core mineralogy volume measurements for **a** quartz, **b** clay, **c** limestone, and **d** dolomite for Barnett Shale wells A and B (data for 2000 depth intervals derived from Fig. 9 of Verma et al. 2016)

mineral volumes in its denominator and quartz plus dolomite in the numerator. Both BI and ΔLogR along with the mineralogical and petrophysical data for both wells are displayed as log-depth curves. For this study 11 variables (Gr, Pb, Rs, Np, DT, Qz, Ls, Cl, DI, BI and measured TOC) were digitised and initially each curve is sampled 1000 times across the depth intervals of interest. For the two wells studied the depth intervals of interest encompass the Lower Barnett Shale (Well A 8000–8470 ft; Well B 6490–6790 ft). This provides a resampling of each log curve at every 0.47 ft for well A and a sample for each well-log curve for well B at every 0.30 ft. That sampling density is sufficient to provide a meaningful starting point resolution for TOC prediction analysis. For thicker zones of interest, clearly, a greater

number of resamples would be necessary to achieve that resolution. This dataset was then assessed in various combinations and sampling densities by interpolating between the original 1000 points sampled on each log curve. This made it possible to test the ability of the TOB algorithm to predict TOC for the LBS, evaluating both wells separately and together.

2.2. Transparent-Open-Box Optimized Data Matching

A step-by-step implementation procedure for TOB is detailed in Wood (2019a). The workflow for this algorithm adapted for TOC estimation from well logs is displayed in Fig. 1. A key feature of the

Table 3
Summary of TOC prediction models evaluated for Barnett Shale wells A and B

Model #	Barnett Shale well models evaluated with TOB and solutions compared			
	Model description	Sampling scale of data records	Solutions compared	Q
1	Well A: 5 basic well logs only 8000–8470 ft	1 sample per 0.47 ft	TOB Stage 1	10
		0.47	Stage 2 Best:1F1	2
		1000 data records	Average of 45 Stage 2 runs	3.8
2	Well A: 10 input variables 5 logs + 4 mineral volumes + BI	1 sample per 0.47 ft	TOB Stage 1	10
		1000 data records	Stage 2 Best:5G2	2
		1 sample per 0.47 ft	Average of 45 Stage 2 runs	2.5
3	Well A: 10 input variables 5 logs + 5 mineral to log ratios BI not included as an input	1 sample per 0.47 ft	TOB Stage 1	10
		1000 data records	Stage 2 Best:4G1	2
		1 sample per 0.047 ft	Average of 45 Stage 2 runs	2.8
4	Well A: 10 input variables 5 logs + 4 mineral volumes + BI Zoomed (8376 to 8470 ft)	1 sample per 0.047 ft	TOB Stage 1	10
		0.047	Stage 2 Best:1E2	2
		2000 data records	Average of 49 Stage 2 runs	4
5	Well A: 5 input variables 5 basic logs only Zoomed (8376–8470 ft)	1 sample per 0.047 ft	TOB Stage 1	10
		0.047	Stage 2 Best:2F1	4
		2000 data records	Average of 50 Stage 2 runs	4.2
6	Well B: 5 basic well logs only 6490 to 6790 ft	1 sample per 0.3 ft	TOB Stage 1	10
		0.3	Stage 2 Best:2F3	6
		1000 data records	Average of 45 Stage 2 runs	6.1
7	Well B: 10 inputs 5 logs + 4 mineral volumes + BI	1 sample per 0.3 ft	TOB Stage 1	10
		1000 data records	Stage 2 Best:3G1	3
		1 sample per 0.3 ft	Average of 45 Stage 2 runs	2.5
8	Well A + B: 10 inputs 5 logs + 4 mineral volumes + BI	Well A (1 sample/0.47ft)	TOB Stage 1	10
		Well B (1 sample/0.3ft)	Stage 2 Best:4E3	2
		2000 samples	Average of 45 Stage 2 runs	2.9
9	Well A + B: 6 inputs 5 logs + Stratigraphic height 10,000 data records	Well A (1 sample/0.094ft)	TOB Stage 1	10
		Well B (1 sample/0.06ft)	Sensitivity Best:T14	3
		10,000 data records	Average 49 B Stage 2 solutions	2.79

algorithm is that its evaluation is divided into two stages, each providing an informative and useful estimate that work together to expose cases of overfitting. Each data record evaluated contains all the variable values for a simple depth interval.

TOB Stage 1 assesses matches between records in a small, evenly distributed subset designed to tune the algorithm with a larger subset available for solution training. It does so, seeking the best ten ($Q = 10$) matches, which are those with the lowest sum of the squared difference (VSE_{jk}) between the input variables. Step seven (Fig. 1) is the key procedure for TOB Stage 1 and is calculated with Eq. (1)

$$\sum VSE_{jk} = \sum_{n=1}^{n=N} VSE(Xn)_{jk} \times (Wn), \quad (1)$$

where n is one of N input variables, $VSE(Xn)_{jk}$ is the variable-squared error (VSE) or difference for

variable Xn , j refers to the j th record in the tuning subset, k refers to the k th record in the training subset. Each record j is matched against all records in the training subset, $\sum VSE_{jk}$ is sum of (VSE) differences for all N input variables for each (j) record in the tuning subset matched against all data records (k) in the training subset.

Wn weights ($0 < Wn \leq 1$) are multiplied with the VSE for each of N variables. Each input variable weight = 0.5 for TOB stage 1. This removes bias towards any input variable. For most applications, such as TOC prediction, it is appropriate to set the weight for the dependent variable (i.e. variable $N + 1$) to zero so it does not influence data matching or solution tuning.

The training subset consists of data records not selected in a particular run for small tuning subset or to be held separately (testing subset) to independently

Table 4

Statistical measures of TOC prediction accuracy achieved by the nine models evaluated with the TOB algorithm involving Barnett Shales wells A and B

Model #	Solutions compared	Statistical measures of TOC prediction accuracy for best TOB stage 2 solution for each model (i.e., best of 45 solutions derived for each model applied to all data records involved in the model)						
		RMSE (%)	MSE	APD%	AAPD%	SD (%)	R	R ²
1	TOB Stage 1	0.7238	0.5239	− 8.20	20.11	0.7241	0.8722	0.76077
	Stage 2 Best:1F1	0.6726	0.4524	− 6.10	16.57	0.6729	0.8902	0.79250
	Average of 45 Stage 2 runs	0.7022	0.4936	− 6.69	17.21	0.7022	0.8798	0.77416
2	TOB Stage 1	0.4321	0.1867	− 2.41	9.99	0.4319	0.9569	0.91570
	Stage 2 Best:5G2	0.3342	0.1117	− 0.99	5.15	0.3343	0.9740	0.94869
	Average of 45 Stage 2 runs	0.3530	0.1249	− 0.97	5.44	0.3530	0.9709	0.94267
3	TOB Stage 1	0.4843	0.2346	− 3.70	11.89	0.4846	0.9456	0.89411
	Stage 2 Best:4G1	0.3796	0.1441	− 1.38	6.66	0.3798	0.9664	0.93383
	Average of 45 Stage 2 runs	0.4071	0.1662	− 1.86	7.56	0.4073	0.9611	0.92367
4	TOB Stage 1	0.0252	0.0006	− 0.05	0.23	0.0252	0.9998	0.99963
	Stage 2 Best:1E2	0.0132	0.0002	− 0.01	0.06	0.0132	0.9999	0.99989
	Average of 49 Stage 2 runs	0.0229	0.0006	− 0.04	0.24	0.0229	0.9999	0.99973
5	TOB Stage 1	0.1053	0.0111	− 0.16	0.60	0.1053	0.9967	0.99332
	Stage 2 Best:2F1	0.0184	0.0003	− 0.01	0.12	0.0184	0.9999	0.99980
	Average of 50 Stage 2 runs	0.0334	0.0013	− 0.03	0.19	0.0334	0.9996	0.99919
6	TOB Stage 1	0.6424	0.4126	− 4.54	13.43	0.6421	0.8639	0.74626
	Stage 2 Best:2F3	0.5082	0.2583	− 2.05	8.72	0.5085	0.9167	0.84035
	Average of 45 Stage 2 runs	0.5364	0.2881	− 1.94	8.76	0.5366	0.9074	0.82349
7	TOB Stage 1	0.3560	0.1267	− 2.29	6.66	0.3550	0.9611	0.92374
	Stage 2 Best:3G1	0.1939	0.0376	− 0.58	2.93	0.1940	0.9884	0.97686
	Average of 45 Stage 2 runs	0.2166	0.0471	− 0.73	2.53	0.2166	0.9854	0.97097
8	TOB Stage 1	0.4105	0.1685	− 2.41	8.63	0.4106	0.9568	0.91552
	Stage 2 Best:4E3	0.2967	0.0880	− 0.91	3.74	0.2968	0.9774	0.95538
	Average of 45 Stage 2 runs	0.3117	0.0974	− 1.06	4.70	0.3118	0.9750	0.95062
9	TOB Stage 1	0.1180	0.0139	− 0.31	1.37	0.1180	0.9964	0.99291
	Sensitivity Best:T12	0.0419	0.0018	− 0.03	0.48	0.0419	0.9996	0.99910
	Average 49 Stage 2 solutions	0.0634	0.0045	− 0.04	0.53	0.0634	0.9988	0.99770

test and verify the solutions generated by the tuning subset. The suitable sizes of the tuning and testing subsets are established by trial-and-error sensitivity analysis. For the Barnett shale well datasets used in this study, such sensitivity analysis suggested tuning datasets with about 135 data records ($\sim 13.5\%$ of the dataset) and testing subsets with about 100 data records ($\sim 10\%$ of the dataset) were sufficient to provide repeatable levels of accuracy from five different sample-selection cases. That left about 765 data records ($\sim 76.5\%$ of the dataset) in the training subset to be used for data matching. The samples included in the tuning and testing subset are selected without replacement, so that the same data record cannot appear in more than one of the subsets in any specific sample selection case. The samples selected for the tuning and testing subsets are also spread as widely across the dependent variable range as

possible. Data records for those subsets are not selected randomly as this could cluster data records in certain sections of the dependent variable distribution and leave other areas unsampled in some cases. Also, sampling is configured for five distinct sample selection cases in a way that avoids the same data records being selected for more than one tuning or testing subset involved in those cases.

The computed fraction that each of the highest-ranked matches contributes to the calculated stage 1 prediction (Step 9 Fig. 1) is calculated using the relative magnitude of their $\sum VSE$ values from Eq. 1. This is established with Eq. 2.

$$f_{jq} = \sum VSE_{jq} / \left[\sum_{r=1}^{r=Q} \sum VSE_{jr} \right] \quad (2)$$

Table 5

TOB Stage 2 optimum solutions for TOC prediction derived by the nine models evaluated involving Barnett Shales wells A and B

Weights applied to squared differences of input variables for TOB stage 2 optimized solutions (i.e., best of 45 or more solutions derived for each model)

Model #	1	2	3	4	5	6	7	8	9
Solutions compared	Stage 2 Best:1F1	Stage 2 Best:5G2	Stage 2 Best:4G1	Stage 2 Best:1E2	Stage 2 Best:2F1	Stage 2 Best:2F3	Stage 2 Best:3G1	Stage 2 Best:4E3	Sensitivity Best:T12
Q	2	2	2	2	4	6	3	2	3
GR wt	0.06470	0.45976	0.00000	0.05485	0.41302	0.00000	0.00000	0.00023	0.00000
Pb wt	0.22394	0.00000	0.15105	0.00147	0.00000	0.00000	0.01141	0.00000	0.00000
Rs wt	0.99941	0.00000	0.01091	0.00325	0.00000	0.54382	1.00000	0.89613	0.00000
Np wt	0.04835	0.00000	0.12199	0.00399	0.53091	0.01845	0.02016	0.05175	0.25000
DT wt	0.24755	1.00000	0.19805	0.04781	1.00000	0.04972	0.00000	0.13307	0.20000
Qz wt	N/A	0.00000	N/A	0.00364	N/A	N/A	0.01423	0.95501	N/A
Ls wt	N/A	0.17776	N/A	0.00221	N/A	N/A	0.00310	0.00081	N/A
Cy wt	N/A	0.00000	N/A	0.40988	N/A	N/A	0.00000	0.03654	N/A
Dl wt	N/A	0.00000	N/A	0.98416	N/A	N/A	0.00895	0.27057	N/A
Bl wt	N/A	0.47068	N/A	0.00089	N/A	N/A	0.00000	0.00980	N/A
Gr/Pb	N/A	N/A	0.00000	N/A	N/A	N/A	N/A	N/A	N/A
Gr/Rs	N/A	N/A	0.71571	N/A	N/A	N/A	N/A	N/A	N/A
Cy/Pb	N/A	N/A	0.00000	N/A	N/A	N/A	N/A	N/A	N/A
Qz/Pb	N/A	N/A	0.00983	N/A	N/A	N/A	N/A	N/A	N/A
Cy/Np	N/A	N/A	1.00000	N/A	N/A	N/A	N/A	N/A	N/A
Sth	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.60000

$q = q$ th highest-ranked training record related to the j th tuning record, $r = r$ th highest-ranking training record related to the j th tuning record.

There are Q total specified high-ranking data matches considered for each j th tuning record $f_q =$ fractional contribution to sum of the squared errors of q th highest-ranked training record related to the j th tuning record

These fractional contributions are adjusted to sum to 1 ($\sum_{q=1}^{q=Q} f_q = 1$).

The highest-ranked training record match is that with the lowest $\sum VSE_{jk}$.

That highest ranked match also has the smallest value of f_q .

The training record with lowest value of f_q must provide the largest share to the TOC prediction of the j th tuning record. For this to occur, it is necessary to multiply by $(1 - f_q)$ in Eq. (3):

$$(X_{N+1})_j^{predicted} = \sum_{q=1}^{q=Q} [(X_{N+1})_q \times (1 - f_q)], \quad (3)$$

where $(X_{N+1})_q =$ TOC of q th high-matching training record, $(X_{N+1})_j^{predicted} =$ predicted TOC value for j th tuning record.

TOB Stage 2 involves optimization to identify the most favourable weights (each W_n is allocated a real number between 0 and 1) that reduce the VSE values for each variable. Equation (1) is re-evaluated for the top Q of the ten best-matching records selected for the stage 1 predictions. In Stage 2, the value of Q can fluctuate. An optimum Q value ($2 \leq Q \leq 10$) is identified that minimizes root-mean-squared error (RMSE) between predicted and measured TOC. This is step 11 (Fig. 1). W_n values are adjusted to determine what fraction each input variable will contribute to the TOC prediction of each (j) tuning record. Equation (1) to Eq. (3) are re-evaluated to derive the TOB Stage 2 predictions.

2.3. Optimizers Applied for TOB Stage 2 Prediction

A customized firefly optimizer (Wood 2019b) facilitating metaheuristic profiling (Wood 2016) and coded in visual basic for applications (VBA) is used

for TOB Stage 2 evaluations. Also, two of Microsoft Excel's Solver optimizers, generalized reduced gradient (GRG) and evolutionary (Frontline Solvers 2020) are run to verify the results. This combination of optimizers has proved to be effective at providing a complementary range of near-optimum solutions around the global minimum for each TOB Stage 2 case evaluated.

Based on sensitivity analysis, tuning subsets with between 135 and 150 data records, spread across the entire data set range of TOC, were found to generate statistically repeatable TOC predictions (TOB stage 1 and 2) for the well-log datasets evaluated in this study. These were assessed with testing subsets of 100 data records that were held separately and not involved in the tuning or training subsets or in any way contributing to the solution-tuning process. For each model evaluated 5 distinct cases were considered. In each case, different sets of data records are assigned to the tuning and testing subset, and what records remain in each case go into the training subset. For each case three optimizers were executed three times, resulting in 45 TOB Stage 2 tuned solutions. Those 45 optimized solutions were then applied to the entire dataset to establish an average prediction accuracy and to identify the case that produced the best overall TOC predictions for each model. Also, those 45 solutions for the testing subset were compared with the five TOB Stage 1 solutions (one for each case) for the testing subset. That comparison is focused on distinguishing the few optimized solutions that exhibited overfitting.

2.4. TOC Prediction Accuracy Statistical Measures Assessed

It is revealing to assess prediction accuracy of machine learning methods using several distinct statistical measures. The measures recorded for this study are:

- Root mean square error (RMSE)—objective function of the TOB optimizers.
- Mean square error (MSE).
- Percent deviation between measured and predicted values (PD).
- Average percent deviation (APD).

- Absolute average percent deviation (AAPD).
- Standard deviation (SD).
- Correlation coefficient (R) between measured and predicted values.
- Coefficient of determination (R^2).

The precise formulas used to calculate these statistical accuracy measures are well known (Wood 2019a).

3. Characterization of the Barnett Shale Sampled by Wells A and B

Tables 1 lists the value ranges, means, symbols and units of ten input variables available for TOC prediction. Five of those input variables are derived from the standard well log data curves available for many oil and gas wells drilled (i.e., Gr, Pb, Rs, Np, Dt, see Table 1 for symbol definitions). Four of the other five input variables are mineralogical data (volume percent) derived from core analysis (i.e., Qz, Ls, Cy, Dl, see Table 1 for symbol definitions). The final input variable is the Brittleness Index (BI) originally defined using mineralogical variables only (Jarvie et al. 2007) and later modified to include TOC in the denominator (Wang and Gale 2009) as expressed in Eq. (4):

$$BI = (Qz + Dl) / (Qz + Dl + Ls + Cl + TOC). \quad (4)$$

It is Eq. (1) that Verma et al. (2016) used to calculate BI and express it as a distinct log curve for Wells A and B (their Fig. 4). BI is clearly not independent of TOC, but because TOC in wells A and B is less than 7 volume percent, it makes a very minor contribution to BI calculated using Eq. (4).

The correlation matrix between these variables (Table 2) reveal that none of the standard well log metrics correlates strongly with TOC (i.e., the highest correlation coefficient is -0.38 between bulk density, Pb, and TOC). On the other hand, three of the mineralogical variables (Qz, Ls and Cl) display better correlations with TOC, with Ls having the highest correlation coefficient (0.66) with TOC. The low correlation coefficient (0.11) between BI and TOC testifies to the very small influence that TOC in the

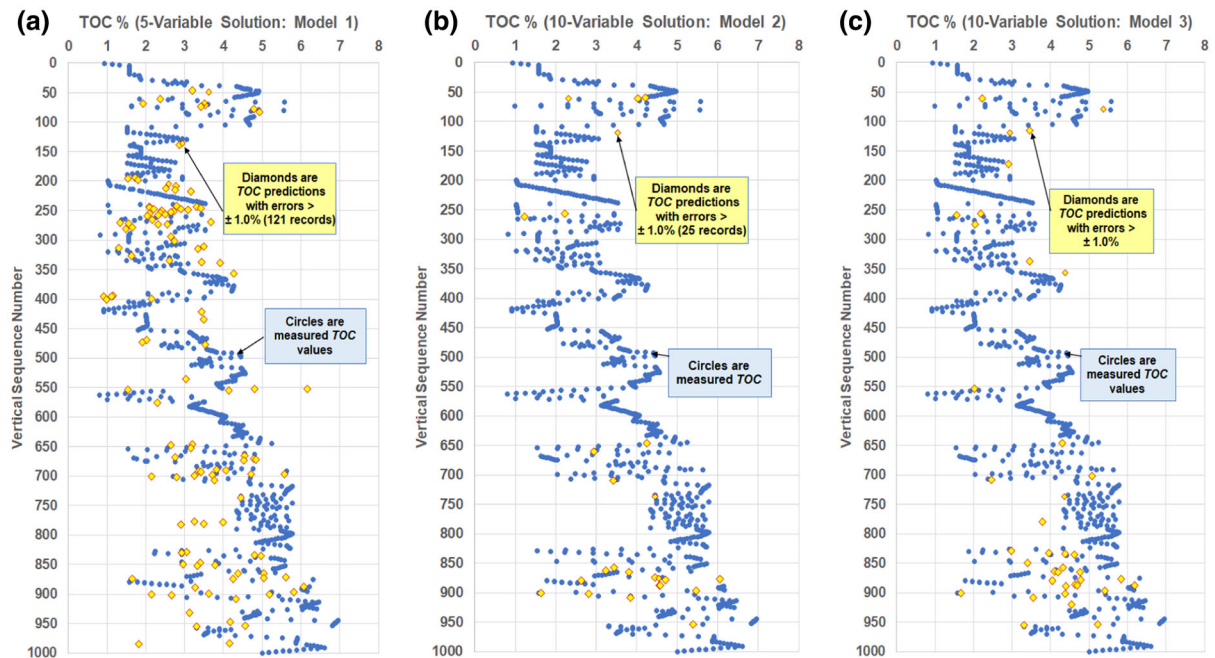


Figure 4

TOC versus sequence number for well A Models, 1, 2 and 3. The data records associated with prediction errors of $\pm > 1\%$ TOC are distinguished

denominator of the BI formula, as applied, has on the BI value.

Variable relationships with TOC are non-linear displaying substantial scatter. There is occasional clustering associated with some of the input variables compared to TOC. These relationships are displayed in Figs. 2 and 3. These generally poor correlation relationships make it difficult for correlation-based machine learning methods (e.g. ANN, LSSVM, multi-linear and non-linear regression) to derive accurate predictions of TOC. The TOB method, by avoiding the involvement of correlations is able to exploit other aspects of the input variable distributions to derive its TOC predictions. For some datasets, it is appropriate to consider the resistivity variable in logarithmic terms, particularly where data values range over several orders of magnitude. This serves to maximise the separation of actual and normalized values in the lower end of the lower resistivity scale and diminish the impact of very high values. As the resistivity for the lower Barnett shale in the two wells studied varies only between about 5 and 2075 Ω -m, a normal resistivity scale is sufficient.

Sensitivity analysis comparing results with normal- and logarithmic-scaled resistivity showed insignificant differences. In datasets where resistivity values vary between fractions of ohm-m and tens of thousands of ohm-m, it would be more appropriate to use logarithm resistivity values for TOC prediction analysis.

4. Results

4.1. TOB Models Applied to Barnett Shale Wells A and B to Predict TOC

Nine models with different input feature selection and/or record sampling density were constructed for evaluation with the TOB algorithm with a view to establishing accurate predictions of TOC from well log data with or without assistance from core mineralogy data. These nine models are summarized in Table 3 and distinguished as follows:

Model 1—Well A five input features only (GR, Rs, Pb, Np, DT) 1000 data records (covering the interval 8000–8470 ft);

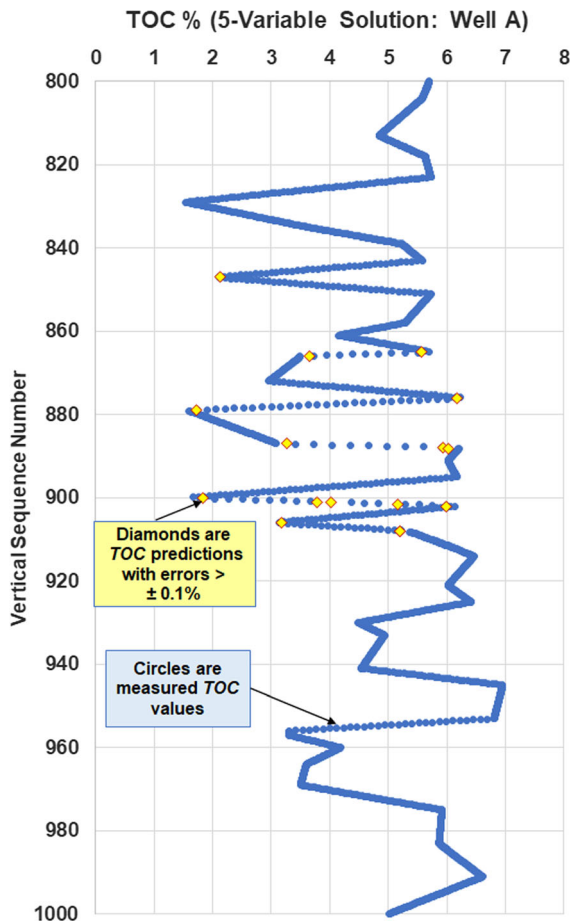


Figure 5

TOC versus sequence number for well A Model 4 (zoomed-in case for interval 800–1000 using 5 well log inputs). The 16 of 2000 data records associated with prediction errors of $\pm > 0.1\%$ TOC are distinguished

Model 2—Well A ten input features selected (GR, Rs, Pb, Np, DT, Qz, Ls, Cy, DI, BI) 1000 data records (covering the interval 8000–8470 ft);

Model 3—Well A ten inputs features selected including ratios between well-log variables and core mineralogy volumes (GR, Rs, Pb, Np, DT, Gr/Pb, Gr/Rs, Cy/Pb, Qz/Pb, Cy/Np) 1000 data records (covering the interval 8000–8470 ft);

Model 4—Well A ten input features selected (GR, Rs, Pb, Np, DT, Qz, Ls, Cy, DI, BI) 2000 data records (covering the interval 8376–8470 ft), a zoomed-in section of the Model 2 variables across the lower Well A depth interval, sampling all ten input

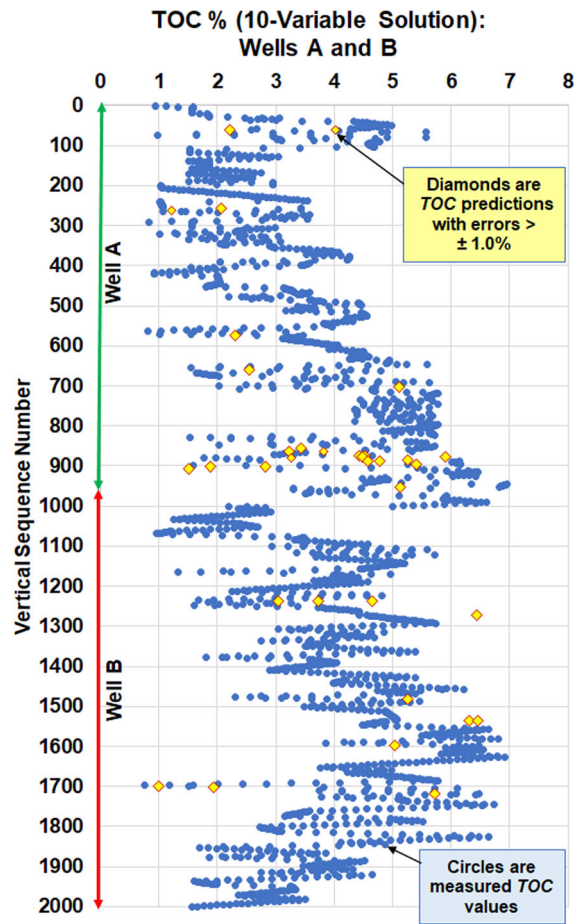


Figure 6

TOC versus sequence number for the combined Well A and B dataset of Model 8. The 34 of 2000 data records associated with prediction errors of $\pm > 0.1\%$ TOC are distinguished

variables for that section ten times more densely than in Model 2;

Model 5—Well A five input features selected (GR, Rs, Pb, Np, DT) 2000 data records (covering the interval 8376–8470 ft), a zoomed-in section of the Model 1 variables across the lower Well A depth interval, sampling all five input variables for that section ten times more densely than in Model 1;

Model 6—Same as *Model 1* but for well B (6490–6790 ft);

Model 7—Same as *Model 2* but for well B (6490–6790 ft);

Model 8—Same as *Models 2* and *7*, but combining the 1000 records of each well to form a combined

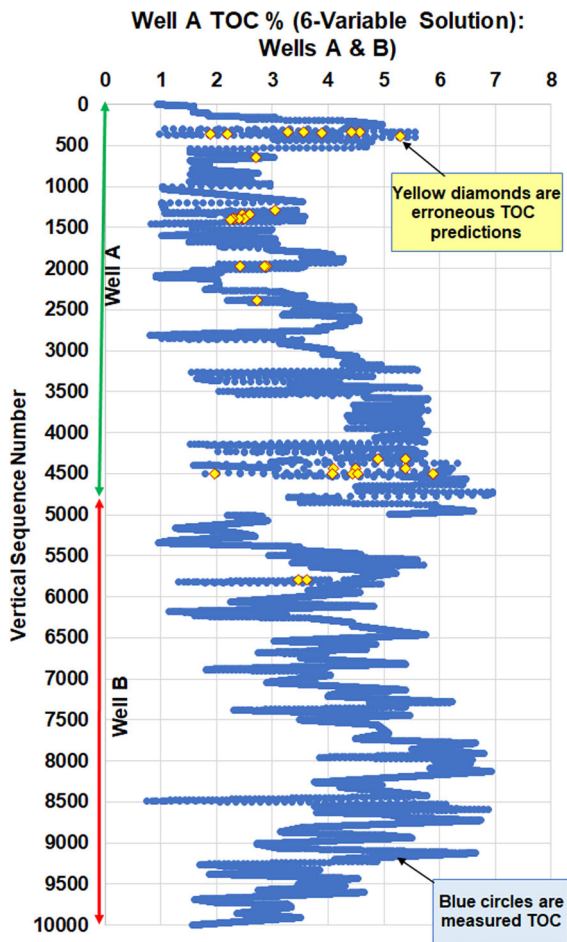


Figure 7

TOC versus sequence number for the combined well A and B dataset of Model 9. The 34 of 10,000 data records associated with prediction errors of $\pm > 0.316\%$ TOC are distinguished

2000-record integrated model incorporating both wells A and B;

Model 9—Combining 5000 samples from each of wells A and B using six input variables, the five basic well logs and stratigraphic height, with both wells sampled 5 times more densely than *Model 8* to form a combined 10,000-record integrated model.

A summary of the prediction accuracies achieved for specified optimum solutions used to evaluate Models 1 to 10 with the TOB algorithm is provided in Tables 4 and 5.

For each of the nine models more than 45 optimized (TOB Stage 2) solutions were established and TOC prediction accuracies derived. Table 4

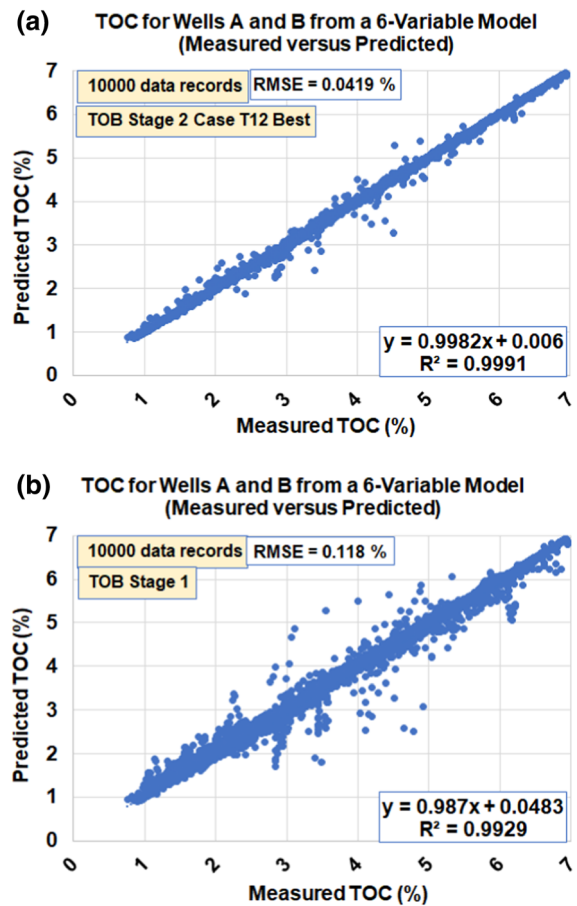


Figure 8

TOC predictions versus interpolated measured data for Model 9. The prediction accuracy is high for this denser data record sampling (one sample per 0.09ft for well A and 0.06ft for well B). The 34 significant prediction errors mentioned in the text for TOB Stage 2 solution (a) are clearly visible. Only two of those errors are associated with well B. The TOB Stage 1 solution (b) itself achieves high accuracy, but with 236 records displaying TOC prediction errors of $\pm > 0.316\%$ TOC

displays the average TOC prediction accuracies achieved by all these solutions together with the TOB stage 1 and best TOB Stage 2 solutions.

Comparison of the best TOB Stage 2 solutions for Models 1 and 6 indicates that using only the five basic well log variables as inputs TOB achieves TOC prediction accuracies that are useable (RMSE of 0.51% (Well B) and 0.67% (Well A); R^2 of 0.84 (Well B) and 0.79 (Well A). The higher accuracy achieved for Well B is likely to be a consequence of the more closely spaced sampling. one sample per

0.3ft for Well B versus one sample per 0.47ft for Well A. However, TOC prediction accuracy is much improved for these sampling densities for the 10-variable evaluations provided by Models 2 and 7 (RMSE of 0.19% (Well B) and 0.33% (Well A); R^2 of 0.98 (Well B) and 0.95 (Well A)). Model 3 for well A uses a different set of 10 variables (i.e., basic well logs plus some ratios between selected mineralogical variables and the basic well logs) but fails to match the TOC prediction accuracy achieved by Model 2. The absence of BI as an input variable in Model 3 may be contributing to the slight reduction of TOC prediction accuracy for Model 3 (RMSE of 0.38%; R^2 of 0.93) versus Model 2.

Figure 4 identifies the TOC prediction errors of $\pm > 1\%$ for the 5-variable and 10-variable models for well A with one sample per 0.47ft (121 data records for Model 1; just 25 data records for Model 2; and, 39 records for Model 3). This highlights the improvements in accuracy achieved by adding the mineralogical data as input variables. For the 10-variable models for Well A it is apparent (Fig. 4b, c) the majority of the significant TOC prediction errors are located in the lower part of the Barnett Shale (i.e. in sequence interval 800–1000). In particular, it is within the rapidly changing TOC laminated section of that lower interval that the 10-variable models struggle to predict accurately at one sample per 0.47 ft.

Models 4 and 5 demonstrate that by substantially increasing the sample density for the lower interval of Well A (i.e., the sequence interval from 800 to 1000 is resampled for all variables at one sample for all variables every 0.047 ft) the TOC prediction accuracy for that interval is dramatically improved for 5-variable Model 4 and 10-variable Model 5. The finer sampling density achieved by interpolation between all log values used in Models 1 and 2, enables the TOB algorithm applied to the zoomed-in Models 4 and 5 to generate no prediction errors of $\pm > 1\%$ for sequence 800 to 1000. This is particularly encouraging for the 5-variable Model 4, as it suggests that with finer sampling, using just the basic well log inputs, highly accurate TOC predictions can be achieved (RMSE of 0.02%; R^2 of 0.9998). Figure 5 identifies the TOC prediction errors of $\pm > 0.1\%$ for the 5-variable Model 4 (well A)

with one sample per 0.047ft. The best solution for that model involves just 16 prediction errors of $\pm > 0.1\%$ in the 2000 data records predicted (Fig. 5), with the maximum prediction error of 0.28% TOC associated with record number 888.

4.2. Models Considering Wells A and B Jointly to Predict TOC

Model 8 combines the data for well A (from Model 2) and well B (from Model 7) for the 10-variable inputs and normalizes the records as a single 2000-record dataset. That combined dataset is then evaluated with the TOB algorithm to predict TOC achieving an accuracy intermediate between that achieved by Models 2 and 7. The accuracy achieved by model 8 is RMSE of 0.03%; R^2 of 0.95584. Model 8 generates only 34 TOC prediction errors of $\pm > 1\%$ TOC and these are primarily located in the Well A interval 800 to 1000 (Fig. 6), as they are in Model 2. Only eleven of the records with errors of prediction errors of $\pm > 1\%$ TOC are in the well B section of the combined model.

Model 8 demonstrates that a single TOB stage 2 solution can accurately predict TOC in a dataset incorporating the 10-variable inputs of both wells A and B. This is encouraging as it suggests that other wells drilled through the Lower Barnett Shale could also have their TOC predicted with reasonable accuracy using that TOB Stage 2 solution.

It is relatively straight forward to use the data in model 8 to test records from Well A against the Well B dataset and vice versa. However, the TOB predictions generated in this way are poor; RMSE $\sim 1.2\%$; $R^2 \sim 0.35$ on average for both wells predicted in that way. This is not surprising as the two wells are drilled into quite distinct part of the Fort Worth Basin. The wells are 48 km apart with well A located in the deeper northeast part of the basin, and well B located towards the shallower western flank of the basin. As those wells have quite distinctive well log curves and lithology when compared in detail, it is not surprising that the log curves from one of those wells are not very accurate in their predictions of TOC in the other well. However, for wells drilled in similar basin settings to each of those wells, it should be expected that more accurate TOC predictions could be made in

this way. Multi-well studies are required to confirm the prediction accuracy that could be achieved by the TOB algorithm for nearby wells or wells located comparable lithological and basin settings.

If predictions of other wells are to be possible then it would be more relevant to attempt to do this with the basic well log suite with or without additional stratigraphic information. This is so because there are just a few wells with comprehensive core analysis to provide the mineralogical data required for the 10-variable model 8. Those few well tend to already have adequate TOC measurements from cores. On the other hand, there are many wells drilled into the Barnett shale for which only basic well log data is available and no mineralogical and TOC measurements from cores. It is these wells that need to have their TOC contents predicted from their basic log data curve, if possible.

Model 9 builds on the knowledge gained from the zoomed in 5-input variable Model 4 and develops a more modest zoomed-in sampling for both wells. It samples the complete section of each well with 5000 data records and combines them into a 10,000-record model. For well A that is a density of one data record every 0.094 ft. For well B that is a density of one data record every 0.060 ft. One additional variable is added to the basic well logs ((GR, Rs, Pb, Np, DT, StH), which is stratigraphic height (StH) measured as a fraction between 0 and 1 (0 being the base of the Barnett shale; 1 representing the top). StH can be calculated for all Barnett shale wells, so it is a useful general indicator of stratigraphic position within the section that is readily exploited by the TOB algorithm. Thus, Model 9 is a 6-input variable model zoomed in five times from the scale used for Model 8 and excluding all mineralogical variables. The TOB Stage 2 TOC prediction accuracy achieved by model 9 is RMSE of 0.04%; R^2 of 0.9991. The best TOB Stage 2 solution for Model 9 generates only 34 TOC prediction errors of $\pm > 0.316\%$ TOC (equivalent to a squared error of 0.1) with all but two of them within the thinly laminated or rapidly changing zones of well A (Fig. 7).

The measured versus predicted values for model 9 are displayed in the Fig. 8a, b. The TOB stage 1 solution (Fig. 8b) for model 9 does most of the work in generating accurate TOC predictions, itself

achieving accuracy of RMSE of 0.12%; R^2 of 0.9929. It is this model (i.e., the 10,000 records and the best TOB solution) that has the most potential to be used in the prediction of TOC in those surrounding wells for which only basic well log data is available and no core mineralogy or TOC measurements exist.

5. Discussion

The objective of this study is to demonstrate that the TOB optimized data matching algorithm is functionally capable of predicting TOC from well log data to high degrees of accuracy on a supervised learning basis. Clearly, the analysis presented show that it is effective at accurately predicting TOC from comprehensive well log and mineralogical data, based on the analysis of two individual wells (A and B) spaced 48 km apart penetrating the Lower Barnett Shale. The greater the sampling density of the input log curves, the greater the TOC prediction accuracy achieved within each individual well on a supervised basis.

However, for the method to be of practical use it needs to be able to predict TOC in wells on an unsupervised basis, i.e., provide credible estimates of TOC for those many wells in which there is only a basic suit of well log data available, but without measured TOC values available from core analysis. To do this, the data records of those “known” wells with detailed core measurements and comprehensive well log curves, e.g. wells A and B, together with their optimized TOB solution(s), can potentially be used as reference sets against which the “unknown” wells can be matched to provide TOC solutions.

TOB is a data matching technique, so the unknown wells need to be reasonable similar to the known wells in terms of lithology and stratigraphic sequence sampled for meaningful data matches to be established. For certain basic log responses to be comparable in known and unknown wells, the logged sections also need have experienced similar burial and compaction histories. This means that each known well, or reference set of wells, is only likely to provide TOC predictions of meaningful accuracy for unknown wells within a certain limited distance from

it, based on lateral geological environment variabilities.

As explained in Sect. 4, wells A and B are in quite different basin settings. The well sections logged by those wells differ in current depth by about 1500 feet and the LBS is 470 ft thick in well A compared to 300ft thick in well B. Consequently, matching those wells against each other does not result in TOC prediction of sufficient accuracy. However, combining wells A and B data in a single TOB reference set (e.g. Models 8 and 9) has greater potential to provide credible predictions for unknown wells located between them, or at least in the vicinity of each of those wells.

Model 9 has the greatest potential to be of practical use in predicting TOC for unknown wells penetrating the LBS in the north-eastern part of the Fort Worth Basin. Unknown wells without TOC measurements are unlikely to have the mineralogical data available that is required to match with the 10-variable solution provided in Model 8. On the other hand, the 6-variable dataset for both wells A and B (Model 9), involving just the basic well log curves and stratigraphic height metric, is a dataset that many unknown surrounding wells would possess the necessary basic log data required for record matching. This study has focused on TOC. However, the same TOB methodology could potentially be applied to predict other source rock attributes, such as the Rock-Eval S2 peak values, from well logs, with or without mineralogical data inputs. Future multi-well studies are required to demonstrate that this can be conducted effectively and provide credible TOC (and other source rock attribute) predictions for the unknown wells.

6. Conclusions

The optimized data-matching, transparent open box (TOB) algorithm, can successfully establish highly accurate total organic carbon (TOC) predictions from suites of well logs, with or without the support of detailed mineralogical data, on a supervised learning basis for one or more wells. Applied to two well sections (A and B), 48 km apart, penetrating the Lower Barnett Shale (LBS) in the north-eastern

part of the Fort Worth Basin, TOB achieves increasing TOC prediction accuracy (RMSE from 0.7% improving to < 0.01% TOC; R^2 from 0.793 improving to 0.999) as the sampling density of the log curves increases from one data record for each ~ 0.5 -ft interval to one record for each ~ 0.04 -ft interval. TOB can also be successfully tuned to accurately predict TOC for a combined dataset of both wells A and B together for a 10-input-variable model (RMSE = 0.3% TOC; $R^2 = 0.955$) sampled with data records every 0.4 ft, and a 6-input-variable model (RMSE = 0.04% TOC; $R^2 = 0.999$) sampled at with data records every 0.075 ft. The 6-input-variable, combine-well model involves the five basic well logs (Gr, Pb, Rs, Np, DT), with fractional stratigraphic height making up the sixth input variable from the base of the LBS. It is the 6-variable model that has potential to be used practically to predict TOC in surrounding wells, in comparable basin settings, for which only the basic suite of logs is available. This non-correlation, data-matching technique has the potential to offer a credible alternative to the various multi-variate regression models and correlation-based machine learning methods for predicting TOC from well logs in an accurate, consistent, reliable and transparent manner. Future multi-well studies are required to confirm that potential and to establish the TOC prediction accuracy that might be achieved when applied in unsupervised circumstances.

7. Availability of Data and Material

The data and information used come from published studies cited in the manuscript. Listings of the data are available on request.

Author Contributions As sole author of this work, I have conducted all the modelling, interpretation and writing of this manuscript.

Funding

No funding was involved in supporting this research.

Code Availability

The models presented were conducted in Excel with VBA macros executing the equations presented in the manuscript. The VBA code is not included in this manuscript.

Conflicts of interest The author has no conflicts of interest or competing interests.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

- Abouelresh, M., & Slatt, R. (2012). Lithofacies and sequence stratigraphy of the Barnett Shale in east-central Fort Worth Basin, Texas. *AAPG Bulletin*, 96(1), 1–22. <https://doi.org/10.1306/04261110116>.
- Alizadeh, B., Maroufi, K., & Heidarifard, M. H. (2018). Estimating source rock parameters using wireline data: An example from Dezful Embayment, South West of Iran. *Journal of Petroleum Science and Engineering*, 167, 857–868.
- Alizadeh, B., Najjari, S., & Kadkhodaie, A. (2011). Artificial neural network modeling and cluster analysis for organic facies and burial history estimation using well log data: A case study of the South Pars gas field, Iran. *Computers and Geosciences*, 45, 261–269.
- Al-Mudhafar, W. J. (2017). Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. *Journal of Petroleum Exploration and Production*, 7, 1023–1033.
- Alshakhs, M., & Rezaee, R. (2017). A new method to estimate total organic carbon (TOC) content, an example from Goldwyer Shale Formation, the Canning Basin. *The Open Petroleum Engineering Journal*, 10, 118–133. <https://doi.org/10.2174/1874834101710010118>.
- Arbogast, J.S., & Utley, L. (2003). Extracting hidden details from mixed-vintage log data using automatic, high-speed inversion modeling of resistivity and SP logs: Supplementing incomplete log suites and editing poor quality log data using neural networks. In *AAPG annual convention*, May (poster presentation).
- Beers, R. F. (1945). Radioactivity and organic content of some Paleozoic shales. *American Association of Petroleum Geologists Bulletin*, 26, 1–22.
- Bolandi, V., Kadkhodaie-Ilkhchi, A., Alizadeh, B., Tahmorasi, J., & Farzi, R. (2015). Source rock characterization of the Albian Kazhdumi formation by integrating well logs and geochemical data in the Azadegan oilfield, Abadan plain, SW Iran. *Journal of Petroleum Science and Engineering*, 133, 167–176.
- Boyer, C., Kieschnick, J., Rivera, R. S., Lewis, R. E., & Waters, G. (2006). Producing gas from its source. *Schlumberger Oilfield Review*, pp. 36–49.
- Cheng, D., Yuan, X., Tan, C., & Wang, M. (2016). Logging-lithology identification methods and their application: A case study on the Chang 7 Member in the central-western Ordos Basin, NW China. *China Petroleum Exploration*, 21(5), 1–10.
- El Sharawy, M. S., & Gaafar, G. G. (2012). Application of well log analysis for source rock evaluation in the Duwi Formation, Southern Gulf of Suez, Egypt. *Journal of Applied Geophysics*, 80, 129–143.
- Fertl, W. H., & Chilingar, G. V. (1988). Total organic carbon content determined from well logs. *SPE Formation Evaluation*, 3, 407–419. <https://doi.org/10.2118/15612-PA>.
- Herron, M. M., Grau, J., Herron, S. L., Kleinberg, R. L., Machlus, M., Reeder, S. L., Vissapragada, B., Burnham, A., Day, R., & Allix, P. (2011). Total organic carbon and formation evaluation with wireline logs in the Green River oil shale. In *SPE-147184-MS SPE annual technical conference and exhibition*, 30 October–2 November, Denver (p. 19). Society of Petroleum Engineers. <https://doi.org/10.2118/147184-MS>.
- Herron, S. L. (1988). Source rock evaluation using geochemical information from wireline logs and cores (abs). *American Association of Petroleum Geologists Bulletin*, 72, 1007.
- Hester, T., & Schmoker, J. (1987). Determination of organic content from formation-density logs, Devonian-Mississippian Woodford Shale, Anadarko Basin, Oklahoma, United States. Geological Survey USGS Open file report #87-20.
- Huang, R., Wang, Y., Cheng, S., Liu, S., & Cheng, L. (2015). Selection of logging-based TOC calculation methods for shale reservoirs: A case study of the Jiaoshiba shale gas field in the Sichuan Basin. *Natural Gas Industry B*, 2, 155–161.
- Huang, Z., & Williamson, M. A. (1996). Artificial neural network modeling as an aid to source rock characterization. *Marine and Petroleum Geology*, 13, 227–290.
- Jarvie, D. M., Hill, R. J., Ruble, T. E., & Pollastro, R. M. (2007). Unconventional shale-gas systems: The Mississippian Barnett Shale of North-Central Texas as one model for thermogenic shale-gas assessment. *AAPG Bulletin*, 91, 475–499. <https://doi.org/10.1306/12190606068>.
- Kadkhodaie, A., Rahimpour, H., & Rezaee, M. R. (2009). A committee machine with intelligent systems for estimation of total organic carbon content from petrophysical data. *Computers and Geosciences*, 35, 457–474.
- Kadkhodaie, A., & Rezaee, M. R. (2017). Estimation of vitrinite reflectance from well log data. *Journal of Petroleum Science and Engineering*, 148, 94–102.
- Kamali, M. R., & Mirshady, A. A. (2004). Total organic carbon content determined from well logs using $\Delta\log R$ and neuro fuzzy techniques. *Journal of Petroleum Science and Engineering*, 45, 141–148.
- Khoshnoodkia, M., Mohseni, H., Rahmani, O., & Mohammadi, A. (2011). TOC determination of Gadvan Formation in South Pars Gas field, using artificial intelligent systems and geochemical data. *Journal of Petroleum Science and Engineering*, 78, 119–130.
- Liu, Y., Chen, Z., Hu, K., & Liu, C. (2013). Quantifying total organic carbon (TOC) from well logs using support vector regression. In *Integration GeoConvention*, Calgary (p. 6).
- Luffel, D. L., Guidry, F. K., & Curtis, J. B. (1992). Evaluation of Devonian shale with new core and log analysis methods. *SPE 21297-PA 1192–1197*. <https://doi.org/10.2118/21297-PA>.

- Mallick, R. K., & Raju, S. V. (1995). Thermal maturity evaluation by sonic log and seismic velocity analysis in parts of Upper Assam Basin, India. *Organic Geochemistry*, 23(10), 871–879.
- Mann, U. (1986). Relation between source rock properties and wireline log parameters, an example from Lower Jurassic Posidonia Shale, NW Germany. *Organic Geochemistry*, 10(4–6), 1105–1112.
- Mendelson, J. D., & Toksoz, M. N. (1985). Source rock characterization using multivariate analysis of log data. In *Transactions of 26th Annual Logging Symposium*, 17–20 June, Dallas, Texas (p. 26). Paper SPWLA-1985-UU.
- Meyer, B. L., & Nederlof, M. H. (1984). Identification of source rocks on wireline logs by density/resistivity and sonic transit time/resistivity cross plots. *American Association of Petroleum Geologists Bulletin*, 68, 121–129.
- Passey, Q. R., Bohacs, K., Esch, W. L., Klimentidis, R., & Sinha, S. (2010). From oil-prone source rock to gas-producing shale reservoir-geologic and petrophysical characterization of unconventional shale gas reservoirs. In *International oil and gas conference and exhibition in China* (p. 29). Society of Petroleum Engineers. <https://doi.org/10.2118/131350-MS>.
- Passey, Q. R., Moretti, F. U., & Stroud, J. D. (1990). A practical modal for organic richness from porosity and resistivity logs. *American Association of Petroleum Geologists Bulletin*, 74, 1777–1794.
- Pollastro, R. M., Jarvie, D. M., Hill, R. J., & Adams, C. W. (2007). Geologic framework of the Mississippian Barnett Shale, Barnett-Paleozoic total petroleum system, Bend arch–Fort Worth Basin, Texas. *AAPG Bulletin*, 91(4), 405–436. <https://doi.org/10.1306/10300606008>.
- Schmoker, J. (1979). Determination of organic content of Appalachian Devonian Shales from formation-density logs: Geologic notes. *AAPG Bulletin*, 63, 1504–1509.
- Schmoker, J. W. (1981). Determination of organic-matter content of Appalachian Devonian shales from gamma-ray logs. *American Association of Petroleum Geologists Bulletin*, 65, 2165–2174.
- Schmoker, J. (1993). Use of formation-density logs to determine organic-carbon content in Devonian Shales of the Western Appalachian Basin and an additional example based on the Bakken Formation of the Williston Basin. In J. B. Roen & R. C. Kepferle (Eds.), *Petroleum geology of the Devonian and Mississippian black shale of eastern North America* (pp. J1–J14). U. S. Geological Survey Bulletin 1909, U.S. Government Printing Office.
- Schmoker, J. W., & Hester, T. C. (1983). Organic carbon in Bakken formation, United States portion of Williston basin. *American Association of Petroleum Geologists Bulletin*, 67, 2165–2174.
- Sfidari, E., Kadkhodaie, A., & Najjari, S. (2012). Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems. *Journal of Petroleum Science and Engineering*, 86–87, 190–205.
- Shi, X., Wang, J., Liu, G., Yang, L., Ge, X., & Jiang, S. (2016). Application of extreme learning machine and neural networks in total organic carbon content prediction in organic shale with wire line logs. *Journal of Natural Gas Science and Engineering*, 33, 687–702.
- Singh, P., Slatt, R. M., & Coffey, W. (2008). Barnett Shale - Unfolded: Sedimentology, sequence stratigraphy, and regional mapping. *Gulf Coast Association of Geological Societies Transactions*, 58, 777–795.
- Frontline Solvers. (2020). Standard excel solver—Limitations of nonlinear optimization. <https://www.solver.com/standard-excel-solver-limitations-nonlinear-optimization>. Accessed 16 March 2020.
- Swanson, V. E. (1960). Oil yield and uranium content of black shales. USGS Professional Paper 356-A, pp. 1–44.
- Tabatabaei, S. M. H., Kadkhodaie, A., Hosseinib, Z., & Asghari Moghaddam, A. (2015). A hybrid stochastic-gradient optimization to estimating total organic carbon from petrophysical data: A case study from the Ahwaz oilfield, SW Iran. *Journal of Petroleum Science and Engineering*, 127, 35–43.
- Tan, M., Song, X., Yang, X., & Wu, Q. (2015). Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: A comparative study. *Journal of Natural Gas Science and Engineering*, 26, 792–802.
- Verma, S., Zhao, T., Marfurt, K. J., & Devegowda, D. (2016). Estimation of total organic carbon and brittleness volume. *Interpretation*, 4(3), 373–385. <https://doi.org/10.1190/INT-2015-0166.1>.
- Walper, J. L. (1982). Plate tectonic evolution of the Fort Worth Basin. In C. A. Martin (Ed.), *Petroleum geology of the Fort Worth Basin and bend arch area* (pp. 237–251). Dallas Geological Society.
- Wang, F. P., & Gale, J. F. W. (2009). Screening criteria for shale-gas systems. *Gulf Coast Association of Geological Societies Transactions*, 59, 779–793.
- Wang, Y. (2013). The method of application of gamma-ray spectral logging data for determining clay mineral content. *Journal of Oil and Gas Technology*, 35(2), 100–104.
- Wood, D. A. (2016). Metaheuristic profiling to assess performance of hybrid evolutionary optimization algorithms applied to complex wellbore trajectories. *Journal of Natural Gas Science and Engineering*, 33, 751–768. <https://doi.org/10.1016/j.jngse.2016.05.041>.
- Wood, D. A. (2018). A transparent Open-Box learning network provides insight to complex systems and a performance benchmark for more-opaque machine learning algorithms. *Advances in Geo-Energy Research*, 2(2), 148–162. <https://doi.org/10.26804/ager.2018.02.04>.
- Wood, D. A. (2019a). Transparent open box learning network provides auditable predictions for coal gross calorific value. *Modeling Earth Systems and Environment*, 5, 395–419. <https://doi.org/10.1007/s40808-018-0543-9>.
- Wood, D. A. (2019b). Sensitivity analysis and optimization capabilities of the transparent open box learning network in predicting coal gross calorific value from underlying compositional variables. *Modeling Earth Systems and Environment*, 5, 753–766. <https://doi.org/10.1007/s40808-019-00583-1>.
- Zhao, P., Mao, Z., Huang, Z., & Zhang, C. (2016). A new method for estimating total organic carbon content from well logs. *AAPG Bulletin*, 100(8), 1311–1327.