Modeling Earth Systems and Environment (2019) 5:395–419
https://doi.org/10.1007/s40808-018-0543-9

ORIGINAL ARTICLE

Transparent open-box learning network provides auditable predictions for coal gross calorific value

David A. Wood[1]

Received: 24 September 2018 / Accepted: 3 November 2018 / Published online: 16 November 2018

Abstract

Auditing and forensic analysis of how each prediction is calculated are key attributes of transparent open-box learning networks (TOB). It provides the full calculation and input metric contributions for each of the predictions it derives. There are two stages in executing TOB predictions (stage 1 matches and ranks using squared-error analysis; stage 2 optimizes and conducts sensitivity analysis). Neither stage involves generating or extrapolating correlations between the input variables. Both stages of the calculation generate accurate predictions for datasets with multiple, highly-dispersed and non-linear influencing inputs. The transparent way in which generates predictions leads to better understanding of the interplays between the input variables. Such attributes have direct relevance to the complex systems modelled in the coal industry [e.g., gas calorific value (GCV) prediction and coal petrology–grindability relationships]. The algorithm is applied here to predict GCA for a large published database (6339 records) of US coals including proximate and ultimate analysis metrics. The TOB predicts GCV with accuracy (RMSE $\leq$ 0.3 MJ/kg; $R^2 > 0.99$). The transparency of the TOB method contrasts with the hidden relationships involved in many neural-network based prediction systems. Worked examples are provided to show the detailed prediction calculations associated with individual data points. The TOB approach applied to predicting coal GCV can help to verify the source of specific samples (e.g. specific mines or coal basins) using readily understandable underlying calculations available for audit and display. The TOB is therefore also suitable for identifying the provenance of specific coal samples based on proximate and/or ultimate analysis.

Keywords Transparent GCA prediction · Learning network for coal GCA · Predicting non-linear systems without deriving correlations · Auditing prediction calculations · Forensic interrogation of coal learning network

Introduction

Predictions of dependent-variable values from complex systems of multiple non-linear influencing variables with highly dispersed distributions are key requirements for the coal industry. Predicting gross calorific value of various grades of coal (Mesroghli et al. 2009) and relationships between its petrological factors and grinding properties (Bagherieh et al. 2008) are common examples. Empirical correlations and artificial intelligence algorithms that are widely used to derive predictions for dataset of variable size covering regional and/or local coal sources. Many of these methods provide meaningfully accurate predictions where the underlying inputs are related in highly non-linear and irregular ways. Often, the metric of interest for commercial valuation is expensive to measure repeatedly by laboratory testing, making such prediction tools cost effective. This is the case with proximate and ultimate coal analysis. Artificial intelligence tools are therefore growing in their deployment for such applications (Schmidhuber 2015).

A problem with many machine-learning algorithms is that they lack transparency. They do not reveal easily how each prediction they generate is derived. This is because they typically involve hidden, complex and multi-dimensional correlations. This means that they typically do not provide straightforward and auditable input–output relationships between the variables involved their predictions. For

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40808-018-0543-9) contains supplementary material, which is available to authorized users.

✉ David A. Wood
dw@dwasolutions.com

[1] DWA Energy Limited, 25 Badgers Oak Bassingham, Lincoln LN5 9JP, UK

this reason, some are reluctant to rely on machine-learning algorithms, where such information is of critical importance (e.g., commercial valuation or error analysis for specific value intervals; both of which are relevant to GCV analysis of coal samples).

This opaqueness leads to many practitioners being sceptical about the predictions derived from neural-network methods, particularly their claims to accuracy when applied to relatively limited data sets. They are often viewed as black boxes for this reason (Heinert 2008) and their inability to reveal the details of their underlying calculations can be frustrating. This is despite their ability, based on a range of statistical-accuracy measures, to achieve impressive levels of prediction accuracy for a wide range of complex systems. Indeed, some algorithms are prone to the pitfalls of overfitting (Lever et al. 2016) i.e., their hidden correlations are too dependent on a particular set of data, introducing doubt regarding their ability to fit additional data records as they become available. This is particularly a problem for datasets covering a range of the prediction-metric intermittently, i.e., significant gaps in the value range covered by the underlying dataset.

Locally-weighted learning methods (Atkeson et al. 1997) combined with lazy learning principles (Birattari et al. 1999), originating from the much earlier recognition of the benefit of nearest-neighbour prediction methods (Fix and Hodges 1951; Cover and Hart 1967), can be configured to provide transparency. However, these approaches tend to be applied more to pattern recognition algorithms (Garcia et al. 2012; Chen and Shah 2018) rather than non-linear regression predictions, where the application of the more-opaque neural networks now dominate. Moreover, such approaches often seek to linearize highly non-linear systems on a localized or neighbourhood basis (Bontempi et al. 1999). Nevertheless, there is the potential for such approaches to be configured with transparency in mind (Shakhnarovich et al. 2006).

The transparent open-box (TOB) learning-network algorithm (Wood 2018) overcomes many of the issues mentioned by not relying upon hidden correlations to calculate its predictions. It applies a matching technique between a tuning and training data subsets. The degree of match between the data records is quantified using squared-error analysis. The TOB stage 1 prediction establishes a set of high-ranking

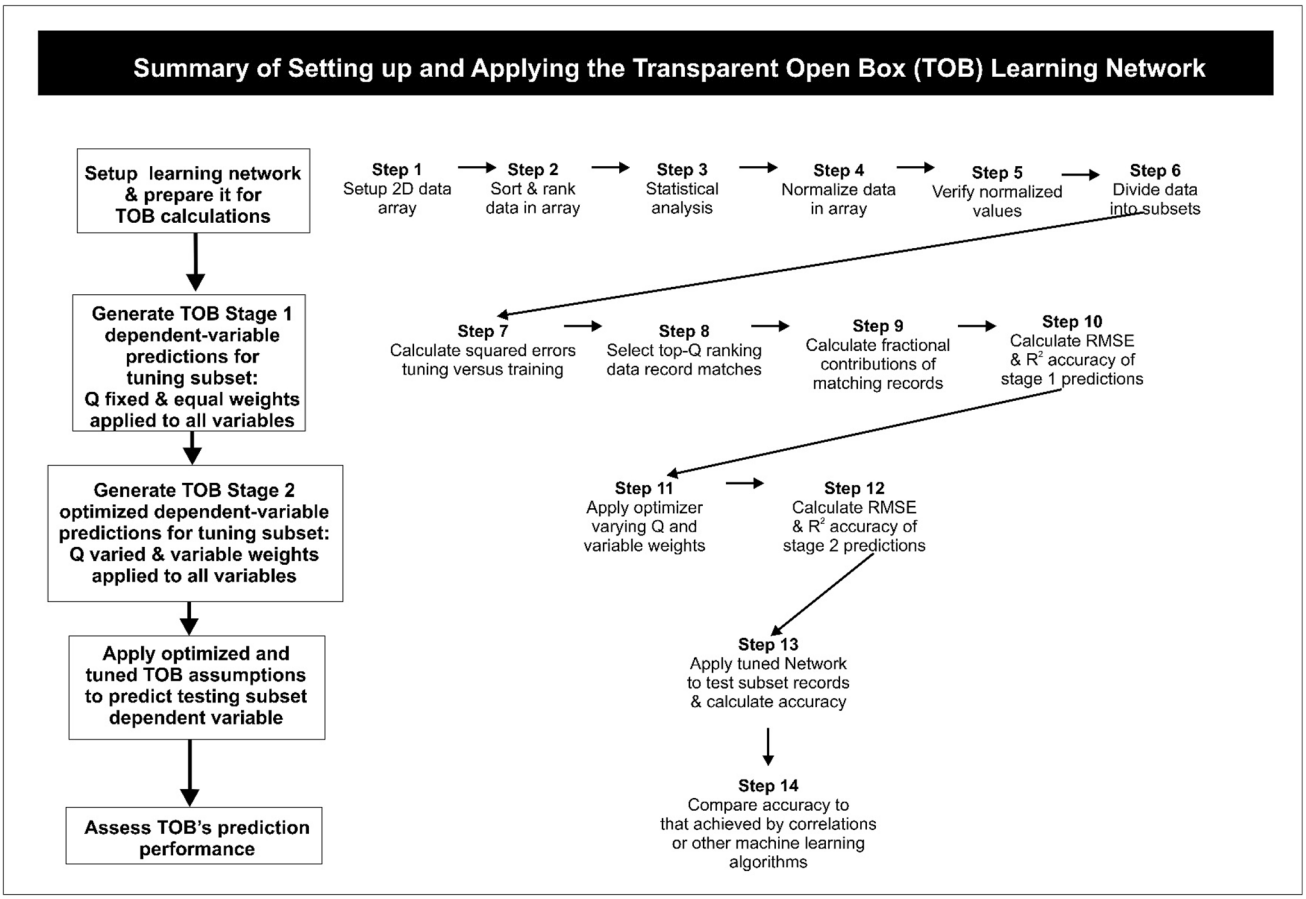

Fig. 1 Diagrammatic representation of the steps and stages in applying the transparent open-box (TOB) learning network algorithm (Wood 2018). See Appendix 1 (TOB method) for a detailed description of each stage and step including the mathematical formulations

matching records from which an initial prediction is made. The TOB stage 2 then applies an optimizer to identify the optimum weights to assign to each input to improve its prediction accuracy. The transparent calculations used in TOB stages 1 and 2 are readily available to audit, and the level of accuracy it achieves compares favorably with the more-opaque artificial-intelligence (AI) techniques (e.g., adaptive neuro-fuzzy inference systems, multi-layer perceptron and radial-basis function artificial neural networks, least squares support vector machines, and hybrids of those with evolutionary optimization algorithms).

The AI methods mentioned do not need to be totally opaque. Simulation methodologies can provide a degree of transparency (Elkatatny et al. 2016). Variable importance algorithms can also establish the covariances between the influencing variables of AI methods. Auret and Aldrich (2012) achieve this with a random-forest algorithm. The TOB method goes further than this because it facilitates drilling down into the underlying variables to obtain the exact calculations involved in each of its predictions.

Here, the TOB method is applied to a 6339-record data set for US coals including both proximal and ultimate influencing variables to predict GCV (Appendix 1, supplementary file). Highly-accurate predictions are achieved using a small tuning data subset (~ 1.5% of the full dataset). Matches are achieved through error analysis in a training subset that constitutes about 97% of the entire database. The remaining 1.5% of the data records are not involved in the training or tuning process. They are used as a testing-data subset to independently test the TOB's prediction performance. The dataset involves nine influencing variables that contribute through various applied weightings to predict GCV as the dependent variable.

Although a coal GCV dataset is used to demonstrate the benefits of the TOB learning network to the coal industry, there are other coal-related systems for which ANN is frequently used as prediction tool that could equally benefit from its application. For example, coal petrography and petrology influencing variables in relation to a measure of coal grindability as a dependent variable (Trimble and Hower 2003; Bagherieh et al. 2008).

TOB method

TOB stages 1 and 2 comprise of 14 steps (Wood 2018). These steps are summarized in a flow diagram (Fig. 1). Stage 1 builds upon lazy learning (Birattari et al. 1999) and nearest neighbour (Chen and Shah 2018) principles but with very specific error drivers. Stage 2 goes far beyond such principles by linking the selection of variable weightings to an optimizer, providing a more flexible and versatile weighting regime than typically associated with k-nearest neighbour classifiers (Samworth 2012).

The details and mathematical formulations involved in each of the 14 steps required to establish and implement a TOB learning network are described in Appendix 1.TOB Stage-1 predictions (steps 1–10) are often found to be quite accurate (e.g., comparable to those provided by typical k-learning algorithms). However, they typically can be much improved upon by applying TOB stage 2.

The TOB learning network can be successfully applied using spreadsheets (e.g. Excel workbooks) for mid-sized data sets. Fully-coded algorithm formats or hybrid VBA plus spreadsheet setups can speed up deployment for such datasets. For large datasets it is appropriate to deploy the TOB

Table 1 Statistical summary of dataset compiled for gross calorific value (GCV) of 6339 US coals with each record linking measured proximate and ultimate analysis variables to their measured GCV value (MJ/kg)

COALQUAL dataset used for gross calorific value (GCV) interval 6–35 MJ/kg prediction										
Compiled dataset: 6339 data records	Moisture (%)	Volatiles (%)	Fixed carbon (%)	Ash (%)	H (%)	C (%)	N (%)	O (%)	S (%)	GCV (MJ/kg)
Variable descriptor	#1	#2	#3	#4	#5	#6	#7	#8	#9	Dependent
Min	0.40	3.80	4.10	0.90	1.70	18.80	0.20	0.20	0.07	6.54
Max	57.20	55.70	87.00	24.90	9.50	89.60	5.60	59.90	12.80	35.46
Mean	8.79	32.40	48.38	10.43	5.34	65.58	1.29	15.50	1.86	27.07
Standard deviation	10.63	6.15	11.10	5.20	0.68	12.30	0.34	11.96	1.66	5.39
20-percentile	2.00	27.90	40.20	5.80	4.87	56.47	1.00	7.00	0.60	22.97
40-percentile	2.92	31.90	46.37	8.27	5.17	65.71	1.25	9.00	0.90	27.45
60-percentile	5.10	34.68	51.02	11.01	5.40	70.82	1.40	12.43	1.66	29.54
80-percentile	14.20	37.32	56.90	14.90	5.70	75.77	1.59	24.02	3.11	31.46

The dataset is that compiled by Matin and Chelgani (2016) with records filtered from the US Geological Survey Coal Quality (COALQUAL) database (Version 2.0), open file report 97–134 (Bragg et al. 1997). Variables #1 to #9, inclusive, represent the input variables to the TOB model

Includes all data records in the Training, Tuning and Testing subsets

algorithm in a fully-coded configuration, e.g., in Octave, Python, MatLab, R, VBA, etc.

A hybrid VBA-Excel spreadsheet configuration is used here to predict the gas calorific value (GCV) from a published dataset of coals from the United States (6339 data records).

Dataset compiled to predict coal gross calorific value (GCV)

There are numerous well-established published correlations based on linear and multi-variable regressions, particularly for coals from the United States (US), based on proximate and/or ultimate analysis (Given et al. 1986; Neavel et al. 1986; Singh and Kakati 1994; Channiwala and Parikh 2002; Majumder et al. 2008; Mathews et al. 2014). Several of these provide predictions with low absolute errors and correlation coefficients (R^2) > 0.9 between measured and predicted GCV. However, as many of the variables involved in proximate and ultimate analysis vary in a non-linear manner, prediction improvements have been achieved by applying non-linear regression or machine-learning algorithms such as artificial neural networks (ANN), support vector regression (SVR) or adaptive network based fuzzy inference system (ANFIS) (Patel et al. 2007; Mesroghli et al. 2009; Chelgani et al. 2010, 2011; Yalcin Erik and Yilmaz 2011; Kavsek et al. 2013; Tan et al. 2015; Feng et al. 2015). These are mainly based on proximate analysis of relatively small data sets of coals from India and China but achieve high levels of prediction accuracy with $R^2 > 0.99$ between measured and predicted GCV in some cases.

Tan et al. (2015) also demonstrate a GCV prediction performance with $R^2 > 0.99$ for their SVR algorithm using the many thousands of samples of US coal analysis provided by the US Geological Survey Coal Quality (COALQUAL) database (Bragg et al. 1997). Matin and Chelgani (2016) demonstrated that the random forest algorithm (Breiman 2001; Auret and Aldrich 2012) establishing covariances between the influencing variables, could produce highly accurate predictions of GCV ($R^2 > 0.97$ for proximate data; $R^2 > 0.99$ for ultimate data) using a filtered version of the COALQUAL (version 2) database. Matin and Chelgani (2016) filtered out coal records with > 25% ash as being unsuitable for use in power production and those with analysis that did not sum to 100.

This left 6339 data records of the COALQUAL database which they used to test their random forest algorithm. It is

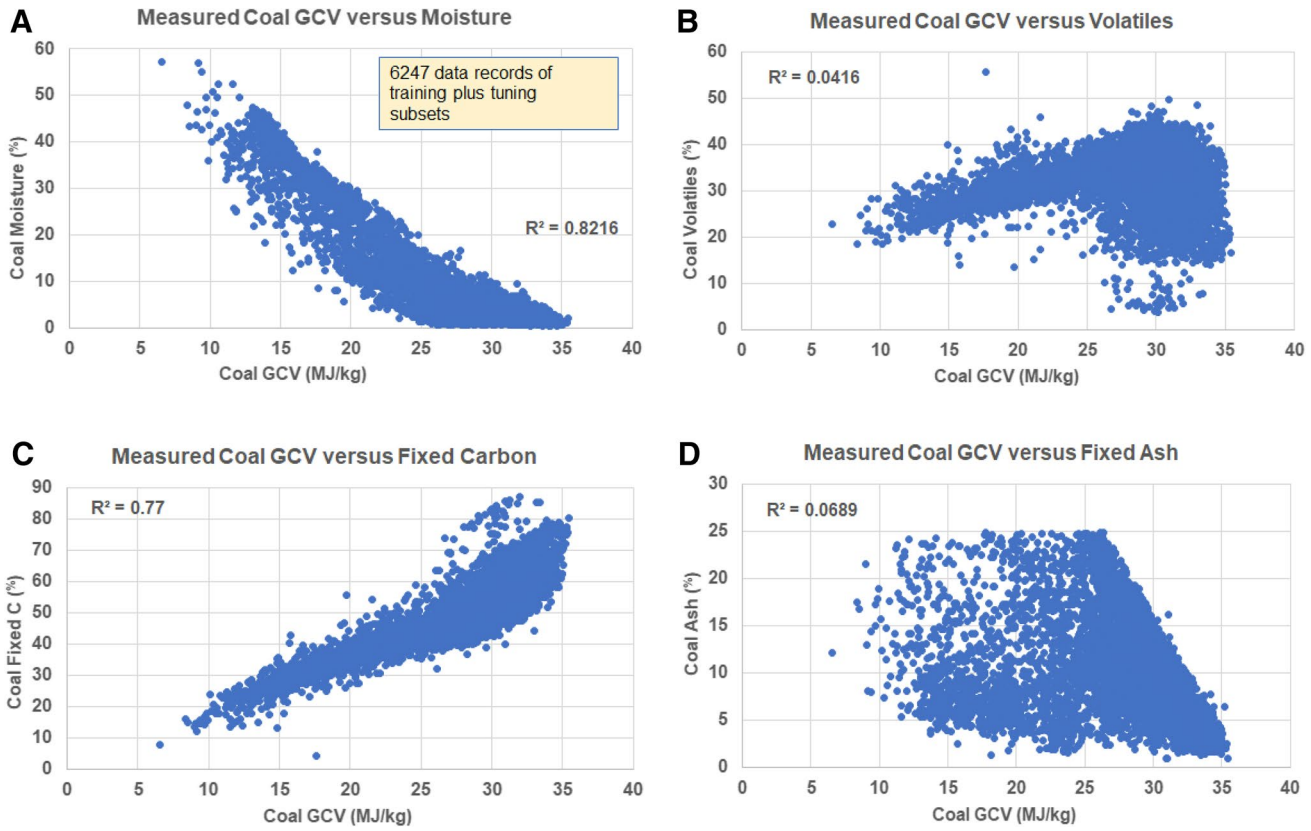

Fig. 2 **a–d** Proximate analysis variable relationships with gross calorific value for the 6339 data records for US coals compiled from the US Geological Survey Coal Quality (COALQUAL) database version 2.0, open file report 97–134 (Bragg et al. 1997)

these 6339 data records that are used here to test the TOB algorithm. It should be noted that the COALQUAL dataset has now been updated and extended (> 13,000 coal records) by the issue of version 3 (Palmer et al. 2015), but for the current purpose it is deemed more useful to use the filtered version 2 dataset for which published GCV-prediction performances are available for comparison. Matin and Chelgani omitted fixed carbon (FC) and oxygen (O) values from the GCV-influencing variables they included in their prediction model, because they were derived from other variables in the proximate or ultimate analysis. These variables are included in the TOB analysis as they are valuable for data-record matching purposes.

The compiled-US-coal-GCV dataset used here to demonstrate the prediction capability of the TOB method spans a significant range of GCV. It also incorporates significant ranges of input-variable distributions (proximate and ultimate analysis) as shown in (Table 1). The details for each of the 6339 data records are provided in a supplementary file (see "Appendix"). This includes the values of all variables and links to the sample numbers listed in the extracts from the COALQUAL version 2 data base

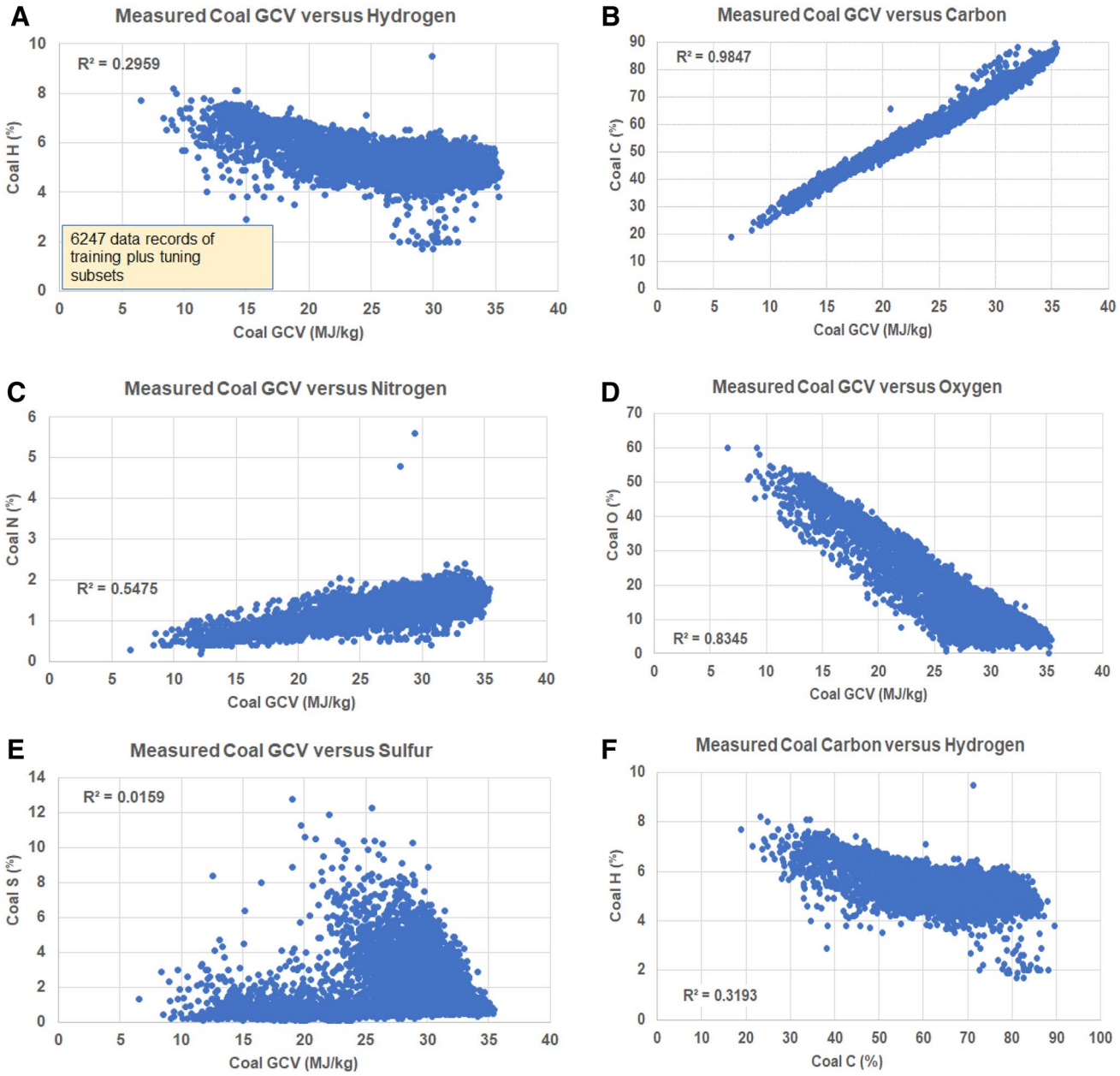

Fig. 3 **a**–**f** Ultimate Analysis variable relationships with gross calorific value for the 6339 data records for US coals compiled from the US Geological Survey Coal Quality (COALQUAL) database version 2.0, open file report 97–134 (Bragg et al. 1997)

Table 2 Prediction accuracy versus TOB optimization control metrics (Q and Wn) for the complete GCV data set

Transparent open box (TOB) learning network results and variable weightings in the prediction of gas calorific value (GCV) of coal from proximate and ultimate data variables (dataset with 6339 records)													
Variable description	Variable number	Pre-optimization equal weightings	Best solution solver GRG multistart	Best solution solver evolutionary algorithm	Sensitivity analysis with Q constrained to integers progressively from 10 to 2 (All cases runs with for the 92 records of the tuning subset with the Solver GRG optimizer configured in the same way)								
Q constrained to	Integer constraints		2 to 10	2 to 10	10	9	8	7	6	5	4	3	2
Q selected for solution	Integer #	10	9	8	10	9	8	7	6	5	4	3	2
Prediction performance of optimum and constrained optimum solutions applied to the tuning subset (92 records: ~ 1.5% of total dataset)													
RMSE	MJ/kg	0.48381	0.33462	0.34544	0.33932	0.33492	0.34181	0.34312	0.35975	0.35137	0.35786	0.39624	0.41393
R^2	fraction	0.9923	0.9963	0.9961	0.9962	0.9963	0.9961	0.9961	0.9958	0.9959	0.9958	0.9949	0.9944
Weightings (0 < = w < =1) Applied to constrained optimum solutions for the tuning subset													
Moisture (%)	#1	0.5	5.304E−04	1.937E−02	5.472E−05	0.000E+00	6.758E−03	1.560E−02	6.645E−02	2.564E−02	1.600E−02	2.208E−02	9.190E−02
Volatiles (%)	#2	0.5	8.317E−03	1.024E−02	8.106E−03	5.577E−03	6.760E−03	7.888E−03	0.000E+00	0.000E+00	0.000E+00	1.789E−02	1.993E−02
Fixed carbon (%)	#3	0.5	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.067E−01
Ash (%)	#4	0.5	1.576E−03	0.000E+00	6.160E−07	1.772E−03	3.061E−03	0.000E+00	0.000E+00	0.000E+00	0.000E+00	4.516E−03	0.000E+00
H (%)	#5	0.5	0.000E+00	9.669E−01	0.000E+00	2.676E−03	0.000E+00	0.000E+00	1.006E−01	0.000E+00	0.000E+00	0.000E+00	0.000E+00
C (%)	#6	0.5	1.000E+00	1.042E−02	1.000E+00	1.000E+00	1.000E+00	1.000E+00	7.089E−01	1.000E+00	4.267E−01	1.000E+00	1.000E+00
N (%)	#7	0.5	2.062E−03	1.430E−02	2.709E−03	0.000E+00	1.383E−03	2.555E−04	0.000E+00	1.082E−03	0.000E+00	0.000E+00	0.000E+00
O (%)	#8	0.5	0.000E+00	6.583E−02	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	2.057E−02	1.251E−02	0.000E+00	3.003E−03
S (%)	#9	0.5	4.348E−03	7.733E+00	9.531E−03	0.000E+00	5.076E−03	8.859E−02	0.000E+00	6.196E−03	4.363E−03	0.000E+00	0.000E+00
Ratio of #4 weight to #1 weight		1	2.97	0.00	0.01	N/A	0.45	0.00	0.00	0.00	0.00	0.20	0.00
Prediction performance of optimum solution variable weightings and Q value applied to the testing subset (91 records: ~ 1.5% of total dataset)													
RMSE	MJ/kg	0.41134	0.30556	0.30552	0.29482	0.30923	0.31114	0.32784	0.33437	0.34563	0.37973	0.41478	0.45600
R^2	fraction	0.9949	0.9970	0.997	0.9972	0.9969	0.9969	0.9966	0.9964	0.9961	0.9954	0.9945	0.9934

(Bragg et al. 1997) as complied by Matin and Chelgani (2016). From the supplementary file it is possible to locate the exact COALQUAL sample number and US State of origin of each.

It is clear from Table 1 that the data set is skewed towards the higher end of the GCV range; nearly 80% of the samples fall in the GCV range 23–35.5 MJ/kg. This means that the lower end of the GCV scale (i.e., < 15 MJ/kg) is sparsely sampled; representing about 4.5% of the data records. This feature of the dataset is addressed in the learning networks constructed.

The highly dispersed and non-linear relationships between each of the input variables (#1 to #9) and the dependent variable GCV are illustrated in Figs. 2 and 3. For the proximate analysis variables (Fig. 2), moisture content and fixed carbon show the best correlations GCV; the former a negative correlation ($R^2 = 0.8216$), and the latter a positive correlation ($R^2 = 0.77$). On the other hand, Ash and Volatiles show poor correlations with GCV and greater dispersal (Fig. 2). The dispersal in the Ash versus GCV relationship gradually reduces in this dataset for GCV values > 25 MJ/kg. For the ultimate analysis variables (Fig. 3), carbon and oxygen show the best correlations GCV; the former a positive correlation ($R^2 = 0.9847$), and the latter a negative correlation ($R^2 = 0.8345$). The other variables in Fig. 3 show significant dispersion and non-linearity, particularly S.

TOB predictions of coal gross calorific value (GCV) from the compiled 6339-record GCV dataset

The GCV dataset is divided into subsets (*training* = 6155 records; *tuning* = 92 records; and, *testing* = 91 records). Some 97% of the 6339 records (full data set evaluated) reside in the training subset. Allocation of records for tuning and testing is spread arbitrarily across the entire range of data values. It is best to avoid random allocations as this is likely to lead to clustering and gaps in specific data ranges.

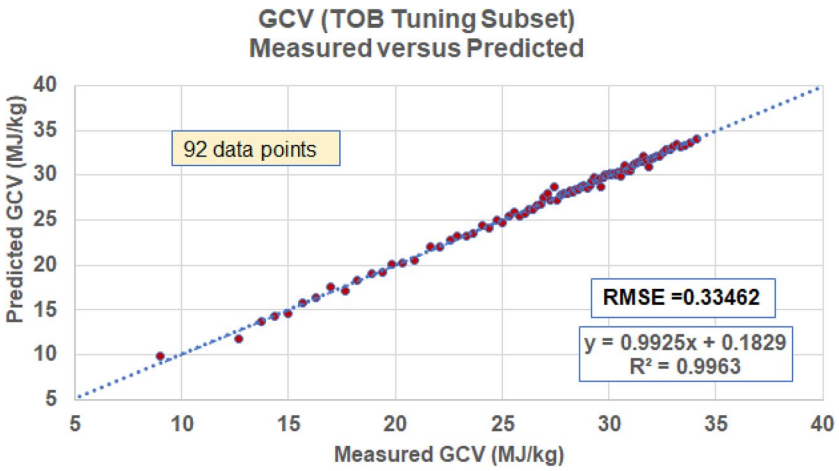

Fig. 4 Predicted versus measured GCV (MJ/kg) for 92 TOB-tuning subset records with 6155 records in the training subset

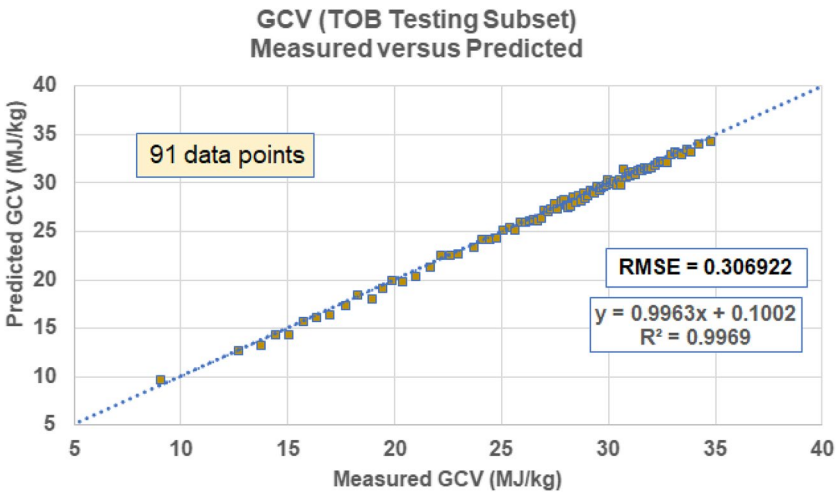

Fig. 5 Predicted versus measured GCV (MJ/kg) for 91 TOB-tuning subset records with 6155 records in the training subset. The data records of the testing subset were excluded from the tuning and training subsets

Table 3 Example audit of the calculation details for the TOB stage 1 GCV prediction associated with specific data records.

Example of how the GCV prediction calculation of each data record in the TOB network can be audited in detail after stage 1														
Stage 1	Evenly weighted sum of squares errors (wSSE)													
Rank of matches (in training subset)	Top-ranking matched records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent variable GCV (MJ/kg)			
Test record	2738	-0.9331	0.3950	0.0109	-0.1000	-0.1026	0.4124	-0.6667	-0.7621	-0.0683	-0.1026	The same top ten matches established by TOB stage 1 are used in TOB stage 2 with different weights and Q value applied		
Ranking	Records	#1	#2	#3	#4	#5	#6	#7	#8	#9	GCV			
1	2065	-0.9701	0.3410	0.0656	-0.0867	-0.1308	0.4582	-0.6481	-0.8137	-0.1123	0.6264			
2	2696	-0.9405	0.4200	-0.0207	-0.0275	-0.1410	0.3816	-0.5889	-0.7611	-0.0636	0.5546			
3	2703	-0.9507	0.3064	0.0615	-0.0417	-0.1282	0.4068	-0.6667	-0.7655	-0.1312	0.5663			
4	2705	-0.9282	0.3480	0.0203	-0.0433	-0.1487	0.3712	-0.6407	-0.7166	-0.1658	0.5370			
5	2770	-0.9437	0.4836	-0.0253	-0.1417	-0.1538	0.3842	-0.6667	-0.7286	0.0102	0.5732			
6	4228	-0.9521	0.3568	0.0598	-0.1417	-0.1231	0.4446	-0.5519	-0.7876	-0.1013	0.6288			
7	2730	-0.9401	0.5183	-0.0639	-0.0917	-0.1026	0.3955	-0.6667	-0.7588	-0.0212	0.5703			
8	2712	-0.9412	0.4170	-0.0075	-0.0650	-0.1282	0.4136	-0.5778	-0.7571	-0.2019	0.5925			
9	2133	-0.9327	0.3264	0.0745	-0.1733	-0.1385	0.4997	-0.6556	-0.8214	-0.1406	0.6517			
10	2792	-0.9577	0.3911	0.0519	-0.1750	-0.1282	0.4605	-0.7407	-0.7554	-0.1940	0.6269			
		Normalized values (Scale: -1 to + 1)												
		0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	input variables equally weighted			
Ranking	Records	wSSE#1	wSSE#2	wSSE#3	wSSE#4	wSSE#5	wSSE#6	wSSE#7	wSSE#8	wSSE#9	Sum of weighted squared errors (SumE)	Y = X/SumE	F = Y/Z	Contributions to prediction F*GCV (of record)
1	2065	6.835E-04	1.455E-03	1.500E-03	8.889E-05	3.978E-04	1.047E-03	1.715E-04	1.331E-03	9.676E-04	7.642E-03	15.1315	0.1420	0.0890
2	2696	2.734E-05	3.137E-04	4.994E-04	2.628E-03	7.396E-04	4.740E-04	3.025E-03	5.050E-07	1.111E-05	7.719E-03	14.9814	0.1406	0.0780
3	2703	1.550E-04	3.928E-03	1.283E-03	1.701E-03	3.287E-04	1.596E-05	0.000E + 00	5.612E-06	1.975E-03	9.393E-03	12.3113	0.1156	0.0654
4	2705	1.215E-05	1.105E-03	4.426E-05	1.606E-03	1.065E-03	8.505E-04	3.361E-04	1.038E-03	4.744E-03	1.080E-02	10.7061	0.1005	0.0540
5	2770	5.579E-05	3.928E-03	6.548E-04	8.681E-04	1.315E-03	3.990E-04	0.000E + 00	5.612E-04	3.085E-03	1.087E-02	10.6410	0.0999	0.0573
6	4228	1.808E-04	7.277E-04	1.199E-03	8.681E-04	2.104E-04	5.185E-04	6.591E-03	3.241E-04	5.443E-04	1.116E-02	10.3575	0.0972	0.0611
7	2730	2.480E-05	7.603E-03	2.797E-03	3.472E-05	0.000E + 00	1.436E-04	0.000E + 00	5.612E-06	1.111E-03	1.172E-02	9.8670	0.0926	0.0528
8	2712	3.279E-05	2.412E-04	1.681E-04	6.125E-04	3.287E-04	6.384E-07	3.951E-03	1.263E-05	8.917E-03	1.426E-02	8.1067	0.0761	0.0451
9	2133	6.199E-08	2.353E-03	2.028E-03	2.689E-03	6.443E-04	3.810E-03	6.173E-05	1.758E-03	2.611E-03	1.595E-02	7.2476	0.0680	0.0443
10	2792	3.038E-04	7.425E-06	8.410E-04	2.812E-03	3.287E-04	1.153E-03	2.743E-03	2.245E-05	7.899E-03	1.611E-02	7.1773	0.0674	0.0422
		Q=	10	Initial (Stage 1) Q Value and Weight Assumptions							1.156E-01	106.5273	1	0.5893
											Total (X)	Total (Z)	Sum of F	Normalized Prediction
													Min h	6.5361

Table 3 (continued)

Example of how the GCV prediction calculation of each data record in the TOB network can be audited in detail after stage 1

Stage 1	Evenly weighted sum of squares errors (wSSE)										
Rank of matches (in training subset)	Top-ranking matched records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent variable GCV (MJ/kg)
										Max h	35.4645
									Stage 1 Provisionally Predicted GCV for data record: #2738		29.5238
									Actual Measured GCV for data record: #2738		29.0936

This calculation is for the TOB prediction for data record 2738 (part of the testing subset)

Steps 2 and 6 of the method (Appendix 1) help to distribute the records for tuning and testing systematically (e.g., using a specified ranking-interval spacing to select from the ranked and sorted dataset), but avoids being subjectively selective.

This allocation was achieved initially ranking the full dataset in ascending (or descending) order of GCV values and then by selecting every 70th data record from the full data set to be put to one side during the tuning process and allocated to the testing subset. Two additional records were also added to the testing subset: one close to the lower GCV limit; and, one close to the upper GCV limit. A similar approach was then taken in the selection of the tuning subset records from the remaining dataset, i.e., every 70th data record from the combined training and tuning subset (ranked in order of GCV values) was allocated to the tuning subset. Two additional records are then added to the tuning subset: one close to the lower GCV limit; and, one close to the upper GCV limit. This approach ensures a full spread of GCV values, representative of the entire data range, in both the testing and tuning subsets. Such a spread cannot be guaranteed by using random sampling. For the learning network to be tuned across its entire dependent variable range it is essential that the tuning subset is distributed relatively evenly across that range.

Table 2 provides TOB-prediction results for coal GCV dataset. The optimum GCV-prediction performance is assessed by comparing actual and predicted GCV values. This is initially calculated for the tuning-subset records and achieves RMSE = 0.33462 MJ/kg; $R^2 = 0.9963$ with optimized Q = 9 (i.e. the nine-highest ranking data record matches from the training subset are used in the TOB stage-2 predictions). The TOB stage-2 variable weights established for the optimum solution (most accurate predictions) were: *w#1* = 5.304E−04, *w#2* = 8.317E−03, *w#3* = 0, w#4 = 1.576 E−03′ w#5 = 0; w#6 = 1.0; w#7 = 2.062E−03, w#8 = 0, and w#9 = 4.348E−03. The very small weights applied to several of the input variables have significant influence on the prediction accuracy achieved. Section 6 addresses this point. Suffice it to say here that it does not follow that the higher the weight applied to a variable in the TOB method, the more significant that variable is in determining the predicted values.

Table 2 compares the prediction achieved with sub-optimal values of Q (i.e., Q = 2 to 10) with the optimal value of Q = 9. The results show that accurate GCV predictions can be achieved using most Q values in that range ($R^2 = 0.9944$ for Q = 2; compared to 0.9963 for Q = 9). This suggests that for this learning network the value of Q plays a subordinate role, as high degrees of accuracy are achieved for all values of Q tested. This is further emphasized by the low range of root mean square (RMSE) values of 0.33462 (Q = 9) to 0.41393 MJ/kg (Q = 2) MJ/kg. These

Table 4 Example audit of the calculation details for the TOB stage 2 GCV prediction associated with specific data records. This calculation is for the TOB prediction for data record 2738 (part of the testing subset)

Example of how the GCV prediction calculation of each data record in the TOB network can be audited in detail after stage 2														
Stage 2	Optimized and independently weighted sum of squares errors (wSSE)													
	Variables	#1	#2	#3	#4	#5	#6	#7	#8	#9		The variable details of each of the top ten matching records for the record being predicted are available to inspect		
Rank of matches (from stage 1)	Top-ranking matched records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent variable GCV (MJ/kg)			
Test record #	2738	−0.9331	0.3950	0.0109	−0.1000	−0.1026	0.4124	−0.6667	−0.7621	−0.0683	−0.1026			
1	2065	−0.9701	0.3410	0.0656	−0.0867	−0.1308	0.4582	−0.6481	−0.8137	−0.1123	0.6264			
2	2696	−0.9405	0.4200	−0.0207	−0.0275	−0.1410	0.3816	−0.5889	−0.7611	−0.0636	0.5546			
3	2703	−0.9507	0.3064	0.0615	−0.0417	−0.1282	0.4068	−0.6667	−0.7655	−0.1312	0.5663			
4	2705	−0.9282	0.3480	0.0203	−0.0433	−0.1487	0.3712	−0.6407	−0.7166	−0.1658	0.5370			
5	2770	−0.9437	0.4836	−0.0253	−0.1417	−0.1538	0.3842	−0.6667	−0.7286	0.0102	0.5732			
6	4228	−0.9521	0.3568	0.0598	−0.1417	−0.1231	0.4446	−0.5519	−0.7876	−0.1013	0.6288			
7	2730	−0.9401	0.5183	−0.0639	−0.0917	−0.1026	0.3955	−0.6667	−0.7588	−0.0212	0.5703			
8	2712	−0.9412	0.4170	−0.0075	−0.0650	−0.1282	0.4136	−0.5778	−0.7571	−0.2019	0.5925			
9	2133	−0.9327	0.3264	0.0745	−0.1733	−0.1385	0.4997	−0.6556	−0.8214	−0.1406	0.6517			
10	2792	−0.9577	0.3911	0.0519	−0.1750	−0.1282	0.4605	−0.7407	−0.7554	−0.1940	0.6269			
		Normalized values (scale: −1 to +1)												
	Weights	0.00030	0.00824	0.00000	2.004E−03	0.000E+00	1.000E+00	1.872E−03	0.000E+00	3.985E−03	Optimized			
Rank of matches (from stage 1)	Top−ranking matched records	wSSE#1	wSSE#2	wSSE#3	wSSE#4	wSSE#5	wSSE#6	wSSE#7	wSSE#8	wSSE#9	Sum of weighted squared errors (SumE)	Y = X/SumE	F = Y/Z	Contributions to prediction F*GCV (of record)
Test Record #	2738	Q=	9	Optimized Q value	Independently weighted sum of squares errors (wSSE)									
1	2065	4.139E−07	2.399E−05	0.000E+00	3.562E−07	0.000E+00	2.094E−03	6.419E−07	0.000E+00	7.712E−06	2.127E−03	7.13E+00	0.0183	0.0115
2	2696	1.656E−08	5.171E−06	0.000E+00	1.053E−05	0.000E+00	9.481E−04	1.132E−05	0.000E+00	8.853E−08	9.752E−04	1.55E+01	0.0399	0.0221
3	2703	9.385E−08	6.475E−05	0.000E+00	6.818E−06	0.000E+00	3.192E−05	0.000E+00	0.000E+00	1.574E−05	1.193E−04	1.27E+02	0.3264	0.1848
4	2705	7.358E−09	1.822E−05	0.000E+00	6.434E−06	0.000E+00	1.701E−03	1.258E−06	0.000E+00	3.781E−05	1.765E−03	8.59E+00	0.0221	0.0119
5	2770	3.379E−08	6.475E−05	0.000E+00	3.479E−06	0.000E+00	7.980E−04	0.000E+00	0.000E+00	2.459E−05	8.908E−04	1.70E+01	0.0437	0.0251
6	4228	1.095E−07	1.200E−05	0.000E+00	3.479E−06	0.000E+00	1.037E−03	2.467E−05	0.000E+00	4.338E−06	1.082E−03	1.40E+01	0.0360	0.0226
7	2730	1.502E−08	1.253E−04	0.000E+00	1.392E−07	0.000E+00	2.873E−04	0.000E+00	0.000E+00	8.853E−06	4.216E−04	3.60E+01	0.0924	0.0527
8	2712	1.986E−08	3.977E−06	0.000E+00	2.455E−06	0.000E+00	1.277E−06	1.479E−05	0.000E+00	7.107E−05	9.359E−05	1.62E+02	0.4161	0.2466
9	2133	3.754E−11	3.878E−05	0.000E+00	1.078E−05	0.000E+00	7.619E−03	2.311E−07	0.000E+00	2.082E−05	7.690E−03	1.97E+00	0.0051	0.0033
10	2792	0	0	0	0	0	0	0	0	0	0	0	0	0
											1.516E−02	3.89E+02	0.4864311	0.5805
											Total (X)	Total (Z)	Sum of F	Normalized Prediction
													Min h	6.5361
													Max h	35.4645

Table 4 (continued)

Example of how the GCV prediction calculation of each data record in the TOB network can be audited in detail after stage 2

Stage 2	Optimized and independently weighted sum of squares errors (wSSE)	
	Stage 2 Optimized Prediction of GCV for data record: #2738	29.3975
	Actual Measured GCV for data record: #2738	29.0936

values indicate that the dataset is not noticeably underfitted when Q = 2.

The TOB stage 1 predictions ($Q = 10$ and $Wn = 0.5$; see the left side of Table 2) also show credible accuracy (RMSE = 0.48381 MJ/kg; $R^2 = 0.9923$). The data-record-matching conducted by TOB Stage 1 is, clearly, an essential contributing component to the optimum GCV predictions.

Testing-subset predictions (Table 2: lower two rows) applying optimum Q and Wn values achieves very slightly higher accuracy in GCV prediction than the tuning set GCV predictions. Testing subset accuracy metric values are: RMSE = 0.30556 MJ/kg and $R^2 = 0.9970$ (Table 2). The accuracy of the TOB method's predictions compares favourably with those achieved by other machine learning and empirical correlation methods applied to this dataset (Matin and Chelgani 2016) and other data sets (Feng et al. 2015; Tan et al. 2015).

Figures 4 and 5 display the optimum TOB predictions for coal GCV the tuning and testing subsets, respectively. It is apparent from these graphs that the data coverage for coals in the GCV < 15 range is only sparsely sampled by both tuning and testing subsets. In order to verify that the TOB network can produce consistent predictions of meaningful accuracy in this sparser data area a separate TOB network is sampled to analyse and tune that section of the GCV distribution more extensively.

Auditing and interrogating TOB predictions

Matching of data records rather than establishing correlations between input variables is the basis of the TOB methodology. It is constrained in its predictions by the range covered by the lowest and highest GCV values; it cannot extrapolate beyond that range (in contrast to many other AI methods).

For many machine-learning algorithms and empirical correlations it makes sense to apply their optimum coefficients to the data records in the training set as well as the tuning and/or testing sets used to verify their accuracy. When applying the TOB algorithm predictions are made only for data records that are not already present in the training subset. Predictions for records already in the TOB training set would achieve exact matches (RMSE = 0) and provide no insight to the accuracy of the method. This is a fundamental and distinguishing difference between the TOB methodology and most other AI methods (except K-learning-type methods).

As there are no hidden or difficult-to-access intermediate correlations involved in TOB the underlying calculations in each prediction are accessible Indeed, a key benefit of the TOB algorithm is that it allows each prediction calculation step to be readily analysed. Tables 3, 4, 5 and 6 provide examples of how this is achieved and the

Table 5 Example audit of the calculation details for the TOB stage 1 GCV prediction associated with specific data records. This calculation is for the TOB prediction for data record 193 (part of the testing subset)

Example of How the GCV Prediction Calculation of Each Data Record in the TOB network can be Audited in Detail after Stage 1														
Stage 1	Evenly Weighted Sum of Squares Errors (wSSE)													
Rank of matches (in training subset)	Top-ranking matched records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent variable GCV (MJ/kg)			
Test record	193	− 0.2641	0.1908	− 0.2449	− 0.3583	0.1538	− 0.0847	− 0.6667	0.0687	− 0.8696	− 0.0065	The same top ten matches established by TOB stage 1 are used in TOB stage 2 with different weights and Q value applied		
Ranking	Records	#1	#2	#3	#4	#5	#6	#7	#8	#9	GCV			
1	1652	− 0.2852	0.1757	− 0.2125	− 0.3875	0.0538	− 0.0757	− 0.7259	0.0958	− 0.9214	− 0.0525			
2	216	− 0.2923	0.1175	− 0.1653	− 0.4083	0.1026	− 0.0452	− 0.6296	0.0452	− 0.8853	0.0260			
3	3517	− 0.2570	0.1715	− 0.2521	− 0.3083	0.0256	− 0.1215	− 0.7037	0.1089	− 0.8696	− 0.1192			
4	5173	− 0.2342	0.2671	− 0.2893	− 0.4408	0.1641	− 0.0483	− 0.6963	0.0744	− 0.9529	0.0076			
5	158	− 0.1831	0.1484	− 0.3004	− 0.2667	0.2051	− 0.1384	− 0.6667	0.0921	− 0.8853	− 0.0388			
6	196	− 0.2254	0.1175	− 0.2232	− 0.3667	0.1538	− 0.0621	− 0.5185	0.0352	− 0.9010	0.0178			
7	189	− 0.1440	0.1977	− 0.3334	− 0.3517	0.1179	− 0.1489	− 0.6778	0.1357	− 0.8288	− 0.1308			
8	1551	− 0.3204	0.2062	− 0.2063	− 0.3917	0.1282	0.0113	− 0.7778	− 0.0251	− 0.8539	0.0774			
9	181	− 0.3275	0.3218	− 0.2883	− 0.3417	0.1282	− 0.0678	− 0.5556	0.0385	− 0.9010	− 0.0255			
10	1640	− 0.3011	0.2536	− 0.2820	− 0.2783	0.0641	− 0.0746	− 0.7667	0.0245	− 0.7879	− 0.0293			
		Normalized values (Scale: − 1 to + 1)												
		0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	input variables equally weighted			
Ranking	Records	wSSE#1	wSSE#2	wSSE#3	wSSE#4	wSSE#5	wSSE#6	wSSE#7	wSSE#8	wSSE#9	Sum of weighted squared errors (SumE)	Y = X/SumE	F = Y/Z	Contributions to prediction F*GCV (of record)
1	1652	2.232E−04	1.129E−04	5.226E−04	4.253E−04	5.000E−03	4.086E−05	1.756E−03	3.682E−04	1.344E−03	9.793E−03	15.1796	0.1430	−0.0075
2	216	3.967E−04	2.680E−03	3.169E−03	1.250E−03	1.315E−03	7.820E−04	6.859E−04	2.750E−04	1.234E−04	1.068E−02	13.9219	0.1311	0.0034
3	3517	2.480E−05	1.856E−04	2.619E−05	1.250E−03	8.218E−03	6.743E−04	6.859E−04	8.081E−04	0.000E + 00	1.187E−02	12.5201	0.1179	−0.0141
4	5173	4.479E−04	2.911E−03	9.853E−04	3.403E−03	5.260E−05	6.640E−04	4.390E−04	1.622E−05	3.467E−03	1.239E−02	12.0019	0.1130	0.0009
5	158	3.279E−03	8.984E−04	1.539E−03	4.201E−03	1.315E−03	1.440E−03	0.000E + 00	2.750E−04	1.234E−04	1.307E−02	11.3715	0.1071	−0.0042
6	196	7.501E−04	2.680E−03	2.357E−04	3.472E−05	0.000E + 00	2.554E−04	1.097E−02	5.612E−04	4.937E−04	1.599E−02	9.2994	0.0876	0.0016
7	189	7.208E−03	2.406E−05	3.920E−03	2.222E−05	6.443E−04	2.056E−03	6.173E−05	2.245E−03	8.343E−04	1.702E−02	8.7364	0.0823	−0.0108
8	1551	1.587E−03	1.188E−04	7.450E−04	5.556E−04	3.287E−04	4.612E−03	6.173E−03	4.399E−03	1.234E−04	1.864E−02	7.9736	0.0751	0.0058
9	181	2.009E−03	8.583E−03	9.429E−04	1.389E−04	3.287E−04	1.436E−04	6.173E−03	4.545E−04	4.937E−04	1.927E−02	7.7154	0.0727	−0.0019
10	1640	6.835E−04	1.973E−03	6.902E−04	3.200E−03	4.027E−03	5.171E−05	5.000E−03	9.778E−04	3.337E−03	1.994E−02	7.4550	0.0702	−0.0021
		Q=	10	Initial (Stage 1) Q Value and Weight Assumptions							1.487E−01	106.1747	1	−0.0288
											Total (X)	Total (Z)	Sum of F	Normalized Prediction
													Min h	6.5361

Table 5 (continued)

Example of How the GCV Prediction Calculation of Each Data Record in the TOB network can be Audited in Detail after Stage 1

Stage 1	Evenly Weighted Sum of Squares Errors (wSSE)													
Rank of matches (in training subset)	Top-ranking matched records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent variable GCV (MJ/kg)			
													Max h	35.4645
												Stage 1 Provisionally Predicted GCV for data record: #193		20.5842
												Actual Measured GCV for data record: #193		20.9061

information it provides. These tables detail the TOB predictions (stages 1 and 2 presented separately) for record # 2738 (Tables 3, 4) and record # 193 (Tables 5, 6). Tables 3 and 5 focus on the TOB-stage-1 calculations ($Q = 10$; $Wn = 0.5$). Tables 4 and 6 focus on TOB-stage-2 predictions ($2 \le Q \le 10$; $0 \le Wn \le 1$). The left side of the tables displays the ten high-matching records established by TOB stage 1 for data records #193 and #2738. The upper half of the tables displays the nine input-variable values (#1 to #9) and the dependent variable (GCV). These values are all expressed in normalized terms (ranging from -1 to $+1$). The matching records are listed in order of the matching rank; the first is the closest match; the one listed 10th is the 10th -best match of all the records, from the 6155 records in the GCV training set.

The lower half of Tables 3, 4, 5 and 6 applies the Q and Wn values to the squared errors calculated for the variables. The column fourth from the right (in the lower half each table) lists the sum of the weighted-squared errors (SumE). For stage 1 (Q = 10) the values of ten matching records are involved in the calculation of SumE; for stage 2 (Optimum Q = 6) the values of only six matching records are involved in the calculation of SumE. In both cases it is the last three columns to the right in the lower half of these tables that detail the actual and relative contributions of each of the high-matching records to the prediction. It is the sum of the contributions in the right end column (lower half of the tables) that provides the prediction in normalized terms. the lowest two rows (right side numbers) in each of Tables 3, 4, 5 and 6 provide a comparison with of the actual predicted GCV (transformed from the normalized scale) and the measured GCV for the data record being analysed.

The last three columns of Tables 3, 4, 5 and 6 (lower half of those tables) are the key calculations to focus upon. These use the SumE values to generate GCV predictions for the data record in question. The formulas provided in the heading of each of those three columns explains the calculations made.

For record #2738 (Tables 3, 4) each of the top-ten high-ranking records contributes small amounts to the prediction, but the higher the ranked match the greater its contribution to that prediction. The rank#1 match contributes 14.2% to the prediction whereas the rank #10 match only contributes 6.7% to the GCV prediction for record 2738. The GCV prediction achieved is quite accurate; < +0.5 MJ/kg above the measured value of 29.1 MJ/kg. In TOB stage 2 (Table 4) for record 2738 only the top nine match records contribute to the GCV prediction because the optimized Q value selected is 9. Also, as the weightings for each variable are not equal for the optimized solution the relative contributions of those top-ranking matches are no longer governed only by the closeness of their match with the record 2738 variables. This is clear

Table 6 Example audit of the calculation details for the TOB stage 2 GCV prediction associated with specific data records. This calculation is for the TOB prediction for data record 193 (part of the testing subset)

Example of how the GCV prediction calculation of each data record in the TOB network can be audited in detail after stage 2														
Stage 2	Optimized and independently weighted sum of squares errors (wSSE)													
	Variables	#1	#2	#3	#4	#5	#6	#7	#8	#9		The variable details of each of the top ten matching records for the record being predicted are available to inspect		
Rank of matches (from stage 1)	Top−ranking matched records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent Variable GCV (MJ/kg)			
Test record #	193	−0.2641	0.1908	−0.2449	−0.3583	0.1538	−0.0847	−0.6667	0.0687	−0.8696	−0.0065			
1	1652	−0.2852	0.1757	−0.2125	−0.3875	0.0538	−0.0757	−0.7259	0.0958	−0.9214	−0.0525			
2	216	−0.2923	0.1175	−0.1653	−0.4083	0.1026	−0.0452	−0.6296	0.0452	−0.8853	0.0260			
3	3517	−0.2570	0.1715	−0.2521	−0.3083	0.0256	−0.1215	−0.7037	0.1089	−0.8696	−0.1192			
4	5173	−0.2342	0.2671	−0.2893	−0.4408	0.1641	−0.0483	−0.6963	0.0744	−0.9529	0.0076			
5	158	−0.1831	0.1484	−0.3004	−0.2667	0.2051	−0.1384	−0.6667	0.0921	−0.8853	−0.0388			
6	196	−0.2254	0.1175	−0.2232	−0.3667	0.1538	−0.0621	−0.5185	0.0352	−0.9010	0.0178			
7	189	−0.1440	0.1977	−0.3334	−0.3517	0.1179	−0.1489	−0.6778	0.1357	−0.8288	−0.1308			
8	1551	−0.3204	0.2062	−0.2063	−0.3917	0.1282	0.0113	−0.7778	−0.0251	−0.8539	0.0774			
9	181	−0.3275	0.3218	−0.2883	−0.3417	0.1282	−0.0678	−0.5556	0.0385	−0.9010	−0.0255			
10	1640	−0.3011	0.2536	−0.2820	−0.2783	0.0641	−0.0746	−0.7667	0.0245	−0.7879	−0.0293			
		Normalized values (Scale: −1 to +1)												
	Weights	0.00030	0.00824	0.00000	2.004E−03	0.000E+00	1.000E+00	1.872E−03	0.000E+00	3.985E−03	Optimized			
Rank of matches (from stage 1)	Top-ranking matched records	wSSE#1	wSSE#2	wSSE#3	wSSE#4	wSSE#5	wSSE#6	wSSE#7	wSSE#8	wSSE#9	Sum of weighted squared errors (SumE)	Y = X/SumE	F = Y/Z	Contributions to prediction F*GCV (of record)
Test Record #	193	Q=	9	Optimized Q Value	Independently Weighted Sum of Squares Errors (wSSE)									
1	1652	1.351E−07	1.862E−06	0.000E+00	1.705E−06	0.000E+00	8.171E−05	6.573E−06	0.000E+00	1.071E−05	1.027E−04	2.13E+02	0.5959	−0.0313
2	216	2.403E−07	4.419E−05	0.000E+00	5.009E−06	0.000E+00	1.564E−03	2.568E−06	0.000E+00	9.837E−07	1.617E−03	1.35E+01	0.0378	0.0010
3	3517	1.502E−08	3.060E−06	0.000E+00	5.009E−06	0.000E+00	1.349E−03	2.568E−06	0.000E+00	0.000E+00	1.359E−03	1.61E+01	0.0450	−0.0054
4	5173	2.712E−07	4.799E−05	0.000E+00	1.364E−05	0.000E+00	1.328E−03	1.643E−06	0.000E+00	2.763E−05	1.419E−03	1.54E+01	0.0431	0.0003
5	158	1.986E−06	1.481E−05	0.000E+00	1.684E−05	0.000E+00	2.881E−03	0.000E+00	0.000E+00	9.837E−07	2.915E−03	7.49E+00	0.0210	−0.0008
6	196	4.542E−07	4.419E−05	0.000E+00	1.392E−07	0.000E+00	5.107E−04	4.108E−05	0.000E+00	3.935E−06	6.005E−04	3.64E+01	0.1019	0.0018
7	189	4.365E−06	3.966E−07	0.000E+00	8.906E−08	0.000E+00	4.112E−03	2.311E−07	0.000E+00	6.650E−06	4.124E−03	5.30E+00	0.0148	−0.0019
8	1551	9.611E−07	1.958E−06	0.000E+00	2.226E−06	0.000E+00	9.225E−03	2.311E−05	0.000E+00	9.837E−07	9.254E−03	2.36E+00	0.0066	0.0005
9	181	1.216E−06	1.415E−04	0.000E+00	5.566E−07	0.000E+00	2.873E−04	2.311E−05	0.000E+00	3.935E−06	4.576E−04	4.77E+01	0.1337	−0.0034
10	1640	0	0	0	0	0	0	0	0	0	0	0	0	0
											2.185E−02	3.57E+02	0.8448018	−0.0392
											Total (X)	Total (Z)	Sum of F	Normalized Prediction
													Min h	6.5361
													Max h	35.4645

Table 6 (continued)

Example of how the GCV prediction calculation of each data record in the TOB network can be audited in detail after stage 2

Stage 2	Optimized and independently weighted sum of squares errors (wSSE)		
		Stage 2 Optimized Prediction of GCV for data record: #193	20.4335
		Actual Measured GCV for data record: #193	20.9061

as the prediction is now dominated by record match ranks #8 and #3 (41.6% and 40.0% contributions, respectively, to the GCV prediction; see lower half Table 4 column second from the right). The optimization adjustment for this record generates a better stage-2 prediction, i.e., ~+0.3 above the 29.1 MJ/kg measured GCV value for record 2738.

The TOB stage 2 prediction is not always better than the TOB stage 1 prediction, but that is the case most of the time as the optimizer is minimizing the RMSE across the entire tuning or testing subsets. The second example, for testing subset data record 193, shows a case where the stage 1 TOB prediction outperforms the stage 2 TOB prediction. The stage 1 prediction (Table 5) is ~-0.32 MJ/kg less than the measured GCV value for that record of 20.9 MJ/kg, whereas the stage 2 prediction (Table 6) is ~0.47 MJ/kg less than the measured value. In the stage#1 prediction all ten top-matching records contribute between ~7.0% and 14.3% to the prediction (rank # 1 contributes the most, but it does not dominate the prediction). In contrast, the stage #2 prediction is dominated by ranked-matches #1 (59.6%), #9(13.4%) and #6 (10.1%). When a prediction has more significant input from the lower of the top-ranking matches, the accuracy of that prediction is sometimes impaired. Both stage 1 and stage 2 predictions are credible predictions for record 193 but influenced by those top-ten ranking matches quite differently.

For both examples, the important role of those variables with very-low-Wn (>0) values in determining the stage 2 predictions is apparent. If those variable weights were zero, they would not contribute to the improved accuracy of the optimized solution.

The details provided by Tables 3, 4, 5 and 6 highlight just how deeply it is possible to interrogate both the stage-1 and stage-2 predictions and its high level of prediction transparency. Moreover, it is able to provide almost forensic insight into the similarity between each record tested against each record in the training subset. This can be useful in the case of trying to identify the exact provenance of a certain coal (e.g. a specific basin or even a specific mine can sometimes be identified by the closeness of the high-ranking matches). Hence, in some cases it is not only the prediction of the dependent variable value that can be derived from this learning network, but this can be accompanied by provenance information. Other machine learning algorithms and empirical relationships that are underpinned by correlations cannot easily deliver this level of detail into the degree of similarity with specific records in their training subsets.

Table 7 Statistical summary of data subset (283 records) compiled for gross calorific value (GCV) between > 6 and < 15 MJ/kg of US coals with each record linking measured proximate and ultimate analysis variables to their measured GCV value (MJ/kg)

COALQUAL dataset used for gross calorific value (GCV) interval > 6 to < 15 MJ/kg prediction										
Compiled dataset: 283 data records	Moisture (%)	Volatiles (%)	Fixed carbon (%)	Ash (%)	H (%)	C (%)	N (%)	O (%)	S (%)	GCV (MJ/kg)
Variable descriptor	#1	#2	#3	#4	#5	#6	#7	#8	#9	Dependent
Min	18.30	18.50	7.80	3.50	2.90	18.80	0.20	32.20	0.10	6.54
Max	57.20	40.00	35.75	24.30	8.20	42.10	1.20	59.90	8.40	14.99
Mean	38.64	25.55	24.46	11.36	6.62	34.62	0.62	45.85	0.93	13.30
Standard deviation	6.11	2.72	4.20	5.39	0.74	3.81	0.16	4.88	0.91	1.46
20-percentile	34.40	23.60	21.92	6.54	6.12	31.90	0.50	42.34	0.35	12.24
40-percentile	37.98	24.80	24.30	8.55	6.60	34.50	0.60	45.80	0.50	13.35
60-percentile	40.72	25.90	26.02	11.70	6.90	36.37	0.60	47.80	0.80	13.98
80-percentile	43.20	27.20	27.65	17.18	7.20	37.61	0.70	49.60	1.30	14.49

This data subset is extracted from the low GCV end of that compiled by Matin and Chelgani (2016) with records filtered from the US Geological Survey Coal Quality (COALQUAL) database (Version 2.0), open file report 97–134 (Bragg et al. 1997)

Includes all data records in the Training, Tuning and Testing subsets

Analysis of a more sparsely populated data subset (GCV > 6 to < 15 MJ/kg)

To provide more insight into the more sparsely populated GCV < 15 MJ/kg range of the compiled dataset, a separate TOB network is evaluated using just the 283 samples in the data base with GCV < 15 MJ/kg. The statistical summary of the measured variables for these 283 records (which are also included in the dataset described in Table 1) is provided in Table 7.

This more-focused TOB divides its data records, using the same methodology as already described (training records = 235; tuning records = 24 records; testing records = 24 records). This TOB tunes this GCV interval with 24 data records, whereas in the larger-dataset TOB previously described only involved 5 data records to tune that GCV interval. Table 8 describes the details of the tuned and optimized prediction performance of this focused TOB, with sensitivities, demonstrating high prediction accuracy as illustrated (Fig. 6) for its testing subset (RMSE = 0.2944; $R^2 = 0.9644$). The accuracy is not as high as for the larger dataset (see Fig. 5) due to the greater spacing of training-subset data records (i.e. more sparsely distributed data) for the GCV interval < 15 MJ/kg. This highlights a key positive feature of the TOB learning network approach, i.e., it is resistant to over-fitting sparse data sets. As the spacing between datapoints in the training set increases, the statistically accessed accuracy of its predictions tends to decrease. Although such behaviour is intuitive that is not necessarily the outcome with empirical calculations or learning networks driven by complex correlations between the variables, which are prone to over-fitting.

Table 8 reveals that the best prediction performance for this focused TOB network is for Q = 9. The sensitivity analysis shows that RMSE increases and R^2 decreases as the value of Q decreases below the value of 9 with R^2 falling below 0.96 for Q values below 8. However, for Q = 3 the prediction accuracy is also good and superior to values of Q = 2 or Q = 4. This is true for both tuning and testing subsets (Table 8). Indeed Q = 3 represents a local minimum, which the Solver evolutionary optimizer selected as its optimum (became trapped at) on most of its runs. This bimodal optimization outcome suggests that for some data records better predictions are achieved for Q = 3. Closer inspection of data record #3452, highlighted on Fig. 6 for the optimum Q = 9 tuned setting as having a relatively low accuracy in comparison to most other records in the testing subset.

Tables 9 and 10 describe the detailed calculation of the GCV TOB stages 1 and 2 predictions for data record #3452, respectively. In the stage 1 prediction (Table 9) the matching record ranked#1 (record #3106) contributes 32.95% to the GCV prediction, with the other high−ranking−matched records contributing progressively less until matched record ranked #10 (record#21) contributes just 6.0% to the GCV prediction. For this stage 1 prediction, the top−three matched records contribute > 50% to that calculation. This achieves a prediction of high accuracy, i.e., −0.3 below the measured GCV value of 12.67 MJ/kg.

For data record #3452 the TOB-stage-2 prediction (Table 10) generates a significantly less-accurate prediction than that achieved by the stage-1 prediction (Table 9). The reason for this is that in the TOB stage 2 solution, with Q = 9 and variable weights applied, the top-3 matched records only contribute about 18% to the prediction. On the other hand,

Table 8 Prediction accuracy versus TOB optimization control metrics (Q and Wn) for the 283-record GCV data set covering the range GCV > 6 and < 15 MJ/kg

Transparent open box (TOB) learning network results and variable weightings in the prediction of gas calorific value (GCV) of coal for interval > 6 to < 15 MJ/kg from proximate and ultimate data variables (dataset with 283 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 24 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 #	10	9	3	10	9	8	7	6	5	4	3	2
Prediction performance of optimum and constrained optimum solutions applied to the tuning subset (24 records: ~ 8.5% of total dataset)													
RMSE	MJ/kg	0.56072	0.27460	0.31032	0.28002	0.27460	0.31344	0.34573	0.34469	0.35414	0.33870	0.30785	0.41020
R^2	fraction	0.8925	0.9686	0.9615	0.9685	0.9686	0.9635	0.9513	0.9489	0.9459	0.9505	0.9614	0.9403
Weightings ($0 \le w \le 1$) Applied to constrained optimum solutions for the tuning subset													
Moisture (%)	#1	0.5	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
Volatiles (%)	#2	0.5	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
Fixed carbon (%)	#3	0.5	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
Ash (%)	#4	0.5	0.000E+00	3.847E-02	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00
H (%)	#5	0.5	2.018E-01	9.458E-01	1.688E-01	1.691E-01	1.944E-01	2.044E-01	0.000E+00	9.525E-02	2.919E-01	1.084E-01	2.787E-01
C (%)	#6	0.5	1.000E+00	2.007E-02	1.000E+00	8.379E-01	9.992E-01	1.000E+00	1.000E+00	7.504E-01	1.000E+00	1.000E+00	1.000E+00
N (%)	#7	0.5	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.828E-03	8.576E-03	4.542E-03	5.440E-03	0.000E+00
O (%)	#8	0.5	0.000E+00	1.110E-01	0.000E+00	0.000E+00	0.000E+00	0.000E+00	4.400E-02	0.000E+00	0.000E+00	0.000E+00	0.000E+00
S (%)	#9	0.5	0.000E+00	3.000E+00	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.763E-02	2.731E-02	3.670E-02	7.809E-02	3.806E-02
Ratio of #6 weight to #5 weight		1	4.95	0.02	5.92	4.95	5.14	4.89	N/A	7.88	3.43	9.23	3.59
Prediction performance of optimum solution variable weightings and Q value applied to the testing subset (24 records: ~ 8.5% of total dataset)													
RMSE	MJ/kg	0.51253	0.29438	0.33076	0.28814	0.29438	0.33067	0.37210	0.37810	0.35196	0.35030	0.34314	0.39727
R^2	fraction	0.9116	0.9644	0.9572	0.9663	0.9644	0.9581	0.944	0.9419	0.9498	0.9503	0.9534	0.9357

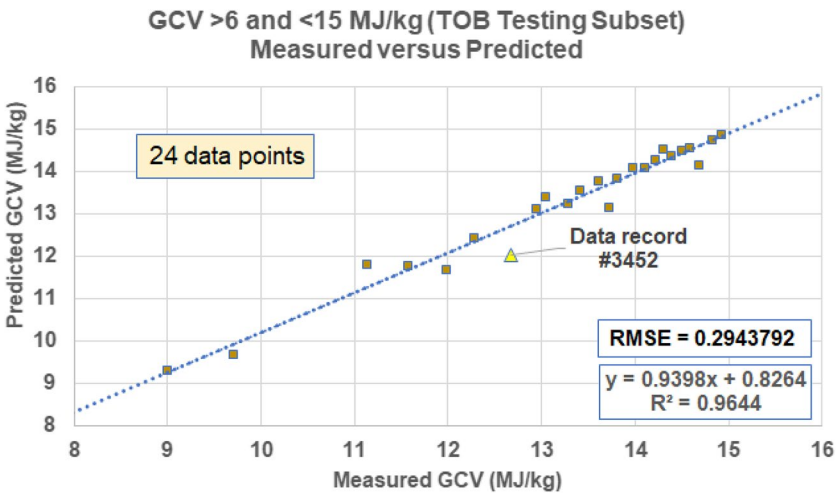

Fig. 6 Predicted versus measured GCV (MJ/kg) for 24 testing subset data records used to test the TOB model with 283 records (GCV > 6 and < 15 MJ/kg) in the training subset. The 24 testing data records were excluded from the TOB training process. Data record #3452 with a relatively poor fit between measured and predicted GCV is highlighted and its TOB prediction is considered in detail in Table 9

the matched record #8 (record # 3110) contributes 68% to the GCV prediction. This achieves a prediction of less impressive accuracy, i.e., −0.64 below the measured GCV value of 12.67 MJ/kg. In this case, considering the analysis just described and the sensitivity analysis of Table 8, a case could be made for applying a Q = 3 cut off for the prediction of this data record. This approach highlights how the transparency of the TOB learning network's calculation aids the analysis of outlier data records (i.e., those for which predictions fall significantly off trend). It makes it possible identify, in detail, the reasons for such outlying prediction values. It also often provides the justification for potential adjustments that might be made to improve /correct the predictions for such problematic data records.

Auditing TOB predictions and conducting sensitivity analysis (e.g. varying Q values from the optimum and changing the data-subset allocation percentages) focused on specific data-records facilitates rigorous outlier analysis; something that is difficult with most other AI methods not easily possible with correlation-based machine learning algorithms or empirical calculations. This TOB strength is particularly beneficial for datasets for which details of specific data-record predictions are important (e.g. for commercial valuation purposes or detailed sample provenance purposes; both of which apply to GCV and commercial coal datasets). This feature could also be usefully applied to other commercially-important characteristics of coal (e.g., predicting coal grindability from multiple input variables based on coal petrological properties).

Although the coal dataset studied here is relatively large, and the TOB algorithm clearly copes well with such numbers of data records, as a "big data" tool, the TOB algorithm may have some limitations with very large datasets. Clearly, the algorithm has to contain and manage a large training data base, whereas the performance (i.e., computational speed) of the algorithm is also likely to progressively deteriorate as the intrinsic dimensionality of the variable space increases. Further studies are required to establish the limits of applicability of the algorithm to such "big data" sets. However, although computational time is likely to deteriorate for very large data sets, the transparency provided by the TOB algorithm may compensate for this. As stage 2 of the algorithm focuses on just a few of the best matches (i.e., up to ten or so) the collective influence of a significant number of variables would remain fully transparent.

The COALQUAL dataset lends itself to further studies on the impacts of sparse data coverage on TOB prediction performance. A future study will conduct sensitivity analysis that progressively excludes percentages of the dataset from the training data subset used for model tuning (i.e. adding those excluded data records to the testing subset). This will quantify how sparse the training data subset can become before it ceases to yield meaningfully accurate predictions for the dependent variable.

Conclusions

The transparent open-box (TOB) learning network algorithm provides credible and reliable predictions of dependent variables, such as coal gross calorific value (GCV), that involve complex, highly dispersed and non-linear datasets for the influencing variables. Its high-prediction accuracy, demonstrated in this study, when applied to predict GCV from nine influencing variables from proximate and ultimate analysis from a large published data set (6339 data records of US coals) testifies to such capabilities. The method could be easily applied to more limited datasets, e.g., those based upon only the easier to obtain proximate analysis variables.

TOB's prediction performance for this published coal data set compares favourably to that reported by other artificial-intelligence algorithms and empirical correlations,

Table 9 Example audit of the calculation details for the TOB stage 1 GCV prediction associated with specific data records. This calculation is for the TOB prediction for data record 193 (part of the testing subset)

Example of how the GCV prediction calculation of each data record in the TOB network can be audited in detail after stage 1														
Stage 1	Evenly weighted sum of squares errors (wSSE)													
Rank of Matches (in training subset)	Top-ranking matched records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent variable GCV (MJ/kg)			
Test record	3452	−0.1928	−0.7581	0.2021	0.6154	0.2075	0.0129	−0.2000	−0.5596	−0.0361	0.4505	The same top ten matches established by TOB stage 1 are used in TOB stage 2 with different weights and Q value applied		
Ranking	Records	#1	#2	#3	#4	#5	#6	#7	#8	#9	GCV			
1	3106	−0.1774	− 0.5907	0.0089	0.6731	0.1321	0.0300	−0.4000	− 0.5162	− 0.2530	0.2030			
2	3453	−0.4036	− 0.1907	0.1163	0.5385	0.1321	0.2790	0.0000	− 0.6895	− 0.1325	0.6447			
3	3449	−0.5116	− 0.2837	0.1592	0.7788	− 0.0189	0.2189	0.0000	− 0.8628	0.1084	0.5567			
4	3440	−0.2391	− 0.6000	− 0.0912	0.9327	0.1321	− 0.1588	− 0.6000	− 0.4801	− 0.5181	0.1144			
5	3459	−0.0129	− 0.6279	0.0662	0.3269	0.2453	0.1245	− 0.6000	− 0.2635	− 0.5904	0.2882			
6	5622	−0.0643	− 0.2186	− 0.1199	0.2500	0.2830	− 0.0215	0.0000	− 0.2202	− 0.2289	0.2156			
7	3460	−0.3285	− 0.6167	0.4962	0.3260	0.1585	0.5116	− 0.2400	− 0.5653	− 0.6627	0.8058			
8	3107	−0.0848	− 0.6930	0.0376	0.5673	0.2453	0.0043	− 1.0000	− 0.3791	− 0.4217	0.3311			
9	3110	−0.1260	− 0.4884	0.0018	0.4808	0.2830	0.1588	− 0.6000	− 0.3646	− 0.7590	0.4851			
10	21	−0.3882	− 0.5349	0.2522	0.6827	0.1321	0.3562	− 0.4000	− 0.6390	− 0.8313	0.6337			
		Normalized values (Scale: −1 to + 1)												
		0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	input variables equally weighted			
Ranking	Records	wSSE#1	wSSE#2	wSSE#3	wSSE#4	wSSE#5	wSSE#6	wSSE#7	wSSE#8	wSSE#9	Sum of weighted squared errors (SumE)	Y = X/SumE	F = Y/Z	Contributions to prediction F*GCV (of record)
1	3106	1.190E−04	1.402E−02	1.866E−02	1.664E−03	2.848E−03	1.474E−04	2.000E−02	9.384E−04	2.352E−02	8.191E−02	42.1273	0.3295	0.0669
2	3453	2.222E−02	1.610E−01	3.687E−03	2.959E−03	2.848E−03	3.540E−02	2.000E−02	8.445E−03	4.645E−03	2.612E−01	13.2115	0.1033	0.0666
3	3449	5.081E−02	1.125E−01	9.217E−04	1.336E−02	2.563E−02	2.122E−02	2.000E−02	4.598E−02	1.045E−02	3.009E−01	11.4681	0.0897	0.0499
4	3440	1.071E−03	1.250E−02	4.304E−02	5.034E−02	2.848E−03	1.474E−02	8.000E−02	3.154E−03	1.161E−01	3.238E−01	10.6567	0.0833	0.0095
5	3459	1.619E−02	8.480E−03	9.242E−03	4.161E−02	7.120E−04	6.226E−03	8.000E−02	4.382E−02	1.536E−01	3.599E−01	9.5896	0.0750	0.0216
6	5622	8.261E−03	1.455E−01	5.184E−02	6.675E−02	2.848E−03	5.894E−04	2.000E−02	5.758E−02	1.858E−02	3.720E−01	9.2764	0.0725	0.0156
7	3460	9.212E−03	9.996E−03	4.325E−02	4.188E−02	1.203E−03	1.244E−01	8.000E−04	1.668E−05	1.963E−01	4.270E−01	8.0822	0.0632	0.0509

Table 9 (continued)

Example of how the GCV prediction calculation of each data record in the TOB network can be audited in detail after stage 1

Stage 1	Evenly weighted sum of squares errors (wSSE)													
Rank of Matches (in training subset)	Top-ranking matched records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent variable GCV (MJ/kg)			
8	3107	5.829E−03	2.120E−03	1.354E−02	1.156E−03	7.120E−04	3.684E−05	3.200E−01	1.629E−02	7.432E−02	4.340E−01	7.9511	0.0622	0.0206
9	3110	2.234E−03	3.639E−02	2.007E−02	9.061E−03	2.848E−03	1.065E−02	8.000E−02	1.900E−02	2.613E−01	4.415E−01	7.8155	0.0611	0.0297
10	21	1.909E−02	2.492E−02	1.254E−03	2.265E−03	2.848E−03	5.894E−02	2.000E−02	3.154E−03	3.162E−01	4.486E−01	7.6920	0.0602	0.0381
		Q=	10	Initial (Stage 1) Q Value and Weight Assumptions							3.451E+00	127.8703	1	0.3695
											Total (X)	Total (Z)	Sum of F	Normalized Prediction
													Min h	6.5361
													Max h	14.9934
										Stage 1 provisionally predicted GCV for data record: #3452				12.3272
										Actual measured GCV for data record: #3452				12.6697

with the added benefit that it is more easily audited and generally more transparent. The TOB algorithm does not develop any correlations when calculating its predictions. Instead, it establishes (in TOB stage 1) the closest matches with ten data records in its large associated training subset. In TOB stage 2 the algorithm improves its prediction, based on statistical measures of accuracy for tuning and testing data subsets (i.e., minimizing root mean squared error between predicted and measured GCV values). It achieves this by applying an optimizer to select the number of those matches ($2 \leq Q \leq 10$) and applying tuned weights to the errors associated with each input variable.

The calculations involved in the predictions derived from the TOB algorithm are individually auditable. Standard Solver optimizers or customized evolutionary or non-linear optimization algorithms can be used to successfully and transparently achieve the TOB stage 2 optimized predictions. Such flexibility and access to the underlying calculations is not possible with most other artificial-intelligence prediction methods or empirical calculations.

An additional valuable feature of the TOB algorithm is the ease with which sensitivity analysis can be conducted by modifying its Q value. In particular, the Q-value sensitivities can help to identify whether the algorithm is over-fitting or underfitting a dataset. These positive attributes make the TOB algorithm a suitable prediction-performance benchmark with which to compare the predictions of other machine-learning and empirical correlation algorithms. It typically provides complementary results to other algorithms with respect to insight to the underlying dataset. Indeed, in some cases, where the dataset covers coals from many different regions and mines the TOB algorithm has the ability, through its record matching stage 1 routine, to identify the provenance of specific samples.

The detailed calculations shown for example data records demonstrate exactly how the predictions of the TOB algorithm can be audited and assessed. These detailed calculations are not complex, rather they highlight the prediction mechanisms involved and the key roles played by the optimized Q value and the input variable weights in producing the stage 2 optimized predictions. The ability to interrogate and verify in detail specific predictions is increasingly important for providing user confidence in prediction algorithms. By revealing useful information about the relative importance of identified training-subset records in terms of their contributions to specific predictions, and the problematic nature of other data records (e.g., outlying values of certain metrics not replicated in other data records), the TOB method provides such user confidence. In some applications it may be worth sacrificing a small degree of accuracy in order to obtain such insight and confidence associated with the predictions to be deployed.

Table 10 Example audit of the calculation details for the TOB stage 2 GCV prediction associated with specific data records. This calculation is for the TOB prediction for data record 193 (part of the testing subset)

Example of How the GCV Prediction Calculation of Each Data Record in the TOB network can be Audited in Detail after Stage 2														
Stage 2	Optimized and Independently Weighted Sum of Squares Errors (wSSE)													
	Variables	#1	#2	#3	#4	#5	#6	#7	#8	#9		The variable details of each of the top ten matching records for the record being predicted are available to inspect		
Rank of Matches (from Stage 1)	Top−ranking Matched Records	Moist	Vols	FixC	Ash	H	C	N	O	S	Dependent Variable GCV (MJ/kg)			
Test Record #	3452	− 0.1928	− 0.7581	0.2021	0.6154	0.2075	0.0129	− 0.2000	− 0.5596	− 0.0361	0.4505			
1	3106	− 0.177378	− 0.590698	0.0089445	0.6730769	0.1320755	0.03004292	− 0.4	− 0.516245	− 0.253012	0.2030			
2	3453	− 0.403599	− 0.190698	0.1162791	0.5384615	0.1320755	0.27896996	− 1.11E−16	− 0.689531	−0.13253	0.6447			
3	3449	− 0.511568	− 0.283721	0.1592129	0.7788462	− 0.018868	0.21888412	− 1.11E−16	− 0.862816	0.1084337	0.5567			
4	3440	− 0.239075	− 0.6	− 0.091234	0.9326923	0.1320755	− 0.15879828	− 0.6	− 0.480144	− 0.518072	0.1144			
5	3459	− 0.012853	− 0.627907	0.0661896	0.3269231	0.245283	0.12446352	− 0.6	− 0.263538	− 0.590361	0.2882			
6	5622	− 0.064267	− 0.218605	− 0.119857	0.25	0.2830189	− 0.02145923	− 1.11E−16	− 0.220217	− 0.228916	0.2156			
7	3460	− 0.328535	− 0.616744	0.4962433	0.3259615	0.1584906	0.51158798	− 0.24	− 0.565343	− 0.662651	0.8058			
8	3107	− 0.084833	− 0.693023	0.0375671	0.5673077	0.245283	0.00429185	− 1	− 0.379061	− 0.421687	0.3311			
9	3110	− 0.125964	− 0.488372	0.0017889	0.4807692	0.2830189	0.15879828	− 0.6	− 0.364621	− 0.759036	0.4851			
10	21	− 0.388175	− 0.534884	0.2522361	0.6826923	0.1320755	0.35622318	− 0.4	− 0.638989	− 0.831325	0.6337			
		Normalized values (Scale: −1 to +1)												
	Weights	0.00000	0.00000	0.00000	0.000E+00	2.018E−01	1.000E+00	0.000E+00	0.000E+00	0.000E+00	Optimized			
Rank of Matches (from Stage 1)	Top−ranking Matched Records	wSSE#1	wSSE#2	wSSE#3	wSSE#4	wSSE#5	wSSE#6	wSSE#7	wSSE#8	wSSE#9	Sum of weighted squared errors (SumE)	Y = X/SumE	F = Y/Z	Contributions to prediction F*GCV (of record)
Test Record #	3452	Q=	9	Optimized Q Value	Independently Weighted Sum of Squares Errors (wSSE)									
1	3106	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.150E−03	2.947E−04	0.000E+00	0.000E+00	0.000E+00	1.444E−03	3.07E+02	0.1695	0.0344
2	3453	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.150E−03	7.081E−02	0.000E+00	0.000E+00	0.000E+00	7.196E−02	6.17E+00	0.0034	0.0022
3	3449	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.035E−02	4.244E−02	0.000E+00	0.000E+00	0.000E+00	5.279E−02	8.41E+00	0.0046	0.0026
4	3440	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.150E−03	2.947E−02	0.000E+00	0.000E+00	0.000E+00	3.062E−02	1.45E+01	0.0080	0.0009
5	3459	0.000E+00	0.000E+00	0.000E+00	0.000E+00	2.874E−04	1.245E−02	0.000E+00	0.000E+00	0.000E+00	1.274E−02	3.48E+01	0.0192	0.0055
6	5622	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.150E−03	1.179E−03	0.000E+00	0.000E+00	0.000E+00	2.328E−03	1.91E+02	0.1052	0.0227
7	3460	0.000E+00	0.000E+00	0.000E+00	0.000E+00	4.857E−04	2.487E−01	0.000E+00	0.000E+00	0.000E+00	2.492E−01	1.78E+00	0.0010	0.0008
8	3107	0.000E+00	0.000E+00	0.000E+00	0.000E+00	2.874E−04	7.368E−05	0.000E+00	0.000E+00	0.000E+00	3.611E−04	1.23E+03	0.6781	0.2246
9	3110	0.000E+00	0.000E+00	0.000E+00	0.000E+00	1.150E−03	2.129E−02	0.000E+00	0.000E+00	0.000E+00	2.244E−02	1.98E+01	0.0109	0.0053
10	21	0	0	0	0	0	0	0	0	0	0	0	0	0
											4.439E−01	1.81E+03	0.3099582	0.2990
											Total (X)	Total (Z)	Sum of F	Normalized Prediction
													Min h	6.5361

Table 10 (continued)

Example of How the GCV Prediction Calculation of Each Data Record in the TOB network can be Audited in Detail after Stage 2

	Max h	14.9934
Stage 2 Optimized Prediction of GCV for data record: #3452		12.0289
Actual Measured GCV for data record: #3452		12.6697

Appendix 1: TOB learning network method details

TOB stage 1 (data matching and provisional prediction)

Step 1 Set up a 2-D array of N input variables and one dependent variable to be predicted for each of M data records.

Step 2 Arrange the data records in a systematic order defined by the prediction variable's values (e.g. ascending or descending value order).

Step 3 Derive maximum and minimum values (and other standard statistics, such as mean and standard deviation) for all records in the dataset (Table 1).

Step 4 Normalize the data in the array so each variable spans a range from minus 1 to plus 1 (−1, +1). This is achieved by using Eq. (1)

$$X_i *= 2 * \left[\left(X_i - Xmin\right)/(Xmax - Xmin)\right] - 1 \quad (1)$$

where: X_i: variable X value for the ith data record, $Xmin$: minimum value of variable X, $Xmax$: maximum value of variable X.

Step 5 Generate statistical analysis of the normalized values to check that the variables are all correctly normalized.

Step 6 Distribute the data records between training, tuning and testing subsets. Sensitivity analysis is conducted to establish the optimum percentage of data records to allocate to each data subset. Firstly, the data records to be used for testing are extracted from the complete data set and placed to one side. Sensitivity analysis then helps to divide the remaining data records between the training and tuning subsets in proportions that achieve an acceptable prediction accuracy. For most data sets the training subset is likely to hold more than 75% of the data records. For large datasets of several thousand data records the sensitivity analysis often reveals that the training subset can be a much larger percentage without compromising prediction accuracy.

Step 7 The variable squared error (VSE) between each variable in the J data records of the tuning-data subset and the K data records in the training-data subset are calculated using Eq. (2):

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

where: $X_k(tr)$ = variable X value for the kth training-subset data record, $X_j(tu)$ = variable X value for the jth tuning-subset data record, $VSE(X)_{jk}$ = squared error value for variable X for the jth tuning-subset data record versus the kth training-subset data record.

$\sum VSE$ is then established as the sum of the VSE values for each variable for each data record match using Eq. (3):

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

where: $VSE(Xn)_{jk}$ = squared error for variable Xn for the jth tuning-subset data record versus the kth training-subset data record. $\sum VSE_{jk}$ = sum of the squared errors for all $N+1$ variables for that data record match.

Wn = weight ($0 < Wn <= 1$) applied VSE of each of the N + 1 variables involved. These weights are all set to the same values (e.g. 1) in TOB stage 1 to avoid any bias in the initial training of the prediction network.

Step 8 Select and rank (lowest in $\sum VSE$ is ranked number 1*)* the top-Q-matching data records in the training subset for each tuning subset data record. Q = 10 is typically sufficient for TOB stage 1. However, Q could be adjusted to higher or lower values, if necessary to improve prediction accuracy.

Step 9 The Q-selected training-subset data records (i.e. best matches) for the jth tuning-subset data record each contribute a fraction to the prediction of the dependent variable. That fraction is proportional to the relative $\sum VSE$ scores of those Q records for the jth data record That fraction is calculated with Eq. (4) to Eq. (6) and

$$f_q = \sum VSE_{jq} \Big/ \left[\sum_{r=1}^{r=Q} \sum VSE_{jr} \right] \tag{4}$$

where: $q = q$th top-ranking training-subset record for the jth tuning-subset data record. f_q = fractional contribution of q^{th} top-ranking records for the jth tuning-subset data record.

The constraint defined by Eq. (5) applies the sum of the f values applied to each matching data record.

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

The matching training-subset data record with the lowest $\sum VSE_{jk}$ value should contribute most to the dependent-variable prediction for the jth tuning-subset data record. To achieve this $(1-f)$ is the multiplier applied in Eq. (6) to each of the Q top-matching records.

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

Where:

$\left(X_{N+1}\right)_j^{predicted}$ = dependent variable for the qth data record in the training subset.

$\left(X_{N+1}\right)_j^{predicted}$ = Stage − 1 TOB predicted value for the dependent variable for the jth tuning-set data record.

This prediction is provisional because equal weights (Wn) are applied to the variables in TOB stage 1.

Step 10 Measures of statistical accuracy are calculated for the TOB stage 1 predictions. The measures used include: coefficient of determination (R^2); mean square error (MSE); and, root mean square error (RMSE). These are calculated with Eq. (7) to Eq. (9), respectively.

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

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

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

where: X_j = dependent variable (i.e. $\left(X_{N+1}\right)_j$ in Eq. (6)) for the jth tuning-subset data record; X_j^{actual} = actual (or directly measured) value of the dependent variable for the jth tuning-subset data record; $X_j^{predicted}$ = predicted value of the dependent variable for the jth tuning-subset data record; X_{ave}^{actual} = average actual value of the dependent variable for all J data records in the tuning subset.

TOB stage 2 (optimization)

Step 11 Optimization is performed to minimize RMSE (Eq. 9) collectively for the J data records in the tuning subset. This is achieved by adjusting optimization control metrics while applying certain constraints.

The two optimization control metrics are:

1. Varying the values applied to the N input-variable weights (W_n). Small non-zero values to weights applied to certain variables can and do have a significant impact on the accuracy of the predictions derived.
2. Varying the number (Q) of top matching records in Eqs. (4), (5) and (6). For most data sets: $2 \le Q \le 10$. The optimizer is allowed to select the best integer value of Q to minimize RMSE. It does this by systematically changing the value of Q in the three equations mentioned and by comparing the RMSE value for the predictions generated for each integer value of Q evaluated in the range $2 \le Q \le 10$. For examples, if Q is set to "4", the predictions for all of the tuning subset data records only use the top-4 matching records from the training subset related to each tuning subset record in making their predictions. In this way the optimization algorithm identified which value of Q leads to the most accurate predictions for the tuning subset as a whole.

Here, the Generalized Reduced Gradient (GRG) algorithm option of the standard "Solver" optimizer in Microsoft Excel (Frontline Solvers 2018) is used, in conjunction

with visual basic for application (VBA) code, to conduct the optimization process. Other evolutionary optimizers could be applied to achieve similar outcomes. For mid-sized dataset calculating the TOB predictions in Excel facilitates the display all the intermediate calculations in a convenient format.

The top-matching data records in the training subset for each tuning-subset data record are carried forward from TOB stage 1 for selection by TOB stage 2. Equation (3) is re-evaluated by varying W_n in each iteration of the optimizer. Additionally, TOB stage-2 $\sum VSE_{jq}$ scores are derived with Eq. (4) by varying Q $(2 < Q \leq 10)$ in each iteration of the optimizer, contrasting with the fixed value of Q used in TOB stage 1.

Step 12 Calculate TOB stage-2 *RMSE* and R^2 values for the predictions provided by the optimum step 11 solution. Compare the TOB stage-2 predictions with the TOB stage-1 predictions to assess the prediction improvements achieved, if any. Running sensitivity analysis with different values of Q (i.e. Q = 2 to 10) often provides insight to potential underfitting or overfitting issues with the data set.

Step 13 Calculate TOB stage-1 and stage-2 predictions for the independent testing data subset using the optimum values established for *Wn* and Q in step 11. Calculate and evaluate the *RMSE* and R^2 values for the predictions calculated for the testing data. Reviewing the intermediate steps in the calculations often provides useful insight to the variables that have the most influence on prediction accuracy (it is often not those with the highest *Wn* values). It also helps perform outlier analysis (i.e., understanding why some data records lead to less-accurate predictions).

Step 14 Consider whether the prediction accuracy achieved by the method is sufficiently meaningful for it to be relied upon. Also, evaluate how its prediction accuracy compares with other machine-learning tools.

Appendix 2: Details of data records in the dataset

Supplementary data associated with the coal proximate and ultimate analysis data set to which the TOB network is applied (Matin and Chelgani 2016; Bragg et al. 1997) can be found, in the online version. The data in the supplementary Excel file is listed in one sheet as the complete dataset (6339 data records) and another with those records sorted in ascending order of GCV. To further aid transparency, other sheets in that file list the actual data records assigned to the training, tuning and testing subsets used for the analysis presented. This enables readers to view exactly how the TOB network was configured for the analysis described.

References

Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. Artif Intell Rev 11(1–5):11–73, 1997

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

Baghericeh AH, Hower JC, Baghericeh AR, Jorjani E (2008) Studies of the relationship between petrography and grindability for Kentucky coals using artificial neural network. Int J Coal Geol 73:130–138

Birattari M, Bontempi G, Bersini H (1999) Lazy learning meets the recursive least squares algorithm. Advances in neural information processing systems, vol 11. MIT Press, Cambridge, pp 375–381

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

Bragg LJ, Oman JK, Tewalt SJ, Oman CJ, Rega NH, Washington PM, Finkelman RB (1997) US geological survey coal quality (COALQUAL) database: version 2.0. US geological survey open-file report 97–134. https://pubs.er.usgs.gov/publication/ofr97134. Accessed 15 Nov 2018

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Channiwala SA, Parikh PP (2002) A unified correlation for estimating HHV of solid, liquid and gaseous fuels. Fuel 81:1051–1063

Chelgani SC, Mesroghli SH, Hower JC (2010) Simultaneous prediction of coal rank parameters based on ultimate analysis using regression and artificial neural network. Int J Coal Geol 83:31–34

Chelgani SC, Hart B, Grady WC, Hower JC (2011) Study relationship between inorganic and organic coal analysis with gross calorific value by multiple regression and ANFIS. Int J Coal Prep Util 31:9–19

Chen GH, Shah D (2018) Explaining the success of nearest neighbor methods in prediction. Found Trends R Mach Learn 10(5–6):337–588

Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27

Elkatatny S, Tariq Z, Mahmoud M (2016) Real time prediction of drilling fluid rheological properties using artificial neural networks visible mathematical model (whitebox). J Pet Sci Eng 146:1202–1210

Feng Q, Zhang J, Zhang X, Wen S (2015) Proximate analysis-based prediction of gross calorific value of coals: a comparison of support vector machine, alternating conditional expectation and artificial neural network. Fuel Process Technol 129:120–129

Fix E, Hodges JL Jr (1951) Discriminatory analysis, nonparametric discrimination: consistency properties. Technical report, USAF School of Aviation Medicine

Frontline Solvers (2018) Standard Excel solver—limitations of nonlinear optimization. https://www.solver.com/standard-excel-solver-limitations-nonlinear-optimization. Accessed May 2018

Garcia S, Derrac J, Cano J, Herrera F (2012) Prototype selection for nearest neighbor classification: taxonomy and empirical study. IEEE Trans Pattern Anal Mach Intell 34(3):417–435

Given PH, Weldon D, Zoeller JH (1986) Calculation of calorific values of coals from ultimate analyses: theoretical basis and geochemical implications. Fuel 65:849–854

Heinert M (2008) Artificial neural networks – how to open the black boxes? In: Reiterer A, Egly U (eds) Application of artificial intelligence in engineering geodesy. Proceedings of AIEG Vienna, Vienna, pp 42–62 (ISBN 978-3-9501492-4-1)

Kavšek D, Bednárová A, Biro M, Kranvogl R, Vončina DB, Beinrohr E (2013) Characterization of Slovenian coal and estimation of coal heating value based on proximate analysis using regression and artificial neural networks. Cent Eur J Chem 11(9):1481–1491

Lever J, Krywinski M, Altman N (2016) Model selection and overfitting. Nat Methods 13:703–704. https://doi.org/10.1038/nmeth.3968 **(Published online)**

Majumder AK, Jain R, Banerjee P, Barnwal JP (2008) Development of a new proximate analysis based correlation to predict calorific value of coal. Fuel 87:3077–3081

Mathews JP, Krishnamoorthy V, Louw E, Tchapda AH, Castro-Marcano F, Karri V, Alexis DA, Mitchell GD (2014) A review of the correlations of coal properties with elemental composition. Fuel Process Technol 121:104–113

Matin SS, Chelgani SC (2016) Estimation of coal gross calorific value based on various analyses by random forest method. Fuel 177:274–278

Mesroghli S, Jorjani E, Chelgani SC (2009) Estimation of gross calorific value based on coal analysis using regression and artificial neural networks. Int J Coal Geol 79:49–54

Neavel RC, Smith SE, Hippo EJ, Miller RN (1986) Interrelationships between coal compositional parameters. Fuel 65:312–320

Palmer CA, Oman CL, Park AJ, Luppens JA (2015) The US Geological Survey coal quality (COALQUAL) database version 3.0: US geological survey data series 975, p 43 with appendixes. https://doi.org/10.3133/ds975

Patel SU, Jeevan Kumar B, Badhe YP, Sharma BK, Saha S, Biswas S, Chaudhury A, Tambe SS, Kulkarni BD (2007) Estimation of gross calorific value of coals using artificial neural networks. Fuel 86:334–344

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

Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117

Shakhnarovich G, Darrell T, Indyk P (2006) Nearest-neighbor methods in learning and vision: theory and practice (neural information processing). The MIT Press, Cambridge (ISBN 026219547X)

Singh KP, Kakati MC (1994) New models for prediction of specific energy of coal. Fuel 73:301–303

Tan P, Zhang C, Xia J, Fang Q-Y, Chen G (2015) Estimation of higher heating value of coal based on proximate analysis using support vector regression. Fuel Process Technol 138:298–304

Trimble AS, Hower JC (2003) Studies of the relationship between coal petrology and grinding properties. Int J Coal Geol 54:253–260

Wood DA (2018) 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 2(2):148–162

Yalcin Erik N, Yilmaz I (2011) On the use of conventional and soft computing models for prediction of gross calorific value (GCV) of coal. Int J Coal Prep Util 31(1):32–59