A new method for TOC estimation in tight shale gas reservoirs

Hongyan Yu[1,2], Reza Rezaee[3], Zhenliang Wang[1], Tongcheng Han[4], Yihuai Zhang[3], Muhammad Arif[3], Lukman Johnson[3]

[1]State Key Laboratory of Continental Dynamics, Department of Geology , Northwest University, Xi'an, 710069 China.

[2]Research Institute of BGP, CNPC, Zhuozhou 072750, China.

[3]Department of Petroleum Engineering, Curtin University, 26 Dick Perry Avenue, 6151 Kensington, Australia.

[4] China University of Petroleum (East China), School of Geoscience,Qingdao 266580, China

Corresponding author: Hongyan Yu, email: amelia-yu@hotmail.com

Abstract

Total organic carbon (TOC) estimation is significantly crucial for shale reservoir characterization. Traditional TOC estimation methods (such as Passey and Schmoker method) do not provide accurate TOC predictions in shale gas reservoirs especially for the self-generated and self-stored reservoirs. This study proposes, for the first time, a new TOC prediction method based on Gaussian Process Regression (GPR) bridging geostatistics and machine learning technique. The method utilizes a non-parametric regression approach in shale TOC predictions, and not only provides the expert solutions in high-dimension processing, small samples and non-linear problems, but also has a better adaptation and generalization ability compared with other machine learning methods. The approach accounts for all the well logging attributes and chooses the relevant logs to build TOC estimation model, and 7 different kernel functions and 5 attributes groups are analyzed to get the optimized hyperparameters in practice. Application of the developed model to two shale gas reservoirs showed that the model predicted TOC matched well with that from the laboratory measurements. The proposed model based on GPR method provides an accurate way for the TOC prediction in the tight shale gas reservoirs.

Key words: Total organic carbon; Gaussian Process Regression; shale gas; wireline logs; machine learning.

1. Introduction

Tight shale gas/oil has gained massive attention all over the world in the past decades as supplement energy in the energy shortages (Jarvie 2012, Jarvie et al. 2007). Total Organic Carbon (TOC) is one of the crucial parameters in shale gas reservoir assessments and is also regarded as one of the key variables that directly influences the rock quality, hydraulic fracturing design, and shale gas-in-place estimations (Passey et al. 2010, Sondergeld et al. 2010). As a part of the rock matrix, organic matter also strongly affects the geomechanical properties of shales (Altowairqi et al. 2015). In addition, organic carbon content and maturity are important factors impacting organic porosity, and controlling the absorbed gas in the shale gas reservoirs (most of the absorbed gas is occurred on organic matter) (Montgomery et al. 2005, Ross and Bustin 2007). Moreover, TOC controls the micro structure, texture, porosity, permeability and wettability of the shale reservoirs (Altowairqi et al. 2015, Sondergeld et al. 2010, Sone and Zoback 2013, Zhang et al. 2012). Thus, a reliable method for the characterization of shale organic matter and for the accurate prediction of TOC is crucial for hydrocarbon exploration and production from these unconventional reservoirs (Ding et al. 2015, Wang et al. 2016).

Generally, the presence of organic matter could be identified from well logging and several methods have been proposed in the former study for TOC calculation. For instance, Schmoker and Hester (1983) established a model which calculates TOC using the reciprocal of bulk density. Although the method requires small amount of input data, it may not work well in the situations when the bulk density is affected by reservoir/geological parameters (Schmoker 1979, Schmoker and Hester 1983). Passey et al. (1990) proposed a TOC estimation model based on porosity logs (e.g. sonic, neutron and density) and resistivity log (Passey et al. 1990). This method is relatively universal and can be used once the representative baselines for the logs are properly defined. However, the associated shortcoming is that the log-baseline may significantly vary from well-to-well and also across the formations and different depositional environments. Fortunately, such limitations nowadays could be overcome by machine learning approaches. The Machine learning approach, such as Neural Network method has been applied for shale TOC estimation (Alizadeh et al. 2012, Khoshnoodkia et al. 2011, Tan et al. 2015). The method utilizes laboratory measurement of shale organic matter, and combines the measured TOC with well logs, followed by the data calibration for machine learning. Recently, Tan et al. (2015) used Support Vector Regression Machine approach to estimate TOC in various organic shales using a variety of Kernel Functions. However, Neural Network ignores the generalization and always results in overfitting. Moreover, the possible functions of prior probability of Support Vector Regression Machine are an unlimited dataset, which will cost long time for training and calculating.

Thus, in this paper we propose a new method based on Gaussian Process Regression (GPR), which utilizes a non-parametric regression approach in shale TOC predictions. Such method can provide expert solutions in high-dimension processing, small samples, and non-linear problems, and has a better adaptation and generalization ability. We first developed a workflow for the GPR to estimate TOC. In order to find the best way to estimate TOC, we tested a total of 7 kernel functions for 5 attribute groups to obtain the optimized function and attributes in the machine training. The method was then applied to predicting TOC in two different shale gas reservoirs (one with high TOC shale from Ordos basin, China; and the other with low TOC shale from Canning Basin, Western Australia). The obtained TOC was finally compared with the results from the traditional methods to show the effectiveness of our present method.

2. Background

2.1. The traditional TOC estimation methods

2.1.1. Schmoker's method

Schmoker and Hester (1983) proposed that TOC has a positive linear correlation with the reciprocal of bulk density (Schmoker 1979, Schmoker and Hester 1983):

$$TOC = (A \times \frac{1}{\rho}) - B \qquad (1)$$

The values of *A* and *B* are calculated based on the organic matter density, matrix density and the ratio of weight percentages of organic matter to organic carbon. Hence, equation (1) is simplified as:

$$TOC = (154.497 \times \frac{1}{\rho}) - 57.261 \qquad (2)$$

2.1.2. Modified Schmoker's method

The constants *A* and *B* in the traditional Schmoker's method (equation 1) are sometimes not suitable for some basins. Therefore, the constants are usually obtained from the linear regression of TOC test and density log, with the equation given as:

$$TOC = (A^* \times \frac{1}{\rho}) - B^* \qquad (3)$$

2.1.3. Passey's method

Passey et al (1990) developed a practical method which uses the overplaying of sonic logs and deep resistivity log in a proper scale. They suggested the logs should be overlain in the water-saturated and organic lean interval, and this overlain is defined as the baseline. If the organic matter is present, a separation from the two curves will be observed. Thus, the separation can be calculated as follows:

$$D\log \mathrm{R} = lg\left(RD / RD_{baseline}\right) + 0.02 \times \left(\Delta t\text{-}\Delta t_{baseline}\right) \qquad (4)$$

Then, TOC can be calculated using the equation below:

$$TOC = D\log R \times 10^{2.297\text{-}0.1688LOM} \qquad (5)$$

Where, *RD* is the deep resistivity of rock and RD_{baseline} is the deep resistivity of baseline, Ω.m; and $\triangle t$, $\triangle t_{baseline}$ is the transit time of rock and baseline respectively, us/m; LOM is the level of organic maturity.

2.2. Gaussian Process Regression (GPR) method

2.2.1. Principle

The Machine learning tools are now popular in the petroleum exploration and production (Ahmadi et al. 2014, Al-Anazi and Gates 2010, Hasebe and Nagayama 2002, Kuo et al. 2007, Lukoševičius and Jaeger 2009, Rasmussen 2006, Witten and Frank 2005). The Machine learning method can be based on either of the two possible processes (Chen et al. 2005, Hammen 2003, Kotsiantis et al. 2007, Michalski et al. 2013): (1) Parametric regression, which is based on the determination of a suitable set of parameters that can show the mapping (like Polynomial Regression and Neural Network), and (2) the Bayesian Regression, which defines one function distribution and gives a prior probability to each possible function to compensate for the first approach that ignores the generalization and thus results in overfitting. However, these possible functions in Bayesian Regression are an unlimited dataset, e.g. the infinite possible functions lead to a new question – how to select the possible function in a limited time? For the functions selection in such unlimited dataset, GPR is the best choice (Rasmussen, 2006; Silversides and Melkumyan, 2016).

Gaussian Process (GP) theory is well established for predictions in various research areas including reservoir engineering (Silversides and Melkumyan 2016), electric engineering (Yuan *et al.*, 2008), and Spectroscopic (Chen et al. 2007). GP is very powerful and leverages on many convenient properties of the Gaussian distribution to enable tractable inference. Gaussian Process Regression (GPR) is a professional regression method in processing high-dimension, small samples, and non-linear problems (Dudley 2010). Compared to Neural Network and Support Vector Machine, GPR is simple to implement, and is flexible and fully probabilistic using hyperparameters, and hence, has higher adaptation and generalization ability (Bonilla et al. 2008, Datta et al. 2016, Lawrence 2004, Tonner et al. 2017, Wang et al. 2008).

Mathematically, GPR is a collection of random variables. Any finite number of variables have a joint Gaussian distribution. GPR is completely specified by a mean function and a positive definite covariance function (Paciorek and Schervish 2004).

For a given set of inputs: $D=\left\{\left(x_i,\ y_i\right),\ i=1,\ 2,\ \ldots,n\right\}, x_i \in R^d$ and $y_i \in R$.

The mean function is given by:

$$m(x)=E\left[f(x)\right] \tag{6}$$

The covariance function is given by:

$$k(x,x^{'})=E\left[(f(x)-m(x))(f(x^{'})-m(x^{'}))\right] \text{ where } x,x^{'} \in R^d \tag{7}$$

We need to predict $f(x_*)$ for the test data x_*, first, the process is defined as:

$$f(x) \sim GP\left[m(x),k(x,x^{'})\right] \tag{8}$$

For the regression problem, the model is as follows:

$$y=f(x)+\xi \tag{9}$$

Once hypothesized that the noise $\xi \sim N(0,\sigma_{noise}^2)$ where, σ_{noise}^2 is the variance of the noise.

As distribution, we defined μ and Σ,

$$\mu_i = m(x_i) \tag{10}$$

$$\Sigma_{ij} = k(x_i,x_j) \tag{11}$$

Hence, we get priori distribution of observed value y as:

$$y \sim N\left(\mu,\Sigma+\sigma^2 I\right) \tag{12}$$

Where, I is the identify matrix.

And the combination priori distribution of noisy y and predicted $f(x_*)$ is:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} (\Sigma+\sigma^2 I)\Sigma_* \\ \Sigma_*^{\mathrm{T}} & \Sigma_{**} \end{bmatrix}\right) \tag{13}$$

Where: $\mu = m(x_i)$, i=1, 2, ... , n for the training means; μ_*, the test means; Σ, the covariances for training set; Σ_*, training-test set covariances; Σ_{**}, test set covariances.

Then the Posteriori Distribution of predicted value $f(x_*)$ is obtained as follows:

$$f_* \mid y \sim N\left(\mu_* + \Sigma_*^{\mathrm{T}}(\Sigma^{-1}+\sigma^2 I)(y-\mu), \Sigma_{**} - \Sigma_*^{\mathrm{T}}(\Sigma+\sigma^2 I)^{-1}\Sigma_*\right) \tag{14}$$

μ_*, Σ_* is the mean and covariance of $f(x_*)$.All above equations are the main equations for Gaussian Process Regression prediction.

2.2.2. Kernel Function Selection

GPR is parameterized by a mean function and a Kernel (covariance) Function. The Kernel Functions are powerful tools which control the algorithm of GPR's accuracy. They provide a bridge to manipulate data as though they are projected into a higher dimensional space, instead of operating on their original space (Mierswa and Morik 2005). The function transfers from linearity to non-linearity for algorithms which can be expressed in terms of dot products between two vectors (Sahami and Heilman 2006). Some sets of data hardly build regression in the lower dimensional space with linear algorithm (Figure 1a), while, it is very easy to build regression model in the higher dimensional space with non-linear algorithm (Figure 1b). Hence, Kernel functions make the regression more efficient during the model building.

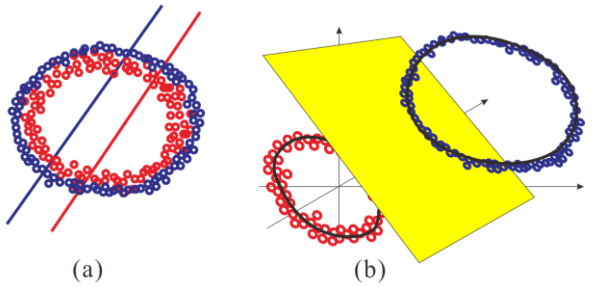


Figure 1. Schematic of the kernel functions: (a) Original space; (b) Projected space (higher dimensional).

3. Material and Methodology

3.1. Characteristics of the shale formations used

Two shale formations were considered for TOC estimation in this work: 1) High TOC content (4 wt% to 10 wt%, average is 6 wt%) shale gas reservoir in Yanchang Formation of Ordos Basin at northwest of China (see Figure 2a), and 2) Low TOC (0.1 wt% to 4 wt%, average is 0.7 wt%) shale gas reservoir of Goldwyer Formation of Canning Basin in Western Australia (see Figure 2b). The locations of the basins are marked in the dotted rectangle in Figure 2. These two shale gas reservoirs are significantly different in organic matter characteristic showing decent range of TOC variation (Figure 3). The important geological features of the two formations are described in Table 1.

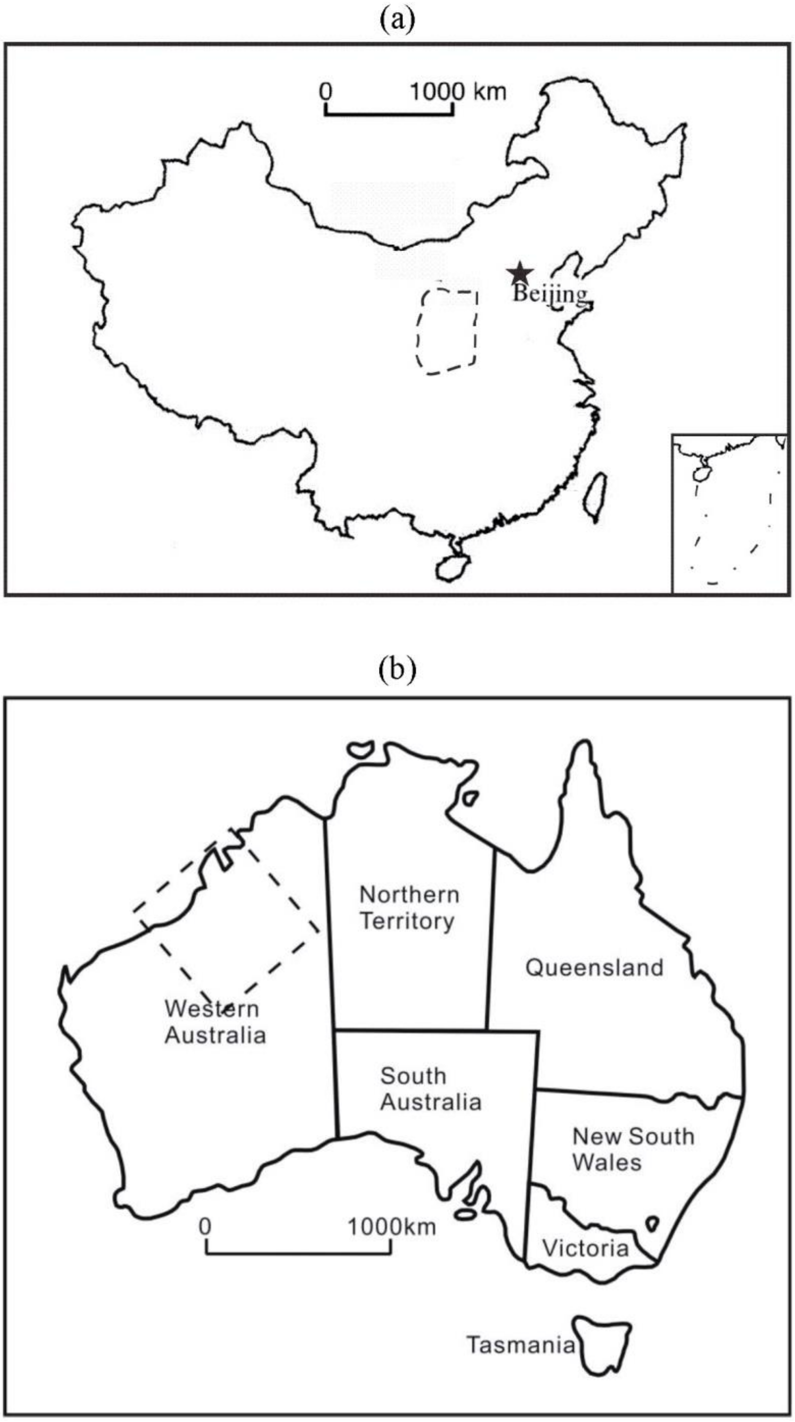


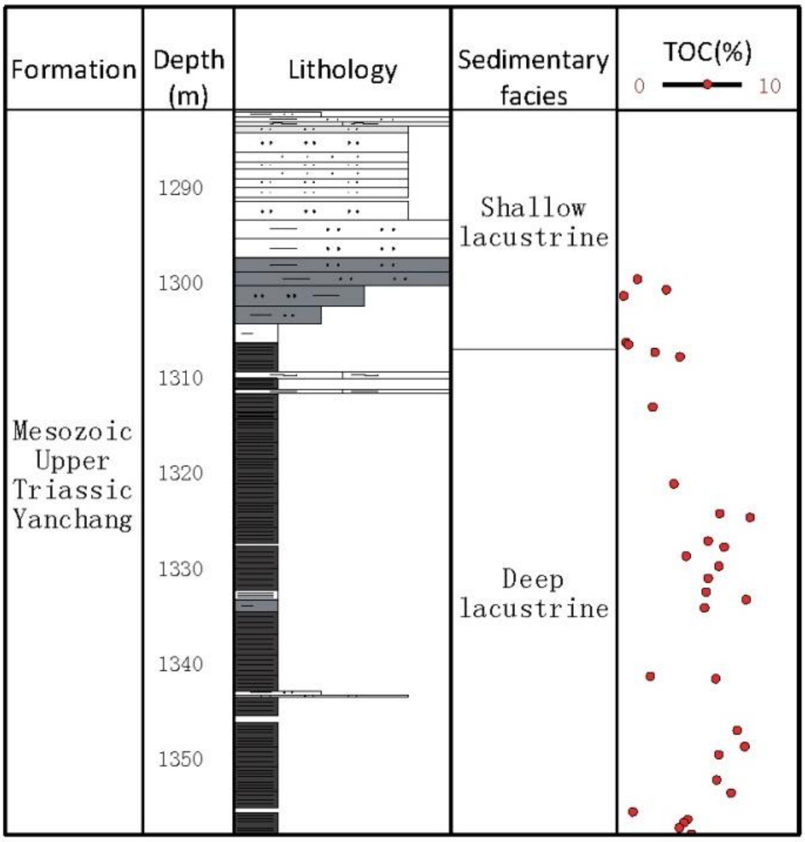
Figure 2. Research area location map: (a) Ordos Basin China (high TOC shale formation); (b) Canning Basin, Western Australia (low TOC shale formation)

Table 1. Geological features of the two shale formations considered.

Feature	Yanchang Formation, China	Goldwyer Formation, WA
Basin name	Ordos basin	Canning basin
Deposit environment	Deep lake or Semi-Deep	shallow marine and

	lake	subtidal
System	Triassic	Ordovician
Biological reproduction	High	Fair
Organic matter content	Organic rich	Low to fair

(a)



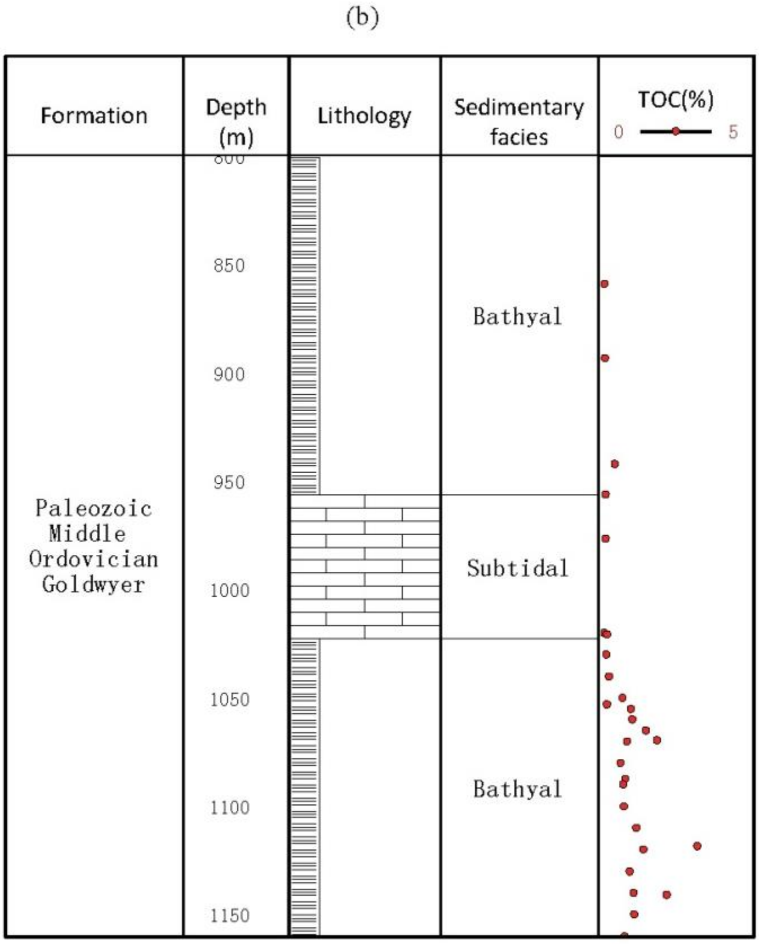


Figure 3. Stratigraphy for the research areas: (a) Ordos basin shale with TOC (wt %); (b) Canning basin shale with TOC (wt %).

3.2. Gaussian Process Regression flow chart

We developed a work flow for Gaussian Process Regression, involving six basic steps as shown in Figure 4.

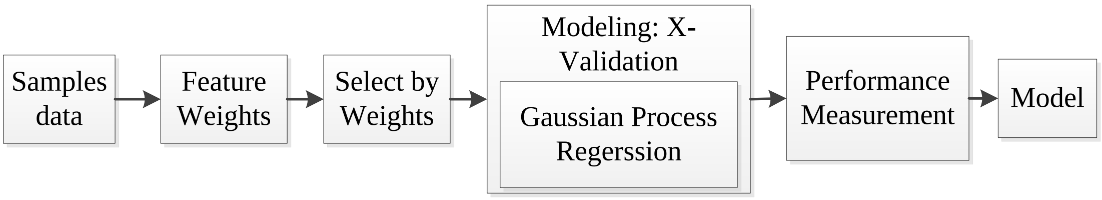


Figure 4. Gaussian Process Regression flow chart.

(1) The first step is the data collection which includes the well-log responses and TOC

pyrolysis experiment data.

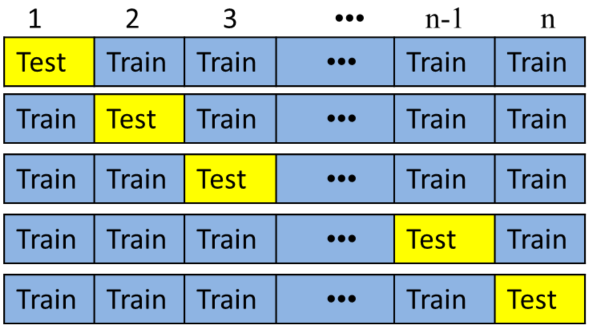
(2) The next step, seen as ‘Feature Weights’, computes the weights of the wireline logs relative to the core derived TOC – based on the weighting algorithms. The quality of features has a significant impact on the performance of a learning algorithm for the regression tasks. For Machine learning, adding, cutting or exchanging unlabeled (regular) attributes will alter the structure of the dataset (Miller 2002). The accuracy can be reduced if there are irrelevant or redundant features in the training dataset. Hence, different subsets will establish different models. We select 4 types of feature weighting algorithms including: Correlation, Support Vector Machine (SVM), Principal Component Analysis (PCA), and Deviation weighting algorithm in order to get a more accurate subset.

(3) Next, we select the relevant attributes using the cut-threshold (set as 0.1) in the weights ranking which is calculated in step 2 for future regression. I. e., if the weight of the attribute is higher than 0.1, then this attribute will be selected in the regression modeling.

(4) The fourth step is ‘modeling’ where GPR is building model and X-validation is used to train and test the model. X-validation (cross-validation), which is a model evaluation method, is also important during the modeling (Picard and Cook, 1984). Without using X-validation, the model would have a perfect score with the training data but would fail to predict other data which it hasn’t seen before. For validation, the newest and most common approach is “Leave-One-Out Cross Validation (LOO CV)” which is a special case of k-fold cross validation (Rodriguez et al. 2010, Triba et al. 2015). LOO CV uses all the samples except the one in creating the model, and the remaining one is employed for testing. For a ‘*n* points’ dataset, each of the data should be used to test the performance of the learned model on the new data once (Figure 5). Then, *n* different training sets are generated. Thus, the GPR learner can learn *n* times from different type of datasets to correct the regression models.

(5) ‘Performance Measurement’ tests the accuracy and error of the model.

(6) In the final step ‘model formation’, the model is established.



1	2	3	•••	n-1	n
Test	Train	Train	•••	Train	Train
Train	Test	Train	•••	Train	Train
Train	Train	Test	•••	Train	Train
Train	Train	Train	•••	Test	Train
Train	Train	Train	•••	Train	Test

Figure 5. Schematic of Leave-one-out cross-validation.

Usually, different Kernel functions provide different transferred ways from low dimensional space to high dimensional space. Also, the distribution of different class samples is different; therefore it is uncertain to use a single scale kernel function for all samples. The performance of kernel method will be improved by selecting different scale kernel functions based on the sample features. In this study, a number of 7 Kernel functions (as shown in Table 2) were chosen in order to get a more accurate model. These Kernel functions are major functions for continuous target data.

Table 2. Kernel functions used in this work.

Kernel function title	Equation	Equation number
Exponential (RBF)	$K(x,y)=\exp(-\frac{\lVert x-y\rVert}{2\sigma^2})$	(15)
Cauchy Kernel	$K(x,y)=\frac{1}{1+\frac{\lVert x-y\rVert^2}{2\sigma^2}}$	(16)

Laplace Kernel	$K(x,y)=\exp(-\frac{\lVert x-y\rVert}{\sigma})$	(17)
Polynomial Kernel	$K(x,y)=(\alpha x^T y+c)^d$	(18)
Sigmoid Kernel	$K(x,y)=\tan(\alpha x^T y+c)$	(19)
Gaussian Kernel	$K(x,y)=\exp(-\frac{\lVert x-y\rVert^2}{2\sigma^2})$	(20)
Multiquadric Kernel	$K(x,y)=\sqrt{\lVert x-y\rVert^2+c^2}$	(21)

4. Results and discussion

4.1. TOC computation by conventional methods

4.1.1. Schmoker and modified Schmoker methods

The Schmoker method, equation (2), was used to generate continuous TOC data which were validated with Rock Eval analysis data from two research fields respectively. For the high TOC gas shale of the Ordos basin, TOC from core test and from the model prediction are in general good fit with each other (as shown in Figure 6) with the root mean square error (RMSE) of 1.5472. The greater predicted TOC than the core results is basically due to the fact that the variables *A* and *B* are not suitable for the Ordos basin. Hence, we recalculated A^* and B^* with the tested TOC and well logging data using linear regression method. Then, the equation was modified as following:

$$TOC=(133.44\times\frac{1}{\rho})-49.679 \tag{22}$$

Equation (22) was also applied to the Ordos basin for predicting TOC. The resulted RMSE (with the value of 1.1163) was significantly improved compared to the original Schmoker equation (Figure 7).

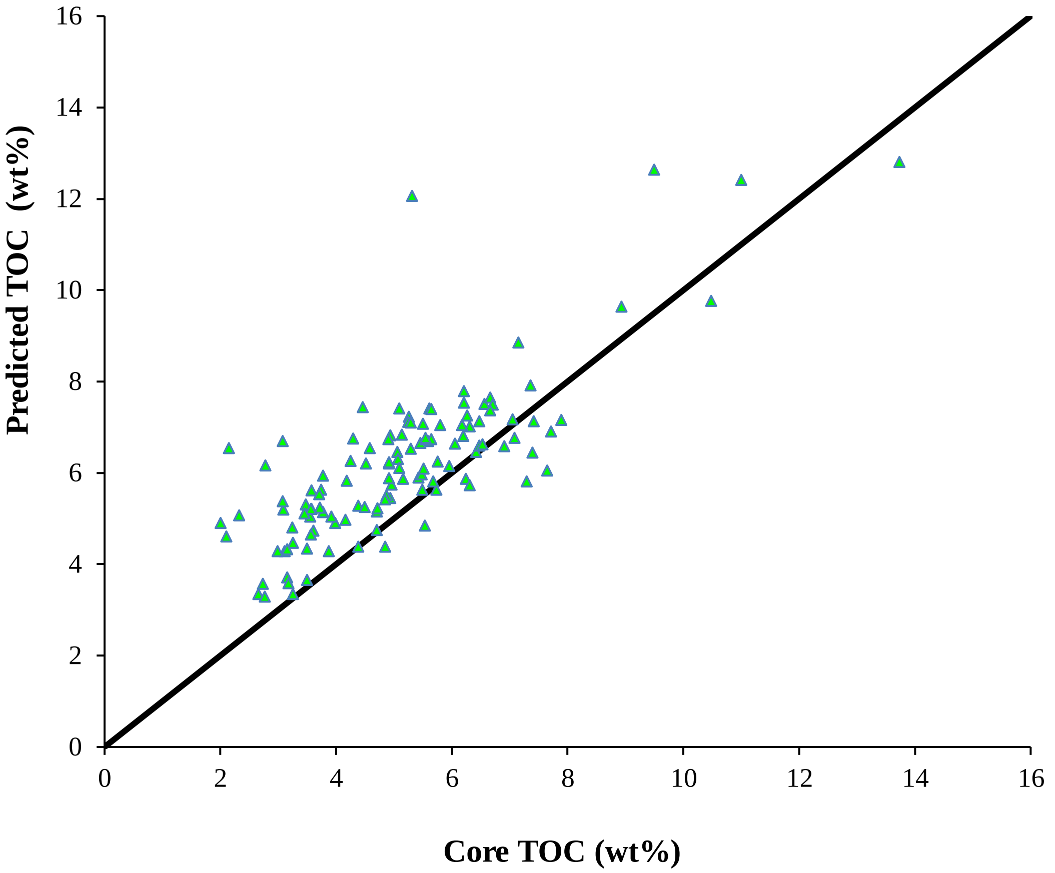


Figure 6. Correlation between core-derived TOC and original Schmoker prediction for Ordos basin with RMSE value of 1.5472.

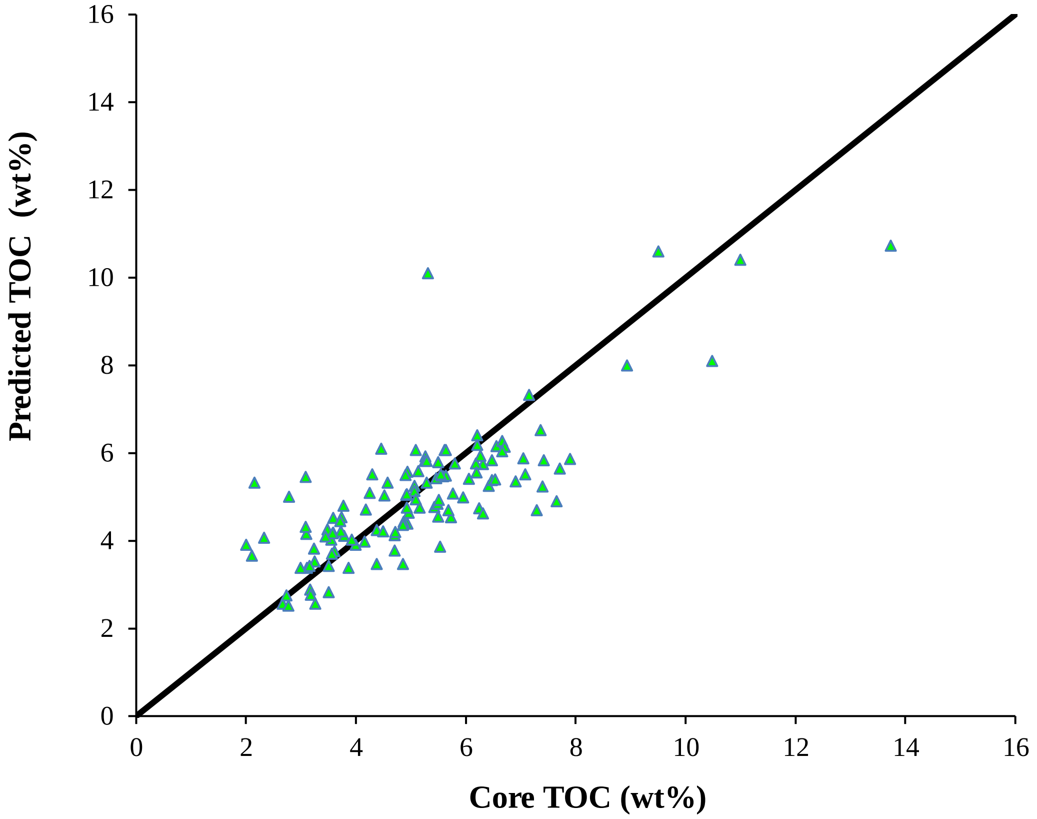


Figure 7. Correlation between core-derived TOC and modified Schmoker prediction for Ordos basin with RMSE value of 1.1163.

We also calculated TOC using Schmoker's method in the low TOC gas shale of Canning basin. The results are show in Figure 8, where the predicted TOC is significantly different from the laboratory-core derived TOC with poor RMSE of 7.8, thus such Schmoker's method is not suitable for TOC estimation in low TOC shales of the Canning Basin. The possible reason for this could be that the lower organic matter is not usually captured by the density logs. The relationship between the core derived TOC and the reciprocal of density was found to be not very clear. And even worse we could not re-calculate A^* and B* with the TOC core data and well logging data because of their poor correlation.

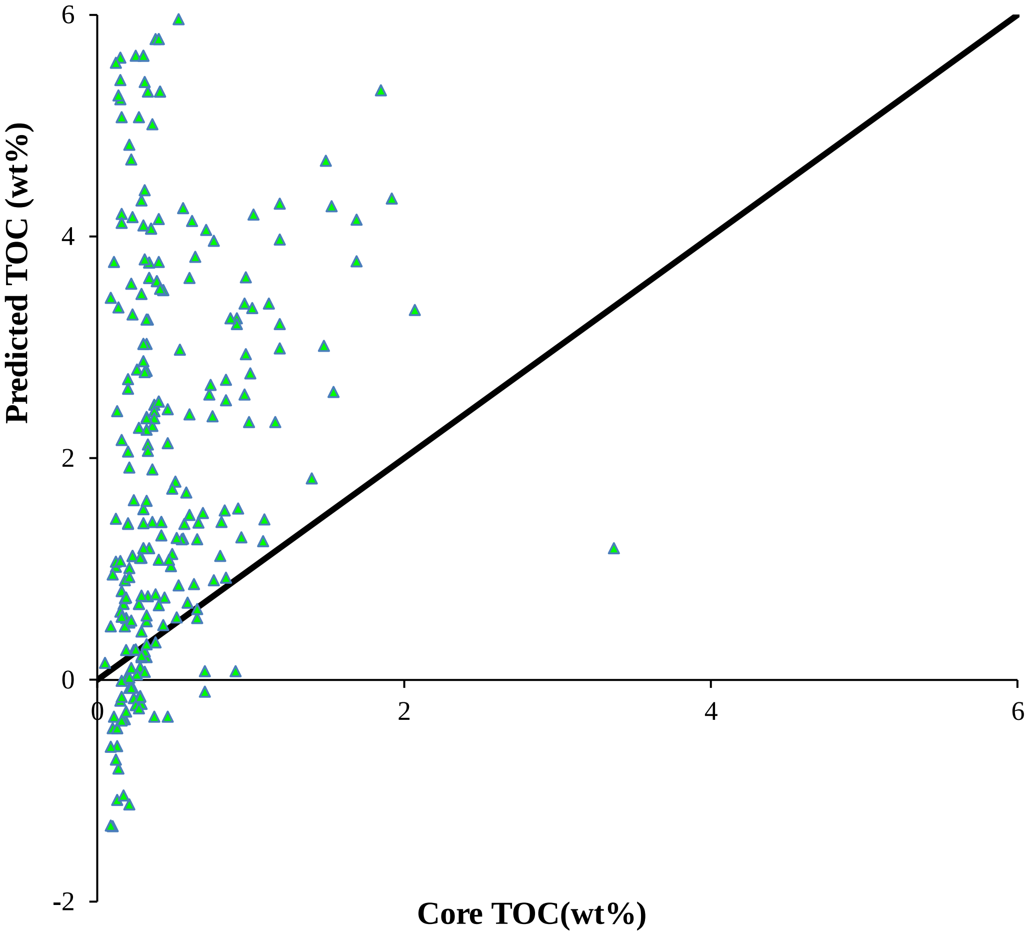


Figure 8. Correlation between core-derived TOC and Schmoker's prediction for Canning basin with RMSE value of 7.8.

4.1.2. Passey method

Passey method has more widely selected parameters than Schmoker method (Bolandi et al. 2017, Kim et al. 2017, Zhao et al. 2016) Firstly, we defined the baseline of resistivity and sonic logs, then we got *D*LogR from equation (4). Here, LOM (level of organic maturity) in equation (5) is a parameter about maturity which can be obtained from the TOC-DlogR plot. The average of LOM in the high TOC shale of Ordos basin was 10.5 (Figure 9). The calculated TOC using Passey method is in RMSE of 1.0579 with the core data (Figure 10), which shows a little bit improved accuracy compared to the Schmoker's method.

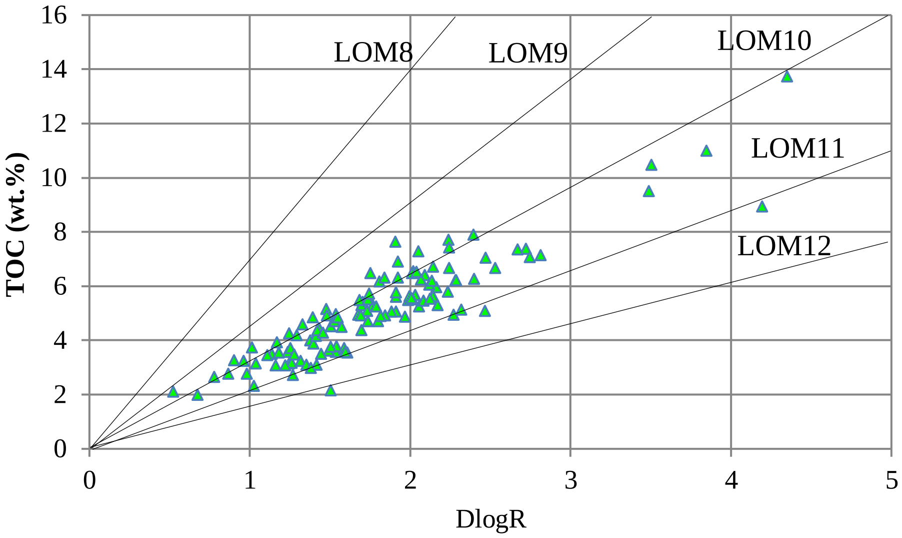


Figure 9. Relation between TOC and DlogR (LOM Chart) for Ordos basin.

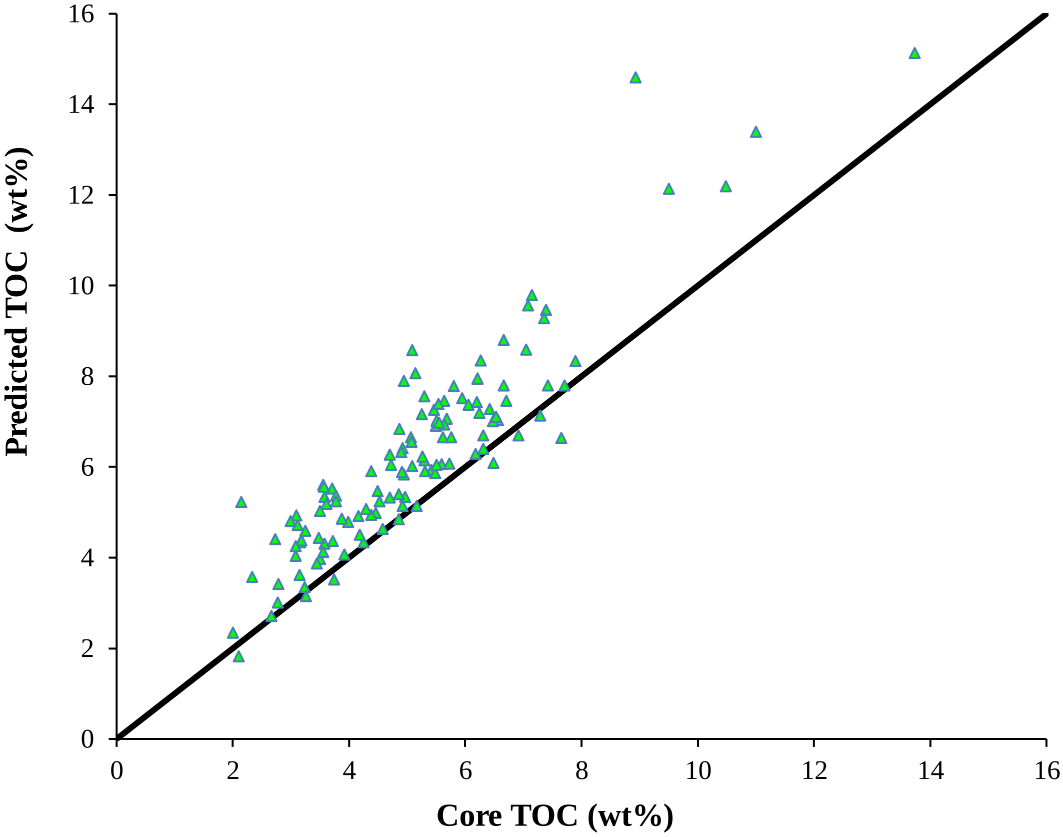


Figure 10. Correlation between core-derived TOC and Passey prediction for Ordos basin with RMSE value of 1.0579.

For the low TOC shale (Canning basin), the plot of TOC content with DlogR fromthe

Passey method is shown in Figure 11, with the average LOM of 13 for the Canning basin. The results of the predicted TOC from equation (5) versus that from the core measurement are given in the Figure 12. Although the correlation between the predicted and the core derived TOC is better than the corresponding results from the Schmoker method as shown in Figure 8 for Canning basin, yet the match is still not good enough, with RMSE of 1.37.

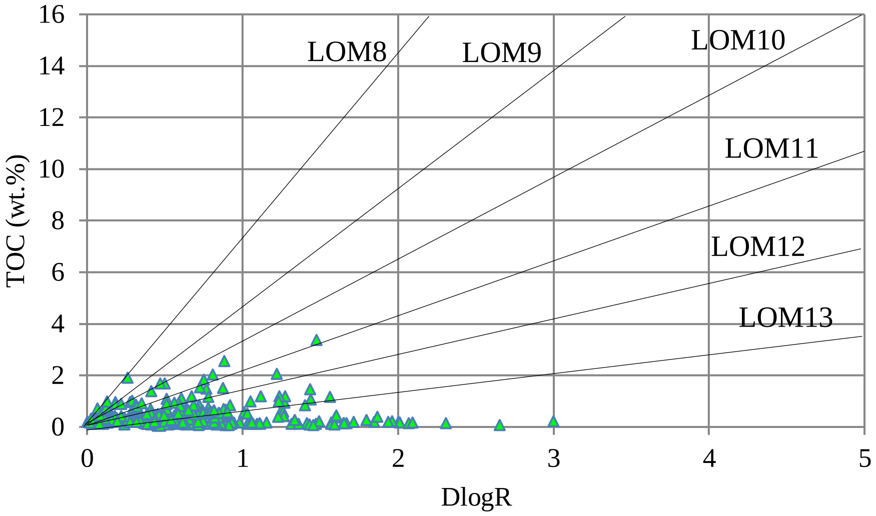


Figure 11. Relationship between TOC and DlogR (LOM chart) for Canning basin

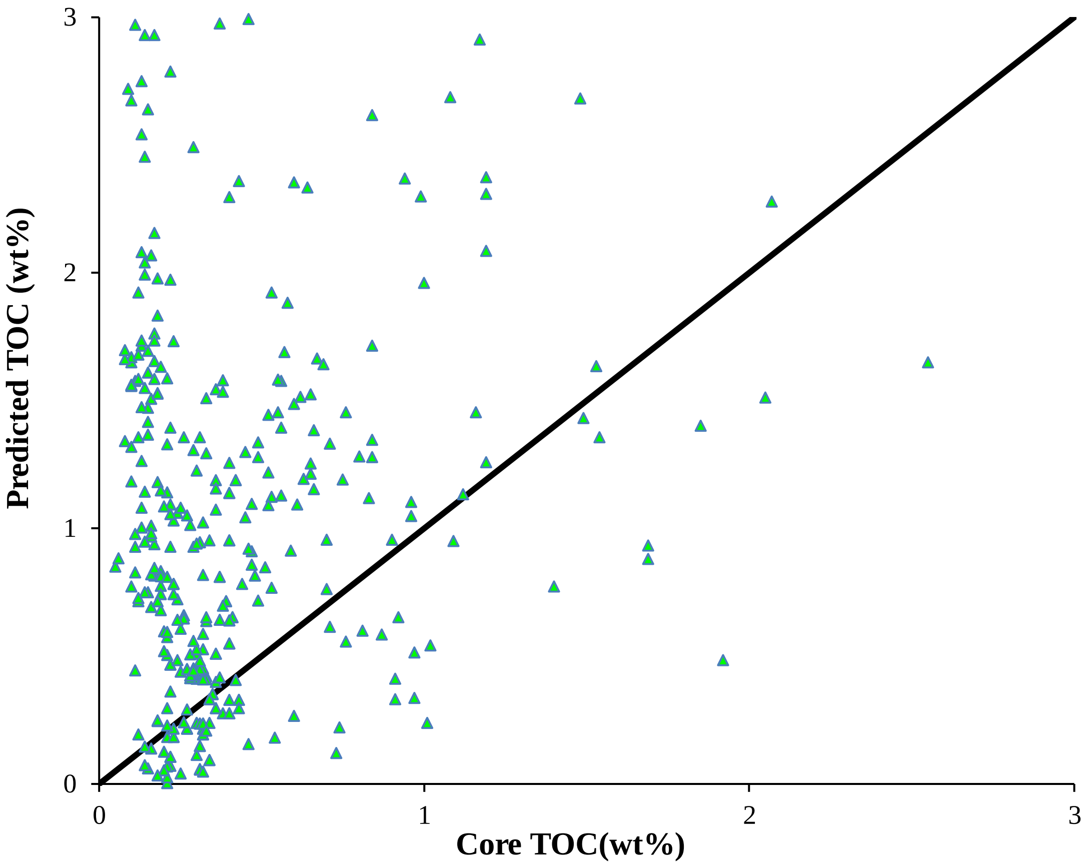


Figure 12. Correlation between core-derived TOC and Passey prediction for Canning basin with RMSE value of 1.37.

4.2. TOC computation by Gaussian Process Regression method

To calculate TOC using the GPR method, the first step is the training Dataset Preparation. A total of 10 types of well logging data were chosen in the low TOC shale formation of the Ordos basin, including natural gamma-ray (GR), spectrum gamma-ray logs (Uranium (U), Thorium (Th), and Potassium (K)), K-Th (the uranium-free gamma ray curve), sonic log (DT), density log (RHOB), photoelectric absorption factor (PEF), neutron log (NPHI), and the deep resistivity (RD). According to the weights calculated by the 4 algorithms mentioned above, 4 attributes groups are generated based on a cut-threshold of 0.1 (Table 3). Hence, 5 attributes groups including the group with all of the attributes were trained by the GPR flow (as shown in Figure 4) with the 7 kernel functions. Table 4 shows the RMSE of the final regression model performed on the training data, from which, we can see that the subset set up by PCA algorithm has the highest accuracy among all the 5 groups, in the meanwhile, Cauchy Kernel function has better accuracy compared to the other 6 Kernel Functions. Therefore, we found that the subset chosen by PCA weights with Cauchy Kernel demonstrated a best performance, with an RMSE of 0.344. As a consequence, the correlation between the predicted TOC and the Rock Eval TOC values was significantly improved when compared to the traditional methods (Figure 13).

Table 3. Attributes selection based on weight algorithms for Ordos basin.

Weights algorithm	Attributes groups
ALL	GR, DT, RHOB, NPHI, PEF, RD,KTh, U, Th, K
Correlation	GR, DT, RD, RHOB, NPHI, KTh, U, Th
SVM	GR, DT, RD, RHOB, NPHI, KTh, U
PCA	GR, DT, KTh
Deviation	GR, DT, RD, PEF, NPHI, KTh, U, Th, K

Table 4. RMSE of Kernel Functions of different distribution groups from final models performance for Ordos basin.

Kernel	RBF	Cauchy	Laplace	Polynomial	Sigmoid	Gaussian	Multiquadric
ALL	1.013	0.440	0.891	72.764	5.605	5.605	2.052
Correlation	1.275	0.520	1.124	singular	5.621	5.621	1.968
SVM	1.054	0.398	0.878	singular	5.513	5.513	1.891
PCA	0.924	**0.344***	0.717	singular	5.443	5.443	1.868
Deviation	1.047	0.446	0.924	73.473	5.682	5.682	1.906

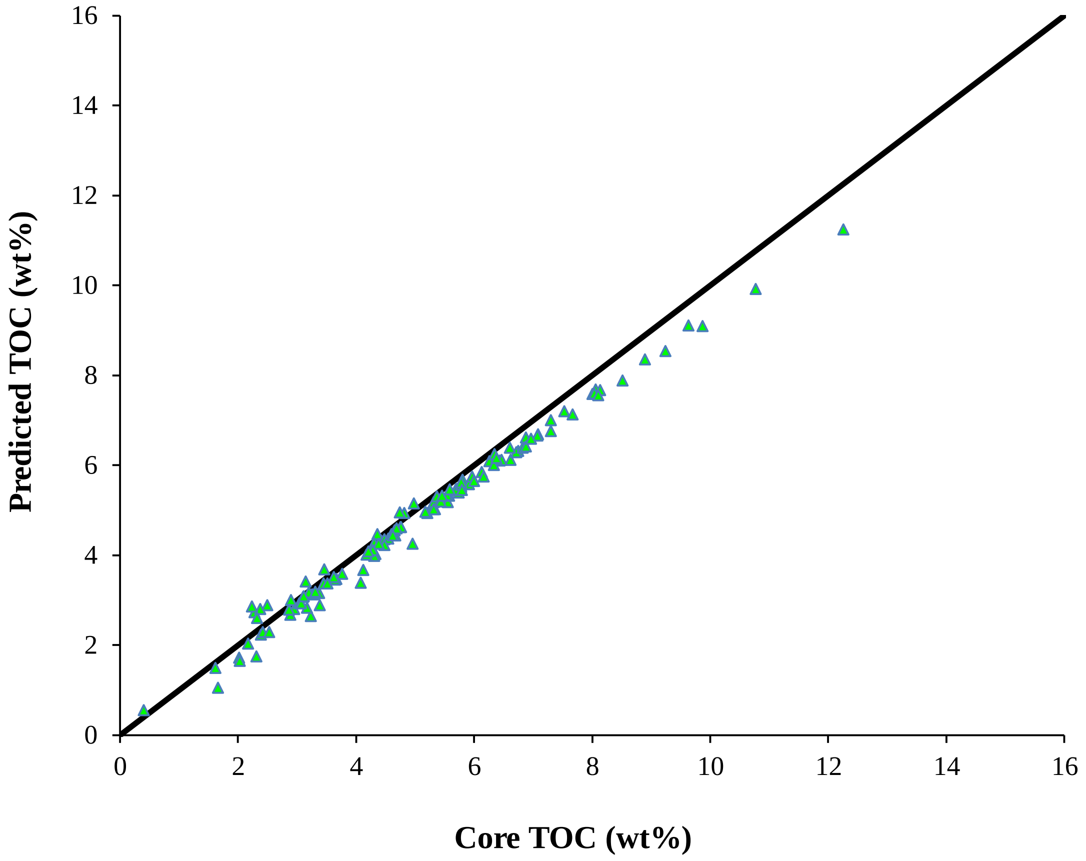


Figure 13. Correlation between core derived TOC and GPR prediction for Ordos basin with RMSE value of 0.344.

The dataset for Canning basin includes natural gamma-ray (GR), caliper log (CAL), sonic log (DT), density log (RHOB), neutron log (NPHI), deep lateral resistivity (LLD), shallow lateral resistivity (LLS), and micro spherically focused log (MSFL). The attribute groups selected by the 4 weights algorithms are shown in Table 5. Table 6 shows the RMSE of the GPR model performance with the 7 kernel functions of the 5 groups on the training data. From Table 6, we can see that the subset set up by SVM algorithm has the highest accuracy among the 5 groups, and Cauchy Kernel function still shows better accuracy than the other 6 Kernel Functions. Thus, the subset chosen by SVM weights with Cauchy Kernel demonstrated a best performance, with a RMSE of 0.079. We also plot the predicted TOC and the core derived TOC in Figure 14, which shows that these two data have high correlation with GPR method. Cauchy Kernel, as one of the typical kernel Function, has high performance in both high TOC reservoir and low TOC reservoir. Cauchy Kernel function, coming from the Cauchy distribution, has multi-scale representation ability. It is a long-tailed kernel and can be used to give long-range influence and sensitivity over the high dimension space. Also, it is suitable to classify the samples with smooth distribution no matter in small or big variance of the noise.

Table 5. Attributes selection based on weight algorithms for Canning basin.

Weights algorithm	Attributes groups
All	GR, DT, RHOB, NPHI, LLD, LLS, MSFL, CALI
Correlation	GR, DT, RHOB, CALI
SVM	GR, DT, RHOB, NPHI, LLD, LLS
PCA	GR, DT, LLD, LLS
Deviation	GR, DT, LLD, LLS, MSFL

Table 6. RMSE of Kernel Functions of different distribution groups from final models performance for Canning basin.

Kernel	RBF	Cauchy	Laplace	Polynomial	Sigmoid	Gaussian	Multiquadric
ALL	0.17	0.087	0.152	7.994	0.5	0.184	0.495
Correlation	0.127	0.108	0.092	singular	0.5	0.194	0.329
SVM	0.165	**0.079***	0.140	singular	0.5	0.185	0.496
PCA	0.221	0.218	0.155	singular	singular	0.710	0.499
Deviation	0.169	0.087	0.151	8.734	0.5	0.184	0.495

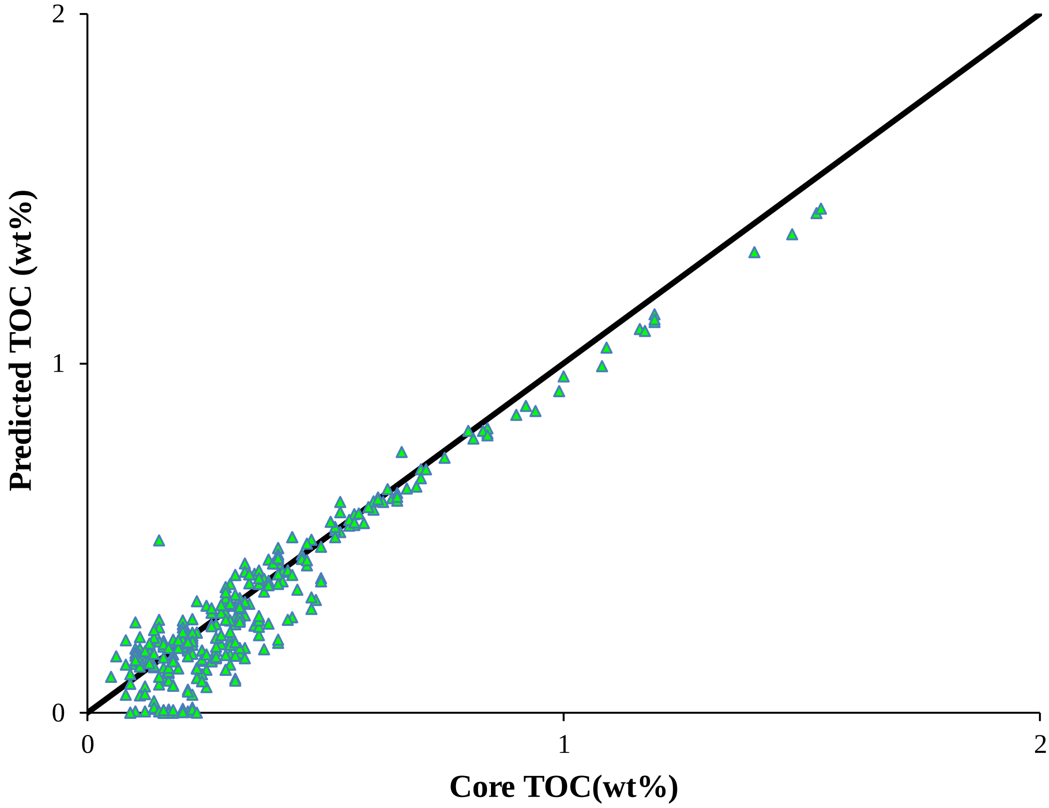


Figure 14. Comparison of core TOC and GPR predicted TOC for Canning basin with RMSE value of 0.079.

4.3. Validation

The results presented above are obtained when the model is trained and compared with the same set of data, and this sometimes leads to overfitting as the model only suits the modeling data but does not has the ability to predict the data it has not seen before. Therefore, the model should be validated with new data. Then, GPR method was also applied in new wells of Ordos basin and Canning Basin (these wells are not in the training data) for validation and the results including Traditional methods are shown in Figure 15 and Figure 16. It is observed that the results from GPR showed better correlations than the other methods in both high TOC and the low TOC gas shale reservoirs. Passey and Schmoker methods gave the results that are larger than the laboratory data, especially in the low TOC reservoir. The reason could be explained by the fact that they only utilize parameters from one or two well logs and thus the responses of these logs do not representatively reflect the TOC trend in the whole reservoir. We also found that the gas content and TOC may have similar impact on the logs, especially in formation with high gas content, therefore, traditional methods based on the simple logs may give incorrect results, e.g. the logs showing low density and high resistivity in the interval from 900 m to 950 m (Figure 16) do not represent high TOC, but are a results of high gas content. However, GPR approach fixes this problem and chooses well logs which mainly reflect TOC, and then these well logs will be used for training the model between mean and covariance functions. Hence, GPR method is much more accurate in TOC prediction.

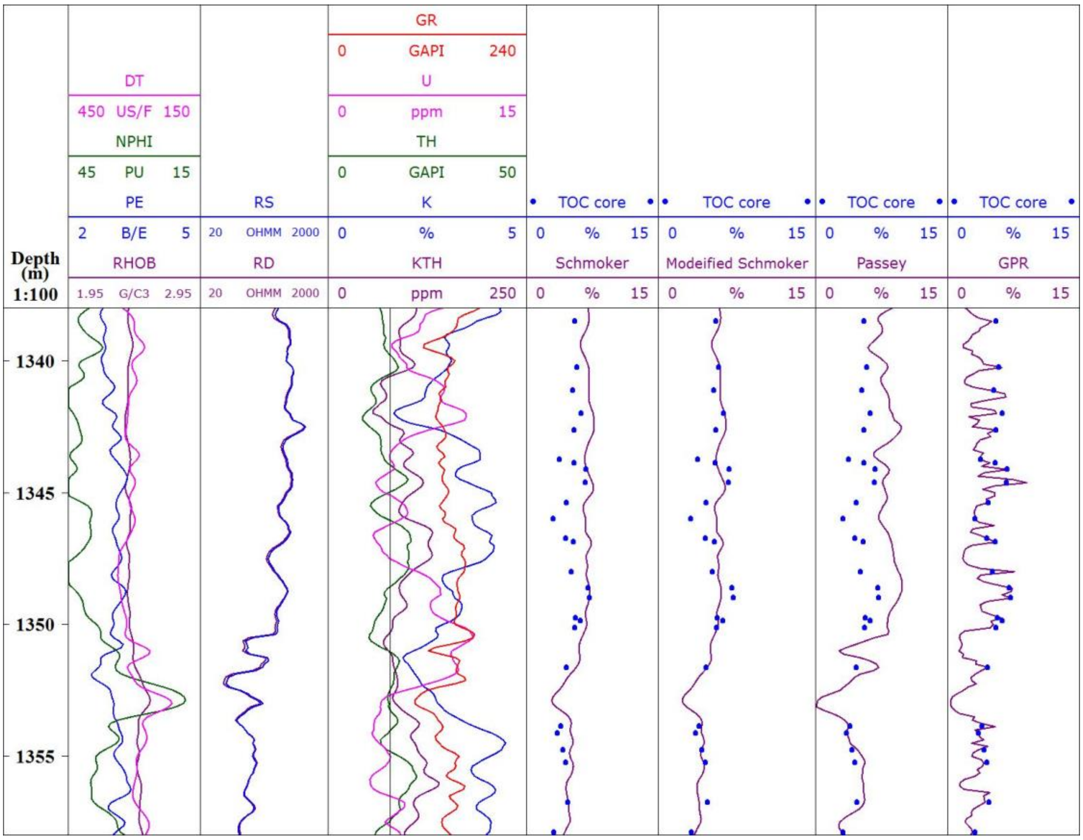


Figure 15. A comparison of prediction TOC using different methods of Yanchang Formation of Ordos basin.

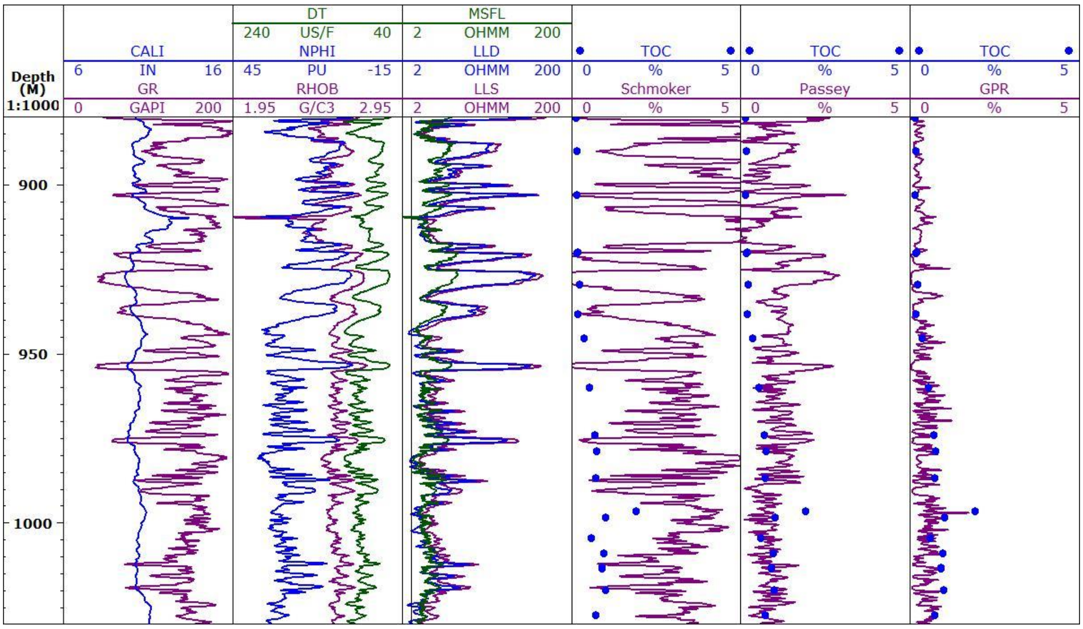


Figure 16. A comparison of prediction TOC using different methods of Goldwyer Formation of Canning basin.

4.4. Discussion

In this work, we have presented a method to apply Gaussian Process Regression technique in calculating TOC in tight shale gas reservoirs. Traditional methods only consider one or two logs, thus can't reflect the true TOC characteristics in the complicated shale gas reservoirs, especially in the low TOC shale gas reservoirs. GPR method, on the other hand, accounts for all the well logging attributes and chooses the relevant logs to build TOC estimation model. Compared to the artificial neural network and support vector machine approach, GPR is more liable to process the complicated regression problem as high dimension, nonlinear and small size samples. Further, the prediction of GPR is probabilistic so that empirical confidence intervals can be calculated, and based on these information, the prediction can be refitted in some region of interest. Finally, this method can specify different kernels; as an example, we selected 7 different kernel functions and 5 attributes groups to get the optimized hyperparameters for GPR techniques to be of value in practice. Overall, the above points make the GPR model more accurate than traditional methods.

5. Conclusions

We have proposed a new method for the TOC prediction based on the machine learning technique – Gaussian Process Regression (GPR) and compared the modeling results with those from the traditional method (Schmoker and Passey method) in two different shale gas reservoirs (Ordos basin – high TOC and canning basin – low TOC). A workflow to predict TOC using GPR was developed that chose a total of 7 kernel functions and 5 attribute groups coming from 4 weighting algorithms. The results showed that the traditional methods can not accurately estimate TOC for shale gas reservoirs, especially for low the TOC formation. Schmoker and Passey methods tended to overestimate the TOC in high gas content reservoirs. Machine learning results of TOC obtained from GPR for Canning basin and Ordos basin were much close to the laboratory test results even in the low TOC content reservoir. Further, we also found that the Cauchy Kernel function showed lower error than the others kernel functions for each attributes groups in both reservoirs. Because such GPR method accounted for a variety of well log data of the corresponding formation for TOC computation, thus, high accuracy and low error model was obtained. It was found that the proposed GPR method demonstrated high accuracy and generalization compared to the traditional methods.

We thus conclude that the GPR method is an efficient and accurate tool for TOC estimations in tight shale gas reservoir and the result is more reliable in comparison with the traditional methods.

Acknowledgements

This study was supported by Special Fund from the State Key Laboratory of Continental Dynamics at Northwest University in China, and the Donors of the Shaanxi Province Specialized Research Fund of Higher Education (grant No. 14JK1740), also acknowledges Opening Foundation of Shaanxi Key Laboratory of

Exploration and Comprehensive Utilization of Mineral resources.

References

Ahmadi, M. A., Ebadi, M. and Hosseini, S. M. (2014) Prediction breakthrough time of water coning in the fractured reservoirs by implementing low parameter support vector machine approach. *Fuel, 117*, pp. 579-589.

Al-Anazi, A. and Gates, I. (2010) A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. *Engineering Geology, 114*(3), pp. 267-277.

Alizadeh, B., Najjari, S. and Kadkhodaie-Ilkhchi, A. (2012) 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, Persian Gulf, Iran. *Computers & Geosciences, 45*, pp. 261-269.

Altowairqi, Y., Rezaee, R., Evans, B. and Urosevic, M. (2015) Shale elastic property relationships as a function of total organic carbon content using synthetic samples. *Journal of Petroleum Science and Engineering, 133*, pp. 392-400.

Bolandi, V., Kadkhodaie, A. and Farzi, R. (2017) Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation, the Persian Gulf basin, offshore Iran. *Journal of Petroleum Science and Engineering*.

Bonilla, E. V., Chai, K. M. and Williams, C. (2008) Multi-task Gaussian process prediction. in *Advances in neural information processing systems*. pp. 153-160.

Chen, H.-H., Hunter, L., Poteat, H. T. and Snow, K. K. (2005) Machine learning method. in: Google Patents.

Chen, T., Morris, J. and Martin, E. (2007) Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems, 87*(1), pp. 59-71.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association, 111*(514), pp. 800-812.

Ding, J., Xiaozhi, C., Xiudi, J., Bin, W. and Jinmiao, Z. (2015) Application of AVF Inversion on Shale Gas Reservoir TOC Prediction. in *2015 SEG Annual Meeting*: Society of Exploration Geophysicists.

Dudley, R. M. (2010) Sample functions of the Gaussian process. in *Selected Works of RM Dudley*: Springer. pp. 187-224.

Hammen, D. G. (2003) Machine learning method. in: Google Patents.

Hasebe, M. and Nagayama, Y. (2002) Reservoir operation using the neural network and fuzzy systems for dam control and operation support. *Advances in Engineering Software, 33*(5), pp. 245-260.

Jarvie, D. M. (2012) Shale resource systems for oil and gas: Part 2—Shale-oil resource systems.

Jarvie, D. M., Hill, R. J., Ruble, T. E. and 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*(4), pp. 475-499.

Khoshnoodkia, M., Mohseni, H., Rahmani, O. and 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*(1), pp. 119-130.

Kim, T., Hwang, S. and Jang, S. (2017) Petrophysical approach for S-wave velocity prediction based on brittleness index and total organic carbon of shale gas reservoir: A case study from Horn River Basin, Canada. *Journal of Applied Geophysics, 136*, pp. 513-520.

Kotsiantis, S. B., Zaharakis, I. and Pintelas, P. (2007) Supervised machine learning: A review of classification techniques. in.

Kuo, J.-T., Hsieh, M.-H., Lung, W.-S. and She, N. (2007) Using artificial neural network for reservoir eutrophication prediction. *Ecological modelling, 200*(1), pp. 171-177.

Lawrence, N. D. (2004) Gaussian process latent variable models for visualisation of high dimensional data. in *Advances in neural information processing systems*. pp. 329-336.

Lukoševičius, M. and Jaeger, H. (2009) Reservoir computing approaches to recurrent neural network training. *Computer Science Review, 3*(3), pp. 127-149.

Michalski, R. S., Carbonell, J. G. and Mitchell, T. M. (2013) *Machine learning: An artificial intelligence approach,* Springer Science & Business Media.

Mierswa, I. and Morik, K. (2005) Automatic feature extraction for classifying audio data. *Machine learning, 58*(2-3), pp. 127-149.

Miller, A. (2002) *Subset selection in regression,* CRC Press.

Montgomery, S. L., Jarvie, D. M., Bowker, K. A. and Pollastro, R. M. (2005) Mississippian Barnett Shale, Fort Worth basin, north-central Texas: Gas-shale play with multi–trillion cubic foot potential. *AAPG bulletin, 89*(2), pp. 155-175.

Paciorek, C. and Schervish, M. (2004) Nonstationary covariance functions for Gaussian process regression. *Advances in neural information processing systems, 16*, pp. 273-280.

Passey, Q., Creaney, S., Kulla, J., Moretti, F. and Stroud, J. (1990) A practical model for organic richness from porosity and resistivity logs. *AAPG bulletin, 74*(12), pp. 1777-1794.

Passey, Q. R., Bohacs, K., Esch, W. L., Klimentidis, R. and 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*: Society of Petroleum Engineers.

Rasmussen, C. E. (2006) Gaussian processes for machine learning.

Rodriguez, J. D., Perez, A. and Lozano, J. A. (2010) Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence, 32*(3), pp. 569-575.

Ross, D. J. and Bustin, R. M. (2007) Impact of mass balance calculations on adsorption capacities in microporous shale gas reservoirs. *Fuel, 86*(17), pp. 2696-2706.

Sahami, M. and Heilman, T. D. (2006) A web-based kernel function for measuring the similarity of short text snippets. in *Proceedings of the 15th international conference on World Wide Web*: AcM. pp. 377-386.

Schmoker, J. W. (1979) Determination of organic content of Appalachian Devonian shales from formation-density logs: Geologic notes. *AAPG bulletin, 63*(9), pp. 1504-1509.

Schmoker, J. W. and Hester, T. C. (1983) Organic carbon in Bakken formation, United States portion of Williston basin. *AAPG bulletin, 67*(12), pp. 2165-2174.

Silversides, K. L. and Melkumyan, A. (2016) A Dynamic Time Warping based covariance function for Gaussian Processes signature identification. *Computers & Geosciences, 96*, pp. 69-76.

Sondergeld, C. H., Ambrose, R. J., Rai, C. S. and Moncrieff, J. (2010) Micro-structural studies of gas shales. in *SPE Unconventional Gas Conference*: Society of Petroleum Engineers.

Sone, H. and Zoback, M. D. (2013) Mechanical properties of shale-gas reservoir rocks—Part 1: Static and dynamic elastic properties and anisotropy. *Geophysics, 78*(5), pp. D381-D392.

Tan, M., Song, X., Yang, X. and 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*, pp. 792-802.

Tonner, P. D., Darnell, C. L., Engelhardt, B. E. and Schmid, A. K. (2017) Detecting differential growth of microbial populations with Gaussian process regression. *Genome research, 27*(2), pp. 320-333.

Triba, M. N., Le Moyec, L., Amathieu, R., Goossens, C., Bouchemal, N., Nahon, P., Rutledge, D. N. and Savarin, P. (2015) PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Molecular BioSystems, 11*(1), pp. 13-19.

Wang, J., Chen, L., Kang, Q. and Rahman, S. S. (2016) Apparent permeability prediction of organic shale with generalized lattice Boltzmann model considering surface diffusion effect. *Fuel, 181*, pp. 478-490.

Wang, J. M., Fleet, D. J. and Hertzmann, A. (2008) Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence, 30*(2), pp. 283-298.

Witten, I. H. and Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques,* Morgan Kaufmann.

Zhang, T., Ellis, G. S., Ruppel, S. C., Milliken, K. and Yang, R. (2012) Effect of organic-matter type and thermal maturity on methane adsorption in shale-gas systems. *Organic geochemistry, 47*, pp. 120-131.

Zhao, P., Mao, Z., Huang, Z. and Zhang, C. (2016) A new method for estimating total organic carbon content from well logs. *AAPG bulletin, 100*(8), pp. 1311-1327.