Data mining and machine learning for identifying sweet spots in shale reservoirs

Pejman Tahmasebi, Farzam Javadpour, Muhammad Sahimi

A R T I C L E   I N F O

Article history:
Received 12 February 2017
Revised 20 June 2017
Accepted 11 July 2017
Available online 14 July 2017

Keywords:
Neural networks
Genetic algorithm
Big data
Shale formation
Fracable zones
Brittleness index

A B S T R A C T

Due to its complex structure, production form a shale-gas formation requires more drillings than those for the traditional reservoirs. Modeling of such reservoirs and making predictions for their production also require highly extensive datasets. Both are very costly. In-situ measurements, such as well-logging, are one of most indispensable tools for providing considerable amount of information and data for such unconventional reservoirs. Production from shale reservoirs involves the so-called fracking, i.e. injection of water and chemicals into the formation in order to open up flow paths for the hydrocarbons. The measurements and any other types of data are utilized for making critical decisions regarding development of a potential shale reservoir, as it requires hundreds of millions of dollar initial investment. The questions that must be addressed include, does the region under study can be used economically for producing hydrocarbons? If the response to the first question is affirmative, then, where are the best places to carry out hydro-fracking? Through the answers to such questions one identifies the sweet spots of shale reservoirs, which are the regions that contain high total organic carbon (TOC) and brittle rocks that can be fractured. In this paper, two methods from data mining and machine learning techniques are used to aid identifying such regions. The first method is based on a stepwise algorithm that determines the best combination of the variables (well-log data) to predict the target parameters. However, in order to incorporate more training, and efficiently use the available datasets, a hybrid machine-learning algorithm is also presented that models more accurately the complex spatial correlations between the input and target parameters. Then, statistical comparisons between the estimated variables and the available data are made, which indicate very good agreement between the two. The proposed model can be used effectively to estimate the probability of targeting the sweet spots. In the light of an automatic input and parameter selection, the algorithm does not require any further adjustment and can continuously evaluate the target parameters, as more data become available. Furthermore, the method is able to optimally identify the necessary logs that must be run, which significantly reduces data acquisition operations.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Due to the recent progress in multistage hydraulic fracturing, horizontal drilling and advanced recovery methods, shales, recognized as unconventional reservoirs, have become a promising source of energy. They are formed by fine-grained organic-rich matters and were previously considered as source and seal rock that, due to gas production and high pressures in the conventional reservoirs were traditionally called the trouble zones. Such regions were usually ignored, which explains why no comprehensive data are available for them. Thus, their characterization is still a massive task (Tahmasebi, Javadpour, & Sahimi, 2015a,b; Tahmasebi, Javadpour, Sahimi, & Piri, 2016; Tahmasebi, 2017; Tahmasebi & Sahimi, 2015). Furthermore, shales exhibit highly variable structures and complexities from basin to basin, and even in small fields. They host very small pores, and have low-matrix permeability and heterogeneity, both at the laboratory and field scales. Due to such difficulties and given the fact that new methods for characterization of shale reservoirs are still being developed, application of the characterization and modeling methods for the traditional reservoir to shales is of great importance. In particular, accurate characterization of such reservoirs entails integrating various information, including petrophysical, geochemical, geomechanical, and reservoir data (Tahmasebi & Sahimi, 2016a,b;
Apart from its type, efficient drilling may be thought of as targeting the most productive zones of a reservoir with maximum exposure. For shale reservoirs, this concept is equivalent to areas with high total organic carbon (TOC) and high fracability, i.e., brittleness, which calls for comprehensive characterization of such complex formations. High TOC and fracable index (FI) reflect high quality of shale-gas reservoirs. The role of the TOC is clear, as it is one of the main factors for identifying an economical shale reservoir. Higher TOC, ranging from 2% to 10%, represents richer organic contents and, consequently, higher potential for gas production. Since natural gas is trapped in both organic and inorganic matters, the TOC denotes the entire organic carbons, and is a direct measure of the volume and maturity of the reservoir.

The FI influences the flow of hydrocarbons in a shale reservoir and any future fracking in it. Thus, identifying the layers in a reservoir with high FI is of great importance. The FI controls a shale reservoir’s production since it strongly influences the wells’ production. It also provides very useful insight into where and how new wells should be placed and spaced. In fact, unlike conventional reservoirs that depend on long-range connectivity of the permeable zones, optimal well locations and spacing control the performance of shale reservoirs and future fracking operations in them. Thus, separating the brittle and ductile zones of rock is a key aspect of successful characterization of shale-gas reservoirs. Moreover, brittle shale has high potentials for being naturally fractured and, consequently, exhibits good response to fracking treatments.

Past successful experience indicated that characterization of shale reservoirs need accurate identification of the so-called sweet spots, i.e., the zones that present the best production of the potential for high production, and the potential fracable zones, which are critical to maximizing the production and future recovery. The placement of most of the wells is closely linked with the sweet spots, as well as the fracable zones for hydraulic fracturing. For example, the TOC represents the ability of a shale reservoir in storing and producing hydrocarbons. Fracability is controlled mainly by mineralogy and elastic properties, such as the Young’s and bulk moduli and the Poisson’s ratio (Sullivan Glaser et al., 2013). Therefore, identification of the sweet spots is of great importance to shale reservoirs. Such spots are characterized through high kerogen content, low water saturation, high permeability, high Young’s modulus and low Poisson’s ratio.

One of the primary, as well as most affordable, methods for characterizing complex reservoirs is coring and collecting petrophysical data, as well as well logs. The latter can be integrated with former in order to develop a more reliable model. In principle, well-log data can be provided continuously and, thus, they represent a real-time resource. Because of a huge number of wells in a typical shale reservoir, we refer to such datasets as big data. Clearly, such information is very useful when it is coupled with some techniques that help better identify the sweet spots and fracable zones. Eventually, the questions that must be addressed are: where one should/should not drill new wells? Where are the zones with high/low fracability index?

Aside from such critical questions, another issue regarding the available big data is the fact that new data are continuously obtained as the production proceeds. Thus, any algorithm for the analysis of big data should be flexible enough for rapid adaptation of new data. Furthermore, another important feature of the algorithm should be its ability to use the available information to create a “training platform” for forecasting the important parameters.

In this paper, a very large database consisting of well logs, x-ray diffraction (XRD) data, and experimental core analysis is used to develop a model that reduces the cost and increases the probability of identifying the sweet spots. First, we describe a method of data mining called stepwise regression (Efromyson, 1960; Montgomery, Peck, & Vining, 2012) for identifying the correlations between the target (i.e., dependent) parameters – the TOC and FI, and the available well-log data (the independent variables). Then, a hybrid method borrowed from machine learning and artificial intelligence is proposed for accurate predictions of the parameters. Both methods can be tuned rapidly, and can use the older database to accurately characterize shale reservoirs.

2. Methodology

As mentioned earlier, two very different methods are used in this paper. The first is borrowed from data-mining field by which the correlation between an independent variable and a series of dependent variables is constructed and used for future forecasting. Next, a method of machine learning for developing a more robust model is introduced that recognizes the complex relations between the variables.

2.1. Linear regression models based on statistical techniques

Due to dealing with a multivariate dataset, multiple linear regression (MLR) is selected to model the relationship between the independent and response variables, the TOC and FI. Suppose that the response variable y is related to k predictor variables. Then, as is well-known,

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon. \]  

(1)

represents a MLR that contains k predictor variables. \( \varepsilon \) is the model error and \( \beta_k \) with \( k = 0, ..., k \) are partial regression coefficients. In practice, \( \beta_k \) and the variance of the model error (\( \sigma^2 \)) are not known a priori. The parameters may be estimated using the sample data.

In the k-dimensional space of the predictor variables, Eq. (1) represents a hyperplane. The expected change in the response \( y \) as a result of a change in \( x_j \) is contained in the coefficients \( \beta_j \). It should be noted that the variation of \( y \) is subject to holding all the remaining variables \( x_i (i \neq j) \) constant. More complex functional forms can still be included and modeled with a slight modification of Eq. (1). For example, consider the following equation:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \varepsilon. \]  

(2)

if we let \( \Phi_1 = x_1 \), \( \Phi_2 = x_2 \), \( \Phi_3 = x_1^2 \) and \( \Phi_4 = x_1 x_2 \) then Eq. (2) is written as

\[ y = \beta_0 + \beta_1 \Phi_1 + \beta_2 \Phi_2 + \beta_3 \Phi_3 + \beta_4 \Phi_4 + \varepsilon. \]  

which is again a MLR model. Thus, Eq. (3) may be used to predict the response variable using new dependent variables. The coefficients of Eq. (1) are estimated using least-square curve fit.

2.1.1. Evaluating sub-models significance

Clearly, in the presence of multiple variables, selecting the most important ones for reducing the complexity and performance of a predictive model is of great importance. In other words, we are interested in identifying the best subset of parameters that maintains the variability while minimizing the complexity. Thus, various subsets must be compared and, then, one must decide which one is more representative and accurate than others. One common criterion for model adequacy is the coefficient of multiple determination, \( R^2 \). Thus, for a model with \( p \) variables, \( R^2 \) is calculated by

\[ R^2_p = \frac{SSR_p(p)}{SS_T} = 1 - \frac{SSRes_p(p)}{SS_T}. \]  

(4)
where $SS_{res}(p)$ and $SS_{F}(p)$ indicate, respectively, the residual sum of the squares and regression sum of squares. A higher $p$ results in a higher $R^2_{adj}$. However, an adjusted coefficient of multiple determination that is more transparent is defined by

$$R^2_{adj,p} = 1 - \left( \frac{n - 1}{n - p} \right) (1 - R^2_p).$$

where $n$ is the number of variables. It should be noted that $R^2_{adj,p}$ does not increase by adding more variables, but $R^2_p$ exceeds $R^2_{adj,s}$ if and only if the partial $F$ statistics exceeds $1$ (Edwards, 1969; Haitovsky, 1968; Seber, George, & Lee, 2003). The partial $F$ statistics is used for evaluating the significance of adding a new variable $s$. Thus, one can use a subset of models that maximize $R^2_{adj,p}$.

2.1.2. Variable selection

Variable selection can be computationally expensive, as one must determine the result of adding/eliminating of each variable one by one. There are, however, some efficient methods for doing so that are briefly described below.

- Forward selection

This method begins by adding the first variable that has the maximum correlation with the response variable. Then, if the statistics $F$ exceeds $F_{in}$, it will be added to the model. Note that $F_{in}$ is used as a threshold and is calculated by, $F = \left( \frac{S_{res1} - S_{res2}}{S_{res2}} \right)$. In a similar fashion, the next variable is added to the model if it carries the maximum correlation after the previously added variable. Once again, the statistics $F$ is compared with $F_{in}$ and added to model, if it exceeds the threshold. The process continues until $F$ does not exceed the predefined threshold or no variable is left.

- Backward selection

Backward variable elimination acts exactly as the opposite of the forward selection. It first incorporates all the $n$ variables in the model and, then, removes them one-by-one according to some criterion. The elimination is carried out using the statistics of elimination, $F_{out}$. In other words, the minimum statistics of the variables are compared with $F_{out}$ and are eliminated if it is less than $F_{out}$. The procedure is then repeated for the remaining $n - 1$ variables until the smallest statistics $F$ is not less than $F_{out}$.

- Stepwise selection

The above algorithms are computationally expensive when one deals with a large dataset, as most of the unconventional reservoirs are. Instead, one may use stepwise selection, which is an efficient method to identify the best combination of the variables (Efron & Tibshirani, 1993). Similar to the backward algorithm, in stepwise selection the independent variables are all incorporated into the model one at a time and are re-evaluated using a pre-defined $F$. After adding a new variable, the statistics $F_{out}$ is compared with the pre-defined criterion to decide if some of the existing variables can be eliminated without losing the accuracy. That this is possible is due to one of the previously incorporated variables losing its significance when one or more new variables are added to model. The algorithm is terminated when either the defined statistics $F$ is maximized, or the improvement from adding any new variable $F$ does not exceed the pre-defined $F$. It also should be noted that both $F_{in}$ and $F_{out}$ are used in this method. Generally, one can consider $F_{out} < F_{in}$ in order allowing a variable to be added easier. It should be noted that one may benefit from the recently-developed algorithms for forward selections (Cernuda et al., 2013; Cernuda, Lughofer, Märzinger, & Kasberger, 2011; Serdio et al., 2017).

2.2. Non-linear models based on soft computing architectures

The method described above can be used when one expects to extract spatial linear relationships between the available variable. These methods are, however, unable of discovering the non-linear relationships, as we encounter frequently in characterization of complex shale reservoirs. Thus, in this section we demonstrate the ability of non-linear models based on soft computing for extracting such nonlinear relationships.

The main purpose of machine leaning algorithms is to learn from the existing data and evidence, to be able to make predictions. Thus, most of such approaches rely on the input data, which is why they are classified as data-driven techniques. Such methods have been used widely in the earth science problems, including permeability/porosity estimation (Karimpouli & Malehmir, 2015), grade estimation (Tahmasebi & Hezarkhani, 2011), contaminant modeling (Rogers & Dowla, 1994), predicting temperature distribution and methane output in large landfills that are essentially large-scale porous media in which biodegradation occur (Li et al., 2011; Li, Qin, Tsotsis, & Sahimi, 2012; Li, Tsotsis, Sahimi, & Qin, 2014), etc. Due to the considerable complexity and unpredictable spatial correlations between the available data for shale reservoirs in this paper, we use the fuzzy logic approach, first proposed by Zadeh (1965), and combine it with two more machine learning algorithms, namely, neural networks (NNs) and the genetic algorithm (GA) in order to increase the model’s predictive power. What follows is a brief description of the techniques.

2.2.1. Fuzzy inference system

Fuzzy set theory, or fuzzy logic (FL), was first introduced by Zadeh (1965) in which, unlike the Boolean logic that has two outcomes, namely, true (1) or false (0), the true value of a variable may be between in between in [0, 1]. In other words, the FL allows one to define a partial set membership (see below). Thus, even if a set of vague, ambiguous, noisy, imprecise, and incomplete input information is made available, the FL can still produce a valid response.

Briefly, the FL tries to identify a connection between the input and output through defining some if-then rules, and is composed of some elements that are described here. Typically, the input and output data are presented using a membership function, represented by a curve. The curve attributes some weights to the input and defines how it can be mapped onto the output. The rules are the instructions that define the conditional statements. They, along with the membership functions, map the input onto the output. Finally, the interdependence of the input and output is defined through some fuzzy preposition, or the so-called FL operators. A recent comprehensive review on the FL is given by Gomide (2016).

The aforementioned elements in the FL are typically difficult to set. For example, membership functions, their distributions and the compositions of the fuzzy rules strictly control the FL performance, but are difficult to set. Thus, as the first solution, such parameters may be defined through an exhaustive trial and error. In addition, prior knowledge and experience of the under-study problem can also help defining more accurate if-then rules. In fact, ultra-tight shale gas reservoirs are currently under extensive studies and, thus, some of their complexities are not even discovered or understood yet. Thus, defining such elements for shales may be very misleading. For this aim, the NNs are employed to alleviate such shortcomings and define the necessary parameters in a more meaningful way. Another solution can be use of the GA in the FL (Cordon, Gomide, Herrera, Hoffmann, & Magdalena, 2004). It should be noted that there are many other techniques that deal with fuzzy systems extraction and learning from data (Abonyi, 2003; Lughofer, 2011; Pedrycz & Gomide, 2007), but they are not discussed in this paper.
2.2.2. Integrated neural networks and fuzzy logic

The NNs have been inspired by human brain, with a significantly less complexity, which can be used for modeling complex and highly nonlinear problems whose solutions by the common statistical or analytical methods are not straightforward (Bishop, 1995) or impossible to drive. Briefly, a NN is composed of some hierarchical nodes called neurons, which are trained using some data. The inputs are mapped onto the output space through some weights (or functions) that are initially generated randomly. Then, the weights are updated using the produced error in the output. The process, called training, continues until the error is smaller than a threshold. Next, the resulting NN, or more precisely the mathematical functions derived by it, are used for making prediction. Back propagation is one of the most popular learning algorithms for the NN that is used in conjunction with the gradient descent or the Levenberg–Marquardt method, two optimization algorithms by which the weights are changed iteratively to minimize the following error, in order to determine their optimal values:

\[ E = \frac{1}{2} \sum \left( \text{outp}_i - \text{prd}_i \right)^2 \]  

where \text{outp}_i is the actual output and \text{prd}_i is the predicted value. Minimizing \( E \) entails setting to zero its derivatives with respect to the parameters \( w_{ij} \) in order to decide how to vary \( w_{ij} \) for two connected nodes \((i, j)\). Thus,

\[ w_{ij} := w_{ij} + \Delta w_{ij} \]

where

\[ \Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \]

in which \( \eta \) is a learning rate. In order to prevent the optimization method from being trapped in local minimum of \( E \), as opposed to its true global minimum, one can use the following approach. Suppose that

\[ \varphi_j = -\frac{\partial E}{\partial NN_j} \]

and

\[ \xi_j = -\frac{\partial NN_j}{\partial w_{ij}} \]

then, \( \Delta w_{ij} \) is evaluated by

\[ \frac{\partial E}{\partial w_{ij}} = \frac{\partial NN}{\partial w_{ij}} \times \frac{\partial E}{\partial NN_j} = \varphi_j \xi_j \]

We then have

\[ \Delta w_{ij} = \eta \varphi_j \xi_j \]

Since,

\[ \frac{\partial E}{\partial \xi_j} = -(\text{outp}_i - \text{prd}_i) \]

and

\[ \frac{\partial \xi_j}{\partial NN_j} = f'(NN_j) \]

\( \varphi_j \) is also evaluated by,

\[ \varphi_j = \frac{\partial E}{\partial NN_j} = \frac{\partial E}{\partial \xi_j} \times \frac{\partial \xi_j}{\partial NN_j} = (\text{outp}_i - \text{prd}_i) f'(NN_j) \]

Thus, an approach based on integrating the NN and FL can address the aforementioned issues in defining the latter’s parameters (Jang, 1992, 1993). The integrated system benefits from the training ability of the NN for minimizing the generated error in the output. Thus, through the optimization loop in the NN the FL parameters are tuned. In this fashion, the strengths of the FL qualification and the NN quantification are efficiently integrated.

Suppose that \( n \) and \( 1 \) variables are available as the input (\text{inp}_i) and output (\text{out}_i), respectively, as is the case in this paper. Then, the first-order Sugeno fuzzy system (Tanaka & Sugeno, 1992) is defined by

\[ r_j : \text{If} \text{inp}_1 = \Psi_1^j, \text{inp}_2 = \Psi_2^j, \ldots, \text{and} \text{inp}_n = \Psi_n^j, \text{then} \ f_r = \sum_{i=1}^{n} \beta_{ij}^r \text{inp}_i + \epsilon_i \]

where \( \Psi_1^j \) and \( \beta_{ij}^r \) are, respectively, the \( i \)th fuzzy membership set and the consequent parameter for the \( j \)th input. \( f_r \) indicates the consequence part of the \( r \)th rule. In this study, the Sugeno model is used in which the membership functions and weight adjustment are denoted by premise and consequent parts, respectively; see Fig. 1. The internal operations for the input mapping, weight adjustment and consequent parameter integration are demonstrated as well.

As already mentioned, the NN is integrated with the fuzzy system shown in Fig. 1. The new integrated architecture is shown in Fig. 2 (Jang, 1993). The resulting hybrid machine learning (HML) contains various elements in each layer that are as follow. The inputs are denoted by \( L_1 \), which in this study are various well logs. \( L_2 \) represents the layer that contains the membership functions \( \Psi_1^j \), and measures the degree to which the \( \text{inp}_i \) satisfies the quantifier:

\[ \text{outp}_1^j = \Psi_1^j(\text{inp}_i) \]

\( L_3 \), the third layer, multiplies the input data and produces the firing weights for each rule \( r \), as follow:

\[ \omega_r = t(\xi_{\Psi_1^j}, \ldots, \xi_{\Psi_n^j}) \]

\( L_4 \), the fourth layer, normalizes each firing rule based on the sum of all the previously calculated weights:

\[ \overline{\omega}_r = \frac{\sum_{i=1}^{n} \omega_r}{\sum_{i=1}^{n} \omega_r} \]

Next, the calculated normalized firing strengths are used in the next layer, \( L_5 \), to generate the adaptive function by

\[ \overline{\omega}_r f_r = \overline{\omega}_r(\beta_{1}^r \text{inp}_1 + \beta_{2}^r \text{inp}_2 + \ldots + \beta_{n}^r \text{inp}_n + \epsilon_r) \]

The overall output is then calculated in the sixth layer, \( L_6 \), as

\[ \sum_{r=1}^{i} \overline{\omega}_r f_r \]

Finally, the response, calculated from the previous layer, appears in the last layer, \( L_7 \).

Clearly, some of the aforementioned parameters, such as the membership functions, are vital for an optimal hybrid network. A hybrid learning procedure is applied to the architecture of Fig. 2 that estimates the premise and the consequent parameters and, consequently, allows capturing the complexity and variability between the input and output data. Thus, based on the back-propagation error, the parameters in the premise part are first kept fixed in order to compute the error. Next, by holding fixed the parameters in the consequent part, the premise’s parameters are updated using the gradient-descent method. Finally, the membership functions and their consequent weights are optimized.

Although the NN assists the FL for optimal parameter selection, there are still some parameters that need to be optimized. For example, the NN cannot decide how many membership functions are needed. In addition, due to using two user-dependent parameters – the learning rate and the momentum coefficient – the NN itself may be trapped in a local minimum. It is worth mentioning that momentum adds a small friction to the previous weights.
and updates to the current state. This parameter prevents the network from converging to local minimum. A high value can overshoot the minimum and, likewise, a lower value cannot avoid local minima and it causes the computations to be too long. To address such issues, a third method, namely, the GA, is integrated with the neural-fuzzy structure that helps determining the parameters. Briefly, due to presence of various user-dependent and time-demanding parameters in the NN, such as its number of layers, number of membership functions in the FL and the learning parameters (e.g., the momentum and rate), the GA can be used to accelerate the above trial-and-error procedure. In other words, defining the aforementioned parameters requires a considerable amount of time, whereas the GA algorithm can efficiently and systematically identify the optimal values.

2.2.3. The genetic algorithm

The GA is one of the most successful optimization algorithms that can deal with a variety of problems that are not feasible, both computationally and in terms of the accuracy, to be addressed using many other optimization techniques. For example, the GA can be used when the objective function (the energy $E$ to be minimized) is discontinuous, non-differentiable, highly nonlinear, and even stochastic (Goldberg & Holland, 1988). The technique is built on an initial population of the candidates that mimic the process, and is inspired by natural selection. Each candidate has some specific properties, i.e., chromosomes, which can be modified, e.g., mutated, based on the existing evidence. Thus, the aim is to modify the initial population to reach an optimal solution that minimizes $E$ globally. At each step, the chromosomes (the parameters to be estimated in the present work) are randomly selected from the current parents – the population. Then, “children” are generated for the next population. After some initial trial and error, the GA identifies the potentially useful chromosomes, so that the children will produce smaller error.

The GA uses three types of rules to create the next generation using the current population, which are as follow (Deb & Agrawal, 1999). (i) Mutation, an operation that changes the value of chromosomes; (ii) selection, which chooses the parents from the current children in order to produce the next generation, and (iii) crossover, an operation that combines the chromosomes (the parameters) to increase the variability and prevent the energy $E$ from getting trapped in a local minimum.

In the first step, all the variables are represented by binary strings of 0 and 1. Each chromosome indicates several genes by which various possible values for each free parameter, which needs to be optimized, are examined. Thus, an initial population is generated that contains various parameters of the FL and NN, such as number of the inputs (i.e., well logs), the number of the membership functions, the learning rate, and the momentum coefficient. The candidate are tested and updated progressively by the GA, until the optimal values are obtained.

---

**Fig. 1.** Demonstration of first-order Sugeno fuzzy model along with various membership functions and inputs.

**Fig. 2.** The integrated neural-fuzzy architecture.
3. Results and discussion

As discussed, due to the extensive variability of shale reservoirs, an extensive amount of information is required for their characterization and, hence, it helps to reduce the uncertainty and improve real-time recovery operations. Thus, since well logs provide useful information, and at the same time are widely available, the objective of this study is to use such data to predict two important properties of shales, namely, the TOC and FI. Clearly, none of the well logs can by itself predict the two indices, as there is always very complex correlation between the well logs (Dashtian, Jafari, Koohi Lai, Masihi, & Sahimi, 2011; Karimpouli & Tahmasebi, 2016), which in some cases are mixed with noise. Another important goal of this study is to develop a methodology for deciding the necessary logs that must be run. In other words, significant reduction in data acquisition efforts can be achieved, once the correlation between the two indices and the available data is identified. Together,
the steps described below quantify the two important properties of shale reservoirs in real time.

There are several ways of defining and measuring brittleness (Zoback, 2007). One of the most common methods is based on using the elastic properties, such as the Young’s modulus and the Poisson’s ratio. The latter can be thought of as the rock ability to fail under stress, whereas the ability of rock to resist fracturing can be quantified using the Young’s modulus. Rickman, Mullen, Pettre, Grieser, and Kundert (2008) proposed equations for estimating the brittleness index using well logs.

In this paper, the FI is defined using mineralogical contents. This is motivated by the fact that fracability is mostly controlled by hard mineral contents of rock, such as quartz, calcite and pyrite and, thus, the XRD of such minerals are used as follows (Jarvie, Hill, Ruble, & Pollastro, 2007):

$$FI = \frac{Qz}{Qz + Cal + Cly}$$

where Qz, Cal and Cly are the quartz, calcite and clay contents, respectively. The FI defined by Eq. (22), along with the TOC was used as the dependent (i.e., target) parameters.

3.1. Linear regression models based on statistical technique

The mineralogical analysis and geochemical data for defining the FI were used; their distribution is depicted in Fig. 4, where the sum of the quartz, clay and carbonate minerals is displayed on each vertex. The plot indicates that the content of the samples has higher carbonate.

On the other hand, well-log data are also available. The spatial relationships between the mineralogy, the FI and the gamma ray (GR) log are presented in Fig. 5. As expected, the GR log and FI positively correlate with the Qz content.

Before applying the MLR analysis, one might also check the linear dependence (i.e., pairwise correlation) between the dependent variables FI and TOC, on the one hand, and the available well logs, on the other hand. Such a relationship was studied, and the results are presented in Table 1. The correlation between the FI and sonic (DT), density (PE), resistivity (RT10) and Or logs are stronger than the other logs, whereas the overall correlation between the TOC and the well logs is strong. For example, there is considerable correlation between the TOC, GR and DT. Strong correlation between the TOC and GR is expected, since most of the gas in the samples is trapped in the organic matters.

One can also investigate more complex correlations between three variables. Such correlations are shown in Figs. 6 and 7 for the FI and TOC evaluations, respectively. A weak correlation between the first and second variables and the FI can be seen. Such weak correlations are not sufficient when fitting a 3D surface. For example, a considerable number of mismatches between the fitted surfaces and the selected variables are visible in both cases. Clearly, the correlations with one/two variables do not reveal significant information, and cannot be relied upon. Thus, in order to develop a more robust and accurate model, the proposed MLR is used, which optimally selects the informative variables. The stepwise algorithm is applied to select the appropriate variables among the initial well-log data to estimate the TOC and FI. For this aim, the threshold $F_{in}$ for entering any new variable was considered to be 0.1 with a confidence interval of 90%. As for removing the variable, $F_{out}$ was also set to 0.15. Thus, the following model for estimating the FI was obtained:

$$FI = 0.004PE + 0.0017Thor + 0.032$$

(23)

The correlation coefficient for Eq. (23) is 0.44, which is very low and indicates that the MLR is not able to accurately predict the FI.

On the other hand, the equation for the TOC has a correlation coefficient of 0.88, hence providing reliable predictions for the TOC based on the well-log data, and is given by

$$TOC = 0.016GR + 0.087DT − 1.380K − 4.020$$

(24)

3.2. Non-linear models based on soft computing architectures

Before feeding the input data to the proposed HML method, the data are normalized to prevent any superficial scaling effect. Thus, all the data were normalized between zero and one. The well logs are used as the input, while the TOC and FI represent the output. To test the performance of the proposed method, the data were randomly divided into two groups of training (80%) and testing (20%). The training data should be selected carefully to cover the upper and lower possible boundaries of the input space.

Due to its flexibility for controlling the shape, the Bell membership function,

$$f(x; a, b, c) = \frac{1}{1 + \left| \frac{x-c}{a} \right|^{2b}}$$

(25)

instead of the Gaussian distribution, was used because it can be adjusted easily when the proposed hybrid network needs to tune the variability space between the well-log and the TOC/FI output. In Eq. (25), $a$, $b$, and $c$ are three parameters that can be varied in order to change the shape of the function. In this paper, the momentum and pure-line axon were used as the learning algorithm and the transfer function, respectively. Clearly, the output of pure-line function is the same as the input. Some other parameters that need to be set include the step size and the momentum rate as the learning variables and the input processing elements as the topological parameter. Such parameters can be obtained using trial and error, which is expensive computationally. The GA algorithm can, however, provide optimal solutions, so as to avoid long and cumbersome steps.

An important feature of the GA is its chromosomes (parameters in the present study) that are composed of genes. The chromosomes are each composed of four genes that are the number of inputs, membership functions for each input, the momentum, and the learning rate, which is the same as the number of the variables that should be optimized. First, some random values are generated and assigned to each chromosome. Then, the network
Table 1
Correlation between the FT, TOC and well-log data.

<table>
<thead>
<tr>
<th>Well-log/Dep. Var</th>
<th>GR</th>
<th>Rhob</th>
<th>Nphi</th>
<th>DT</th>
<th>PE</th>
<th>RT10</th>
<th>SP</th>
<th>Ur</th>
<th>Thor</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>FI</td>
<td>−0.014</td>
<td>0.245</td>
<td>−0.34</td>
<td>0.458</td>
<td>0.443</td>
<td>−0.413</td>
<td>−0.260</td>
<td>0.443</td>
<td>0.366</td>
<td>0.179</td>
</tr>
<tr>
<td>TOC</td>
<td>0.736</td>
<td>−0.563</td>
<td>0.703</td>
<td>0.745</td>
<td>−0.675</td>
<td>−0.266</td>
<td>0.428</td>
<td>0.538</td>
<td>0.597</td>
<td>0.410</td>
</tr>
</tbody>
</table>

Fig. 5. Spatial distribution of (a) the FI, and (b) the GR, based on the mineralogical information.

Fig. 6. (a) and (b) represent the scatter plots with the third variable, the FI, while (c) and (d) show 3D fitted surface using the available data.
is evaluated and the fitness function, represented by the following mean-squares error (MSE), is computed for the chromosomes:

\[ \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (T_i - P_i)^2 \]  \hspace{1cm} (26)

where \( T_i \) and \( P_i \) are the target and predicted values, respectively, and \( n \) indicates the number of the training data. The MSE can take on any positive value, of course.

Next, the chromosomes (parameters) are updated based on the magnitude of the last computed MSE to obtain a better fit. Thus, the proposed MSE, Eq. (26), is evaluated in each successful run of the training. The individuals are selected, combined, and mutated in such a way that they minimize the error. Then, the new parents are replaced with the previous erroneous ones that help produce a more accurate generation. The steps that result in the final optimal network need to be repeated several times. The algorithm is summarized in Fig. 8.

The GA parameters were set in variable ranges. For example, after some preliminary calculations the range of the crossover rate was set between 0.05 and 0.95, while the mutation was between 0.01 and 0.3. Each of the probabilities was tested in a round of 54 iterations and the optimal values were used in the GA. The population size was also set to be between 12 and 60. Using the algorithm, several networks with various configurations of the GA operators were tested in order to set the final optimal parameters. For example, as demonstrated in Fig. 9, the optimal networks for the TOC and FI were obtained, respectively, in the 25th and 17th generations. Unlike the MLR, the GA algorithm recognized that not all the input parameters have to be incorporated in the model. Therefore, the final models are based on the contributions of a few of the welllog data. The resulting networks contain the optimal parameters since various configurations were tested using the GA. Such models must achieve the lowest estimation error. Thus, the following parameters were determined to provide the optimal solution:

(i) For the FI evaluation: a mutation rate of 0.15, a network with five nodes in the MFs layer, a learning rate of 0.55, and a momentum of 0.64.
(ii) For the TOC evaluation: a mutation rate of 0.12, a network with four nodes in the MFs layer, a learning rate of 0.62, and a momentum of 0.71.

The results of application of the MLR and HML algorithms are shown in Fig. 10. They both provide acceptable results for the TOC estimation, but the HML delivers predictions that are more accurate. To fit the TOC curve, the MLR tries to mimic the overall distribution of the selected variables such as, for example, the GR and DT logs, whereas the HML represents a different form, and its predictions are in very good agreement with the GR log, indicating that the proposed TOC evaluation is physically valid. The results of the FI estimation, however, represent very different variations, which are mainly due to the very complex correlation between the FI and the well logs (see also Table 1). For example, the MLR method represents a strongly biased FI curve. On the other hand, the HML method determines a more accurate curve for the FI evaluation. The results are numerically compared using the MSE in Table 2.

The distribution of the GR log versus FI is presented in Fig. 11(a), which indicates a direct relationship between the two.
variables and the TOC. Indeed, as can be seen, the TOC increases when both the GR and FI do. It also indicates significant positive correlations between the GR log and the FI. Furthermore, since the MLR method recognized strong correlation between the TOC and GR, DT and K (permeability), a more informative multidimensional representation of such relationship is depicted in Fig. 11(b). The axes represent the GR, DT and the K logs. The other two variables, namely, the TOC and FI, are shown using the size and color, respectively. Simultaneous variation of the size and color indicates the spatial dependence of the selected variables.

For the purpose of better identification of the fracable zones, the same practice was implemented on the GR log and the TOC against the FI. Then, the FI was clustered into four groups using the k-means method (MacQueen, 1967). The classification was performed using the information on the GR log, the FI and the TOC. Thus, one may use the well-log data and instantly determine whether the area under development is fracable or not. The accuracy of the method can also be verified in Fig. 13 as the probability curves represent very different distribution of the FI data.

4. Summary and conclusions

Due to their highly complex structures, shale-gas reservoirs require very accurate modeling. Vertical and lateral variability necessitate more drilling, which consequently leads to significant increase in the cost of the operations. Wireline well logging is one of the most accessible and affordable approaches to continuously monitor such complexities. Shale-gas repositories are associated with extensive/big data. Obviously, analysis of the uncertainty and risk assessment for future development would benefit from big data. However, processing and analyzing such data require advanced techniques.

Aside from acquiring more informative data using various techniques, real-time data integration and modeling is still a significant unsolved problem. Together, both the tools and data analysis techniques aim at identifying the sweet spots that are potentially favorable for successful and economic development of shale-gas reservoir. In shale-gas reservoirs 'sweet spots' are mostly defined as the zones that contain high total organic carbon (TOC) and are easily fracable, i.e., high fracable index [FI]. In this paper, two methods
were presented for simultaneously dealing with such big data and identifying the sweet spots.

The first method was based on data mining techniques that try to determine the spatial correlations between the variables using an efficient stepwise algorithm. The method fits a surface based on the well-log data and predicts both the TOC and FI. The final surface is based only on switching the input parameters and testing various coefficients. Clearly, however, the method is not able to first learn and then adaptively change the coefficients to fit a more complex and higher-order surface.

As an alternative, fuzzy logic, a useful method in machine learning techniques was selected since it can better process and analyze the complex spatial relationship between the shale data. The technique is, however, based on some user-dependent rules that need a vast dataset for shales, their geological settings, wireline well-log data interpretation, and so on, which give rise to another obstacle. From the same school of machine learning, neural networks, due to their excellent performance in providing a learning basis, were used to assist defining the knowledge-based rules. The new technique works very well when the input data are not very complex and with high dimensions. Furthermore, it also has some parameters that must be set carefully, such as the number of the input data, the number of MFs, and learning-related parameters. The data used in this study are, however, highly complex and multidimensional that require extensive trial and error calculation if “brute-force computations” are to be used. Thus, the method is integrated with the genetic algorithm in order to avoid such time demanding and cumbersome parameter adjustment. The new hybrid machine learning (HML) technique was developed, and shown to provide much more accurate predictions for the TOC and FI, when compared with those of the MLR method, and are in very good agreement with the available experimental data for both properties.

The proposed methods in this study can still not capture the whole complexity of shale reservoirs as they manifest highly nonlinear behavior. For example, the HML can hardly follow the trends for the TOC and fracable index. The soft computing method, on the other hand, requires knowledge on various aspects of artificial intelligence. Aside from all such weaknesses, for example, the soft computing method was designed in such a way to minimize the required knowledge. These techniques can be used when one requires use of minimum amount of information in order to identify the sweet spots. The HML method requires least possible effort and user knowledge for shale reservoir characterization and can be rapidly updated using new data. The other available techniques for the application of this study, however, are mostly based on trial and error. They require vast amount of knowledge in geology and well-log interpretation, whereas the proposed methods in this paper minimize such necessities.

The implication of the present work is clear. Methods that are based on trial and error are not useful for characterization of shale reservoirs, and increasing the amount of data, which can be costly, will not improve their performance. On the other hand, advanced computational and analysis techniques of the type that we have discussed and utilized in this paper offer at least two advantages. One is that one does not blindly and based on ad-hoc approaches
try to develop correlations between the various variables. The second advantage of such methods is that they identify the most pertinent variables, so that if one is to invest more resources in order to collect more data, one can concentrate on measuring and gathering information about the most relevant variables, hence reducing the financial burden of such operations.

As the future work, one might consider bringing more geological information and defining more informative inputs. There are some geological signatures that are not reordered in the logging, but can be extracted from other study, such as geophysical surveys. Developing a more robust definition for fracibility can also be studied further. As another future avenue, one can use the elastic properties, instead of the XRD data, which are more feasible and versatile.

Acknowledgements

PT thanks the financial support from the University of Wyoming for this research. FJ would like to thank the support from Nano Geosciences lab and the Mudrock Systems Research Laboratory (MSRL) consortium at the Bureau of Economic Geology, The University of Texas at Austin. MSRL member companies are Anadarko, BP, Cenovus, Centrica, Chesapeake, Cima, Cimarex, Chevron, Concho, ConocoPhillips, Cypress, Devon, Encana, Eni, EOG, EXCO, ExxonMobil, Hess, Husky, Kerogen, Marathon, Murphy, Newfield, Penn Virginia, Penn West, Pioneer, Samson, Shell, Statoil, Talisman, Texas American Resources, The Unconventionals, U.S. Intercrop, Valence, and YPF. XRD data provided by Dr. H. Rowe and wireline well-log data provided by R. Eastwood.

References


