# Machine learning regressors and their metrics to predict synthetic sonic and mechanical properties

Ishank Gupta[1], Deepak Devegowda[1], Vikram Jayaram[2], Chandra Rai[1], and Carl Sondergeld[1]


## Abstract

Planning and optimizing completion design for hydraulic fracturing require a quantifiable understanding of the spatial distribution of the brittleness of the rock and other geomechanical properties. Eventually, the goal is to maximize the stimulated reservoir volume with minimal cost overhead. The compressional and shear velocities ($V_P$ and $V_S$, respectively) can also be used to calculate Young's modulus, Poisson's ratio, and other mechanical properties. In the field, sonic logs are not commonly acquired and operators often resort to regression to predict synthetic sonic logs. We have compared several machine learning regression techniques for their predictive ability to generate synthetic sonic ($V_P$ and $V_S$) and a brittleness indicator, namely hardness, using the laboratory core data. We used techniques such as multilinear regression (MLR), least absolute shrinkage and selection operator regression, support vector regression, random forest (RF), gradient boosting (GB), and alternating conditional expectation. We found that the commonly used MLR is suboptimal with less-than-satisfactory predictive accuracies. Other techniques, particularly RF and GB, have greater predictive capabilities. We also used Gaussian process simulation for uncertainty quantification because it provides uncertainty estimates on the predicted values for a wide range of inputs. Random forest and extreme GB techniques also show low uncertainties in prediction.


## Introduction

The use of machine learning to extract meaningful and actionable information from core and log data is increasingly becoming popular. This is especially relevant to unconventional shales because of their increasing matrix heterogeneity, complex wettability, and anisotropy (Sondergeld et al., 2010a, 2010b). The concepts related to the physics of flow and theoretical models to predict rock properties in shales are still in their infancy stages, and conducting laboratory or field measurements for all the desired properties in all wells is far too expensive. Thus, machine learning is a widely used alternative that is inexpensive, rapid, and capable of identifying hidden correlations in the data that even a trained eye can miss. However, with this being said, not all correlations are true, and it requires knowledge of petrophysics and robust experience with machine learning techniques to generate real value and save millions of dollars in the process.

Several authors have published some unique applications of machine learning to oil and gas reservoirs. Kale et al. (2010) and Gupta et al. (2017a, 2017b) integrate core, log, and production data using machine learning techniques to predict sweet spots within the Eagle Ford, Wolfcamp, Woodford, and Barnett Formations. Other applications include the one reported by Hamzaban and Memarian (2008) who use a sequential clustering technique for drilling and completion parameters to forecast drilling rates and bit wear. Zhao et al. (2015) conduct a survey of classification algorithms in seismic facies classification, whereas Shirzadi et al. (2013) and Ziegel et al. (2011) also highlight the use of data science to several oil and gas applications such as optimizing pipeline inspection, production management systems, and water flooding strategies.

Schuetter et al. (2018) develop a predictive model for the first 12 months of production using several variables. They use different regression techniques such as ordinary least square regression, random forest (RF), gradient boosting (GB), support vector regression (SVR), and decision trees (DTs). Roy et al. (2015) show the excellent performance of generative topographic mapping in predicting estimated ultimate recovery from a combination of geologic, petrophysical, and completion parameters. Several other publications that have applied data mining and analytics for the assessment of unconventional resources are LaFollette et al. (2012), Bhattacharya et al. (2013), and Mohaghegh (2013). These studies cover

[1]The University of Oklahoma, S114, Sarkeys Energy Center, 100 East Boyd Street, Norman, Oklahoma 73019, USA. E-mail: ishank@ou.edu; deepak.devegowda@ou.edu; crai@ou.edu; csondergeld@ou.edu.

[2]Pioneer Natural Resources, Inc., Irving, Texas 75039, USA. E-mail: vikram.jayaram@pxd.com.

Manuscript received by the Editor 3 January 2019; revised manuscript received 3 May 2019; published ahead of production 4 June 2019; published online 30 July 2019. This paper appears in *Interpretation*, Vol. 7, No. 3 (August 2019); p. SF41–SF55, 15 FIGS., 8 TABLES.

http://dx.doi.org/10.1190/INT-2018-0255.1. 

a broad range of techniques, namely tree-based modeling, classification trees, fuzzy logic, time-series analysis, and nonparametric regression techniques.

Akinnikawe et al. (2018) predict synthetic photoelectric and unconfined strength logs from triple-combo and sonic logs using several regression techniques and show that RFs possessed the best predictive capability. Guan (2012) and Ghavami (2011) mention a strong correlation between $V_P$ and density ($\rho_b$), neutron porosity (NPHI), and shale volume ($V_{SH}$) from gamma ray and subsequently use neural networks to generate synthetic sonic logs using conventional triple combo logs ($\rho_b$, NPHI, and $V_{SH}$). Other pertinent studies in this area include prediction of the reservoir properties from measurements while drilling and wireline logs (Bhatt, 2002) and synthetic geomechanical logs (Eshkalak et al., 2014). The studies cited above provide a good understanding of machine-learning-based predictive modeling; however, it is also important to mention that there is a relative lack of discussion centered around the goodness-of-fit and uncertainty quantification.

A large percentage of the studies described earlier and in the current study fall under the broader category of predictive modeling. The two major predictive modeling methods are classification and clustering (James et al., 2013). Classification is supervised learning where the inputs and outputs are known, and a model is trained to predict the output. It includes techniques such as logistical regression (Walker and Duncan, 1967), DTs (Rokach and Maimon, 2008), RF (Breiman, 2001), GB (Breiman, 1997), and artificial neural networks (Schmidhuber, 1992, 2015). Clustering, however, is a form of unsupervised learning for finding structure in the data, commonly to identify groupings within the data based on some similarity measure. $K$-means (Forgy, 1965), hierarchical clustering (Johnson, 1967; Rokach and Maimon, 2005), and density-based clustering (Ester et al., 1996; Jain et al., 2004) are a few clustering algorithms.

This study focuses on two important use cases of machine learning regression techniques. The first application is to predict sonic velocities using other petrophysical variables such as the bulk density, porosity, and mineralogy. This is important because sonic velocities ($V_P$ and $V_S$) are critical for fracture design and modeling the spatial variations in geomechanical properties.

The second application is to predict brittleness indicators such as hardness. Ma and Zoback (2017) show that in a horizontal well, mineralogy controls the proppant placement and hydraulic fracture-related microseismic activity and conclude that there are preferred locations for the placement of fracture stages. The preferred locations can be determined using hardness because it is an indicator of brittleness (Gupta et al., 2018a). Higher hardness corresponds to brittle quartz-rich zones, whereas lower hardness corresponds to total organic carbon (TOC) and clay-rich ductile zones. Maximizing the SRV is strongly linked to optimal fracture stage placement and growth.

## Regression techniques

The data in this study were obtained from 660 as-received core samples from 20 wells in the three shale formations, namely Eagle Ford, Wolfcamp, and Woodford. Cylindrical horizontal plugs (1″ by 1″, parallel to bedding) were cored at all the 660 locations and dried in the oven at 100°C before performing petrophysical measurements. A diverse set of petrophysical measurements such as mineralogy (Sondergeld and Rai, 1993; Ballard, 2007), TOC (Law, 1999), nanoindentation measurements for Young's modulus and hardness (Hay and Pharr, 2000; Shukla et al., 2013), porosity (Karastathis, 2007), and ultrasonic velocities (Mattaboni and Schreiber, 1967; Junck and Benson, 1973) are measured on these horizontal plugs.

The different measurements were processed to remove outliers and normalized using the $z$-score transform to remove any bias related to the magnitude of the petrophysical measurement. The data were also randomly divided into training (90%) and test (10%) data sets. Different regression techniques were applied to the training data set to predict sonic velocities and hardness. Multilinear regression (MLR) (Yan, 2009) is the simplest form of regression with multiple predictors (independent variables $x_1, x_2, x_3, \ldots x_n$) and a single response (dependent variable $y$). The equations for MLR are given below, where $\epsilon$ is the error, and $\hat{y}$ is an estimate of the dependent variable ($y$):

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + \epsilon. \quad (1)$$

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n. \quad (2)$$

$$\epsilon = y - \hat{y}. \quad (3)$$

Multilinear regression assumes a linear relationship between the predictors and the response. The error is also assumed to be normally distributed, independent of the predictors and has a constant variance. The regression is performed using a training data set where the predictors and the response variables are known. The values of the regression coefficients ($\beta_1$, $\beta_2$, $\beta_3$, etc.) are determined by minimizing the prediction errors on the training data set. The quality of the regression is determined using several metrics. The two most commonly used are $R^2$ and the root-mean-square (rms) error. Solving for the unknown $\beta$s in MLR is equivalent to maximizing the value of $R^2$ and minimizing rms error. The equations to calculate both are given below, where $\bar{y}$ is the average value of the response variable:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}. \quad (4)$$

$$\text{rms error} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}. \quad (5)$$

Linear regression has several disadvantages such as being unduly influenced by outliers and, as the name suggests, being restricted to linear models. Although the analysis of variance can be used to control model complexity, overfitting to the data is often a real problem. Several other techniques have overcome some or all these shortcomings and are used in this study. These include least absolute shrinkage and selection operator (LASSO) regression, alternating conditional expectation (ACE), SVR, DT, RF, and GB.

The LASSO (Tibshirani, 1996) technique controls the values of the parameters ($\beta$s) to mitigate overfitting or issues with collinearity when model parameter estimates become unusually inflated. By constraining the parameter estimates, the LASSO technique prevents overfitting and induces resistance to collinear variables. The expression minimized in the LASSO technique is given below, where $\lambda$ is a tuning parameter to scale the penalty, and $p$ is the number of predictors:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}\hat{\beta}_{\mathrm{j}}. \tag{6}$$

Alternating conditional expectations (Breiman and Friedman, 1985) is a nonparametric regression technique that determines optimal transformations of the dependent and independent variables, followed by regressing the optimal transforms to maximize the correlation between the dependent and independent variables. The transformations can be fitted with polynomials to develop the final regression equation.

Support vector regression (Drucker et al., 1997; Smola and Scholkopf, 2004) is a very powerful and highly flexible technique that is not limited to linear models and is robust to outliers. The loss function ignores points with small residuals (based on a predefined threshold $\epsilon$), whereas large residuals contribute linearly. Support vector regression is well suited for linear models (using linear kernels) or nonlinear prediction (using polynomial or radial kernels) and therefore has a higher degree of flexibility. The expression minimized in SVR is given below, where $c$ is the cost parameter and $L_e$ is the loss function:

$$c\sum_{i=1}^{n}L_e(y_i-_i)+\sum_{j=1}^{p}\hat{\beta}_j^2. \tag{7}$$

$$L_e(y_i-\hat{y}_i)=\left\{\begin{array}{ll}0 & \text{if } y_i-\hat{y}_i<\epsilon\\ |y_i-\hat{y}_i|-\epsilon & \text{otherwise}\end{array}\right\}. \tag{8}$$

A DT (Rokach and Maimon, 2008) is a hierarchical structure that consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a target variable. It is particularly useful for data with several predictors of unknown importance. The issues with DT are that they are often insufficient for numerical prediction (or regression). The predictions for continuous variables are often discrete values, and even very complex trees may not be able to provide acceptable prediction accuracies. Decision trees are highly sensitive (less robust) to the data and noise, and small changes in the training data lead to different DTs. The RF (Breiman, 2001) and GB (Breiman, 1997) techniques are a powerful refinement of the DT algorithm and are more robust. Both of these techniques involve creating several independent DTs and then averaging them to improve prediction.

The RF is a bagging algorithm (Breiman, 2001) that resamples the data to make the resulting models more robust, stable, or reliable. Only a subset "$m$" of the "$p$" predictors is used at any specific point to split a given tree, and the choice of "$m$" predictors varies across different trees, thereby inducing a degree of randomness and increasing the robustness of the predictions. However, boosting relies on the use of multiple models (bagging) that are trained to minimize errors made by previous models (boosting). Gradient boosting (Breiman, 1997) is one of the most popular algorithms, which creates successive trees that are sensitive to the residuals of the previous tree. This allows the trees to become sensitive to predictors with large residuals. Boosting is quite popular because of the ease of generalization of the loss functions, even when the derivative is not convex. The RF and GB techniques significantly improve the regression quality and model stability over the DT. They also have several built-in tuning parameters to control the degree of fit and customization based on experience.

## Uncertainty quantification

A Gaussian process produces a mean estimate with a measure of uncertainty. A Gaussian process fitting models the regression function in a probabilistic way by defining a distribution of the functions. The Gaussian distribution of variables calculated from the training data (posterior probability density functions) can be used to calculate the distribution of functions (Kopp, 2018). For instance, Figure 1 shows an example of Gaussian process fitting. The gray-shaded region represents the region of uncertainty.

Mathematically, the Gaussian process describes a distribution over functions, where $m$ is the mean function, and $R$ is the covariance function:

$$f=\mathrm{GP}(m,R). \tag{9}$$

$$m(x)=E[f(x)]. \tag{10}$$

$$R(x,x')=E[(f(x)-m(x))(f(x')-m(x'))]. \tag{11}$$

A function is basically an infinitely long vector; thus, a Gaussian process is a multivariate normal with an in-

finite number of dimensions. According to the property of multivariate Gaussians, a subset of the dimensions will also have a joint Gaussian distribution. This implies that we can calculate mean and covariance functions for the training data and then sample from the respective joint Gaussian distribution to calculate the uncertainty in the regression function.

The covariance function ($R$) is a measure of similarity between $x$ and $x'$. For this to be possible, $R$ must be symmetric and positive definite. The default covariance function used in the Gaussian process fitting is the squared exponential kernel:

$$R(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^T (x - x')}{2l^2}\right), \qquad (12)$$

where $l$ (length scale) and $\sigma$ (variance) are hyperparameters that dictate the range of uncertainty predicted by Gaussian process fitting. In this study, we use the GPfit package in R to fit the Gaussian simulation process. For further details, we refer readers to MacDonald et al. (2015) and Ranjan et al. (2011).

## Study area and methodology

Various core measurements were available at 660 locations in 20 wells in the three different shale plays, i.e., Eagle Ford, Woodford, and Wolfcamp. Table 1 shows the produced oil, gas, and other reservoir properties for the different shale plays. Eagle Ford is late Cretaceous in age. It has four different basins, namely Maverick, Hawkville, San Marcos, and East Texas. Traditionally, it is believed that the East Texas basin is not a part of the Eagle Ford play due to its very high clay content compared to that of the other basins in Eagle Ford. In general, the Eagle Ford shale play is rich in carbonates. The Woodford Formation, on the other hand, is a Devonian-Mississippian shale located in the Anadarko, Arkoma, and Ardmore basins in Oklahoma and Texas. The common lithologies occurring in the Woodford Formation are black shale, chert, sandstone, siltstone, and dolostone. The most productive lithologies are siliceous. Siliceous Formations in the Woodford are highly brittle and contain natural fractures. Finally, the Wolfcamp Formation located in the Permian basin in Texas and New Mexico is a late Cretaceous shale. A lithologic description of the Wolfcamp Formation in the Midland basin is given by Cortez (2012). Wolfcamp consists of four different facies, namely siliceous mudrock, calcareous mudrock, muddy carbonate-clast conglomerates, and skeletal grainstone. Siliceous mudrock has an average clay content of 40%. The clay type is mainly illite. The remaining mineralogy consists of carbonate, quartz, feldspar, pyrite, and apatite. The carbonate content is generally less than 20 wt%.

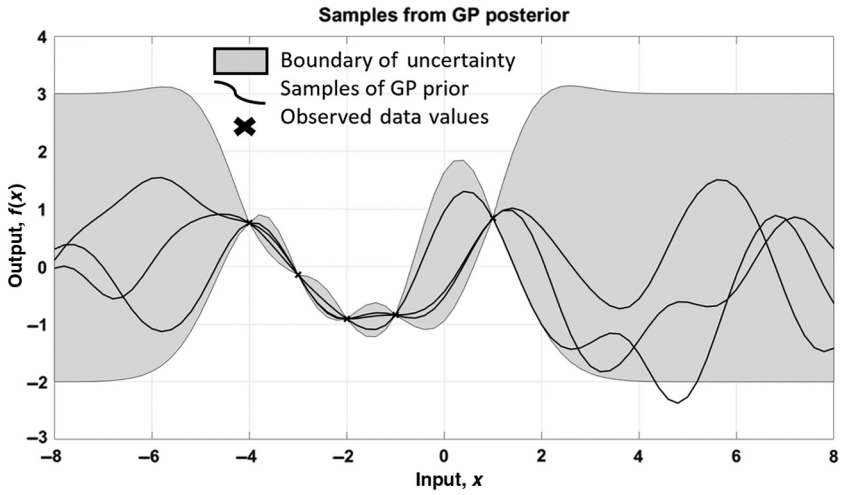



**Figure 1.** Gaussian process fitting for a 1D function. The gray-shaded region shows the range of uncertainty. The uncertainty reduces where actual data points are available (shown by the cross marks).

The horizontal plugs were used for measuring the petrophysical properties. The plugs were cut from as-received core slabs and dried at 100°C before performing the measurements. The first step following data collection was exploratory data analysis using histograms and boxplots (Figure 2) to understand the distribution of individual variables and to visually identify the outliers. We used the EnvStats package (Barnett and Lewis, 1995) in R to detect the outliers prior to bivariate analysis. The Rosner's test function was used to detect the outliers. The function performs the Rosner's generalized extreme Studentized deviate test to identify potential outliers in a data set, assuming the data without any outliers are obtained from a normal

**Table 1. Production figures and other reservoir properties for different shale plays.**

| Formation | Oil produced (bbl) | Gas produced (Tcf) | Thickness (ft) | Depth (ft) | Pressure gradient (psi/ft) |
|---|---|---|---|---|---|
| Eagle Ford | 1.50 | 4.2 | 150–420 (CLR, 2010) | 7000–12,000 (CLR, 2010) | 0.40–0.70 (CLR, 2010) |
| Woodford | 0.09 | 4.6 | 150–400 (CLR, 2010) | 4800–10,000 (CLR, 2010) | 0.60–0.65 (CLR, 2010) |
| Wolfcamp | 0.96 | 4.2 | 2000 (Wilson et al., 2016) | 5500–11,000 (Pioneer, 2013) | 0.55–0.70 (Pioneer, 2013) |

Note: The production numbers were obtained from drilling information (checked 16 December 2016).

(Gaussian) distribution. These detected outliers can be removed using the outliers package, which removes the points farthest away from the sample mean. After outlier removal, data from 600 sampled locations were considered for further analysis. The next step was to assess the relationships between the different input variables. Figures 3 and 4 show examples of various crossplots among sonic $V_P$, hardness, and several petrophysical properties. Based on these crossplots, a few variables were chosen to predict sonic velocities ($V_P$ and $V_S$) such as bulk density, porosity, TOC, clays, and carbonate content. Similarly, variables chosen to predict hardness were ultrasonic velocity ($V_P$), clays, carbonates, TOC, and porosity.

Figures 3 and 4 also suggest that data from different formations cannot be treated as one population with similar characteristics. In this study, we use clustering with $K$-means to divide the data set into different groups for improving the regression fit quality. Detailed steps for clustering with case studies from the Eagle Ford, Woodford, Wolfcamp, and Barnett Formations have been presented in Kale et al. (2010) and Gupta et al. (2017a, 2018b).

In the next step, we apply principal component analysis (PCA) to the input data to convert the set of highly correlated input variables into a set of uncorrelated variables without affecting the information (variance) carried by the data. For instance, ultrasonic velocity is predicted in this study using TOC, porosity, clays, carbonates, and bulk density. These properties are correlated with one another and consequently impact the regression. The uncorrelated predictors also called the principal components can be subsequently used for regression analysis instead of the original predictors (TOC, porosity, clays, carbonates, and bulk density). Again, detailed steps and procedures for PCA are provided in Kale et al. (2010), Gupta et al. (2017a), and Gupta et al. (2018b).

After preprocessing (grouping and PCA), the data are ready for regression analysis using different techniques such as MLR, LASSO regression, SVR, RF, GB, and ACE. The techniques are compared using several error metrics such as $R^2$, adjusted $R^2$, and rms error.

## Results and Discussion

The clustering results are shown in Figure 5. Cluster 1 is rich in clays and has the highest porosity and TOC. A higher TOC is indicative of higher source rock potential, but it also makes the rock type more ductile. Cluster 2 is rich in carbonates, whereas cluster 3 is rich in quartz and is the most brittle rock type, thus making it a good candidate for fracturing. The $K$-means clustering uses TOC, porosity, clays, carbonates, and quartz content as inputs. The variable selection process for clustering and detailed discussion of the results is given in Gupta et al. (2017a, 2018b). Regression analysis is performed separately for each cluster.

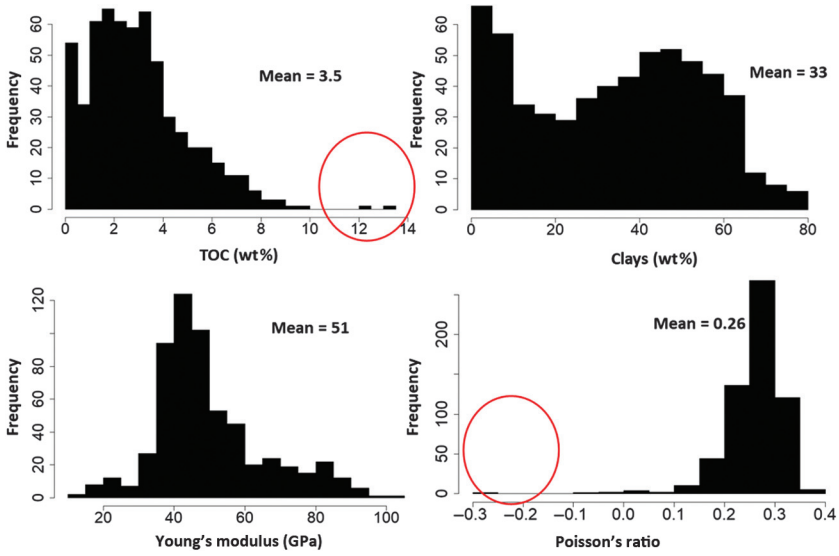



**Figure 2.** Histograms of four different variables to see the distribution and identify the outliers. The red circles represent the outliers.

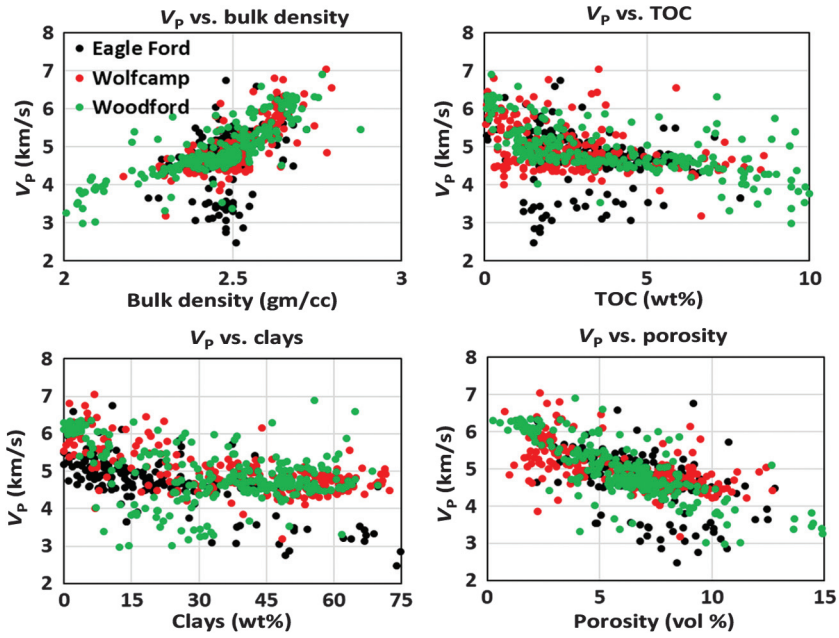



**Figure 3.** Crossplots showing the correlation between the ultrasonic velocity ($V_P$) and several petrophysical laboratory measurements made on horizontal plugs. The crossplots show that $V_P$ and bulk density are directly proportional, whereas higher clay content, TOC, and porosity lead to lower $V_P$.

The goal of regression is to predict all the three velocities in terms of other petrophysical variables, namely the bulk density, porosity, clays, carbonates, and TOC. The data were randomly split into 90% for training and 10% for blind tests. Figure 6 shows the regression results for the different clusters, with RF performing the best. GB and extreme GB techniques also provide reasonably good predictions. Table 2 gives details of the tuning parameters for the different regression techniques used for predicting $V_P$. The tuning parameters were different for each cluster. The results are shown only for cluster 1 as an example.

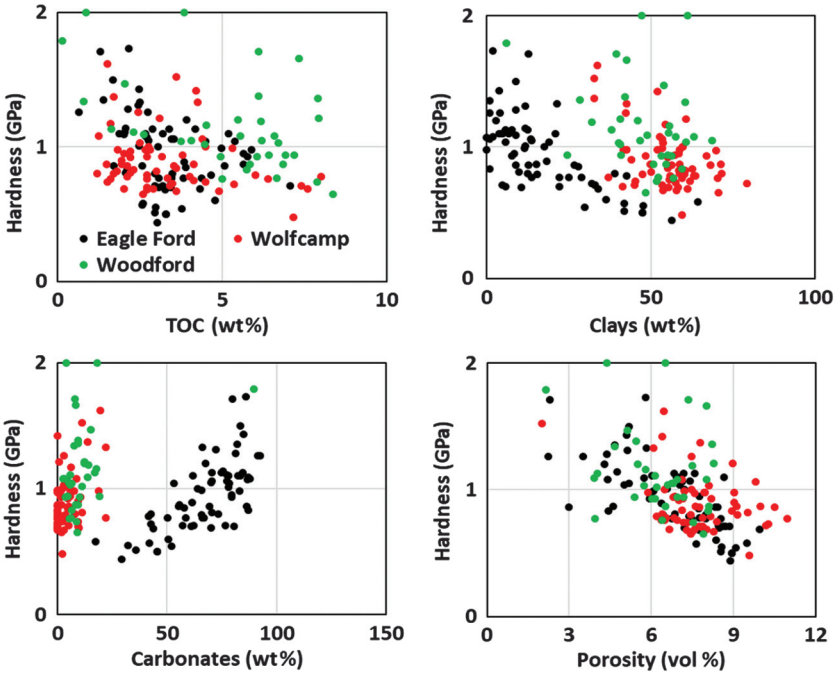



**Figure 4.** Crossplots showing the correlation between hardness and several petrophysical measurements in the laboratory on horizontal plugs. The crossplots show that in the Eagle Ford, the increasing carbonate content leads to higher hardness. In all three shale plays, higher clay content, TOC, and porosity show a lower hardness. It is also evident, particularly from clays and carbonate crossplots, that different formations exhibit different trends.

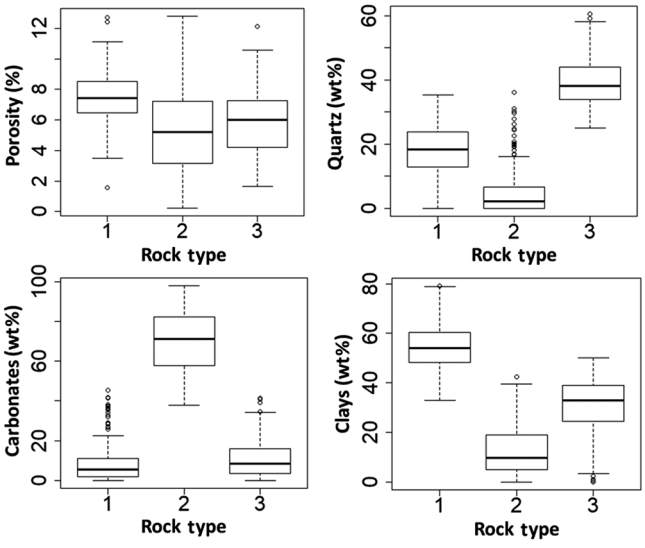



**Figure 5.** Properties of the different clusters created through $K$-means. Cluster 1 is rich in clays and has the highest porosity. Clusters 2 and 3 are rich in carbonates and quartz, respectively.

The tuning parameters allow users to guide the regression to avoid overfitting the data and improve the prediction quality by inducing an optimum amount of complexity in the model. For instance, for the GB method, n.trees represents the total number of trees to be fit. It is equivalent to the total number of iterations. Similarly, n.minobsinnode represents the minimum number of observations in the terminal nodes. The number should be sufficiently high to avoid overfitting to the data. For other tuning parameters, readers are referred to the respective R packages (Table 2) that explain the tuning parameters and explain basic constraints on the values of these parameters.

Qualitatively, Figure 6 shows that the RF technique provided the best prediction. Several quantitative metrics were also calculated to compare the different regression techniques. The metrics are shown in Figure 7. The adjusted $R^2$ is a modified version of $R^2$ that has been adjusted for the number of predictors in the model. The adjusted $R^2$ increases only if the new term improves the model more than what expected by chance. It decreases when a predictor improves the model less than what is expected by chance. On the basis of $R^2$, adjusted $R^2$, and rms error, it is evident that techniques such as RF and extreme GB perform much better than the other techniques.

The different regression techniques discussed thus far can sometimes overfit the data. Several steps were taken to detect and prevent overfitting of the data. First, 10% of the data was used for blind tests, and the blind test data showed a fair agreement with the regression model-predicted values. Figure 8 shows an example of the RF technique. Second, the hypertuning parameters were carefully adjusted to restrict overfitting. For instance, the number of trees or number of iterations (n.trees in GB) and the minimum number of observations in the terminal node (n.minobsinnode in GB) were constrained to prevent overfitting. The maximum number of iterations was limited to 200, considering the limited training data set (600 sampled depths). Moreover, the minimum number of observations in a terminal node was fixed at 5 to avoid creating multiple nodes without adequate representation in the data set. Finally, a 10-fold cross validation was used for regression analysis. The general term for this is $k$-fold cross validation in

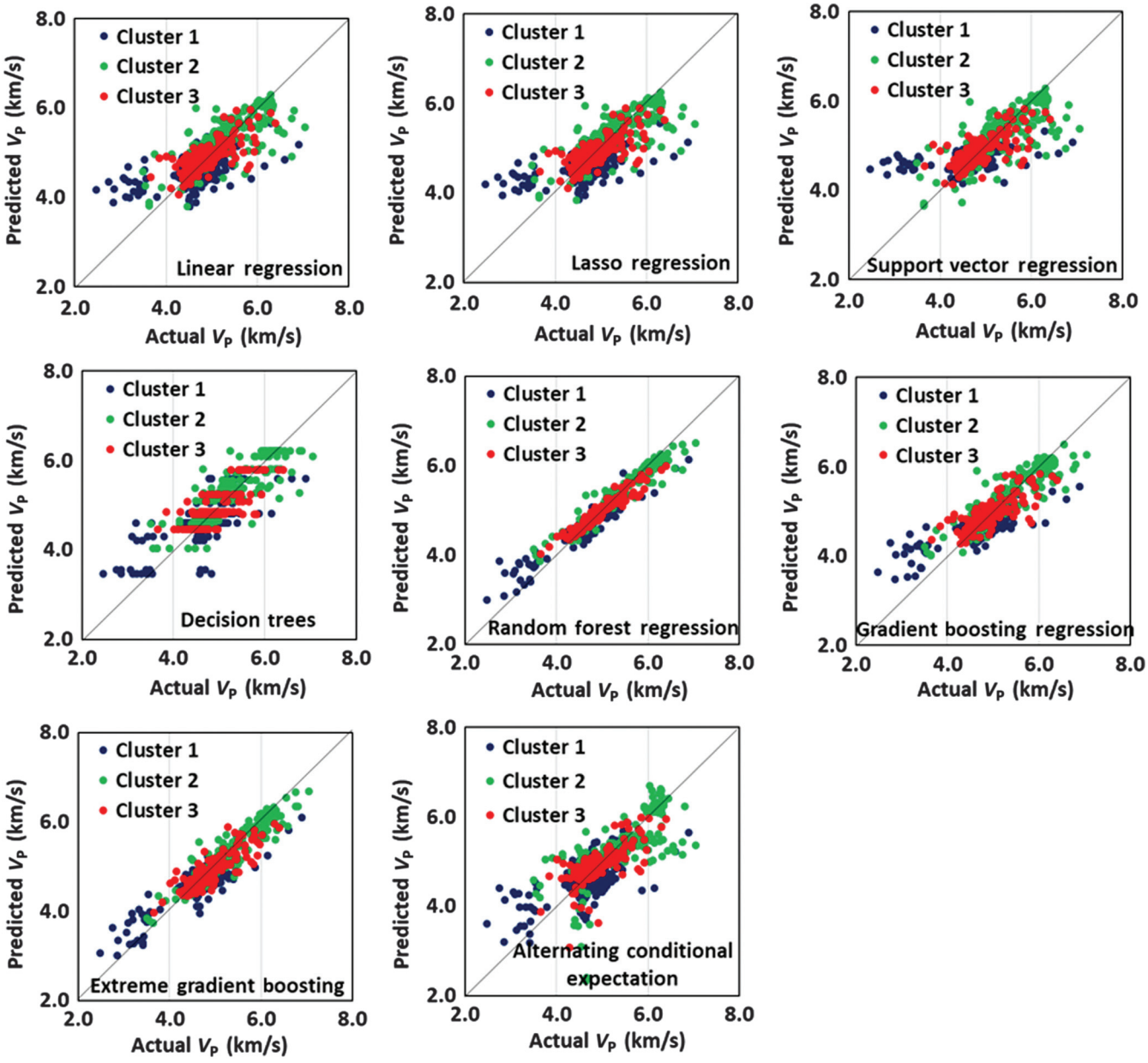



**Figure 6.** Regression results to predict ultrasonic velocity ($V_P$) for different clusters. It is evident that the RF technique provides the best prediction among the eight different regression techniques used.

**Table 2. Tuning parameters for different regression methods used for predicting $V_P$.**

| Method | R packages | Tuning parameter | Tuned values ($K$-means, cluster 1) |
|---|---|---|---|
| LASSO regression | glmnet | Alpha, lambda | 0.01, 0.03 |
| SVR | e1071, kernlab | Cost | 0.23 |
| DTs | Rpart, party | Complexity parameter (cp) | 0.03 |
| RF | randomForest | mtry | 3 |
| GB | Adabag, gbm | n.trees, shrinkage, interaction.depth, n.minobsinnode | 120, 0.11, 1, 7 |
| Extreme GB | xgboost | nrounds, max_depth, eta, gamma, min_chile_weight, subsample | 160, 7, 0.1, 0.5, 1, 0.3 |
| ACE | Hmisc | N/A | N/A |

which the data (90% of the original data in this case) are randomly partitioned into $k$ equal-sized subsample sets. Of the $k$ subsamples, a single sample set is retained as the validation data to test the model and the remaining $k-1$ subsample-sets are used as training data. This process is repeated several times, and the resulting models are averaged. Because the final regression model is an average of several models that fit different parts of the same data set, it prevents overfitting. The three steps together can help detect and restrict overfitting.

The input ($x$) for the Gaussian process fitting is the $V_P$ predictions from the different regression models. The output variable ($y$) for the Gaussian process was the actual measured $V_P$ values in the laboratory. The input and output variables were normalized by the maximum value of $V_P$ measured in the laboratory as the Gaussian process fitting works for input values varying between 0 and 1. We use the exponential kernel with default hyperparameters to compare the results across different models. The parameters used to fit the Gaussian process to different models are listed in Table 3. Figure 9 shows the results for the Gaussian process fitting. In the figure, the blue curve shows the average fit and the red curves represent the uncertainty in prediction (±2*mean-square error). It is evident that RF and extreme GB techniques also provide the lowest uncertainty in prediction.

**Table 3. Parameters used for Gaussian process fitting using the GPfit package in R. Control is a vector of the three tuneable parameters used to optimize the deviance.**

| Control | Nug_thresh | Maxit | Exponential |
|---|---|---|---|
| 200, 80, 2 | 20 | 100 | Power = 1.95 |

Note: Nug_thresh parameter is outlined in Ranjan et al. (2011), and it is used to find the lower bound of the nugget. Maxit represents the maximum number of iterations. In this study, a power exponential correlation function with power = 1.95 was used for Gaussian fitting.

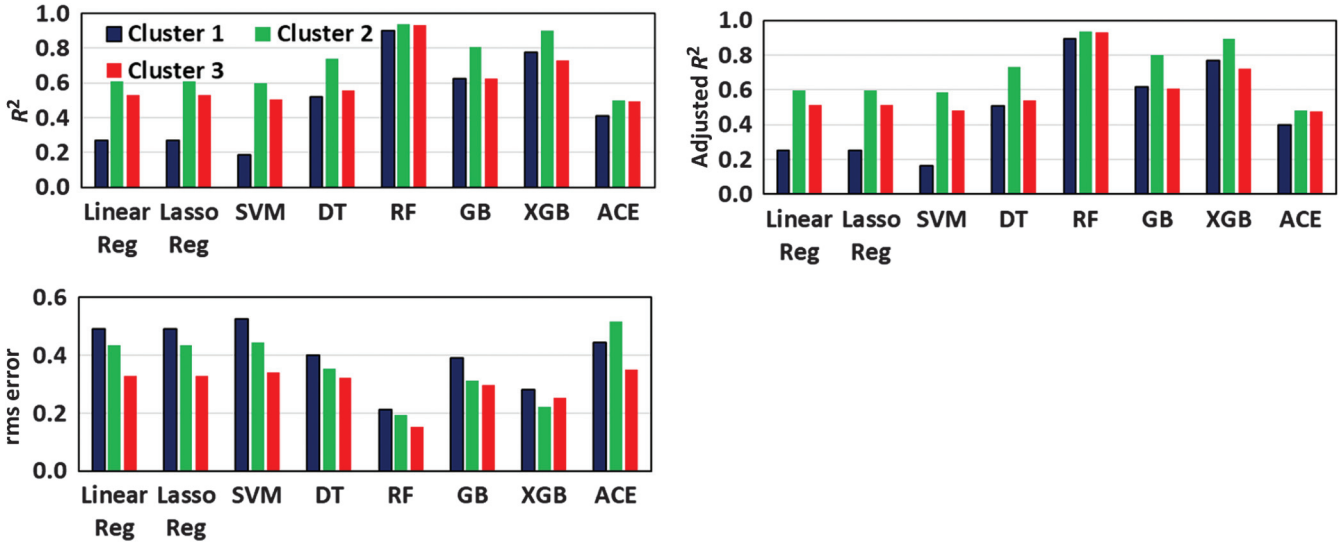



**Figure 7.** Quantitative metrics for the different regression techniques used to predict $V_P$. The results are shown for different clusters derived using K-means. On the basis of $R^2$, adjusted $R^2$, and rms error, it is evident that techniques such as RF and extreme GB perform much better than the other techniques.

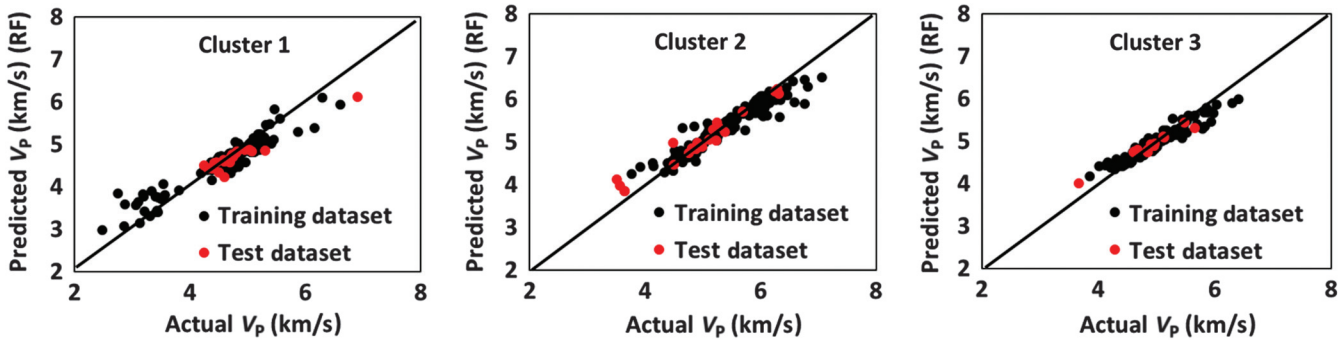



**Figure 8.** Comparison of the predicted sonic $V_P$ versus actual $V_P$ values for the training and test (blind) data sets for the different clusters. The predictions were made using the RF regression model.

The same approach was used for predicting the slow shear wave, fast shear wave, and hardness. Figures 10, 11, and 12 show the prediction results for the different variables. The RF technique and extreme GB techniques yielded the best predictions. Thus, only the linear regression (the base case), RF, and extreme GB regression results are shown to avoid redundancy. Tables 4, 5, and 6 show the tuning parameters for the different regression variables and methods. The results are shown only for cluster 1 as an example.

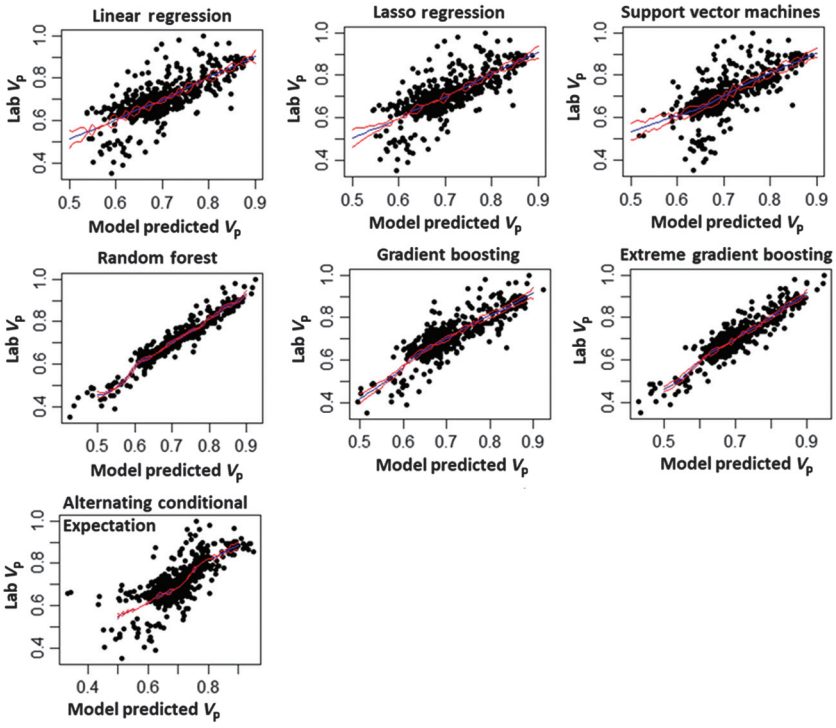



**Figure 9.** The results of Gaussian process fitting for the different model predictions. The $y$-axis corresponds to the actual $V_P$ values measured in the laboratory, and the $x$-axis corresponds to the model-predicted $V_P$ values for the different models. The input and output variables were normalized by the maximum value of $V_P$ measured in the laboratory as Gaussian process fitting works for input values varying between 0 and 1. The blue curve shows the average fit, and the red-dashed curves represent the uncertainty in prediction (±2*mean-square error). It is evident that RF and extreme GB techniques provide the lowest uncertainty in prediction.

In this study, we used 90% of the data as a training data set and 10% of the data as a test data set. A sensitivity study was also carried out to determine the goodness of fit for different splits between the training and test data sets. After outlier removal, data from approximately 600 sampled locations were left for analysis. The outliers were removed using the Rosner's test function in EnvStats package in R. Five different scenarios were considered where the above data set was split into training and test data sets in the ratios 0.1:0.9, 0.3:0.7, 0.5:0.5, 0.7:0.3, and 0.9:0.1. Figure 13 shows the results for predicting $V_P$ using the RF technique. The RF technique was used, as it provided the best predictions on the current data set and therefore would be ideal to study the goodness of fit for different splits. In addition to $R^2$, adjusted $R^2$, and rms error, Akaike information criterion (AIC), and Bayesian information criterion (BIC) are also important metrics of goodness of fit. AIC and BIC are calculated from the log-likelihood, and they represent the information lost by using a specific regression model to represent the actual data. Thus, the best-fit models will lose the least amount of information in the data; i.e., smaller values of AIC/BIC represent the better fit quality. AIC and BIC balance the tradeoffs between complexity of a given model and its goodness of fit, which is the statistical term to describe how well the model "fits" the data or the set of observations. Thus, AIC and BIC impose a penalty on complex models and prevent overfitting of the data. Sometimes, this is also considered a disadvantage that AIC and BIC prefer simple models compared to complex models, thereby preventing the best fit of the data. The equations to calculate AIC and BIC are given below:

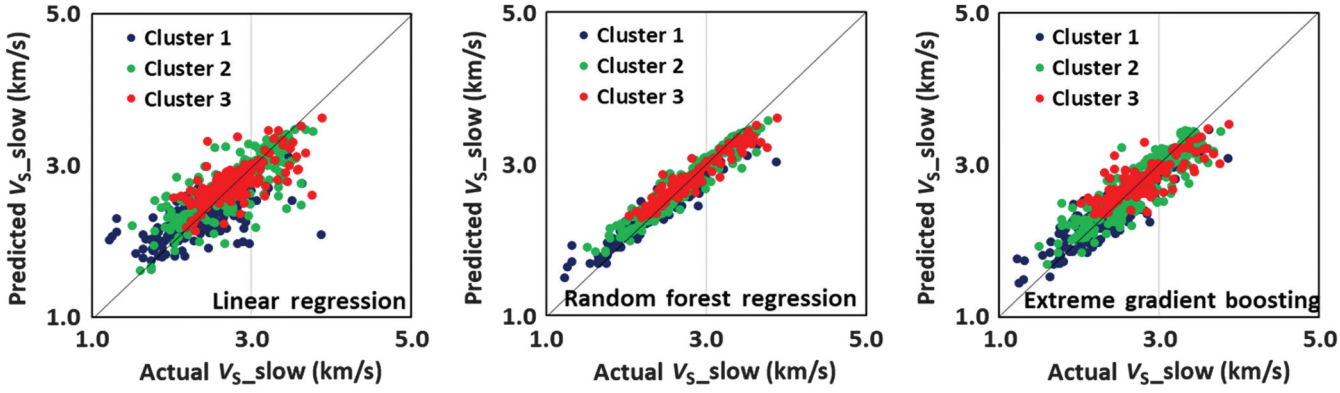



**Figure 10.** Prediction results for the slow S-wave. Linear regression, RF, and extreme GB regression results are shown to avoid redundancy. RF regression technique provides the best prediction.

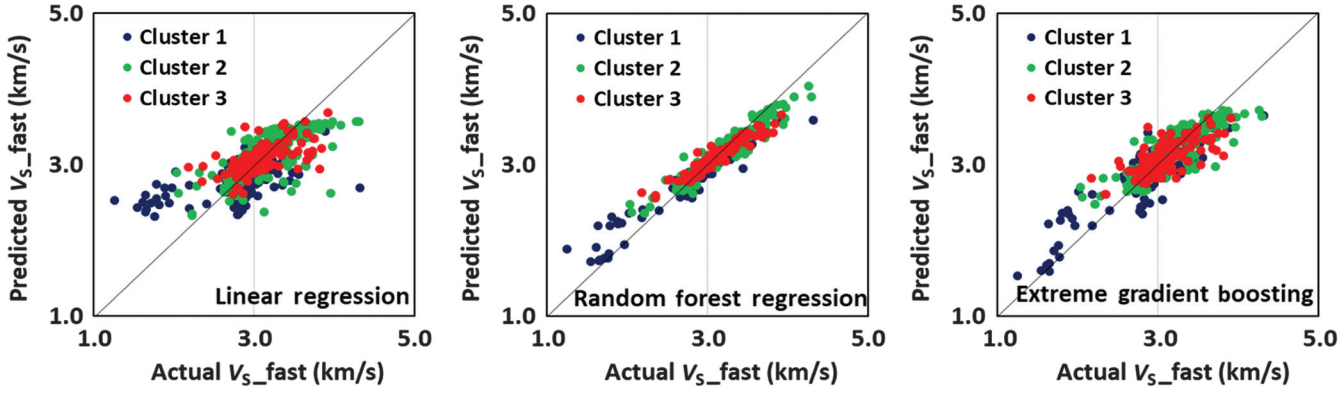



**Figure 11.** Prediction results for the fast S-wave. RF and extreme gb regression show better results than linear regression.

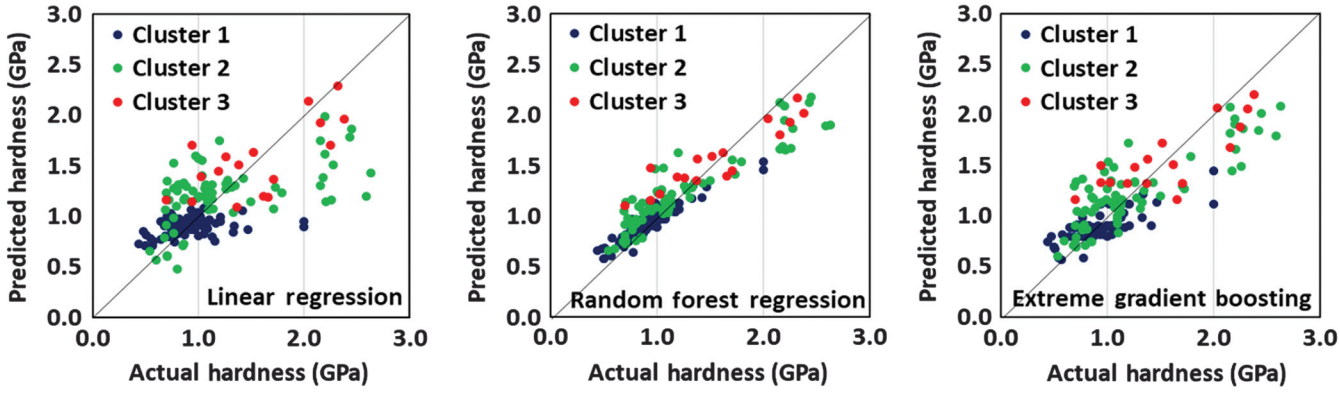



**Figure 12.** Prediction results for the hardness. Random forest and extreme gradient boosting regression show better predictions than linear regression.

**Table 4. Tuning parameters for different regression methods used for predicting $V_S$_slow.**

| Method | R packages | Tuning parameter | Tuned values (*K*-means, cluster 1) |
|---|---|---|---|
| LASSO regression | glmnet | Alpha, lambda | 0.01, 0.01 |
| SVR | e1071, kernlab | Cost | 0.02 |
| DTs | Rpart, party | Complexity parameter (cp) | 0.005 |
| RF | randomForest | mtry | 2 |
| GB | Adabag, gbm | n.trees, shrinkage, interaction.depth, n.minobsinnode | 120, 0.11, 1, 7 |
| Extreme GB | xgboost | nrounds, max_depth, eta, gamma, min_chile_weight, subsample | 100, 7, 0.1, 0.5, 1, 0.4 |
| ACE | Hmisc | N/A | N/A |

**Table 5. Tuning parameters for different regression methods used for predicting $V_S$_fast.**

| Method | R packages | Tuning parameter | Tuned values (*K*-means, cluster 1) |
|---|---|---|---|
| LASSO regression | glmnet | Alpha, Lambda | 0.03, 0.01 |
| SVR | e1071, kernlab | cost | 0.29 |
| DTs | Rpart, party | Complexity parameter (cp) | 0.0065 |
| RF | randomForest | mtry | 3 |
| GB | Adabag, gbm | n.trees, shrinkage, interaction.depth, n.minobsinnode | 100, 0.06, 2, 5 |
| Extreme GB | xgboost | nrounds, max_depth, eta, gamma, min_chile_weight, subsample | 180, 5, 0.1, 0.5, 1, 0.7 |
| ACE | Hmisc | N/A | N/A |

$$\mathrm{AIC} = 2k - 2\ln(L), \tag{13}$$

$$\mathrm{BIC} = k\ln(n) - 2\ln(L), \tag{14}$$

where $n$ is the number of observations (test observations), $k$ is the number of estimated parameters that is the number of predictors $(p) + 1$, and $L$ is the log-likelihood function. Log-likelihood is a function of the parameters of a given statistical (regression) model, for a given specific observed data set. It is defined by the following equation, where $r_i$ is the residual for a given observation and a given regression model. In this study, AIC and BIC were calculated using equations 13, 14, and 15. Although the implementation of AIC and BIC can be directly applied to MLR, for complex tree-based and neural network regression techniques, the implementation can be modified to include the model complexity beyond the number of predictors for better accuracy. For instance, Gao and Jojie (2016) discuss how AIC can be modified for deep neural networks:

$$\mathrm{Log}(L) = \frac{-n}{2}(\log 2\pi + \log\sum_{i=1}^{n} r_i^2 - \log n + 1). \tag{15}$$

**Table 6. Tuning parameters for different regression methods used for predicting hardness.**

| Method | R packages | Tuning parameter | Tuned values ($K$-means, cluster 1) |
|---|---|---|---|
| LASSO regression | glmnet | Alpha, Lambda | 0.04, 0.09 |
| SVR | e1071, kernlab | cost | 0.5 |
| DTs | Rpart, party | Complexity parameter (cp) | 0.001 |
| RF | randomForest | mtry | 3 |
| GB | Adabag, gbm | n.trees, shrinkage, interaction.depth, n.minobsinnode | 200, 0.01, 1, 5 |
| Extreme GB | xgboost | nrounds, max_depth, eta, gamma, min_chile_weight, subsample | 100, 6, 0.3, 0.5, 1, 0.7 |
| ACE | Hmisc | N/A | N/A |

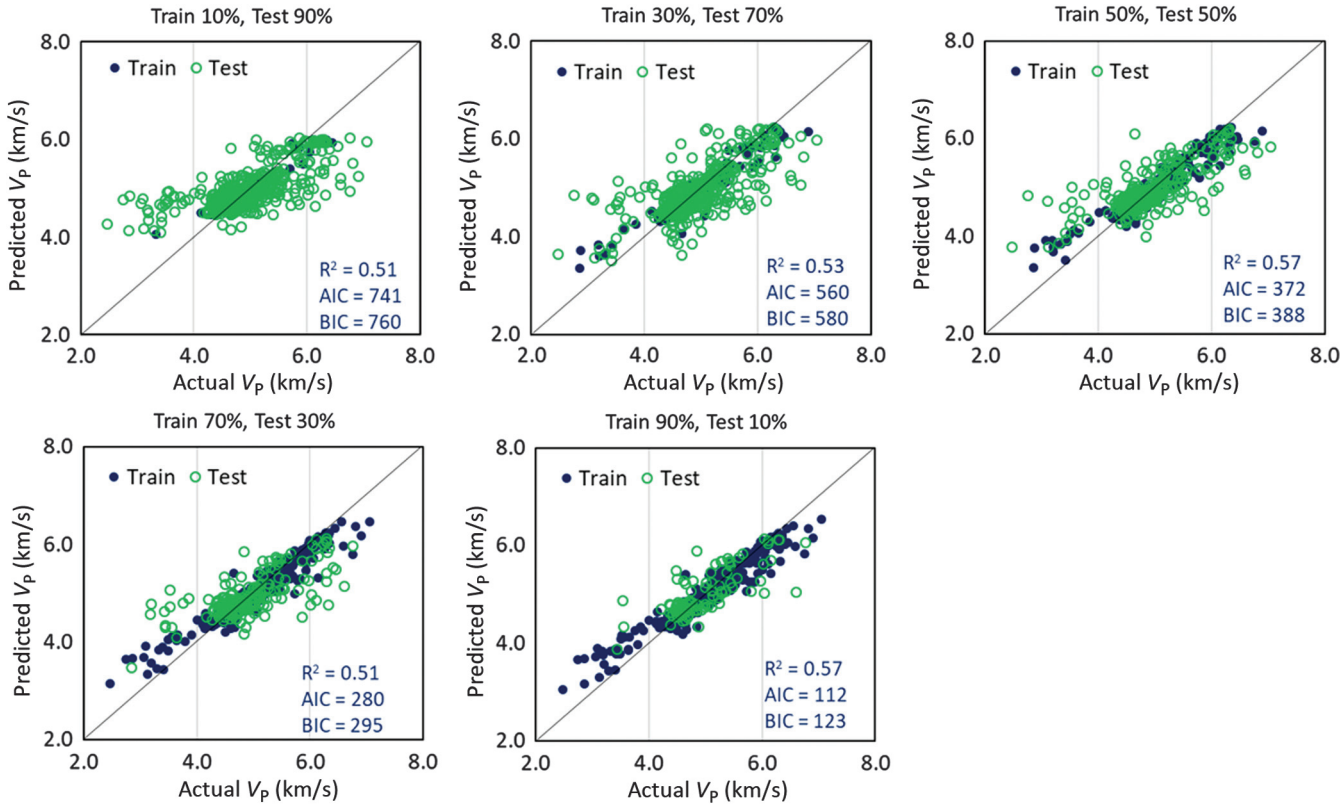



**Figure 13.** Sensitivity analysis was performed on a fraction of the training and test data sets. The figure shows the results for different fits, along with values of different metrics on the test data set, namely $R^2$, AIC, and BIC. It is evident that the $R^2$ values are very similar, but AIC and BIC keep decreasing as we increase the fraction of the training data set. This suggests that the goodness of fit increases with a larger training data set even though it is not explicitly apparent if we look at $R^2$ alone.

Figure 13 shows the results for different fits, along with values of different metrics on the test data set, namely $R^2$, AIC, and BIC. It is evident that the $R^2$ values are very similar, but AIC and BIC keep decreasing as we increase the fraction of the training data set. This suggests that the goodness of fit is increasing with the larger training data set even though it is not explicitly apparent if we look at $R^2$ alone.

The predicted sonic velocities are of great value and can be very useful for characterizing the reservoir rock and the mechanical properties of the rock, therefore, making decisions regarding hydraulic fracture design considerations. Figure 14 shows a possible application in which the Young's modulus can be calculated from predicted sonic velocities using the Birch (1961) equation as shown below:

$$E = \frac{\rho V_S^2 (3V_P^2 - 4V_S^2)}{V_P^2 - V_S^2}. \quad (16)$$

Figure 14 shows the comparison between the Young's modulus calculated from sonic velocities with the actual Young's modulus measured on the cores using nanoindentation measurements. The reasonable agreement between the two shows a possible application of the predicted velocities to predict mechanical properties such as Poisson's ratio, Young's modulus, shear modulus, and bulk modulus.

The predicted hardness can also be used to identify brittle zones across the well length. Figure 15a shows the brittle and ductile regions classified by Grieser and Bray (2007) on Young's modulus and Poisson's ratio crossplot. Other researchers such as Gunther (2017) have also observed similar partitions between ductile and brittle zones on the Young's modulus and Poisson's ratio crossplot. Figure 15b shows the data for this study on the same crossplot. The different points on the crossplot in Figure 15b are colored and sized by the hardness values obtained from the nanoindenter. It is evident that points having higher hardness lie in the brittle region, whereas points having lower hardness lie in the ductile region. Thus, hardness can be used to identify brittle zones and therefore optimize completion zones across the well bore length.

One of the key advantages of using the core data for developing regression models is that most of the petrophysical properties used for regression analysis represent the same volume of rock; i.e., there are no sampling and resolution issues. However, the real practical application lies in predicting the sonic logs (not commonly measured) using the triple combo logs (more commonly measured). The logs are affected by resolution and sampling constraints. This means that in the case of a triple combo log, each log is associated with a different depth of investigation and vertical resolution, and it may not be representative of the same volume of rock. Additionally, the anisotropy in the mechanical properties cannot be captured from the use of triple combo logs; therefore, the predicted sonic logs and the resulting geomechanical properties may not be directly applicable to hydraulic fracturing models. A rigorous upscaling or averaging to

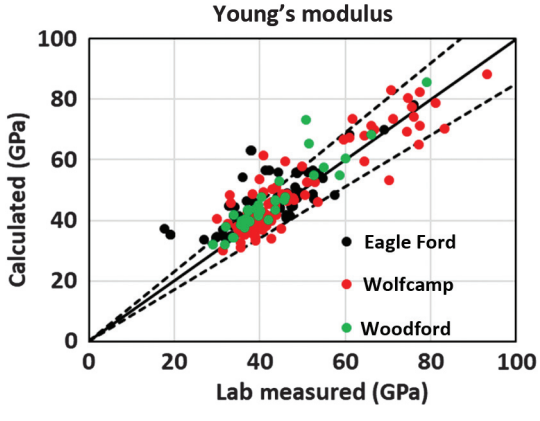



**Figure 14.** Comparison of Young's modulus calculated from core sonic velocities with the actual measured Young's modulus values on the cores. Measurements were made on horizontal plugs. The dashed lines represent 15% uncertainty lines.

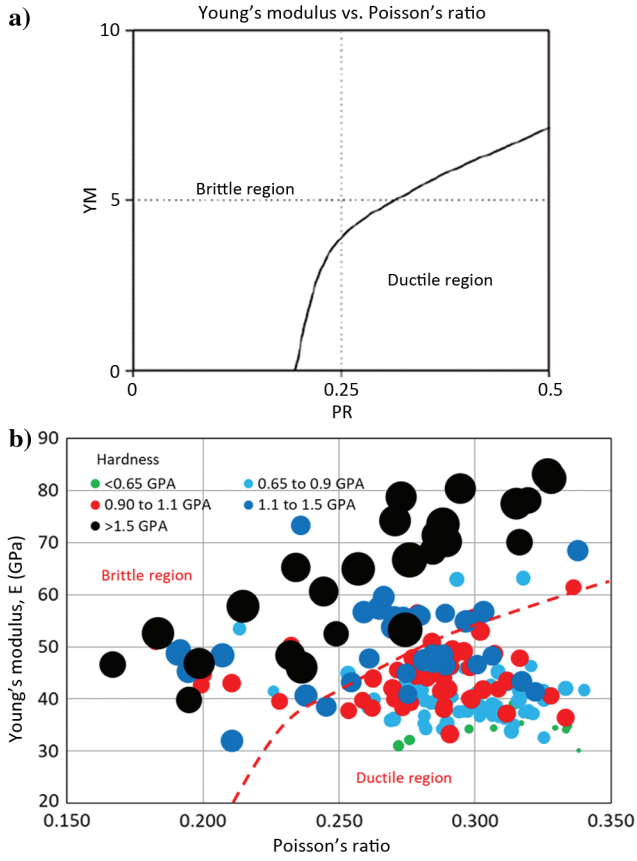



**Figure 15.** (a) Brittle and ductile regions classified by Grieser and Bray (2007) on Young's modulus and Poisson's ratio crossplot. (b) Crossplot of Young's modulus and Poisson's ratio with the bubbles colored and sized by their hardness values. It is evident that points with higher hardness lie in the brittle region, whereas points with lower hardness lie in the ductile region. Thus, hardness can be used to identify brittle zones and therefore optimize completion zones across the well bore length.

bring the different logs to the same vertical resolution and an idea about the formation anisotropy from other independent sources may have to be incorporated to make meaningful application to hydraulic fracture design and well-spacing problems.

## Conclusion

This study demonstrates the use of several different regression techniques to predict synthetic sonic velocities ($V_P$ and $V_S$) and hardnesses using laboratory core data. The regression techniques used in the study were MLR, LASSO and ridge regression, SVR, RF, GB, and ACE.

The study findings indicate that the commonly used MLR approach performs poorly as compared to other techniques such as RFs and GB in terms of predictive capability. Gaussian process simulation showed that these techniques also provided the lowest uncertainty in prediction.

The predicted sonic velocities can be used to calculate the Young's modulus, Poisson's ratio, and other mechanical properties. Similarly, hardness is a brittleness indicator, and it can be used to identify brittle zones across the well length and to optimize completion zones.


## Acknowledgments

This study was conducted as a part of the Shale Gas Consortium and Rock Physics Consortium in IC3 (Integrated Core Characterization Center) laboratory at the University of Oklahoma. We would like to acknowledge the support of all its members.


## Data and materials availability

Data associated with this research are confidential and cannot be released.

## References


Akinnikawe, O., S. Lyne, and J. Roberts, 2018, Synthetic well log generation using machine learning techniques: Presented at Unconventional Resources Technology Conference. URTeC 2877021.

Ballard, B. D., 2007, Quantitative mineralogy of reservoir rocks using Fourier Transform infrared spectroscopy: Presented at SPE Annual Technical Conference and Exhibition, doi: 10.2118/113023.

Barnett, V., and T. Lewis, 1995, Outliers in statistical data, 3rd ed.: John Wiley and Sons, 235–236.

Bhatt, A., 2002, Reservoir properties from well logs using neural networks: Ph.D. dissertation, Norwegian University of Science and Technology.

Bhattacharya, S., M. Maucec, J. M. Yarus, D. D. Fulton, J. M. Orth, and A. P. Singh, 2013, Causal analysis and data mining of well stimulation data using classification and regression tree with enhancements: Presented at SPE Annual Technical Conference and Exhibition, doi: 10.2118/166472-MS.

Birch, F., 1961, The velocity of compressional waves in rocks to 10 kilobars — Part 2: Journal of Geophysical Research, **66**, 2199–2224, doi: 10.1029/JZ066i007p02199.

Breiman, L., 1997, Arcing the edge. Technical Report 486: Statistics Department, University of California, Berkeley.

Breiman, L., 2001, Random forests: Machine Learning, **45**, 5–32, doi: 10.1023/A:1010933404324.

Breiman, L., and J. H. Friedman, 1985, Estimating optimal transformations for multiple regression and correlation: Journal of the American Statistical Association, **80**, 580–598, doi: 10.1080/01621459.1985.10478157.

CLR, 2010, Anadarko Woodford: The SCOOP, Internal Report, Continental Resources Limited, Oklahoma City, Oklahoma, http://media.corporate-ir.net/media_files/irol/19/197380/CLR_Sunday_10_7_12_Anadarko_Woodford_3.pdf, accessed 27 March 2017.

Cortez, M., 2012, Chemostratigraphy, paleoceanography, and sequence stratigraphy of the Pennsylvanian-permian section in the Midland basin of the West Texas, with focus on the Wolfcamp formation: M.S. Geology thesis, The University of Texas at Arlington.

Drucker, H., C. J. C. Burges, L. Kaufman, A. J. Smola, and V. N. Vapnik, 1997, Support Vector Regression Machines, in Advances in Neural Information Processing Systems: MIT Press **9**, NIPS 1996, 155–161.

Eshkalak, M. O., S. D. Mohaghegh, and S. Esmaili, 2014, Geomechanical properties of unconventional shale reservoirs: Journal of Petroleum Engineering, **10**, 961641, doi: 10.1155/2014/961641.

Ester, M., H. P. Kriegel, J. Sander, X. Xu, E. Simoudis, J. Han, and U. M. Fayyad, 1996, A density-based algorithm for discovering clusters in large spatial databases with noise: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226–231.

Forgy, E. W., 1965, Cluster analysis of multivariate data: Efficiency versus interpretability of classifications: Biometrics, **21**, 768–769.

Gao, T., and V. Jojie, 2016, Degrees of Freedom in Deep Neural Networks. arXiv:1603.09260 [cs.LG].

Ghavami, F., 2011, Developing synthetic logs using artificial neural network: Application to Knox County in Kentucky: M.Sc. thesis, West Virginia University.

Grieser, B., and J. Bray, 2007, Identification of production potential in unconventional reservoirs: Presented at the SPE Productions and Operations Symposium.

Guan, S., 2012, Modeling geomechanical property changes using well logging and pressure data in a $CO_2$ enhanced oil recovery reservoir: M.Sc. thesis, Colorado School of Mines.

Gunther, M. J., 2017, Estimating brittleness using seismic data in an unconventional shale reservoir, Fort Worth Basin, North Central Texas: M.S. thesis, University of Oklahoma.

Gupta, I., C. Rai, C. Sondergeld, and D. Devegowda, 2017a, Rock typing in Wolfcamp formation: Presented at the SPWLA 58th Annual Logging Symposium.

Gupta, I., C. Rai, C. Sondergeld, and D. Devegowda, 2017b, Rock typing in Eagle Ford, Barnett, and Woodford formations: Presented at Unconventional Resources Technology Conference.

Gupta, I., C. Sondergeld, and C. Rai, 2018a, Applications of Nano indentation for reservoir characterization in shales: Presented at 52nd US Rock Mechanics/Geomechanics Symposium.

Gupta, I., C. Rai, C. Sondergeld, and D. Devegowda, 2018b, Rock typing in Eagle Ford, Barnett, and Woodford formations: Society of Petroleum Engineers, doi: 10.2118/189968-PA.

Hamzaban, M., and H. Memarian, 2008, Determination of relationship between drilling parameters by clustering techniques: Presented at the 5th Asian Rock Mechanics Symposium.

Hay, J. L., and G. M. Pharr, 2000, Instrumented indentation testing, in ASM Handbook, 232–243.

Jain, A. K., A. L. Topchy, M. H. Law, and J. M. Buhmann, 2004, Landscape of clustering algorithms: Proceedings of the International Conference on Pattern Recognition, 260–263.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013, An Introduction to Statistical Learning with Applications in R: Springer.

Johnson, S. C., 1967, Hierarchical clustering schemes: Psychometrika, **32**, 241–254, doi: 10.1007/BF02289588.

Junck, R., and D. Benson, 1973, A high-resolution pulse transmission technique for determining ultrasonic velocities: Review of Scientific Instruments, **44**, 1044–1048, doi: 10.1063/1.1686297.

Kale, S., C. Rai, and C. Sondergeld, 2010, Rock typing in Gas shales: Presented at SPE Annual Technical Conference, doi: 10.2118/134539-MS.

Karastathis, A., 2007, Petrophysical measurements on Tight Gas shale: Master thesis, University of Oklahoma.

Kopp, W., 2018, Gaussian processes, https://www.youtube.com/watch?v=9hKfsuoFdeQ&index=1&list=PLR6O_WZHBlOGlq9GiYjCZb4nQbqsT11Pl, accessed 12 December 2008.

LaFollette, R. F., W. D. Holcomb, and J. Aragon, 2012, Practical data mining: Analysis of Barnett Shale production results with emphasis on well completion and fracture stimulation: Presented at the SPE Hydraulic Fracturing Technology Conference and Exhibition. SPE 152531-MS, doi: 10.2118.152531-MS.

Law, C., 1999, Evaluating source rocks, Chapter 6, in E. A. Beaumont and N. H. Foster, Exploring for oil and gas traps: AAPG special volumes, Petroleum Geology/Handbook of Petroleum Geology, 6-1–6-41.

Ma, X., and M. D. Zoback, 2017, Lithology variations and cross-cutting faults affect hydraulic fracturing of Woodford shale: A case study: Presented at the SPE Hydraulic Fracturing Technology Conference and Exhibition, doi: 10.2118/184850-MS.

MacDonald, B., P. Ranjan, and H. Chipman, 2015, GPfit: An R package for fitting a Gaussian process model to deterministic simulator outputs: Journal of Statistical Software, **64**, 1–21, doi: 10.18637/jss.v064.i12.

Mattaboni, P., and E. Schreiber, 1967, Method of pulse transmission measurement for determining sound velocities: Journal of Geophysical Research, **72**, 5160–5163, doi: 10.1029/JZ072i020p05160.

Mohaghegh, S., 2013, Shale asset management via advanced data-driven and predictive analysis. SPE Webinar, recorded 13 November 2013.

Pioneer, 2013, Investor Presentation, Wolfcamp, Internal Report, Pioneer Natural Resources, Oklahoma City, Oklahoma, %%www.plsx.com/finder/publicpdf.aspx?key=CksUFGCy3EQrY1U9K2YeqQ%3D%3D, accessed 27 March 2017.

Ranjan, P., R. Haynes, and R. Karsten, 2011, A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data: Technometrics, **53**, 366–378, doi: 10.1198/TECH.2011.09141.

Rokach, L., and O. Maimon, 2005, Clustering methods, in Data mining and knowledge discovery handbook: Springer, 321–352.

Rokach, L., and O. Maimon, 2008, Data mining with decision trees: Theory and applications: World Scientific Pub. Co. Inc.

Roy, A., T. Zhao, V. Jayaram, and D. Devegowda, 2015, Well performance predictions from geologic, petrophysical and completions-related parameters using generative topographic mapping: A field case study: 85th Annual International Meeting, SEG, Expanded Abstracts, 2853–2857, doi: 10.1190/segam2015-5916229.1.

Schmidhuber, J., 1992, Learning complex, extended sequences using the principle of history compression: Neural Computation, **4**, 234–242, doi: 10.1162/neco.1992.4.2.234.

Schmidhuber, J., 2015, Deep learning in neural networks: An overview: Neural Networks, **61**, 85–117, doi: 10.1016/j.neunet.2014.09.003.

Schuetter, J., S. Mishra, M. Zhong, and R. LaFollette, 2018, A data-analytics tutorial: Building predictive models for oil production in an unconventional shale reservoir: Society of Petroleum Engineers, doi: 10.2118/189969-PA.

Shirzadi, S., E. Ziegel, and R. Bailey, 2013, Data mining and predictive analytics transforms data to barrels: Presented at the SPE Digital Energy Conference and Exhibition, doi: 10.2118/163731-MS.

Shukla, P., V. Kumar, M. Curtis, C. H. Sondergeld, and C. S. Rai, 2013, Nanoindentation studies on shales: Presented at 47th US Rock Mechanics/Geomechanics Symposium.

Smola, A. J., and B. Scholkopf, 2004, A tutorial on support vector regression: Statistics and Computing, **14**, 199–222, doi: 10.1023/B:STCO.0000035301.49549.88.

Sondergeld, C. H., R. J. Ambrose, C. S. Rai, and J. Moncrieff, 2010a, Micro-structural studies of gas shales: Presented at the SPE Unconventional Gas Conference. SPE 131771.

Sondergeld, C. H., K. E. Newsham, J. T. Comisky, M. C. Rice, and C. S. Rai, 2010b, Petrophysical considerations in evaluating and producing shale gas resources: Presented at the SPE Unconventional Gas Conference.

Sondergeld, C. H., and C. S. Rai, 1993, A new concept of quantitative core characterization: The Leading Edge, **12**, 774–779, doi: 10.1190/1.1436968.

Tibshirani, R., 1996, Regression shrinkage and selection via the lasso: Journal of the Royal Statistical Society. Series B (Methodological), **58**, 267–288.

Walker, S. H., and D. B. Duncan, 1967, Estimation of the probability of an event as a function of several independent variables: Biometrika, **54**, 167–179, doi: 10.2307/2333860.

Wilson, M. J., M. V. Shaldybin, and L. Wilson, 2016, Clay mineralogy and unconventional hydrocarbon shale reservoirs in the USA. Occurrence and interpretation of mixed-layer R3 ordered illite/smectite: Earth Science Reviews, **158**, 31–50, doi: 10.1016/j.earscirev.2016.04.004.

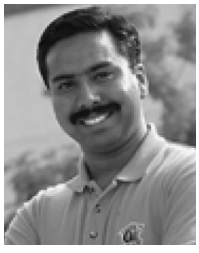
Yan, X., 2009, Linear regression analysis: Theory and computing: World Scientific.

Zhao, T., V. Jayaram, A. Roy, and K. J. Marfurt, 2015, A comparison of classification techniques for seismic facies recognition: Interpretation, **3**, no. 4, SAE29–SAE58, doi: 10.1190/INT-2015-0044.1.

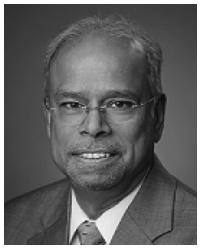
Ziegel, E., R. Bailey, S. Shirzadi, K. Sprague, and W. Hedges, 2011, Using data mining to build a tool to estimate the extent of pipeline corrosion: Presented at the SPE Digital Energy Conference and Exhibition, doi: 10.2118/144181-MS.

**Ishank Gupta** received a B.S. (2009) in petroleum engineering from the University of Petroleum and Energy Studies. He worked as a reservoir engineer at Schlumberger from 2010 to 2015. Currently, he is a Ph.D. candidate at the University of Oklahoma. His research interests include reservoir characterization and unconventional petrophysics.

**Deepak Devegowda** received a B.S. in electrical and electronics engineering from the Indian Institute of Technology. He also received a Ph.D. (2008) in petroleum engineering from Texas A&M University. Currently, he is an associate professor of petroleum engineering at the University of Oklahoma. His research interests include unconventional oil and gas reservoir engineering, high-resolution reservoir characterization, and production optimization.



**Vikram Jayaram** received an M.S. and a Ph.D. in electrical engineering and was awarded the prestigious NASA doctoral fellowship. In his current role at the Pioneer Natural Resources, he is a senior manager for data science and advanced analytics. Before joining Pioneer, he was a principal research scientist for Sabre Corporation, where he architected the next generation revenue management systems prototype using AI/Game Theory models. He has also worked as a senior research geophysicist with Global Geophysical Services and also served as a research faculty with the University of Oklahoma. He did his postdoctoral work at the University of Texas M.D. Anderson Cancer Center working on image reconstruction algorithms in nuclear imaging. He has authored more than 30 peer-reviewed publications in international conferences and journals. He is a senior member of IEEE Signal Processing Society, SPE, AGU, SEG, and technical reviewer/chair for several international conferences and journals in signal processing, machine learning, and petro/geosciences. He is also an associate editor for *Interpretation*. His research interests include developing AI powered production systems for advanced E&P analytics.



**Chandra Rai** received an M.S. (1971) in geophysics from the Indian School of Mines. He also received a Ph.D. (1977) in geology and geophysics from the University of Hawaii. He has worked as a technology director at Amoco for 18 years. Currently, he is director and professor at the University of Oklahoma. He teaches petrophysics, petrophysics lab, seismic reservoir modeling, unconventional reservoirs, and well logging. His research interests include rock physics, petrophysics, and geomechanics with an emphasis on unconventional shale gas and oil reservoirs.

**Carl Sondergeld** received an M.A. (1972) in geology from Queens College. He also received a Ph.D. (1977) in geophysics from Cornell University. Currently, he is professor and Curtis Mewbourne Chair at the University of Oklahoma. He teaches petrophysics, petrophysics lab, technical communications, seismic reservoir modeling, unconventional reservoirs, introduction to petroleum engineering, and well logging. He coaches the SPE Petrobowl team and carries out research in the areas of rock physics, petrophysics, and geomechanics with an emphasis on unconventional shale gas and oil reservoirs.