GEOPHYSICS, VOL. 84, NO. 2 (MARCH-APRIL 2019); P. O39–O47, 12 FIGS., 1 TABLE.
10.1190/GEO2018-0202.1

Correlating geologic and seismic data with unconventional resource production curves using machine learning

Ryan Smith[1], Tapan Mukerji[2], and Tony Lupo[3]

ABSTRACT

Predicting well production in unconventional oil and gas settings is challenging due to the combined influence of engineering, geologic, and geophysical inputs on well productivity. We have developed a machine-learning workflow that incorporates geophysical and geologic data, as well as engineering completion parameters, into a model for predicting well production. The study area is in southwest Texas in the lower Eagle Ford Group. We make use of a time-series method known as functional principal component analysis to summarize the well-production time series. Next, we use random forests, a machine-learning regression technique, in combination with our summarized well data to predict the full time series of well production. The inputs to this model are geologic, geophysical, and engineering data. We are then able to predict the well-production time series, with 65%–76% accuracy. This method incorporates disparate data types into a robust, predictive model that predicts well production in unconventional resources.

INTRODUCTION

A key aspect of optimizing production in unconventional resources is selecting sites to drill that have a relatively low risk of producing below the economic threshold. This process can be more challenging in unconventional resource plays because the science behind what geologic or geophysical traits make a play more likely to be economic are not well-established. Here, we propose using machine-learning methods, which establish relationships between available engineering, geologic, and geophysical data and well performance. This could enable us to quantify the risk of drilling a well that is not economic, avoiding potential losses.

Our objective was to use geologic, seismic, and well-completion data to predict well performance and to develop a more intuitive understanding of the geologic and geophysical drivers for well performance. In our prediction, we used as input traditional geologic data sets, such as porosity, total organic carbon (TOC), and water saturation, as well as seismic attributes. We developed a framework for how seismic attributes could be combined with geologic and well-completion data to predict well performance. This framework could be used in future studies to include additional attributes related to faults, such as coherence.

The curvature attribute, which highlights synclines and anticlines, is thought to be a proxy for fractures (Roberts, 2001). Since the onset of unconventional resource production, natural fractures have been thought to form “sweet spots” for drilling, where enhanced fracture permeability boosts production. However, the hypothesis that curvature attributes can be used to identify enhanced areas for production remains poorly tested. To test this hypothesis, we compared curvature data, geologic data, and completion parameters with well-production data.

To compare these diverse data sets with well production, we needed to summarize well-production time series. Many methods exist that can transform a time series into relatively few values. We chose to use functional principal component analysis (FPCA). This method combines functional data analysis (FDA) with principal component analysis (PCA). We chose to use the method because it can handle noisy data, and has the flexibility to fit time series that do not have seasonal components, which is the case with our time series. FPCA has been performed before on production curves in unconventional shale resources by Grujic et al. (2015), Grujic (2017), and Cai et al. (2017), who used predicted well production with a focus on engineering parameters. We implemented this method, but with a focus on geologic and geophysical parameters while still in-

Manuscript received by the Editor 20 March 2018; revised manuscript received 22 October 2018; published ahead of production 18 December 2018; published online 01 March 2019.

[1]Missouri University of Science and Technology, Department of Geosciences and Geological and Petroleum Engineering, Rolla, Missouri, USA. E-mail: smithryang@mst.edu.
[2]Stanford University, Department of Energy Resources Engineering, Stanford, California, USA. E-mail: mukerji@stanford.edu.
[3]SM Energy Company, Denver, Colorado, USA. E-mail: tlupo@sm-energy.com.

cluding engineering parameters. This method enabled us to summarize each decline curve with only two values, the scores for the first and second functional principal components (FPCs). We then developed a model that predicts these scores (from which decline curves can be generated) given geologic, seismic, and well-completion data using random forests, a machine-learning technique that accounts for nonlinear interactions between multiple variables on the outcome.

Ruths et al. (2017) implement a method that predicted pressure response during well completions using seismic data in random forests, but did not predict well production. This study builds on previous methods by applying a random forest model to predict production curves. Seismic data have not been previously incorporated into predictions of well production in this framework, and show potential for improving the method. Furthermore, we discuss relationships between input parameters and total production, as well as rate of production decline. Because the well-production rate is greatly influenced by the presence of fractures, our research helps to provide a more intuitive understanding of how fractures, estimated with geophysical data, and matrix porosity, interpolated from borehole logs, influence well production.

Using the method outlined above, we were able to predict total production using geologic, geophysical, and well-completion data with a correlation coefficient (r^2) of 0.65–0.76. Many expected relationships, such as the positive impact of high porosity and low water saturation on total production, were observed. Some relationships that were not expected, such as a positive correlation between stratigraphically higher curvature and production, were also observed. Our results suggest that this method, when properly calibrated and used with good training data, could have broad applicability in unconventional resource plays.

In the "Study area" section, we give an overview of the study area; in the "Data preparation" section, we describe the data preparation, including FDA of the production curves and seismic curvature attribute generation; in the "Predicting well production" section, we outline the random forest methods used for predicting production curves. In the "Discussion" section, we interpret the results; and finally, we offer our conclusions in the "Conclusion" section.

STUDY AREA

Our study area is in southwest Texas, and the objective of our analysis is the Late Cretaceous shales and marls of the Eagle Ford Group, a key element of the unconventional shale oil play that has been actively drilled in recent years (Figure 1a). A seismic cross section along the area of interest is shown in Figure 1c, with the location of the cross section shown in Figure 1b.

The portion of the Eagle Ford analyzed in this study resides along the westernmost portion of the known productive fairway in Webb and Dimmit counties. Here, the Eagle Ford was deposited in the Maverick Sub Basin, where the thickest deposits (150–200 m) of productive Eagle Ford exist. The Eagle Ford was deposited in anoxic-to-dysoxic conditions during Cenomanian and Turonian ages of the Late Cretaceous. The producing zone is often subdivided into the Upper and Lower Eagle Ford. The Lower Eagle Ford was deposited conformably upon the Buda formation. The Upper Eagle Ford is overlain by the Austin Chalk, which is overlain by the Anacacho Limestone; both are brittle carbonate formations. The Lower Eagle Ford consists of approximately 60 m of dark black, 2 wt%–8 wt% TOC mudstones that are overlain by approximately 150 m of Upper Eagle Ford marls with TOC between 0.5 wt% and 4.65 wt%. In the Lower Eagle Ford, the average Sw is 0.17 with porosities averaging 11%. The Lower Eagle Ford is underlain by the Buda Limestone, another brittle carbonate formation. The portion of the Eagle Ford Fairway investigated in our study is in the dry gas-to-gas condensate maturity window with high pressure gradients. Changes in the depth of the Lower Eagle Ford could affect the pressure and maturity and thus could be strongly correlated with the productivity of wells drilled within the study area.

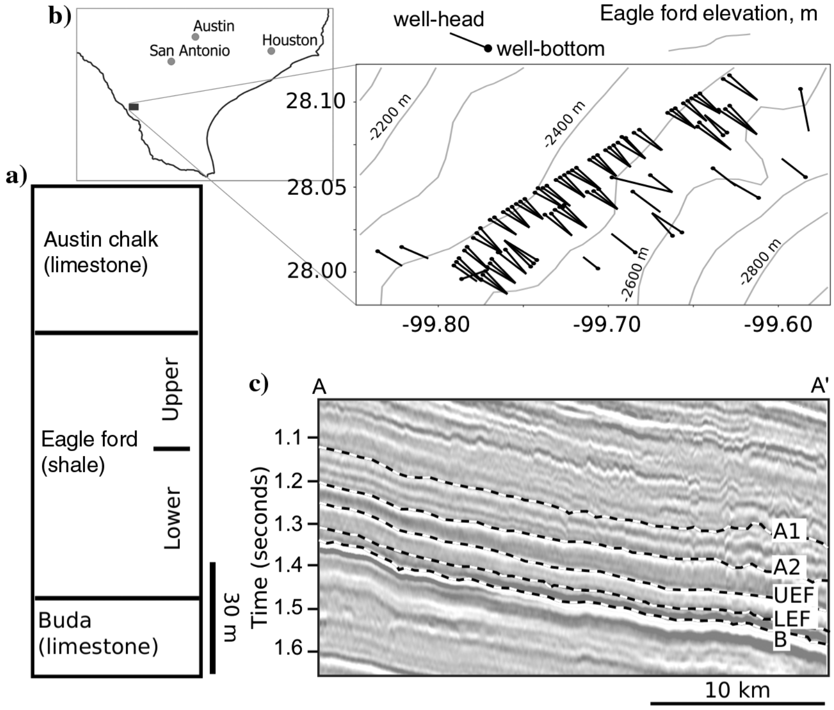

Figure 1. (a) Generalized stratigraphic column showing formation of interest, the Lower Eagle Ford, (b) locations of wells used in study, and (c) seismic section with relevant horizon tops stratigraphically close to the Lower Eagle Ford. Horizon tops shown are Anacacho (A1), Austin Chalk (A2), Upper Eagle Ford (UEF), Lower Eagle Ford (LEF), and Buda (B).

DATA PREPARATION

To perform this analysis, we needed to first prepare, clean, and develop summary statistics for the well-production time series, as well as geophysical and geologic data. The well-production data were initially noisy, and difficult to summarize with one or two statistics. We used a method called FDA to smooth the production data, reducing the noise while keeping the key trends. We then performed PCA on the smoothed data set, which provided two summary statistics for each well that explained 95% of the variation in well production for each well. Geologic information, seismic data, and production decline curves were made available by the company we collaborated with. We performed curvature calculations on the seismic data to transform the data into attributes relevant to our study.

Functional data analysis and principal component analysis

The production of all wells in our study area with at least 1.5 years of data, as well as the mean production across all wells, are shown in Figure 2. The production at each well is complex, time-varying, and contains a significant amount of noise. Comparing geologic and geophysical data (values along a well bore) with production curves (time series) because they are normally viewed is a difficult process because the data types are so different. The production curves are time-series functions, whereas the geologic and geophysical attributes are not time dependent, but are spatially distributed along the well bore.

One could note wells with steeper decline curves, and hypothesize that it is related to permeable fracture flow that drains more quickly than a typical matrix flow. However, identifying wells with these production attributes is normally a subjective process. Our objective was to use a method that could summarize the key trends in well-production time series with relatively few values. We chose to use FPCA, which first smooths the time series, then summarizes its key components, or principal components. By analyzing well-decline curves with FPCA, we reduced the dimensionality of a decline curve time series from approximately 500 (days) to two principal component scores, or coefficients. One of these scores was correlated (with a correlation >0.99) with total production, and the other score was correlated with an initial spike followed by a rapid drop-off in production. This pattern in production is indicative of fracture-dominated flow.

This correlation potentially makes comparison with curvature more natural. With the lower dimensional representation of production curves, one can compare each dimension (principal component) with geologic and geophysical data statistically to determine the correlation, or lack thereof. Here, we used a modification of the FPCA method developed by Grujic (2017).

FDA (Ramsay, 2006) is a data-smoothing technique that uses a series of basis functions $\phi(t)$ to describe a time series $f(t)$:

$$f(t) = \sum_{i=1}^{k} a_i \phi_i(t), \tag{1}$$

where a_i are the coefficients of the basis functions. Thus, in this method, a time series is defined by the coefficients, a_i of the basis functions. To reconstruct the time series, one plugs the coefficients into equation 1. This approximation results in a smoother time series, as the basis functions describing the time series are inherently smooth.

There are several choices for the basis functions. The two most common are Fourier basis functions, for seasonal data, and spline basis functions, for data that are not seasonal. We chose to use spline functions because there is no seasonal or recurring pattern in decline curves. Because production curves can have unexpected jumps that are not meaningful to the overall trend, FDA was helpful in removing data that are less meaningful. For a more in-depth treatment of FDA, see Ramsay (2006).

We performed FDA using the FDA MATLAB package (Ramsay et al., 2009). The number of basis functions that are needed to accurately fit a time series depends upon the nature of the time series and is typically determined with trial and error. We used 20 basis functions to fit the decline curves. We chose this value because it was within the range used successfully by Grujic et al. (2015) and it provided a reasonable fit with the data.

Only wells that had at least 1.5 years of production data were included in our analysis, and of these wells, only the first 1.5 years of production data were included. To prepare the production data for analysis, wells with significant turn-off time and other outliers needed to be removed. We also removed outliers based on histograms of mean choke size during production and total change in choke size during production. We did this because varying the choke size during production, or having an anomalously high or low choke size, alters the timing of well production and makes wells difficult to compare with each other. This brought our data set down from 136 to 125 wells. We also removed additional outliers in later steps, which are outlined below.

After smoothing our data with FDA, we then analyzed the fit of the smoothed production time series to the original time series. From this visual inspection, we identified outliers in approximately 10% of the data. The outliers were generally determined based on unexpected and abnormal jumps in production later in the life of the well, indicating a significant change in choke size or pressure that had been missed by our initial search for outliers. This left us with 97 wells.

After performing FDA, a smoothed version of each time series was described by the coefficients a_i of the basis functions $\phi_i(t)$. We used PCA (Jolliffe, 2002) on the functional data coefficients to reduce the dimensionality of the problem. PCA reduces the dimensionality by transforming a data set into n number of principal components, where n is less than the original dimension of the data set. These principal components are linearly uncorrelated. Each time series is then described by the coefficients, or scores, on each principal component. Similar to FDA, the original time series can be reconstructed by multiplying the coefficients by the principal components, then summing them. For a more in-depth treatment of PCA, see Jolliffe (2002).

The first three principal components explained the 82%, 10%, and 3% of the variance in well production, respectively. Because

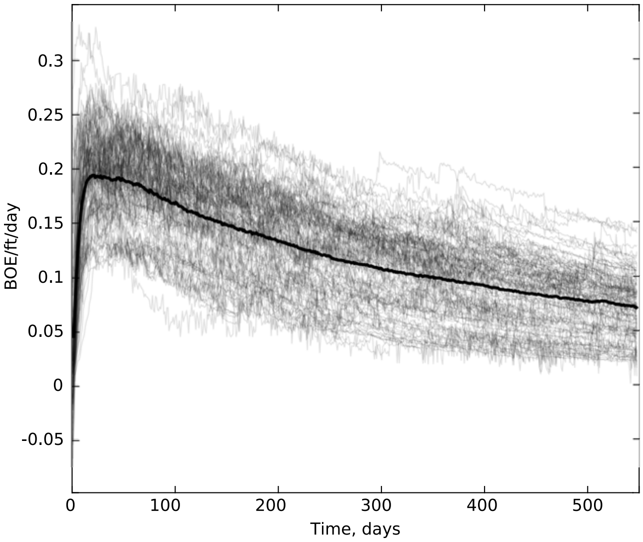

Figure 2. Production of all wells in our study area with at least 1.5 years of production (transparent lines), and the mean production across all wells (the dark thick line).

decline curves have similar patterns from well to well, we were able to account for 91% of the variability in the decline curves with the first two principal components. We did not consider the third principal component in our analysis because it explained so little variance in well production. The first and second principal components are shown in Figure 3.

From the FPCA analysis, each well has two scores, or coefficients, that can be used to reconstruct its time series. The first score s_1 is associated with the first FPC (FPC_1, shown in Figure 3 as a dashed line). The second score s_2 is associated with the second FPC (FPC_2, shown in Figure 3 as a dashed-dotted line). Production decline curves P are reconstructed using FPCs and their scores as follows:

$$P = \text{mean} + s_1\text{FPC}_1 + s_2\text{FPC}_2, \tag{2}$$

where mean is the mean production across all wells (shown in Figure 3 as a solid line). As shown in Figure 3, FPC_1 is positive through the entire time series, so a higher s_1 results in higher overall production. The second principal component shows that for higher scores, the production starts higher but eventually drops to negative values at approximately 200 days, so a higher s_2 results in higher production initially, then lower production after approximately 200 days. Because this is often thought to be the effect of high-permeability, low-storage-fracture networks, this result supports the idea that principal components could be used to indicate areas that are more or less fractured. In addition to natural fractures, however, these production declines could also be related to how the wells were completed, and how closely spaced they are — for example, steeper declines could also be related to a nearby well being turned on. Figure 4 shows the fit of a typical well decline time series with two principal components. Figure 4b shows a clear trend of decreasing FPC scores (first FPC) from southwest to northeast (shallow to deep), representing decreasing production as the basin gets deeper.

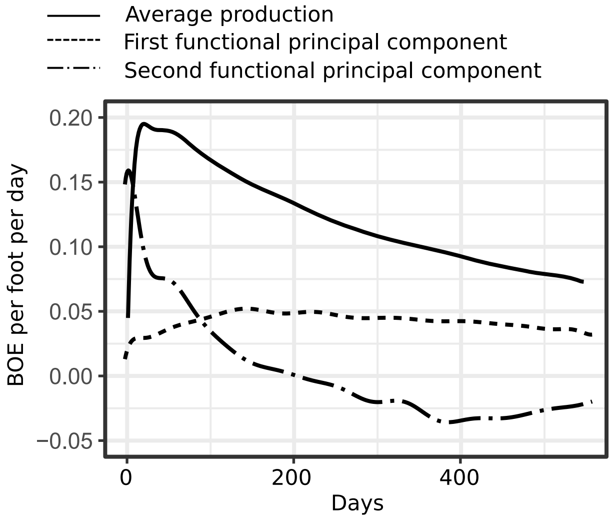

Figure 3. Average production (the solid line), first FPC (dashed line), and second FPC (the dashed- dotted line). Note that the first principal component has mean >0, meaning that positive values result in greater overall production, whereas the second principal component has mean = 0, meaning that positive values have no correlation with total production.

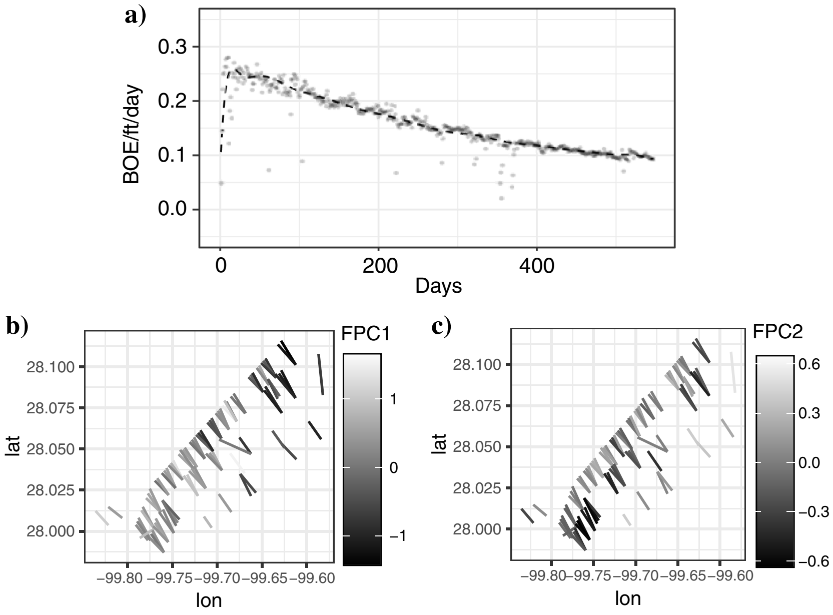

Figure 4. (a) Observed decline compared with that predicted by two principal components. (b) Map of scores on the first FPC. (c) Map of scores on the second FPC. For (b and c), the well trajectories are shown. Each well has two scores (FPC1 and FPC2). (b) The FPC1 score for each well and (c) the FPC2 score for each well.

Curvature attribute generation

Curvature is a seismic attribute that is calculated by taking the second derivative of seismic reflectors (Bergbauer et al., 2003; Al-Dossary and Marfurt, 2006). This highlights zones with anticlines (negative curvature) and synclines (positive curvature). Because areas with anticlines or synclines are prone to have more fractures, curvature has also been used as an approximate proxy for fracture intensity in the subsurface. This technique has been used for more than a decade by the oil and gas industry (Roberts, 2001; Sigismondi and Soldo, 2003). However, there remains a great deal of uncertainty as to whether natural fractures highlighted by curvature volumes enhance production or deter it. Gale et al. (2007) note that natural fractures can either enhance permeability, increasing well production, or connect the reservoir to formations with water, thus hurting well production. The effectiveness of natural fractures is largely dependent on the fracture size and the geologic conditions at the zone of interest.

The curvature attributes have been generated using software developed by the Attribute-Assisted Seismic Processing and Interpretation consortium (Al-Dossary and Marfurt, 2006). The curvature attribute was processed on the amplitude seismic volume within an approximately 0.5 s window of the zone of interest. The seismic data cover an area of approximately 260 km^2. The locations of the wells used in the study are shown in Figure 1. To compute the curvature, we followed the procedure described by Al-Dossary and Marfurt (2006). First, we computed the dip direction of the reflectors (first spatial derivative). The curvature was computed by taking the derivative of the dip direction. Because derivatives increase noise, we applied a long-wavelength filter to reduce noise.

Geologic and geophysical data extraction

We extracted geologic, geophysical, and engineering data along the well bores. A summary table of the variables extracted along each well is shown below. Note that the seismic curvature attributes were extracted along specific horizons, which did not necessarily represent the vertical location of the well bore. For example, along each well bore, the absolute value of the curvature of the Eagle Ford and Anacacho time horizons were extracted (see Figure 5). We extracted the absolute value of curvature because we were interested in the positive (synclines) and negative (anticlines) curvature anomalies because both should be related to fractures. We initially extracted a zone surrounding the well, but we found that the predictions were better when only extracting the data along the well bore. Typically, the well bores closely track the Eagle Ford horizon, so this horizon provided a good estimate of the geophysical properties along the well bore. The Anacacho is located stratigraphically higher than the targeted Eagle Ford, and it could highlight features that are present but subseismic (not detectable) in the Eagle Ford. For instance, a fracture clearly visible from seismic data in the Anacacho could propagate into the Eagle Ford with diminishing magnitude, so that it is not seismically resolved. We initially included other horizons as well, but we found that they did not significantly impact the results. After extracting the values along the well bores, we took their median value as the key summary statistic for each well. In our exploratory studies, we included the 10th and 90th percentiles of the values extracted along the well bore, but we found them generally to be less informative. To simplify the relationships between curvature and production, and to avoid using numerous correlated predictor variables, in our final analysis presented in this paper, we only used the median values.

We included geologic inputs that were either known to or hypothesized to relate to well production: porosity, water saturation, TOC, geothermal gradient, isopach, total vertical depth, and zone. Zone referred to the principal formation that was drilled into (e.g., Upper Eagle Ford, Lower Eagle Ford, and mixed). We did not include the "x" and "y" location of the wells as input because the location alone is not indicative of any physical properties. The total vertical depth represented the "z" location of the wells. The depth of the wells is related to several geologic parameters that are otherwise difficult to estimate, such as pressure and maturity. In this study, we attempted as much as possible to relate physical properties to well production, although in the case of the total vertical depth this was not possible. As with the seismic attributes, we used the median value along each well bore as the summary statistic of each variable for each well.

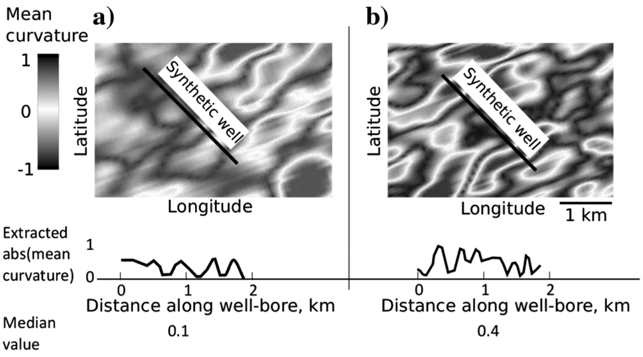

Figure 5. A synthetic example of how summary statistics for curvature attributes on (a) the Eagle Ford and (b) the Anacacho Formations were calculated. The black line overlain on the curvature maps is only a schematic and does not denote an actual well location.

We included engineering parameters that were also either known or believed to impact production: lateral length, which is the length of the drilled well; fluid per foot, which is the amount of fluid used to stimulate the reservoir per foot of well length; stage spacing, which is the distance between each reservoir stimulation; and choke size (starting value and trend over time), which is used to control the well production rate.

PREDICTING WELL PRODUCTION

Random forest models

With the data along the well bores extracted, we created random forest models (Breiman, 2001) to predict the principal component scores from the predictor variables. Random forests or random decision forests are an ensemble-based statistical method for classification and regression. They combine a large number of decision trees that are trained using the training data to predict the response. An example decision tree is shown in Figure 6. In this example, we are predicting barrels of oil equivalent (BOE) produced per lateral foot on a horizontal well after 1.5 years. The predictor variables are water saturation, porosity, and TOC. At each node (box), we move to the right or left based on the value of the variable at that node. For example, if the water saturation is less than 0.4, we move to the left. At the bottom is a prediction of BOE/ft after 1.5 years based on the values of the predictor variables.

Decision trees are popular for machine learning (Hastie et al., 2001) because they are invariant under scaling and other transformations of the predictor values, meaning that variables do not need to be scaled prior to analysis, as is the case with many other machine-learning algorithms. Decision trees are also robust to inclusion of irrelevant features, and the models are inspectable and interpretable. However, single decision trees have high variance and are not very accurate being prone to overfitting. Using an ensemble of decision trees helps to correct for overfitting that can happen when only a single decision tree is trained. Random forests, by averaging over an ensemble of decision trees, help to reduce the variance and increase the accuracy of the final model, but with some loss of interpretability

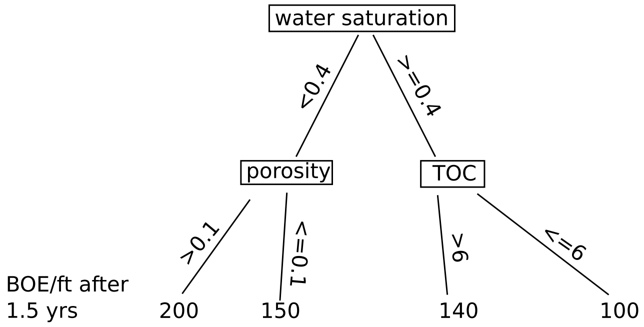

Figure 6. Simple example of a decision tree.

as compared to a single decision tree. We created two random forest models, one for each principal component score. We chose to predict the two principal components separately because they are not correlated (Figure 7).

The predictors in Table 1 were used in both models. Random forests have the advantage of handling data from a variety of sources, and at different scales, efficiently. They also provide a best estimate (mean of all trees), as well as a distribution (ensemble of all trees) for each estimate, allowing us to assess the uncertainty with each prediction.

As the number of trees in the random forest increases, the error decreases, up to a certain number of trees (approximately 200), at which the increase in trees has a negligible effect. We used 500 trees, which in most studies is more than sufficient (Hastie et al., 2001). We tuned the models for the parameter mtry, which is defined as the number of variables randomly sampled as candidates at each node (Breiman, 2002). The tuning process was performed with the *R* package "randomForest," which evaluated the mean-squared error (MSE) on out-of-bag (not used in the analysis) data points, finding the value of mtry that resulted in the lowest possible MSE. The resulting values of mtry for the models predicting the first and second scores on FPCs were 8 and 8, respectively.

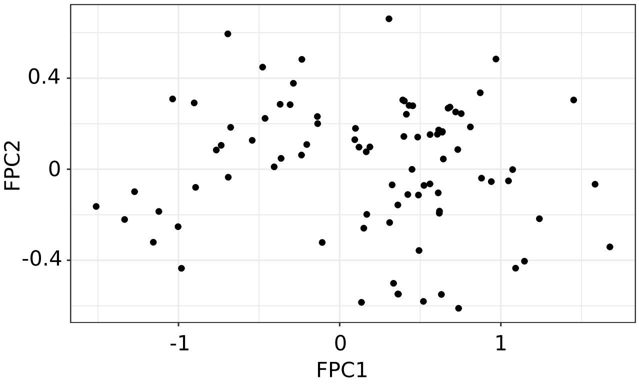

Figure 7. Crossplot of scores on the first and second FPCs. Note that there is no correlation between the two FPCs ($r^2 = 0.01$).

Because these models are complex, their interpretation is not always straightforward. One way of interpreting random forests is through variable importance metrics. A popular variable importance metric is the degree to which permuting (randomly shuffling) any given variable increases the MSE of the prediction. Because permuting the variable is essentially just feeding noisy data into the model, variables which increase the MSE significantly when permuted are considered more important, whereas variables which have little effect on the error are considered less important. Variable importance plots for random forest models predicting the first and second FPCs are shown in Figure 8. Higher values of MSE are associated with more important predictors. Thus, in the first FPC (Figure 8a), which is strongly correlated with production, tvd (depth), EFF_LAT (length of lateral well), porosity, and Fluid. Ft (the amount of fluid used to stimulate the reservoir per foot) are the most important predictor variables.

Another way to analyze the models is with partial dependence plots (Hastie et al., 2001). These plots are a useful way of evaluating the effect of any given variable on the overall prediction. Because

Table 1. Variables used in random forest model.

Variable (abbreviation if different)	Source/description
Geology:	
Porosity	Interpolated from logs, median along lateral length of well
Water saturation (Sw)	Interpolated from logs, median along lateral length of well
Total organic carbon (TOC)	Interpolated from well data, median along lateral length of well
Geothermal gradient (gradient)	Interpolated from well data, median along lateral length of well
Isopach	Interpolated from logs, median along lateral length of well
Total vertical depth (TVD)	Well report data, median along lateral length of well
ZONE	Principal zone that the well was drilled into (Lower Eagle Ford, Upper Eagle Ford, or mixed.
Seismic attributes:	
Eagle Ford curvature (EGFDp50)	Absolute value of mean curvature (of all directions) from AASPI. Median along lateral length of well, Eagle Ford horizon
Anacacho curvature (ANACACHOp50)	Absolute value of mean curvature (of all directions) from AASPI. Median along lateral length of well, Anacacho horizon
Completion parameters:	
Lateral length (EFF_LAT)	Completions database
Fluid per foot (Fluid.Ft)	Completions database
Stage spacing (Stage.Spacing)	Completions database
Starting choke size (choke)	Completions database
Trend of choke size (chokediff)	Best fit of choke size on time series. Raw choke size data from completions database

the variables are rarely independent, partial dependence plots account for the effect that other variables could have on the prediction by averaging the predicted value over all possible values of the predictors. By using partial dependence plots, we can evaluate the impact of key factors on the overall production. For instance, we can see if more curvature anomalies on average predict a high score on the second principal component.

Reconstructing production curves

After predicting the scores on the first and second FPCs, we were able to reconstruct decline curves. We held out approximately 25% of the data, or 26 wells, to validate our method. We compared the estimated total production with the actual total production. We also used as an error analysis the standard deviation of the ensemble of trees. The results of this comparison are shown in Figure 9. As shown, the true total production is within the error of the estimated total production for 23 out of the 26 wells. The r^2 for the relationship of predicted versus actual total BOE/foot is 0.65.

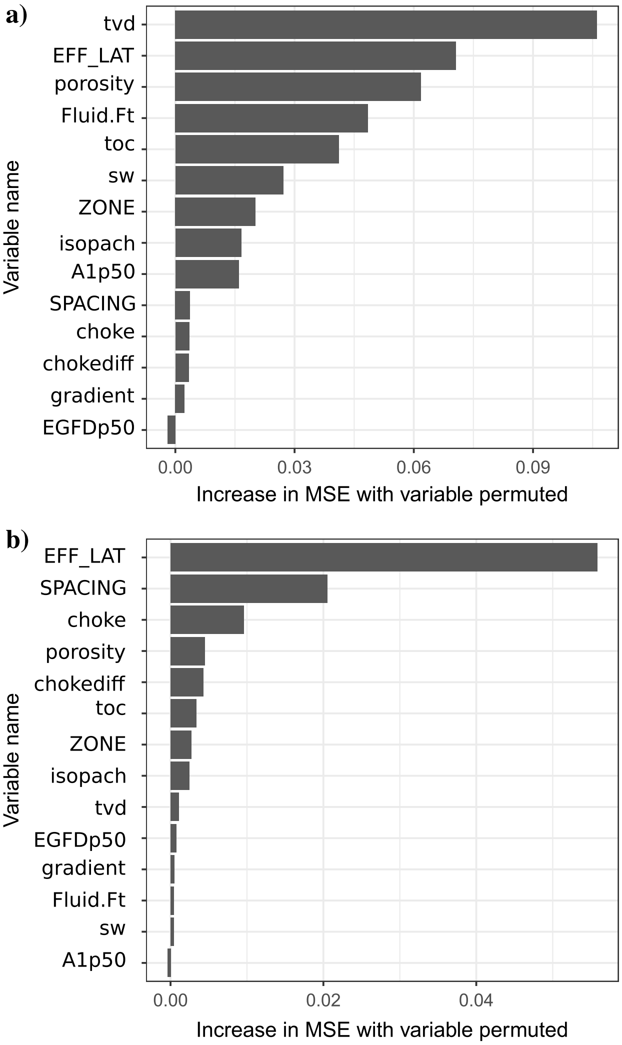

Figure 8. (a) Variable importance for predictions on the first FPC. (b) Variable importance for predictions on the second FPC. Both are shown as the increase in MSE when the variable is permuted or randomly shuffled.

The two wells that were not within one standard deviation of the prediction are shown with a dashed line in Figure 9. At each of these wells, we underpredicted the total production. In spite of the two wells that are underpredicted, this model is able to account for most (65%) of the variation in total well production, as well as characterize much of the temporal distribution of the production. With the two outliers removed, it accounts for 76% of the variation in total well production. These outliers have completion parameters that lie well within the distribution. The most pronounced outlier (the top dashed line in Figure 9, also shown in Figure 10d) has among the lowest values for TOC and relatively high values for water saturation. Both of these parameters are typically associated with worse production, yet the well performs much better than expected. The reasons for this unexpected behavior are currently unknown and require further research to determine.

DISCUSSION

The random forest models produced estimates of the scores on the first and second FPCs, which are directly related to decline curves as shown above. In addition to providing predictions of production, the random forest models can also be used to better understand the key drivers of well production. To assess the impact that each predictor has on the output, we used partial dependence plots. The first principal component was strongly correlated with total production, and the second principal component was not correlated with production; rather, it described the timing of the production — higher scores indicated that more production happened early on, but production dropped off more quickly as the life of the well continued.

We found that the depth of the well had the greatest impact on production, with deeper wells experiencing worse production (Figure 11a). All wells were drilled to the Eagle Ford formation, but the depth of the Eagle Ford Formation varies significantly across the study area. The depth of the well is likely correlated with a great

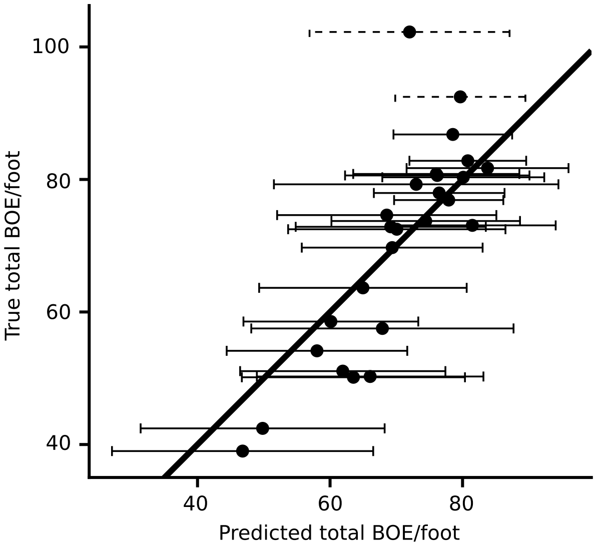

Figure 9. Comparison of the predicted total BOE/foot and the actual total BOE/foot. Outliers are noted with a dashed line. The error bar is a representation of one standard deviation, calculated from the random forest.

number of other parameters (i.e., geologic, geochemical, maturity, pressure, etc.) and further analyses would be required to determine these potentially lurking variables that all greatly affect well performance. As expected, we also found that wells with higher porosity, lower water saturation, and higher TOC performed better than average (Figure 11b–11d).

In addition to geologic parameters, we were also interested in exploring relationships between geophysical data and well production. Some have hypothesized that curvature can be used as a proxy for fractures, which could enhance production in unconventional resource plays (Gale et al., 2007). Testing this hypothesis was one of the objectives of this study. The relationship between curvature and well production, as described by partial dependence plots, is shown below in Figure 12a and 12b. Figure 12a shows that there is very little correlation between curvature in the Lower Eagle Ford and total well production, as indicated by scores on the first FPC. Figure 12b shows that there is a positive relationship between higher curvature on the stratigraphically higher Anacacho horizon and scores on the first FPC.

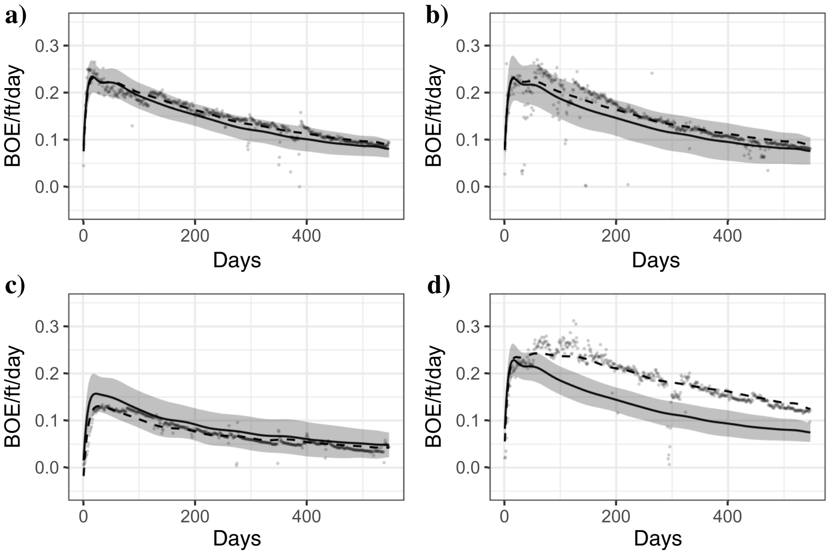

Figure 10. Comparison of decline curves predicted from the random forest model (solid black line) and actual production (dots). The dashed curve is the smooth fit to the data from FPCA. The grayed-out region is one standard deviation around the prediction, calculated from the random forests. (a-c) Prediction and observed data for three typical time series (24 out of the 26 wells had the FPCA decline curve fall within the gray one standard deviation zone). (d) Two out of the 26 wells did not fall within the gray zone. The one shown here is the largest outlier.

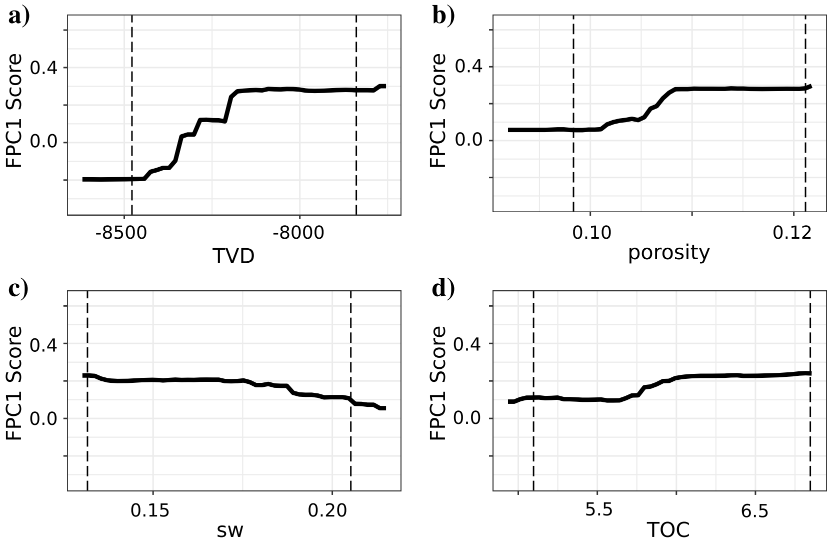

Figure 11. Comparison of geologic parameters with the first FPC, which represents total production. Dotted lines represent the fifth and 95th percentiles of each variable. Areas outside of the dotted lines are not as well-constrained due to fewer training data.

These analyses of the current data indicate that curvature is only weakly useful for identifying areas with higher production potential for the area in our study. The impact of curvature values at different horizons on production is inconsistent. Curvature anomalies in the Anacacho horizon, which is stratigraphically higher, could be a proxy for fractures that are subseismic in the Lower Eagle Ford. Further research is needed to validate this hypothesis, but it is one potential explanation for our observations.

In addition to curvature, we also experimented with including symmetry attributes, which are indicative of faults, as predictors. These attributes did not have a noticeable impact on our predictions. As we noted in the “Introduction” section, the methods developed in this study provide a framework for combining geologic data, seismic attributes, and engineering data to predict well production, but they do not constitute an exhaustive analysis of all seismic attributes that could be useful in predicting production. Additional seismic attributes could prove more helpful in predicting production in our study area. Furthermore, the usefulness of seismic attributes, and any predictors for that matter, is highly dependent upon the geologic setting. From our work, it appears that in our study area, geologic and engineering factors have a more pronounced relationship to production than curvature attributes. Many studies have found seismic data, including seismic attributes, to be closely related to properties controlling reservoir production (Skjervheim et al., 2005; Avseth et al., 2010; Iturrarán-Viveros, 2012; Na’imi et al., 2014). The extent to which they are effective in predicting productive areas is dependent upon the geology, as well as the attributes used. Future work could test the ability of additional seismic attributes to aid in predicting production in our study area or other study areas.

In addition to geologic and geophysical parameters, our random forest models also accounted for engineering parameters. The relationship of engineering parameters to well production is complex, and it often varies based on the surrounding geologic and geophysical properties of the rock. Because partial dependence plots average all geologic variables, they tend to blur out the interrelationships between

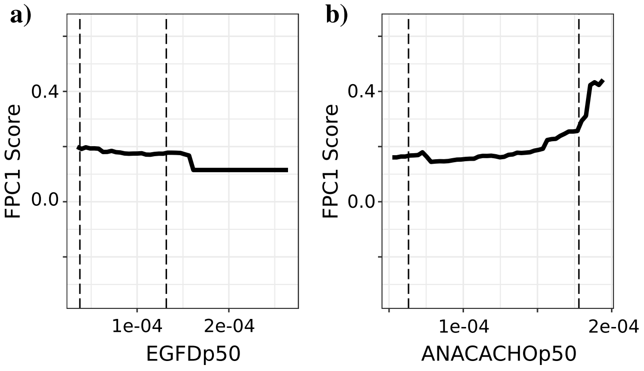

Figure 12. Partial dependence plots showing the influence of (a) Eagle Ford curvature and (b) Anacacho curvature on well production, as indicated by the FPC1 score.

engineering and geologic parameters. Furthermore, the ability of the random forest models to learn relationships between engineering parameters and production is entirely dependent on the training data. Because completion methods are continuously being improved, using relationships inferred from historical completion data to inform future completions decisions has some limitations. We consider our inclusion of engineering parameters as a way to normalize for them, while focusing our discussion on relationships between the production and the geologic and geophysical parameters.

CONCLUSION

This research seeks to use geophysical attributes together with production data to improve predictions about reservoir production. This is a complex problem because the effect of engineering parameters on well production is not well-understood. However, we have demonstrated that machine-learning techniques can use geologic and seismic data, as well as completions data, to predict production time series curves with an r^2 of 65%–76%. Although machine-learning methods can enhance predictions, it is still necessary to use physical models to understand the key mechanisms behind production.

ACKNOWLEDGMENTS

R. Smith was supported by a National Science Foundation Fellowship (grant DGE-114747). We would like to thank SM Energy for sharing their data for this study. We acknowledge support from the sponsors of the Stanford Center for Earth Resources Forecasting (SCERF) and from S. Graham, Dean of the Stanford School of Earth, Energy and Environmental Sciences.

DATA AND MATERIALS AVAILABILITY

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

REFERENCES

Al-Dossary, S., and K. J. Marfurt, 2006, Multispectral estimates of reflector curvature and rotation: Geophysics, **71**, no. 5, P41–P51, doi: 10.1190/1.2242449.

Avseth, P., T. Mukerji, and G. Mavko, 2010, Quantitative seismic interpretation: Applying rock physics tools to reduce interpretation risk: Cambridge University Press.

Bergbauer, S., T. Mukerji, and P. Hennings, 2003, Improving curvature analyses of deformed horizons using scale-dependent filtering techniques: AAPG Bulletin, **87**, 1255–1272, doi: 10.1306/0319032001101.

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

Breiman, L., 2002, Manual on setting up, using, and understanding random forests V3.1 (University of California at Berkeley, Berkeley, CA).

Cai, Q., W. Yu, H. C. Liang, J. T. Liang, S. Wang, and K. Wu, 2017, Development of a powerful data-analysis tool using nonparametric smoothing models to identify drillsites in tight shale reservoirs with high economic potential: SPE Journal, **23**, 719–721.

Gale, J. F., R. R. Reed, and J. Holder, 2007, Natural fractures in the Barnett shale and their importance for hydraulic fracture treatments: AAPG Bulletin, **91**, 603–622, doi: 10.1306/11010606061.

Grujic, O., 2017, Subsurface modeling with functional data: Ph.D. thesis, Stanford University.

Grujic, O., C. Da Silva, and J. Caers, 2015, Functional approach to data mining, forecasting, and uncertainty quantification in unconventional reservoirs: Presented at the SPE Annual Technical Conference and Exhibition.

Hastie, T., R. Tibshirani, and J. Friedman, 2001, The elements of statistical learning: Springer, Springer Series in Statistics.

Iturrarán-Viveros, U., 2012, Smooth regression to estimate effective porosity using seismic attributes: Journal of Applied Geophysics, **76**, 1–12, doi: 10.1016/j.jappgeo.2011.10.012.

Jolliffe, I., 2002, Principal component analysis: John Wiley and Sons Ltd.

Na'imi, S. R., S. R. Shadizadeh, M. A. Riahi, and M. Mirzakhanian, 2014, Estimation of reservoir porosity and water saturation based on seismic attributes using support vector regression approach: Journal of Applied Geophysics, **107**, 93–101, doi: 10.1016/j.jappgeo.2014.05.011.

Ramsay, J. O., 2006, Functional data analysis: John Wiley and Sons Inc.

Ramsay, J. O., G. Hooker, and S. Graves, 2009, Functional data analysis with R and MATLAB: Springer Science and Business Media.

Roberts, A., 2001, Curvature attributes and their application to 3D interpreted horizons: First Break, **19**, 85–100, doi: 10.1046/j.0263-5046.2001.00142.x.

Ruths, T., J. Zawila, S. D. Fluckiger, N. J. Miller, and R. G. Gibson, 2017, New methodology merging seismic, geologic, and engineering data to predict completion performance: The Leading Edge, **36**, 220–226, doi: 10.1190/tle36030220.1.

Sigismondi, M. E., and J. C. Soldo, 2003, Curvature attributes and seismic interpretation: Case studies from Argentina basins: The Leading Edge, **22**, 1122–1126, doi: 10.1190/1.1634916.

Skjervheim, J. A., G. Evensen, S. I. Aanonsen, B. O. Ruud, and T. A. Johansen, 2005, Incorporating 4D seismic data in reservoir simulation models using ensemble Kalman filter: Presented at the SPE Annual Technical Conference and Exhibition.