Please fill in the name of the event you are preparing this manuscript for.	SPE Annual Technical Conference and Exhibition
Please fill in your 6-digit SPE manuscript number.	SPE-206174-MS
Please fill in your manuscript title.	Application Of Machine Learning Methods To Well Completion Optimization: Problems With Groups Of Interactive Inputs

Please fill in your author name(s) and company affiliation.

Given Name	Surname	Company
Peng	Zhou	Department of Petroleum Engineering, Texas A&M University, College Station, TX, United States
Huiyan	Sang	Department of Statistics, Texas A&M University, College Station, TX, United States
Ligang	Lu	Shell International Exploration and Production Inc, Houston, TX, United States
Birol	Dindoruk	Shell International Exploration and Production Inc, Houston, TX, United States

Abstract

In unconventional reservoirs, optimal completion controls are essential to improving well productivity and reducing costs. In this article, we propose a statistical model to investigate associations between shale oil production and completion parameters (e.g., completion lateral length, total proppant, number of hydraulic fracturing stages), while accounting for the influence of spatially heterogeneous geological conditions on hydrocarbon production. We develop a non-parametric regression method that combines a generalized additive model with a fused LASSO regularization for geological homogeneity pursuit. We present an alternating augmented Lagrangian method for model parameter estimations. The novelty and advantages of our method over the published ones are a) it can control or remove the heterogeneous non-completion effects; 2) it can account for and analyze the interactions among the completion parameters. We apply our method to the analysis of a real case from a Permian Basin US onshore field and show how our model can account for the interaction between the completion parameters. Our results provide key findings on how completion parameters affect oil production in that can lead to optimal well completion designs.

Introduction

With the advance of horizontal well drilling and multi-stage hydraulic fracturing completion technologies, production from shale has grown rapidly, making it a major source of hydrocarbon production in the United States. Since the oil price downturn in 2014, the US oil industry has been focusing on reducing the operational expenditures and improving well productivity to improve profit margins and cash flow in the low oil price environment (Curits and Montalbano, 2017). There is a strong demand from the industry to quantify the relationships between completion parameters and hydrocarbon production, which can guide the decision process to more effective completion designs for cost-saving and production enhancement.

There are several major challenges in developing the methodology to quantify the relationships between completion parameters and hydrocarbon production. First, well production is influenced simultaneously by a complex combination of many factors such as geological conditions, reservoir fluid properties, and

well completion parameters. It remains largely elusive how geological factors can be effectively and feasibly estimated and subsequently removed from the quantitative analysis of the relationships between the completion parameters and the production. Second, even with the knowledge on how to effectively remove the effects from the geological factors, the functional associations between completion features and production can be complex in nature and challenging to characterize. Third, the challenge is aggravated by the spatial misalignment between production data and covariates. The database containing vertical deep well logs that penetrate shale layers and measure geological properties often do not have production data, and horizontal wells that have production data often do not penetrate the full formation so the geological measurements at the same location are missing. Finally, the relationships between the completion parameters and hydrocarbon production can be heterogeneous across different sub-regions in a large reservoir area.

Over the years, much research work (Wilson, 2018; Chorn *et al,* 2014; Carpenter, 2018; Dosunmu and Osisanya, 2015; Malayalam *et al*, 2014) has been done on how to optimize completion parameters to enhance oil production from horizontal wells. Numerical simulation is considered as a reliable method that can be applied to define the optimal completion parameters (e.g., amount of proppant for hydraulic fracturing, completion lateral length and stage number of horizontal wells). To run reservoir numerical simulations, one needs to generate geological models that are considered as good representations of the subsurface geological conditions. It requires first to build a geological model that incorporates some prior information (e.g., well logs and geomechanics) collected from the oil field. The resulting geological model will then be calibrated according to the production data by history matching with multiple possible realizations generated for uncertainty quantification (Oliver *et al,* 2008; Gao *et al,* 2016; Chen *et al*, 2018). After these steps, one runs the reservoir simulation under the constraints of different possible completion designs, and then compares the cumulative oil production within a time period (e.g., 12 months) or estimated ultimate recovery (EUR) to decide the optimal completion strategy. The advantage of using numerical simulation is that it is a physics-based approach and takes account of the influence of many factors (e.g., reservoir geological environment, completion method) on the well production. In addition, once a reliable reservoir model becomes available it can address many other complicated engineering problems. Therefore, reservoir simulation is an important tool for reservoir engineers to get insight on field development. However, the disadvantage of numerical simulation is also obvious. It requires tremendous amount of work to build the geological model; the computational cost to run reservoir simulations is also very expensive. Given the complexity of the subsurface geological environment of unconventional reservoirs, the geological model typically requires a large number of simulation grids to characterize the complex geometry of fractures, which makes the computation cost even more unbearable while increasing the uncertainty on the simulation results.

In recent years, artificial intelligence (AI) technologies such as machine learning, deep learning and statistical data mining have been recognized to be useful in solving challenging problems in oil industry that require extensive field data analysis (Chen *et al.* 2020; Yang *et al.* 2020; Yang, Lu, *et al.* 2020; Sen, Ong, *et al*. 2020; Sen, Chen *et al.* 2020; Zhou *et al.* 2018; Pan *et al.* 2021). A large body of work (Lafollette *et al,* 2012; Centurion, S. M., 2011; Al-Alwani *et al*, 2019; Guevara *et al,* 2018; Guevara *et al*, 2019) has been done using data mining to investigate the questions such as how long the lateral wellbore to drill, how many stages to complete and how far apart to place them. Zhong *et.al.* applied and compared several data mining approaches such as Support Vector Machine, Random Forests, Classification and Regression Tree analysis, and Boosted Regression Trees to the data from the Wolfcamp Basin. Their finding reveals that completion lateral length and total proppant amount are important factors driving the first 12-month cumulative oil production. However, the authors did not take into account geological effects, which may lead to bias in their results and findings. Another study by Yuan *et.al.* used a linear model and claimed

through data analytical studies that no distinctive advantage of drilling a well with longer lateral length was found, and no clear correlation trend was observed between longer lateral length and better production performance in the Barnett Basin. The recent progress in the study of completion design using data mining has increased our understanding of key completion controls on shale reservoir productivity. But to the best our knowledge, no research work has been found using data mining to analyze the quantitative relations between production and multiple completion engineering parameters while simultaneously considering the geological confounding effects.

In this paper, we propose a novel method of developing and applying a statistical nonparametric regression model that comprises two parts: a generalized additive model (GAM) (Wood, 2004) to investigate the functional associations between production and key completion parameters, and a random effect to account for geological confounding effects on production. Standard GAMs extend generalized linear models (GLM) (Nelder and Wedderburn, 1972) by replacing linear functions with a sum of smooth functions, which allows the exploration of possible nonlinear relationships between responses and covariates. We then extend the standard GAM models by adding a random effect to capture heterogeneous geological effects via locally homogeneity pursuit regularizations (see, e.g., Li and Sang, 2019), which have gained popularity in recent machine learning and statistics literature for high dimensional data due to their many nice computation and theoretical properties. We propose a regularized likelihood-based inference algorithm to estimate completion control effects and geological effects simultaneously. The method allows us to detect clustered heterogeneity across the Permian Basin automatically without the need to pre-specify the number and shape of clusters.

We will show how the proposed method performs through a use case in the Permian Basin, which is a large oil and natural gas producing area and has a growing number of new wells being drilled and completed. The dataset includes production data, completion data, geological data and other relevant information such as water-oil ratio (WOR) and gas-oil ratio (GOR) that we combine from various sources. The model performance is assessed and compared with other methods in terms of prediction accuracy on results from cross validation. We also present important findings and recommendations regarding completion controls. The method and its results provide a useful tool and guidance to completion engineers for well completion designs that maximize well productivity and saves completion cost.

Data Description and Preliminary Analysis

The data set after preprocessing in this investigation consists of 355 vertical well logging data, and 104 wells with at least 1-year production history in the target zone.

The completion engineering covariates available for analysis include the amount of proppant per stage (lbm/stage), fluid per stage (gal/stage), stage spacing (feet), and completed lateral length (feet). To determine geological effects, we use the logging data from vertical wells that consists of gamma ray (GR), rock density (DEN), deep resistivity (RESDEP) and neutron porosity (NEU_LIM). GR is known to be associated with the rock type in a reservoir. High GR value indicates high shale volume and lower GR value indicates lower shale volume. RESDEP indicates the water saturation level at the well location. High deep resistivity indicates lower water saturation since the hydrocarbon has a high resistivity compared to the fresh water in the formation. NEU_LIM and DEN are associated with the medium porosity at the well location. The density of pure sandstone is 2.65 g/cm^3 and the density of pure limestone is 2.71 g/cm^3, but the hydrocarbon and water in the pore will change the rock density. We can estimate the medium porosity according to NEU_LIM and DEN logs. In addition, we include water-oil ratio (WOR) and shut-in days (Days) as other potentially relevant covariates.

As aforementioned in Introduction, spatial locations of vertical wells that contain geological covariates differ from those of horizontal producing wells containing production and completion data as shown in **Figure 1**. But it is noticeable that vertical wells and horizontal wells are in general distributed in overlapping regions.

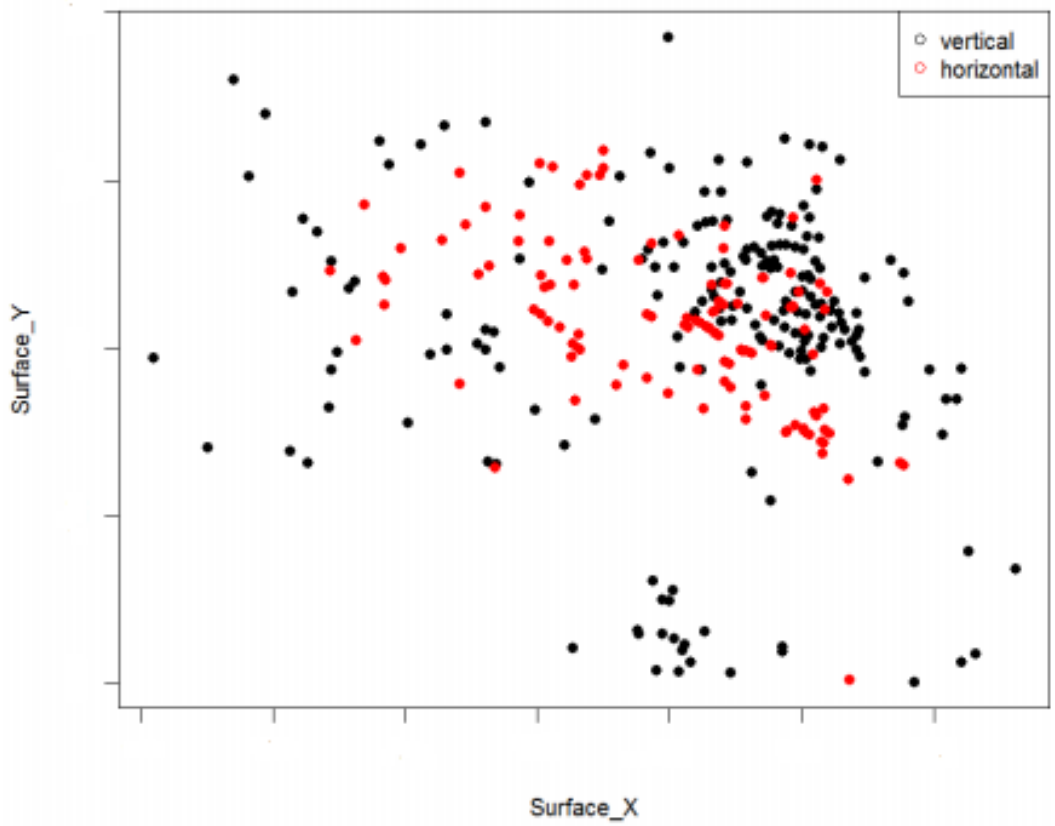


Figure 1: Spatial location distribution of horizontal (red) and vertical (black) wells for completion design analysis

Figure 2 shows the histograms of the response and relevant features. As we can see, the response variable, the 12-month cumulative oil production, approximately follows a log-normal distribution. And we will confirm in the numerical results in Section Real Data Results below that the log transformation on the response variable yields a lower cross-validation error compared to the model using the original response. In addition, from the histogram of completed lateral length, we can see there are very few wells with long lateral length from which we expect that the model will have large uncertainty in the relation between oil production and long completed lateral length.

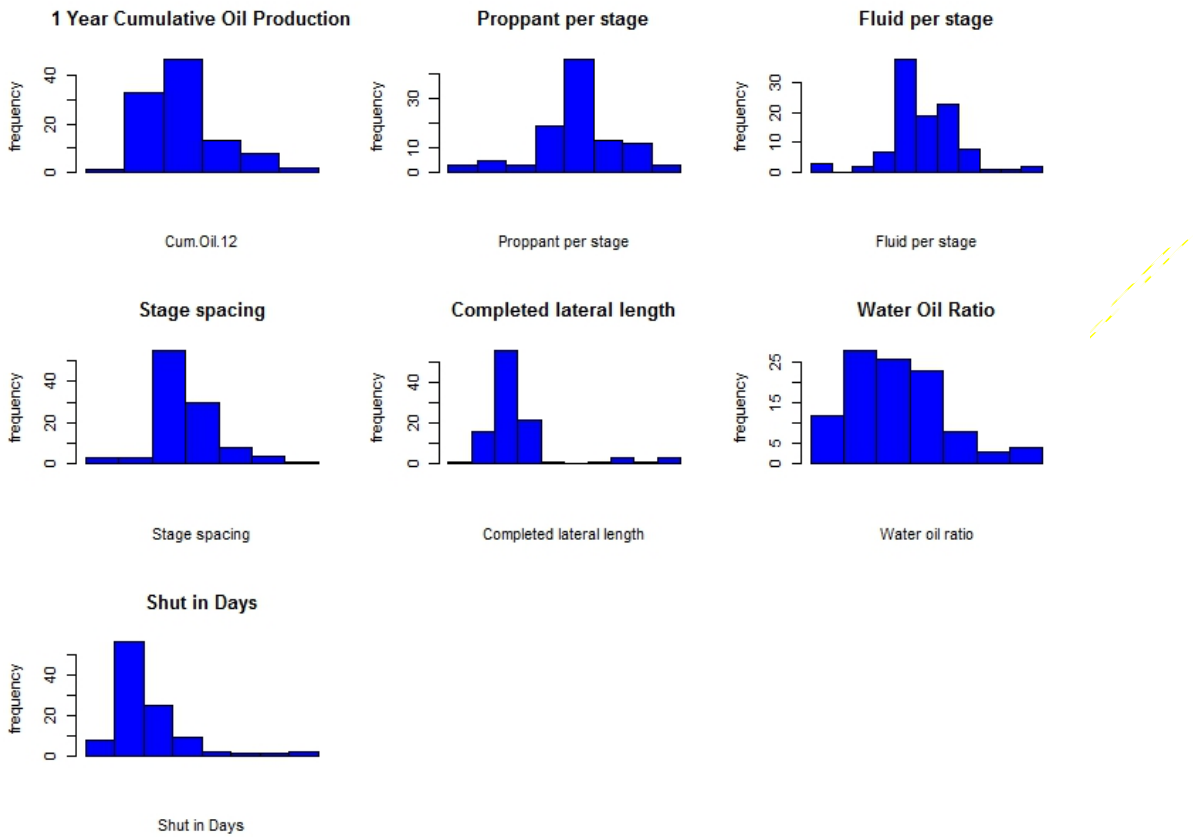


Figure 2: Histograms of response and covariate variables

We conducted a correlation analysis, and the results indicate some considerable relationships among oil cumulative production and the selected predictors (**Table 1**). For example, cumulative oil production is negatively correlated with WOR. In addition, we find that proppant per stage and fluid per stage also exhibit a strong correlation.

	Cum.Oil.12	Proppant per stage	Fluid per stage	Stage spacing	Completed lateral length	WOR	Shut in Days
Cum.Oil.12	1.0	0.2428	0.083	-0.221	0.168	-0.531	-0.082
Proppant per stage	0.2428	1.0	0.654	0.372	0.125	-0.233	0.029
Fluid per stage	0.083	0.654	1.0	0.567	-0.042	-0.120	-0.167
Stage spacing	-0.221	0.372	0.567	1.0	0.029	-0.111	-0.028
Completed lateral length	0.168	0.125	-0.042	0.029	1.0	-0.037	-0.014
WOR	-0.531	-0.233	-0.120	-0.111	-0.037	1.0	-0.269
Shut in Days	-0.082	0.029	-0.167	-0.028	-0.014	-0.269	1.0

Table 1: Correlation matrix of cumulative oil production with relevant features

Methodology

We introduce notations before giving details about our modeling method. Let $(\mathbf{s}_1, \ldots, \mathbf{s}_N)$ be the production well locations where N is the number of production wells, and $(\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_M)$ be the vertical well

locations where M is the number of vertical wells. We let Y_i denote the (log-transformed) production at the production well location $\mathbf{s}_i$, and $\mathbf{Y} = (Y_1, \dots, Y_N)^T$. Let $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,p})$ denote the vector of p covariates observed at production well location $\mathbf{s}_i$. For the Permian data, we have $p = 6$ such covariates including proppant per stage, completed lateral length, stage spacing, fluid per stage, WOR and shut in days at well location $\mathbf{s}_i$, respectively. We let $\mathbf{G}_j$ denote the vector of geological variables collected at vertical well location $\tilde{\mathbf{s}}_j$ for $j = 1, \dots, M$. We denote the L_1, L_2 norms by $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively.

We consider a regression model as follows:

$$Y_i = \mu_i + f(\mathbf{X}_i) + \epsilon_i \ (i = 1,2, \dots, N) \quad (1)$$

where ϵ_i is the residual capturing measurement error, and $f(\mathbf{X}_i)$ is a function describing the effects of covariates $\mathbf{X}_i$ on well production. As completion engineering variables are included in $\mathbf{X}_i$, the estimation and interpretability of this function is of high priority for practical engineering controls. We also include a random effect μ_i to represent the dependence and variability in Y_i that are unexplained by covariates $\mathbf{X}_i$ and random error ϵ_i, for instance, the confounding effects from spatial and geological variables. We let $\mu = (\mu_1, \dots, \mu_N)^T$ and $\mathbf{f} = \big(f(\mathbf{X}_1), f(\mathbf{X}_2), \dots, f(\mathbf{X}_N)\big)^T$. Below, we focus on the model specifications for $\mathbf{f}$ and μ, respectively.

Generative Additive Model

The relationship between completion parameters and production is expected to be a relatively smooth function. Therefore, we choose to model it using a generalized additive model (GAM), which has been acknowledged as an appealing choice to model multivariate functions (Hastie and Tibshirani, 1986; Lin and Zhang, 1999; Wood, 2004). It represents the relationships between the coariates (or predictors) and the dependent variables as a sum of unknown smooth functions, which flexibly allow linear or nonlinear fittings with relaxed assumptions on the actual relationship between the response and the predictors with interpretable results. Specifically, the GAM model for $\mathbf{f}$ takes the form

$$f(\mathbf{X}_i) = \sum_{k=1}^{p} f_k(X_{i,k}; \boldsymbol{\beta}_k) \quad (2)$$

in which each f_k is modeled as a smooth function with parameter $\boldsymbol{\beta}_k$ for $k = 1,2, \dots, p$.

One popular choice for modeling smooth functions (Ahlberg, N. and Walsh, 1967; Ferguson, J. C. 1964) is to use the cubic smoothing splines, where the natural spline basis functions are used with the knots placed at all the observed points to circumvent the problem of knots selection, and the coefficients of these basis functions are regularized to suppress overly wiggly components and to avoid over-fitting (Claeskens and Hjort, 2008).

Specifically, for the k-th predictor X_k, let $\phi_{1,k}, \dots, \phi_{N,k}$ be the truncated power basis functions for natural cubic spline with knots at the observed covariates $X_{1,k}, \dots, X_{N,k}$. Then each individual function f_k can be expressed as $f_k(x, \boldsymbol{\beta}_k) = \sum_{i=1}^{N} \phi_{i,k}(x)\, \beta_{i,k}$ where $\boldsymbol{\beta}_k = \big(\beta_{1,k}, \dots, \beta_{N,k}\big)^T$. To impose a smoothness assumption on f_k, we consider a smoothing penalty term defined as $\lambda_1 \boldsymbol{\beta}_k^T \Omega_k \boldsymbol{\beta}_k$, where Ω_k is the $N \times N$ smoothing penalty matrix whose (i,j)-th element $\Omega_k^{ij} = \int \phi_{i,k}''(t) \phi_{j,k}''(t) dt$. This matrix plays an important role to control for overfitting by penalizing the wiggliness for each f_k. The penalty term involves a so-called smoothing parameter λ_1, controlling the level of penalty; the larger the value of λ_1, the smoother the function.

We remark that for simplicity of illustration, equation (2) does not involve interaction terms between predictors. But it is straightforward to generalize this model to consider interactions among predictors if necessary by using, for instance, tensor products of spline functions (Wood, 2006).

Clustered random effect model

Next, we turn our attention to the model specification for the random effect μ that captures the remaining structures in the data that is unexplained by **f**.

It is reasonable to let μ depend on geological variables because well productivity is heavily tied to local geophysical conditions. Moreover, the subsurface consists of complex and heterogeneous multiple layers, and hence underground geophysical properties often change abruptly across layers. As a result, we expect that geological effects on production may also exhibit non-smoothly varying patterns. In addition, both oil production and geological variables are collected from well locations distributed in space. Previous studies (Tian *et al*., 2018) have suggested the existence of strong spatial patterns in these measurements. It is therefore desired to account for such spatial information when modelling μ. A remaining challenge in modelling μ is to deal with the issue of missing geological covariates at the production well locations.

With the above considerations, we choose to model μ as a clustered random effect determined by both geological covariates and spatial locations. Detecting these clusters allows straightforward interpretations of local associations between response variables and covariates.

Specifically, we propose a flexible regularization model for μ that extends the spatial fused lasso (Li and Sang, 2019) and the k-nearest-neighbor (KNN) lasso for non-parametric regression (Padilla *et al.*, 2018). The method is performed in the following steps:
Construct a graph denoted by $G = (V, E)$ based on information from geological covariates and spatial locations, where $V = v_1, v_2, \dots, v_N$ is the vertex set with N vertices and E is the edge set.
Use the graph from Step 1 to construct the fused lasso penalty for μ as follows:

$$\lambda_2 \sum_{(i,j)\in E} |\mu_i - \mu_j| \tag{3}$$

The regularization in eq. (3), referred to as the fused lasso penalty (Tibshirani and Taylor, 2011), is to encourage homogeneity between the geological effects at two locations if they are connected by an edge in E. λ_2 is a regularization parameter determining the strength of fused lasso penalty and hence the number of clusters. Since the solution of L_1 penalty results in exact fusion or separation between μ_i and μ_j, this regularization automatically leads to a spatially clustered geological random effect.

The edge set E is a key ingredient in the model since it reflects the prior assumption on the homogeneity structure of geological effects. Since similar geological conditions are likely to lead to similar effect on production, it is desirable to construct E such that pairs of locations that have similar values of geological parameters are included to reflect homogeneity among them.

To address the issue of missing geological covariates at the production well locations and take into account spatial information, we first build a KNN graph connecting vertical well locations using their geological measurements. This initial graph is then used to build a new graph connecting horizontal wells based on spatial information. In the intial graph, we include all edges that connect a vertical well location with each of its k nearest neighbors. Here the neighbors are searched by using the distance metric defined on pricinpal component scores of geological parameter values. By principal component analysis, all wells share one space coordinate system spanned by the principal components and for each well, the corresponding score vector can be interpreted as the projection of the original vector (defined by the spatial coordinates and geological parameters at that specific well location) onto each unit principal component coordinate. As a result, we can take the principal component score vector as point coordinates that specify the location in the space spanned by the principal component vectors. Thus, we propose to use the principal component score metric to measure the geological similarity, which is analogous to the method we normally apply to define two-point distance in Euclidean space. Finally, to construct the new graph

connecting horizontal wells, we assume there exists an edge between $\mathbf{s}_i$ and $\mathbf{s}_j$ if the nearest neighbors of $\mathbf{s}_i$ and $\mathbf{s}_j$ are connected by the KNN graph on vertical wells.

The advantages of using fused lasso for cluster detection are in three folds; it provides an integrated approach that allows to detect clusters and estimate model parameters simultaneously, the number of clusters is completely data driven, and the resulting clusters have very flexible shapes as long as they form components of the graph (West *et al*., 1996).

The values of the nearest neighbors k and the number of clusters (or equivalently, λ_2) are two tuning parameters required for our estimation procedure. We determine the optimal number of nearest neighbors and the optimal number of clusters by the leave-one-out (LOO) cross validation criterion described in Section Estimation.

We note that there are other possible ways to construct the edge set E. For example, one may apply spatial interpolation methods to estimate geological parameter values at horizontal locations from measurements at vertical locations (Tian *et al*., 2018), and then construct a KNN graph using the interpolated values. But in practice, subsurface geological parameters often have highly complex dependence structures that can result in large interpolation errors.

Estimation

Using the regularization models for μ and $\mathbf{f}$ presented in Methodology, we have an optimization problem as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\{\mathrm{Y}_i-\mu_i-\sum_{k=1}^{p}\sum_{i=1}^{N}\phi_{i,k}(X_{i,k})\beta_{i,k}\}^2+\sum_{k=1}^{p}\lambda_1\boldsymbol{\beta}_k^T\Omega_k\boldsymbol{\beta}_k+\lambda_2\sum_{(i,j)\in E}|\mu_i-\mu_j| \quad (4)$$

Our goal is to find the estimates of $\boldsymbol{\beta}$ and μ that minimize the above objective function. Below, we will show an algorithm that updates μ and $\boldsymbol{\beta}$ iteratively until convergence.

Given values of μ, the basis regression coefficients $\boldsymbol{\beta}$ are estimated via a penalized least square method that takes a quadratic form, i.e., by finding $\boldsymbol{\beta}$ that minimizes

$$\frac{1}{N}\sum_{i=1}^{N}\{\mathrm{Y}_i-\mu_i-\sum_{k=1}^{p}\sum_{i=1}^{N}\phi_{i,k}(X_{i,k})\beta_{i,k}\}^2+\sum_{k=1}^{p}\lambda_1\boldsymbol{\beta}_k^T\Omega_k\boldsymbol{\beta}_k \quad (5)$$

We use the function **gam** in the R package "mgcv" to solve this optimization (Wood, 2019). And we follow the Generalized Cross Validaton (GCV) (Golub *et al*., 1979) criterion to estimate the smoothing parameter λ_1.

Given values of $\boldsymbol{\beta}$, μ is obtained by solving a regularized convex optimization as below:

$$\frac{1}{N}\sum_{i=1}^{N}\{\mathrm{Y}_i-\mu_i-\sum_{k=1}^{p}\sum_{i=1}^{N}\phi_{i,k}(X_{i,k})\beta_{i,k}\}^2+\lambda_2\sum_{(i,j)\in E}|\mu_i-\mu_j| \quad (6)$$

We can formulate the above equation as a generalized Lasso problem (Tibshirani and Taylor, 2011) as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\{\mathrm{Y}_i-\mu_i-\sum_{k=1}^{p}\sum_{i=1}^{N}\phi_{i,k}(X_{i,k})\beta_{i,k}\}^2+\lambda_2\|\mathbf{H}\mu\|_1 \quad (7)$$

where $\mathbf{H}$ is a $l\times N$ matrix constructed from the edge set E with l edges. For an edge connecting two vertices v_i and v_j, we can represent the penalty term $|\mu_i-\mu_j|$ as $|\mathbf{H}_l\mu|$, where $\mathbf{H}_l$ is a row vector of $\mathbf{H}$ only containing two nonzero elements, 1 at i-th element and -1 at j-th.

The path following type of algorithms (Arnold and Tibshirani, 2016; Shen and Huang, 2010) and alternating direction methods of multipliers (ADMM) (Boyd *et al*. 2011) have been developed to solve the generalized lasso problem. In this article, we use the standard ADMM method which proceeds by first decoupling the likelihood term and the regularization term by introducing new equality contraints $\mathbf{H}\mu-$

$\boldsymbol{\gamma} = \mathbf{0}$, such that the optimization can be written as

$$\underset{\mu,\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{N}\|\mathbf{Y} - \mu - \mathbf{f}(\mathbf{X}, \boldsymbol{\beta})\|_2^2 + \lambda_2\|\boldsymbol{\gamma}\|_1, \text{subject to } \mathbf{H}\mu - \boldsymbol{\gamma} = \mathbf{0} \tag{8}$$

Then, the standard ADMM solves the above equivalent formulation following the iteration steps as below:

$$\begin{aligned} &\text{Step 1: } \mu_{(t+1)} = (\mathbf{I} + \rho\mathbf{H}^T\mathbf{H})^{-1}\{\mathbf{Y} - \mathbf{f} + \rho\mathbf{H}^T(\boldsymbol{\gamma}_{(t)} - \mathbf{u}_{(t)})\} \\ &\text{Step 2: } \boldsymbol{\gamma}_{(t+1)} = S_{\lambda_2/\rho}(\mathbf{H}\mu_{(t+1)} + \mathbf{u}_{(t)}) \\ &\text{Step 3: } \mathbf{u}_{(t+1)} = \mathbf{u}_{(t)} + \mathbf{H}\mu_{(t+1)} - \boldsymbol{\gamma}_{(t+1)} \end{aligned} \tag{9}$$

where $S_{\lambda_2\backslash\rho}$ is the elementwise soft thresholding operator (Tibshirani, 1996) that maps $\mathbf{H}\mu_{(t+1)} + \mathbf{u}_{(t)}$ to $\boldsymbol{\gamma}_{(t+1)}$ in the following way

$$S_{\lambda_2/\rho}(x) = \operatorname{sgn}(x)\max(|x| - \frac{\lambda_2}{\rho}, 0) \tag{10}$$

where ρ denotes the step-size parameter and $\operatorname{sgn}(x)$ denotes the sign function of x. For the detailed derivation of eq. (9) and the general convergence properties of ADMM, we refer to the work from Boyd *et al*. 2011. We use the **admm.genlasso** function in the R package "penreg" (Huling, 2017) to implement these steps.

The two optimizations in eq. (5) and eq. (6) are run iteratively until cluster memberships and model parameters converge. LOO cross validaton is used for model assessment and tuning parameter selection because of the limited number of wells available for the study. Given N horizontal wells, we choose $N - 1$ horizontal wells as training wells to build a production model that incorporates both geological and completion effects and then make a prediction on production for the hold-out well. Each well in our study will serve as a test well once and finally, we will compare the true production values with the prediction obtained from cross validation. The equation of cross validation error is

$$\text{C. V. Error} = \sqrt{\frac{\sum_i (\mathrm{y}_{\mathrm{true},i} - \mathrm{y}_{\mathrm{pred},i})^2}{\sum_i \mathrm{y}_{\mathrm{true},i}^2}} \tag{11}$$

where $\mathrm{y}_{\mathrm{true}}$ and $\mathrm{y}_{\mathrm{pred}}$ denote the true value and prediction value of 1-year cumulative oil production, respectively. We select the combination that yields minimum LOO as the optimal values for the nearest neighbors and the clustering number.

Real Data Results

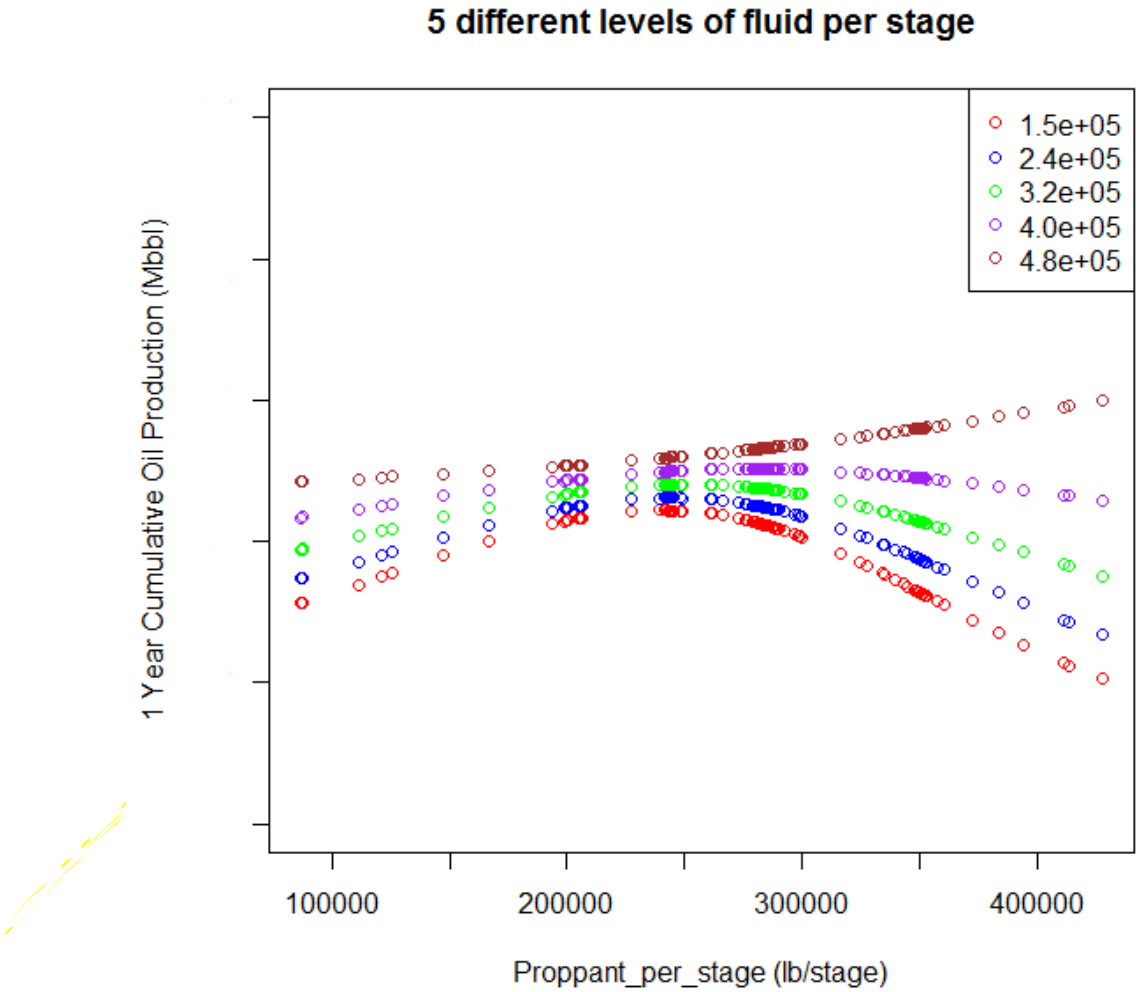
Before running the regularization model in Section Methodology, we perform a preliminary analysis using the GAM model only to select more relevant covariates from **X** according to the LOO cross validation criterion. The results indicate that the most important features for oil production are proppant per stage, fluid per stage, stage spacing, WOR and shut-in days. According to fracture mechanics, the amount of proppant and fluid used in hydraulic fracturing are correlated. Therefore, an additional term is added in the GAM to take account of the interaction between proppant and fluid. Completed lateral length does not show statistical significance in predicting oil cumulative production (**Table** 1). This is because in our dataset most of the wells have completed lateral length concentrated within one interval, and hence we do not see the importance of this feature. However, according to the domain knowledge from production enginnering, we believe completed lateral length is an important feature and in addition, since one of our major goals is to answer the question on whether longer completed lateral length necessarily indicates higher cumulative production, we will include completed lateral length in the final model to examine the effect of long completed lateral length on oil production.

We then analyze the Permian data using the model in (1). **Table 2** reports the LOO cross validation errors for different combinations of the two tunning parameters; the number of neighbors k in the KNN graph

and the number of clusters. From **Table** 2, we can see that the optimal number of neighbors is 4 and the optimal number of clusters is 3, which corresponds to the lowest LOO cross validation error among all combinations. Fixing the tunning parameters at these values, we obtain the parameter estimation results.

Clusters / Neighbors	2	3	4	5	6	7	8	9	10
3	NA	NA	0.260	0.257	0.253	0.253	0.255	0.254	0.253
4	0.256	0.242	0.251	0.251	0.252	0.253	0.254	0.255	0.256
5	0.257	0.256	0.254	0.253	0.254	0.255	0.256	0.258	0.259

Table 2: Leave one out cross validation error (geological clustering)



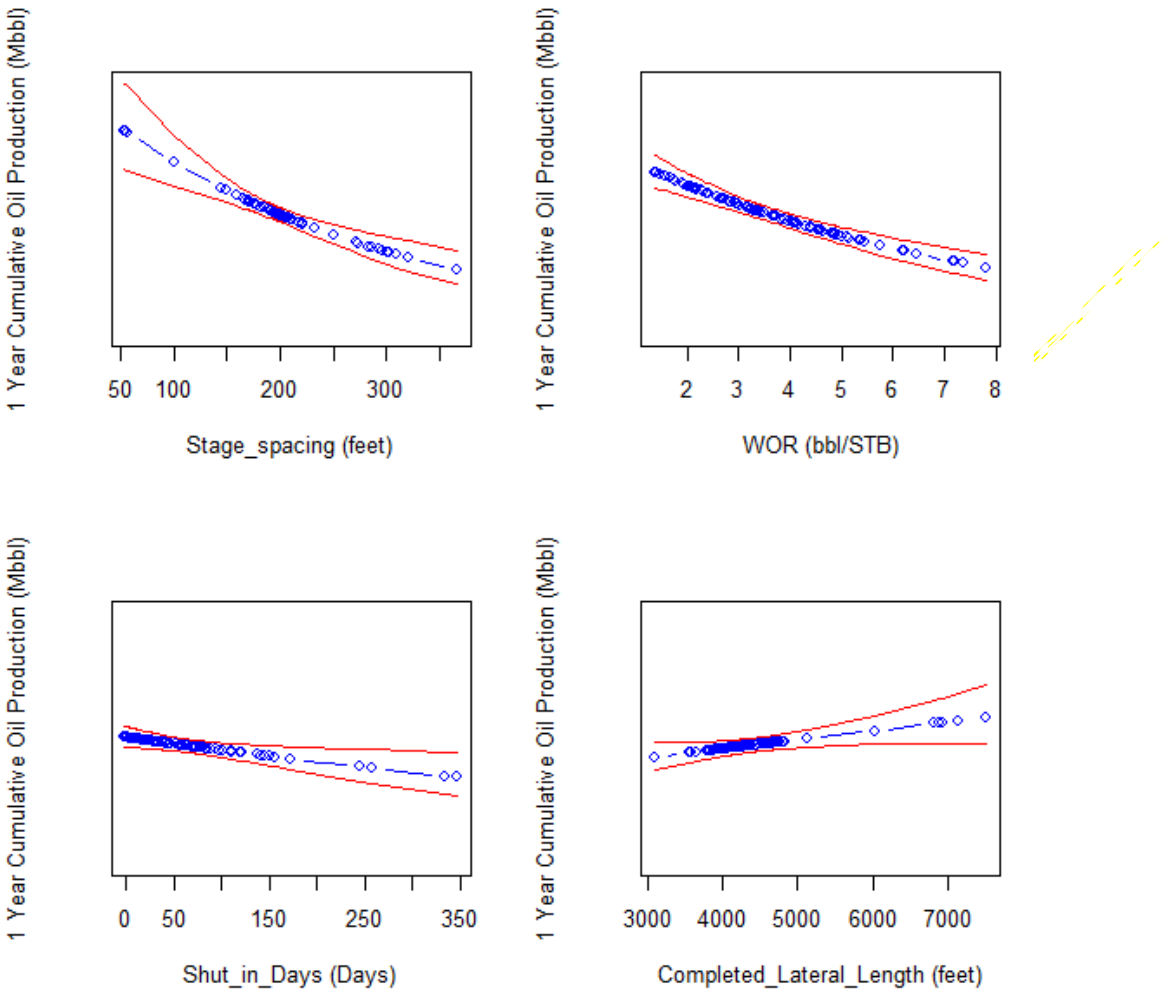


Figure 3: The relation of oil cumulative production with key nongeological features (geological clustering by logging data)

The top subfigure in **Figure 3** reveals the functional association between oil cumulative production and proppant per stage at different levels of fluid per stage. It is important to notice that the optimal proppant per stage changes depending on the level of the fluid per stage. The more fluid we inject, the larger the fracture volume will be. In addition, when the amount of fluid per stage is fixed, injecting more proppant does noe necessarily lead to an increase in oil production. At the average level of fluid per stage, the proppant efficiency starts to decline when proppant per stage reaches beyond a certain high level.

Next, in the bottom subfigure of **Figure** 3, the blue dot line represents the expected contribution to oil cumulative production from each of the other four features and the two red lines in each plot represent the uncertainty on the estimation. The feature stage spacing is negatively correlated with oil cumulative production. The results show a trend that the stage spacing should be kept small, though there are not enough data points below the 150 ft range to support a definitive conclusion; the common choice of the stage spacing can be reduced to 150 ft from 200 ft to enhance the production. As expected, the higher WOR is, the lower the oil cumulative production will be. The same relation holds for the shut-in days. The last plot is a relation between oil production and completed lateral length. From **Figure** 3, when the completed lateral length (CLL) is in the range of 3500~5000 ft, production increases with CLL. For wells with a CLL longer than 5000 ft, the uncertainty grows significantly as the number of the wells decreases dramatically. However, the results support, with relatively high confidence, that the CLL should be closer

to 5000 ft to enhance the production. The method we presented will be helpful to completion engineers to decide the best completion parameters. In the example case above, according to **Figure** 3, we can determine the best value of stage spacing. The optimal value of proppant per stage depends on the amount of fluid per stage. The fluid per stage should be no less than 4.0×10^5 gal/stage. The amount of proppant per stage to be around 3.0×10^5 lbm/stage lbm/stagegiven the large uncertainty that exists at large proppant per stage. The completed lateral length should be around 5,000 ft.

Now we examine the geological clustering results presented in **Figure 4**. Overall, there are three clusters. 11 wells in the red cluster at the bottom right corner have the lowest geological random effect and these 11 wells are all in zone WC_SH_B1. 8 wells in the blue cluster at the top left corner have the highest geological random effect and these 8 wells are all in zone WC_SH_B. Therefore, we can see the detected clusters partially agree with the internal zone information, which is defined as the internal layer where the horizontal lateral is located and is usually derived from a complex and time-consuming analysis of raw geological data by geologists and petro-physicists. In addition, we found the LOO cross validation error directly based on the feature internal zone is 0.251, which is higher than the LOO cross validation error 0.242 achieved from fused lasso geological clustering. This comparison shows us the advantages of the proposed clustering method. If a reservoir is a sweet spot with good geological properties, then we expect the production will be high; if the geological condition is the same, then good completion design will further increase well productivity.

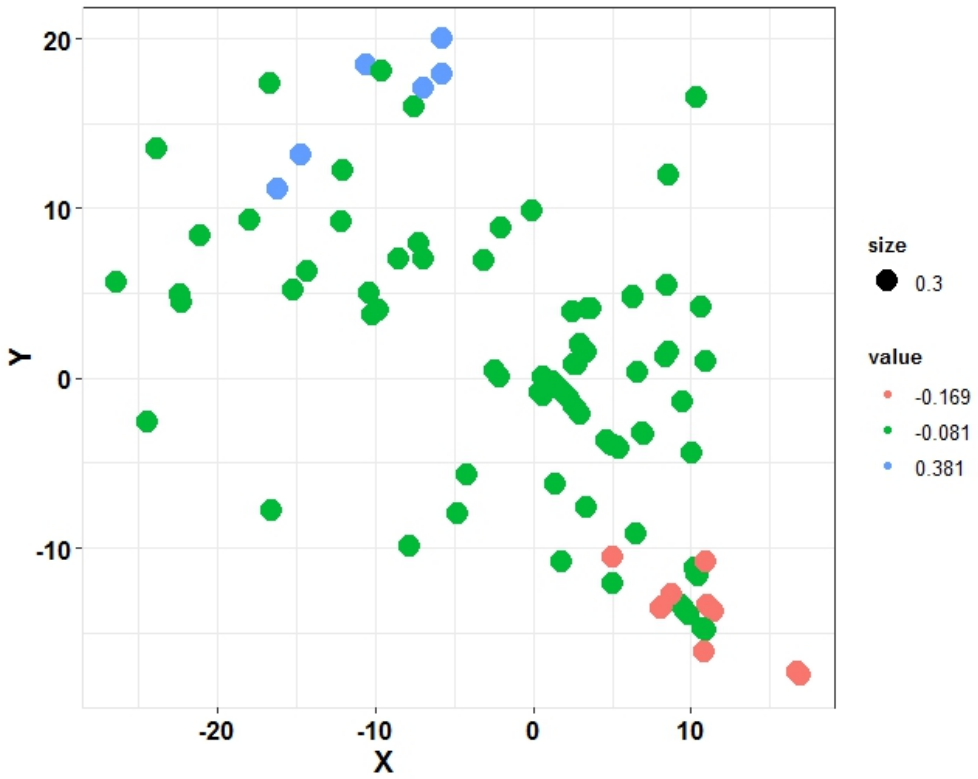


Figure 4: Geological clustering (X,Y coordinate values are hidden for confidentiality purpose)

Last, we compare our method with the model derived from the alternating conditional expectation (ACE) algorithm (Breiman and Friedman, 1985). ACE builds functional transformations of the independent variables $X_k \rightarrow g_k(X_k)$ and the response variable Y $\rightarrow h$(Y) which minimize the regression error variance in the transformed space:

$$e^2(h, g_1, g_2, \dots, g_p) = \frac{E\{[h(Y) - \sum_{k=1}^{p} g_k(X_k)]^2\}}{E[h^2(Y)]} \quad (12)$$

The features used in the ACE model are the same as the features used in the GAM model, but the ACE model does not take account of the interaction between proppant and fluid. We report the result of the ACE algorithm in **Figure 5**, from which we can see the transformed functions of proppant per stage and fluid per stage are not clear in guiding the completion engineers to determine the best parameter values. Also, the LOO cross validation error from the ACE model is 0.29, which is 16% higher than the GAM model. As a result, we recommend the GAM model for the oil production model.

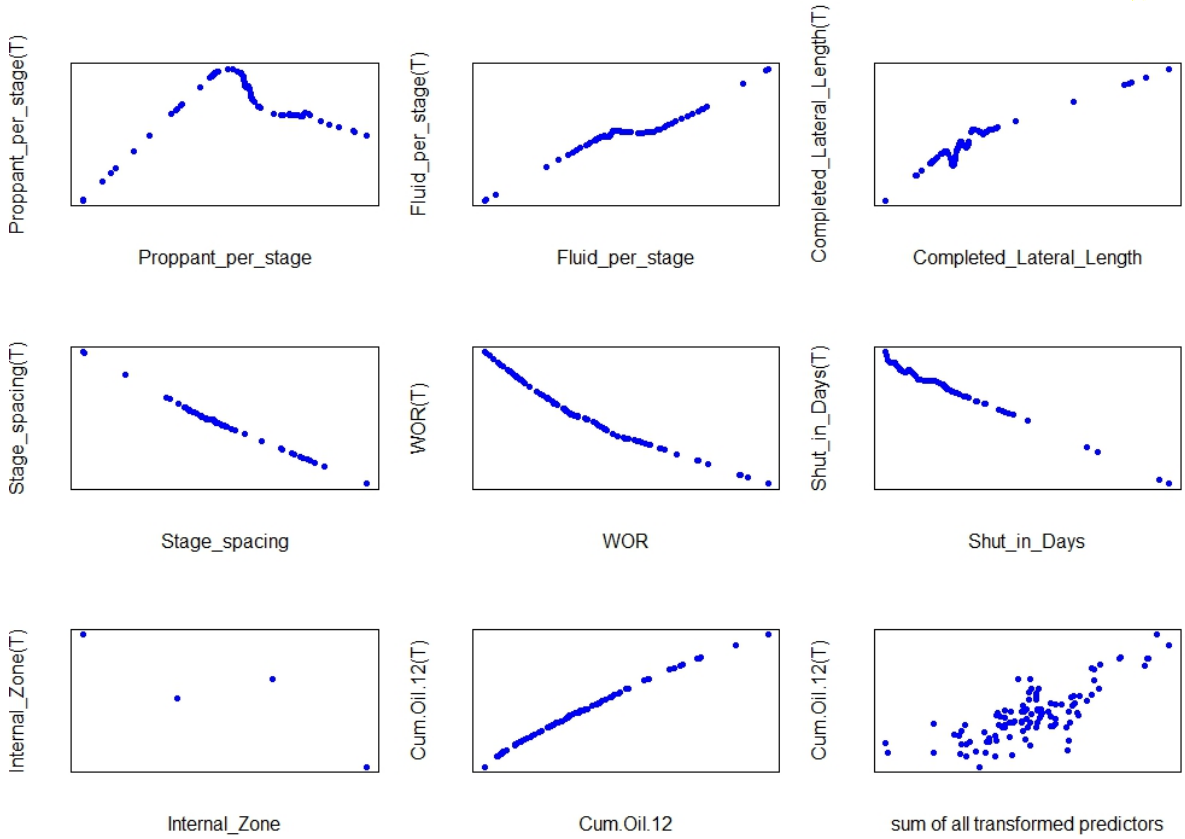


Figure 5: Nonlinear transform of features by the ACE algorithm.

Conclusion

In this paper, we proposed a new approach for well completion design optimization. As well production is a combined reult of geological, completion, and other (for example, operation) effects, we incorporated the geological confounding effect in the model by treating it as a clustered random effect using graphical fused lasso. tTo investigate and quantify the likely nonlinear associations between production and key completion parameters including the interactions between the parameters themselves we applied a generalized additive model (GAM) with a fused LASSO regularization for geological homogeneity pursuit. We have showed how our method can remove the non-completion effects and provide the guidance to the completion engineers for determining the optimal completion parameters by a real case in the Permian asset. The results have shown that the following five features are important for the completion design to maximize oil productions (1) proppant per stage (2) fluid per stage (3) stage spacing (4) WOR (5) shut in days. From the example above, the optimal completed lateral length may be longer than 5000ft. This is because the completed lateral length in the dataset has low variation and is concentrated within

one region. By comparing to ACE, we have also shown the advantages of our method in readily accounting the interaction of the covariats and in LOO lower cross-validation error.

Nomenclature

$\|\mathbf{x}\|_1$	=	$\sum_{i=1}^{n}\|x_i\|$ with $\mathbf{x} = (x_1, x_2, \dots, x_n)^{\mathrm{T}}$
$\|\mathbf{x}\|_2$	=	$\sqrt{\sum_{i=1}^{n} x_i^2}$ with $\mathbf{x} = (x_1, x_2, \dots, x_n)^{\mathrm{T}}$
$\mathbf{X}_i$	=	$(X_{i,1}, X_{i,2}, \dots, X_{i,p})$ the vector of predictors for the i-th horizontal well.
μ_i	=	the geological random effect of the i-th horizontal well
Y_i	=	cumulative oil production of the i-th horizontal well
$f(\mathbf{X}_i)$	=	the unknown function describing the effect of the covariates $\mathbf{X}_i$ on production
ϵ_i	=	the measurement error of the i-th horizontal well
N	=	number of horizontal wells
M	=	number of vertical wells
$\mathbf{s}_i$	=	spatial coordinate of the i-th horizontal well
$\tilde{\mathbf{s}}_i$	=	spatial coordinate of the i-th vertical well
$\mathbf{G}_i$	=	the vector of geological variables for i-th vertical well
$f_k(x; \boldsymbol{\beta}_k)$	=	the k-th unknown function based on the expansion coefficient vector $\boldsymbol{\beta}_k$ in GAM model
$\boldsymbol{\beta}_k$	=	$(\beta_{1,k}, \beta_{2,k}, \cdots, \beta_{N,k})^T$ the expansion coefficient vector in $f_k(x, \boldsymbol{\beta}_k)$
$\phi_{i,k}(x)$	=	the i-th basis function for the unknown function $f_k(x; \boldsymbol{\beta}_k)$
λ_1	=	smoothing parameter controlling overfitting
λ_2	=	strength parameter of graphical fused lasso penalty
V	=	the set of vertices in a graph
E	=	the set of edges in a graph
$\mathbf{H}$	=	a $l \times N$ matrix constructed from edge set E with l edges
$\mathbf{I}$	=	a $N \times N$ identity matrix
$\boldsymbol{\gamma}$	=	the auxiliary vector used in ADMM algorithm
ρ	=	the step size parameter used in ADMM algorithm
S_α	=	the soft thresholding function that depends on the parameter α

Subscripts/Superscripts

T	=	matrix transpose
(t)	=	the t-th iteration

Acknowledgements

This work was performed at Shell Technology Center in Houston, TX. Shell has the full ownership and right of this work.

References

Ahlberg, N. and Walsh. 1967. The Theory of Splines and Their Applications.

Al-Alwani, M. A., Dunn-Norman, S., Britt, L. K., Alkinani, H. H., Al-Hameedi, A. T., Al-Attar, A. M., Trevino, H. A. and Al-Bazzaz, W. H. 2019. Production Performance Evaluation from Stimulation and Completion Parameters in the Permian Basin: Data Mining Approach. URTEC-198192-MS. *SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference, 18-19 November, Brisbane, Australia.*

Arnold, T. B., and Tibshirani, R. J. 2016. Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics,* **25**(1): 1-27.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning,* **3**(1): 1-122

Breiman, L., and Friedman, J. H. 1985. Estimating Optimal Transformations for Multiple Regression and Correlation. *J. Am. Stat. Assoc.* **80**(391): 580-598.

Carpenter, C. 2018. Intelligent Completion in Laterals Becomes a Reality. SPE-0518-0058-JPT, **70**(05), 58-60.

Centurion, S. M. 2011. Eagle Ford Shale: A Multistage Hydraulic Fracturing, Completion Trends and Production Outcome Study Using Practical Data Mining Techniques. SPE-149258-MS. *SPE Eastern Regional Meeting, 17-19 August, Columbus, Ohio, USA*

Chen, C. H., Gao, G. H., Li, R. J., Cao, R., Chen, T. H., Vink, J. C. and Gelderblom, P. 2018. Global Search Distributed Gauss Newton Optimization Method and Its Integration With the Randomized-Maximum-Likelihood Method for Uncertainty Quantification of Reservoir Performance. SPE 182639, *SPE Journal.*

Chen, J., Schiek-Stewart C. and Lu, L. *et al.* 2020. Machine Learning Method to Determine Salt Structures from Gravity Data. Paper presented at the SPE Annual Technical Conference and Exhibition, Virtual, October 2020. SPE-201424-MS. Society of Petroleum Engineers. DOI: 10.2118/201424-MS.

Chorn, L., Stegent, N. and Yarus, J. 2014. Optimizing Lateral Lengths in Horizontal Wells for a Heterogeneous Shale Play. SPE-167692-MS. *SPE/EAGE European Unconventional Resources Conference and Exhibition, 25-27 February, Vienna, Austria*.

Claeskens, G. and Hjort, N. L. 2008. Model Selection and Model Averaging. *Cambridge University Press.*

Curits, T. and Montalbano, B. 2017. Completion Design Changes and the Impact on US Shale Well Productivity. The Oxford Institute for Energy Studies, University of Oxford.

Oliver, D. S., Reynolds, A. C. and Liu, N. 2008. Inverse Theory for Petroleum Reservoir Chracterization and History Matching. *Cambridge University Press*; 1 edition.

Dosunmu, I. and Osisanya, S. 2015. An Economic Approach to Horizontal Well Length Optimization. SPE-177866-MS. *Abu Dhabi International Petroleum Exhibition and Conference, 9-12 November, Abu Dhabi, UAE.*

Ferguson, J. C. 1964. Multi-variable curve interpolation. J. ACM, vol. 11, no. 2, pp. 221-228.

Gao, G, H., Vink, J. C., Chen, C. H., Alpak, F. O. and Du, K. F. 2016. A Parallelized and Hybrid Data Integration Algorithm for History Matching of Geologically Complex Reservoirs. SPE 175039, *SPE Journal*

Golub, G. H., Heath, M. and Wahba, G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics,* **21**(2), 215-223.

Guevara, J., Zadrozny, B., and Buoro, A. *et al*. 2018. A Hybrid Data-Driven and Knowledge-Driven Methodology for Estimating the Effect of Completion Parameters on the Cumulative Production of Horizon Wells. Paper presented at the SPE Annual Technical Conference and Exhibition, Dallas, Texas, USA, September 2018. SPE-191446-MS. Scoiety of Petroleum Engineers. DOI: 10.2118/191446-MS.

Guevara, J., Zadrozny, B., and Buoro, A. *et al.* 2019. A Machine-Learning Methodology Using Domain-Knowledge Constraints for Well Data Integration and Well Production Prediction. *SPE Res Eval& Eng,* SPE-195690-PA, **22**(04), 1185-1200. DOI: 10.2118/195690-PA.

Hastie, T., and Tibshirani, R. 1986. Generalized Additive Models, *Statistics Science,* 1, 297-318.

Huling, J. 2017. *https://github.com/jaredhuling/penreg*

Lafollette, R., Holcomb, W. D. and Aragon, J. 2012. Practical Data Mining: Analysis of Barnett Shale Production Results With Emphasis on Well Completion and Fracture Stimulation. SPE-152531-MS. *SPE Hydraulic Fracturing Technology Conference, 6-8 February, The Woodlands, Texas, USA*.

Li, F., and Sang, H. 2019. Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets. *Journal of the American Statistical Association,* 114*,* 1050-1062.

Lin, X., and Zhang, D. 1999. Inference in generalized additive mixed models using smoothing splines. *Journal of the royal statistical society: Series b (statistical methodology),* **61**(2), 381-400.

Malayalam, A., Bhokare, A., Plemons, P., Sebastian, H. and Abacioglu, Y. 2014. Multi-Disciplinary Integration for Lateral Length, Staging and Well Spacing Optimization in Unconventional Reservoirs. URTEC-1922270-MS, *SPE/AAPG/SEG Unconventional Resources Technology Conference, 25-27 August, Denver, Colorado, USA*.

Nelder, J. and Wedderburn, R. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General).* Blackwell Publishing. **135**(3): 370-384.

Pan, Y., Deng, L. and Zhou, P. *et al*. 2021. Laplacian Echo-State Networks for Production Analysis and Forecasting in Unconventional Reservoirs. *Journal of Petroleum Science and Engineering,* 207, 109068.
Sen, D., Chen, H. and Datta-Gupta, A. *et al.* 2020. Data-Driven Rate Optimization Under Geologic Uncertainty. Paper presented at the SPE Annual Technical Conference and Exhibition, Virtual, October 2020. SPE-201325-MS. Society of Petroleum Engineers. DOI: 10.2118/201325-MS.
Sen, D., Ong, C. and Kainkaryam, S. *et al*. 2020. Automatic Detection of Anomalous Density Measurements due to Wellbore Cave-in. *Petrophysics* **61**(05): 434-449. DOI: 10.30632/PJV61N5-2020a3.
Shen, X. and Huang, H. C. 2010. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105(490), 727-739.
Tian, Y., Ayers, W. B., Sang, H., McCain Jr, W. D., Ehlig-Economides, C. 2018. Quantitative Evaluation of Key Geological Controls on Regional Eagle Ford Shale Production Using Spatial Statistics. *SPE Reservoir Evaluation & Engineering,* **21**(02), 238-256.
Tibshirani, R. J. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267-288.
Tibshirani, R. J. and Taylor, J. 2011. The Solution Path of the Generalized Lasso. *The Annals of Statistics*, 3, 1335-1371.
West, D. B. *et al*. 1996. Introduction to graph theory, Vol. 2 Prentice Hall Upper Saddle River, NJ.
Wilson, A. 2018. Multiphase Flow Simulation Helps Find Optimal Lateral Length for Best Production. SPE-1118-0084-JPT, **70**(11), 84-85
Wood, S. N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association,* 99(467), 673-686.
Wood, S. N. 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics,* **62**(4), 1025-1036.
Wood, S. N. 2019. Package mgcv. *https://cran.r-project.org/web/packages/mgcv/mgcv.pdf*.
Yang, X., Dindoruk, B. and Lu, L. 2020. A comparative analysis of bubble point pressure prediction using advanced machine learning algorithms and classical correlations. *Journal of Petroleum Science and Engineering,* 185, 106598.
Yang, H., Lu, L. and Tsai, K. 2020. Machine Learning Based Predictive Models for CO2 Corrosion in Pipelines with Various Bending Angles. Paper presented at the SPE Annual Technical Conference and Exhibition, Virtual, October 2020. SPE-201275-MS. Society of Petroleum Engineers. DOI: 10.2118/201275-MS.
Yuan, G., Dwivedi, P., Kwok, C. K. and Malpani, R. 2017. The Impact of Increase in Lateral Length on Production Performance of Horizontal Shale Wells. SPE-185768-MS, SPE Europec featured at 79th EAGE Conference and Exhibition.
Zhong, M., Schuetter, J., Mishra, S., and Lafollette, R. F. 2015. Do Data Mining Methods Matter? A Wolfcamp Shale Case Study, URTEC-173334-MS. Unconventional Resources Technology Conference.
Zhou, P., Pan, Y., Sang, H. and Lee, W. 2018. Criteria for Proper Production Decline Models and Algorithms for Decline Curve Parameter Inference. Paper presented at the SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, Texas, USA, July 2018. URTEC-2903078-MS. DOI: 10.15530/URTEC-2018-2903078.

SI Metric Conversion Factors

lbm	x	2.20462	E+00	=	Kg
ft^3	x	3.53147	E+01	=	m^3
gal	x	2.64172	E+02	=	m^3

*Conversion factor is exact