



Society of Petroleum Engineers

FE-187328

Big Data Analytics for Production Data Classification using Feature Detection: Application to Restimulation Candidate Selection

Egbadon Udegbe, Eugene Morgan, Sanjay Srinivasan, The Pennsylvania State University

Copyright 2017, Society of Petroleum Engineers

This paper was prepared for presentation at the 2017 SPE Annual Technical Conference and Exhibition held in San Antonio, Texas, 9-11 October 2017.

This paper was selected for presentation by an SPE program committee following review of information contained in an abstract submitted by the author(s). Contents of the paper have not been reviewed by the Society of Petroleum Engineers and are subject to correction by the author(s). The material does not necessarily reflect any position of the Society of Petroleum Engineers, its officers, or members. Electronic reproduction, distribution, or storage of any part of this paper without the written consent of the Society of Petroleum Engineers is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of SPE copyright.

Abstract

In recent years, there has been a proliferation of massive subsurface datasets from sources such as instrumented wells. This places significant challenges on traditional production data analysis methods for extracting useful information, in support of reservoir management and decision-making. Additionally, with increased exploration interest in unconventional shale gas reservoirs, there is a heightened need for improved techniques and technologies to enhance understanding of induced and natural fracture characteristics in the subsurface, as well as their associated impacts on fluid flow and well performance.

The above challenges have the potential to be addressed by developing Big Data analytics tools that focus on uncovering masked trends related to fracture properties from large volumes of subsurface data, through the application of pattern recognition techniques. We present a new framework for fast and robust production data classification, which is adapted from a real-time face detection algorithm. This is achieved by generalizing production data as vectorized 1-D images with pixel values proportional to rate magnitudes. Using simulated shale gas production data, we train a cascade of boosted binary classification models, which are capable of providing probabilistic predictions. We demonstrate the viability of this approach for identifying hydraulically fractured wells which have the potential to benefit from restimulation treatment. The results show significant improvements over existing type-curve based approaches for recognizing favorable candidate wells, using solely gas rate profiles.

Introduction

In the past decade, there has been a significant increase in the application of streaming monitoring technologies, such as permanent downhole sensors and distributed acoustic sensing tools in instrumented wells. For example, we now have shale fields which contain several wells that generate long-duration production data (such fluid flow rate, density, pressure, and temperature), which are recorded in intervals of minutes to seconds. In the context of flow rate data, Big Data takes the form of abundant fluid volume time series measurements from multiple producing wells within a field, which must be analyzed and inferred for reservoir-related information. However, there has been a lag in the development of efficient techniques to harness the vast amount of reservoir-related information embedded in these well data.

Recently, there has been increased research interest in Big Data analytics for upstream applications, with specific focus on new tools and techniques for addressing problems related to infrastructure, data organization and management, analysis and discovery, and decision support (Feblowitz, 2013). We define the scope of this paper within the context of analysis and knowledge discovery from Big Data, more specifically for applications involving well production data. A number of recent publications within this focus area have applied data mining techniques in an attempt to extract useful information from voluminous subsurface time-series data (Seemann et al. 2013; Aulia et al. 2014; Patri et al. 2014). While these studies attest to the value in developing new analysis tools for extracting information from large volumes of data, there is currently insufficient literature on harnessing subsurface Big Data to improve characterization of hydraulically fractured wells within naturally fractured shale reservoirs.

In the past decade, technological advances in directional drilling and hydraulic fracturing have led to more economically feasible exploitation of unconventional reservoirs such as shale. In order to infer important reservoir properties and performance-related parameters, historical flow rate and pressure data from producing wells are typically analyzed using techniques such as straight line (regime flow) analysis, analytical models and numerical simulation approaches (Clarkson, 2013). Additionally, techniques such empirical decline curve models are commonly used to fit trends to well production data. However, most of these traditional methods (specifically type curve analysis and regime flow methods) are formulated for vertical wells in conventional reservoirs which means that they are generally predisposed to boundary-dominated flow, and only provide average information about bulk reservoir properties (Anderson et al., 2010). In addition, they do not take into account the complex physics of fluid storage and nano-scale fluid flow associated with shale. A number of dual-porosity analytical models have therefore been proposed to characterize production from multistage hydraulically fractured shale wells (Ozkan et al., 2009; Bello and Wattenbarger 2010; Brown et al. 2011). However, these analytical models are often developed in Laplace space corresponding to specific boundary conditions, which means numerical inversion schemes must be used to express the solutions in the time domain. Also, they typically require several matrix, fracture and well input parameters, which are often expensive or impossible to infer accurately.

Restimulation treatment has the potential to improve economic performance of producing shale wells by increasing the conductivity of existing fractures and/or enhancing their contact with the formation (Moore and Ramakrishnan, 2006). However, the influence of matrix and fracture characteristics on the success of restimulation are not completely understood, which has led to uncertainty in determining favorable candidate wells and optimal timing of restimulation treatment. A number of restimulation candidate selection methods have been proposed in literature. For example, favorable candidates have been identified by predicting untapped production capacity (as characterized by real-valued targets such as 5YCum) using artificial neural network (ANN) models, based on detailed data such as geographical, reservoir, production, well completion information. (Shelley, 1999; Mohaghegh et al., 2000). However, under these approaches, there is the potential to impose our understanding of the inputs which influence restimulation success on the model. In other words, there is the potential that some essential unmeasured properties are not captured as model inputs. Additionally, this approach may potentially lead to non-intuitive results (for example, the influence of a parameter such as surface elevation may be assigned non-zero weights). Also, some of the inputs used in training may be unavailable for new wells, which limits the generality of these models. In another study, dimensionless groups related to fracture and other physical parameters have been defined and correlated with restimulation success (Rousell and Sharma, 2011). Numerical simulation studies have also been used to study field production data, in order to define a set of heuristics for candidate selection and restimulation timing (Tavassoli et al., 2013). Overall, many of these existing methods tend to require detailed input data and/or lack generality to other reservoir settings, in addition to being time-consuming to implement, which limits their applicability to problems involving massive datasets.

In this paper, we aim to address the above challenges by developing a new methodology for fast and robust analysis of production data from hydraulically fractured wells. In the context of this work, robustness refers to consistent model performance in support of flexible development strategies (risk-

averse or aggressive), under a wide array of scenarios. We utilize a dual-permeability forward flow modeling approach to generate multiple realizations of rate production profiles by modifying fracture, reservoir and operational parameters. Using this data, we apply pattern recognition tools to help uncover trends associated with favorable and unfavorable restimulation candidates. This is achieved using a binary classification framework adapted from real-time face detection, which uses simple numerical criteria computed directly from raw flow rate data, thus eliminating the need for detailed information and promoting computational efficiency. We show that these features carry information about flow regimes and curvature attributes within well data, at various time scales and time periods during the life of the well. The algorithm also provides probabilistic predictions, which serve as a means to rank candidate wells. While the process of training the classifier can potentially be computationally intensive, the application of the trained classifier on the observed data is extremely fast, thereby rendering the method amenable to real-time classification on the basis of well performance. This in turn has the potential to impact decision-making workflows involving massive datasets collected from producing wells within a field, which helps to mitigate uncertainty in reservoir management and decision-making through continuous analysis.

Approach

In this paper, we propose a probabilistic classification framework which is aimed at addressing analysis challenges involving the massive well production datasets. This approach is demonstrated for a use-case in restimulation candidate selection. In order to achieve this, we have implemented a supervised learning framework. In the context of restimulation candidate selection, this means that we predefine two categories of example wells based on their observed rate response to restimulation – i.e. “favorable” and “unfavorable” restimulation candidates. Using this information, we train a statistical model to distinguish between both sets of examples and estimate the probability of restimulation success. The set of quantitative criteria used in defining favorable and unfavorable candidates will be addressed in upcoming sections.

As an alternative to data-driven techniques which utilize a limited set of reservoir and completion data to forecast real-valued fluid rates after restimulation, we propose an approach which makes categorical predictions on post-refrac well performance (i.e. successful or unsuccessful restimulation treatment) by taking into account subtle decline characteristics embedded within the production rate data. On this basis, the statistical approach in this study is adapted from real-time face detection techniques, which utilize texture-based features in characterizing input data, with the objective of fast and robust classification performance.

In the abundance of available field production records where the outcome of restimulation treatment is known, these well profiles can be directly used in training the classification model under the proposed framework. The trained model can then be used to predict if other candidate wells are likely to benefit from restimulation treatment. On the other hand, in the absence of field datasets, a flow simulation approach can be used to generate an ensemble of well production profiles to serve as training data. However, one important advantage of the proposed classification framework is that instead of explicitly forecasting a real-valued flow rate after restimulation, the classification model outputs a binary categorical response – “favorable” or “unfavorable” restimulation candidate – along with an estimated probability of refrac success. This enables us to generate training data using less detailed flow simulation models, which need only capture the broad reservoir and fracture characteristics in the productive region surrounding the wellbore.

Figure 1 below summarizes the overall workflow we have implemented in the development of the proposed production data classification algorithm.



Figure 1: Summary of overall workflow for developing the proposed production data classification model.

In the next section, we discuss the details of the forward flow simulation approach used in generating training examples and criteria for specifying favorable and unfavorable restimulation candidates. Next, we present the proposed statistical framework for production data classification. Finally we discuss the model training and evaluation before considering a field study.

Input production data

In line with the supervised learning approach proposed in this paper, the first step in implementing a classification scheme for identifying restimulation candidates, is to define examples of wells which have undergone both successful and unsuccessful restimulation. As earlier mentioned, in the abundance of field production data from restimulated wells with accompanying completion information, the proposed classification approach can be applied in order to “learn” the differences, based on rate decline characteristics. However, due to limited availability of field data, we utilize a forward flow simulation approach in this study.

The overall strategy is to simulate multiple realizations of gas flow rate from a hydraulically fractured shale gas well and assign class labels based on post-refrac performance. In the following sections, we discuss the flow simulation model and the reservoir characteristics which contribute to restimulation success.

Flow simulation model

In this paper, we utilize a probabilistic classification approach to develop a set of rules for identifying restimulation candidates, using rate decline characteristics. In other words, in contrast to existing data-driven approaches which explicitly forecast flow rate after restimulation, the proposed classification framework returns a categorical prediction (favorable or unfavorable candidate) and an accompanying probability of restimulation success. Therefore, in generating well data for training, this enables us to implement less detailed flow simulation models which need only capture the aggregate rate response to fracture and reservoir properties in the productive region surrounding hydraulically fractured wells.

On this basis, the flow simulation model used in this study features one multistage hydraulically fractured shale gas well within a 307-acre reservoir. This model is sized to represent the areal extent of a shale gas reservoir of mid-range productivity (Kulga, 2014). We have utilized a $61 \times 61 \times 5$ Cartesian grid with the X- and Y- dimensions fixed at 60 feet, while the vertical grid dimensions have been varied between 20 feet and 60 feet. Within this reservoir model, we randomly sample from a set of uncertainty parameters related to fracture, fluid, reservoir and operational properties, in order to generate an ensemble of well profiles with realistic shale gas rate decline characteristics.

Flow in naturally-fractured media is modeled in CMG using the dual-permeability approach with Gilman and Kazemi (1988) shape factor. In line with the stimulated reservoir volume (SRV) approach (Mayerhofer et al, 2006), we assume that the majority of reservoir productivity comes from a network of hydraulic fractures and rejuvenated natural fractures in the vicinity of the wellbore. Figure 2 below depicts the SRV geometry we have specified in order to model flow within the productive region around the horizontal well.

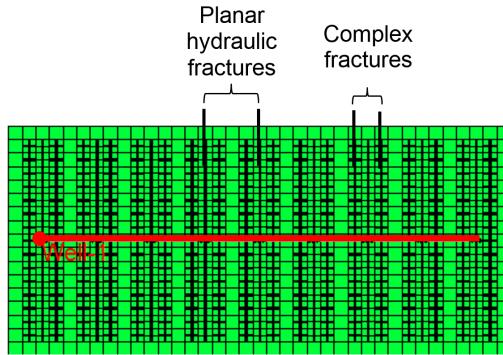


Figure 2: Stimulated reservoir volume (SRV) representation in flow simulator. The productive region around the reservoir is modeled by assuming equally-spaced planar hydraulic fractures each flanked by a network of rejuvenated natural fractures.

In the above figure, the area surrounding the wellbore has been modeled as a series of planar hydraulic fractures, each flanked by a network of rejuvenated natural fractures. We have assumed equally-spaced hydraulic fractures with symmetry about the horizontal wellbore. Both the hydraulic and rejuvenated fractures around the wellbore have been defined using $5 \times 5 \times 1$ logarithmic grid refinement around the fractures, which is intended to accurately capture the fluid flow within fractures, within the shale matrix, and between fracture and shale, as proposed by Rubin (2010). Under this framework, a 1-foot grid cell (central block in refined grid) is used to mimic the effect of a 0.001 foot fracture and the Forchheimer number is scaled to reflect this assumption. In order to generate an ensemble of well production profiles, we have randomly sampled uncertainty parameters related to the SRV fracture characteristics, as displayed in Table 1 below.

Property	Minimum	Maximum
Hydraulic fracture half length (ft)	270	570
Hydraulic fracture spacing (ft)	120	300
Number of hydraulic fracture stages	5	9
Hydraulic fracture conductivity (md-ft)	10	30
Hydraulic fracture tip conductivity reduction (%)	25	90
Complex fracture conductivity reduction (%)	60	75
Natural fracture spacing reduction in SRV (%)	0	90
Pressure dependent fracture conductivity decline constant, d_f	10^{-3}	10^{-5}

Table 1: Flow simulation parameters related to fracture characteristics in SRV region.

In Table 1, all parameters are sampled from uniform distributions, with the uncertainty ranges calibrated using publicly available field production data from West Virginia Marcellus Shale gas wells. The physical extent of the SRV region is dictated by the hydraulic fracture half-length, hydraulic fracture spacing and the number of hydraulic fracture stages. The conductivity at the tip of the fracture is allowed to decrease with increasing distance from the wellbore, using a percent reduction factor. This parameter serves as a proxy to account for mechanisms such as proppant crushing and fluid leak-off. Likewise, complex fracture conductivity is expressed as a percentage of induced hydraulic fracture conductivity. Apart from this, we have included a fractional parameter which accounts for the reduction of natural fracture spacing in the SRV region.

Finally, in order to capture pressure-dependent conductivity reduction in the induced and complex fractures, as proposed in Eshkalak et al (2014), we have applied the Chin and Raghavan (2002) correlation as follows:

$$k_f(\Delta P) = k_{fi} e^{-d_f \Delta P} \quad (1)$$

In Equation 1, d_f denotes pressure-dependent permeability decline rate (i.e. fracture closure rate), k_{fi} represents initial fracture permeability, and ΔP is the pressure drop from initial reservoir pressure. The induced and complex fractures are each assigned different rates of permeability reduction. Since we use constant fracture width in the flow model, d_f is referred to as a conductivity decline constant in Table 1.

Apart from the above fracture-related properties, we have randomly sampled 15 additional reservoir uncertainty parameters as inputs to the flow simulation model. These include diffusion and sorption effects, in order to account for physical phenomena specific to shale gas reservoirs. Also, well production is constrained using a constant pressure specification. The ranges of values from which these parameters have been sampled are shown in Table 2 below.

Parameter	Minimum	Maximum
Formation thickness (ft)	100	300
Non-SRV fracture spacing (ft)	0.1	10
Non-SRV fracture conductivity (md-ft)	0.0001	0.05
Shale porosity (%)	3	13
Shale permeability (md)	1.00E-03	1.00E-04
Shale compressibility (1/psi)	1.00E-06	3.00E-06
Diffusivity (cm ² /s)	2.00E-05	8.00E-05
Langmuir volume (SCF/ton)	50	250
Langmuir pressure (psi)	500	1500
Bulk density (g/cc)	2.1	3
Initial gas saturation	0.65	0.85
Initial pressure (psi)	2000	8000
Reservoir depth (ft)	5000	1000
Initial reservoir temperature (°F)	100	350
Flowing bottomhole pressure (psi)	250	1000

Table 2: Additional reservoir and operational parameters.

Restimulation candidate characterization

Using the flow simulator, the next step is to generate multiple gas production profiles corresponding to both favorable and unfavorable restimulation candidates, to serve as input to our classifier training algorithm. This is achieved using the following steps.

1. Sample multiple combinations of the fracture, reservoir and operational parameters in Table 1 and Table 2.
2. Put the well under constant pressure production for a fixed number of years before simulating a restimulation event
3. Compute post-refrac performance metrics using rate response one year after restimulation
4. Assign a categorical label: favorable candidate (1) or unfavorable candidate (-1), based on the observed production characteristics. Using these labels, we then present the pre-refrac production data to the classification algorithm and train it to predict without having access to the post-refrac outcome.

As discussed in Jayakumar et al. (2013), there are a number of restimulation treatment techniques for multistage hydraulically fractured horizontal wells. Figure 3 below depicts the approach we have implemented in order to model restimulation treatment. As shown in the figure, we insert new hydraulic

fractures between existing planar stages, to increase contact with the formation within the SRV region. In the absence of detailed fracture model inputs, we assume the new hydraulic fractures have the same conductivity, complex fracture geometry (Figure 2) and related fracture properties (Table 1) as existing hydraulic fractures. This is based on an assumption of unchanged operational procedures and rock properties.

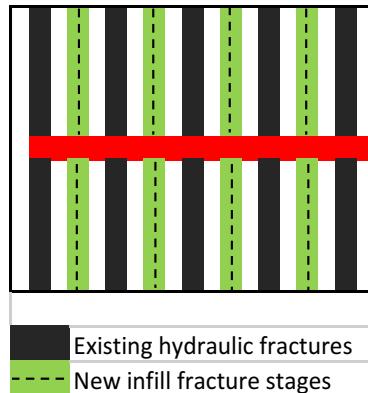


Figure 3: Modeling of infill restimulation treatment (adapted from Jayakumar et al., 2013). The black slabs represent existing hydraulic fracture stages, while the green slabs are the new infill refracturing stages inserted between planar fractures.

Given an ensemble of restimulated well production profiles, the next step is to define favorable and unfavorable restimulation candidate examples, in order to provide a basis for the classification algorithm to learn patterns associated with each category. To achieve this, we fit an empirical decline model (Arps, 1945) to the initial rate data and forecast the gas rate assuming no restimulation. Using this information, we then estimate the percent increase in cumulative production one year after restimulation, relative to the estimated production in the absence of restimulation. Using this approach, we generate a histogram of percent increase in cumulative production one year after restimulation, and apply a threshold at the 50th percentile. This procedure is depicted in Figure 4 below.

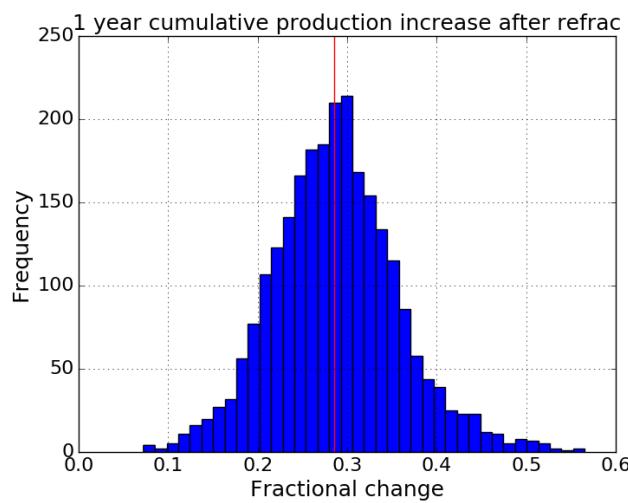


Figure 4: Distribution of cumulative production increase one year after restimulation, with restimulation taking place after 2 years of production. The vertical line represents the 50th percentile (0.285). All wells with less than 0.285 increase in production are considered unfavorable candidates for the purpose of this study, and vice versa.

In the above figure, all wells exhibiting cumulative production improvement greater than 28.5% are deemed good restimulation candidates. Note that the use of a 50th percentile threshold is arbitrary in this case and more care can be taken in defining the threshold to meet specific economic objectives. Figure 5 displays extreme examples of unfavorable and favorable candidates assigned based on this criterion, for wells refractured after 2 years.

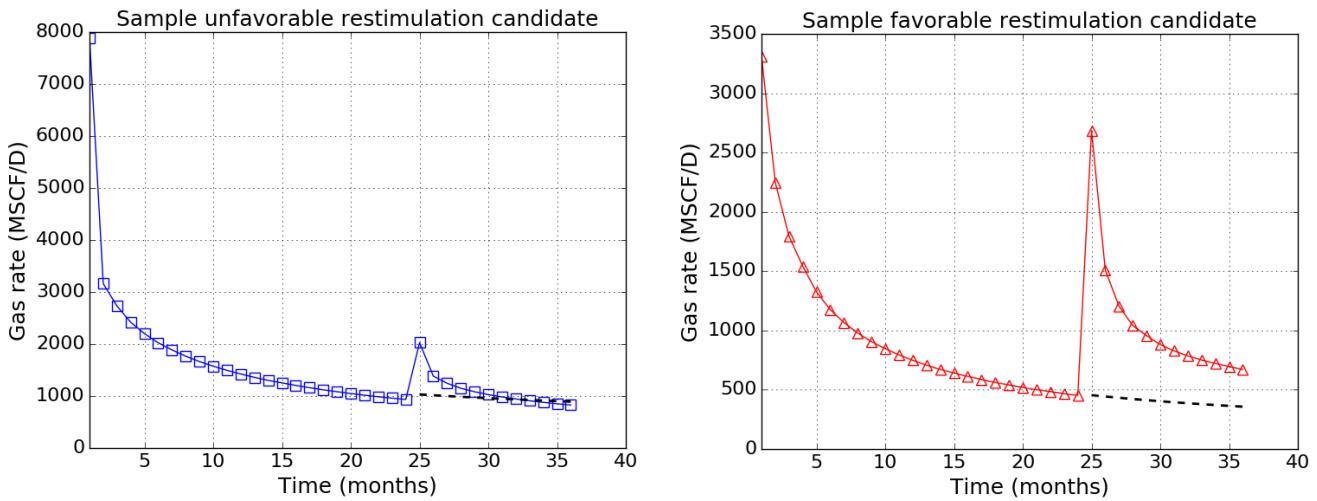


Figure 5: Examples of unfavorable (left) and favorable (right) restimulation candidates based on one-year cumulative production increase. Solid curves represent the gas rate data, while the dashed curves represent the extrapolated production.

In order to capture the operational rate fluctuations typical of field production data, we have introduced perturbations (in percent) to the simulated rate volumes, which are drawn from a Gaussian distribution with a mean of zero and standard deviation of 5%. This serves to provide a more realistic and challenging problem for the pattern recognition algorithm, compared to the direct use of clean synthetic rate responses like those shown in Figure 5. Figure 6 below displays sample input production data after incorporating rate fluctuations.

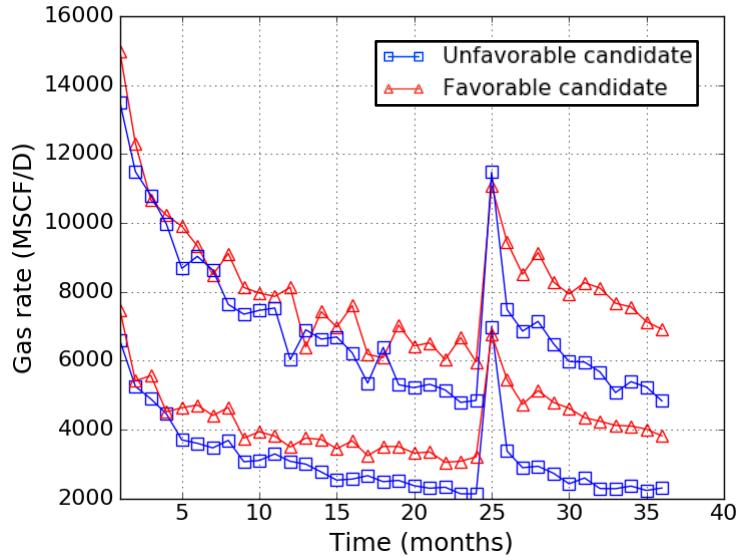


Figure 6: Sample training data with perturbed flow rates. The unfavorable candidates display more drastic decline in flow rate after restimulation.

In the above figure, we see that the simulated flow rate at each time instance has been perturbed, while the overall decline behavior is retained. In line with the criteria for restimulation success depicted in Figure 4 above, the well profiles labeled as unfavorable candidates show significantly steeper decline in production after restimulation. In other words, they have been labeled as unfavorable because they fall below the 50th percentile threshold for cumulative production enhancement (28.5%).

Within the defined framework, what specific combination of flow model parameters contribute to restimulation success? We have first performed manual inspection of univariate and bivariate flow parameter distributions using in simulating the training data, and we have observed no discernable distinction between the distributions of favorable and unfavorable candidates. This suggests that there are no individual or simple combinations of physical parameters that can be modified to guarantee

restimulation success. Thus, we need to consider higher order interrelationship between these variables. In order to establish a link between the observed outcome of restimulation and the influencing reservoir properties, we have therefore performed a multivariate analysis of the input variables used in generating the training dataset. This has been achieved by using principal component analysis to determine the direction in multidimensional space along which we can observe maximal separation between the multivariate distributions of favorable and unfavorable restimulation candidates, based on one-year cumulative production improvement. This is depicted in Figure 7 below.

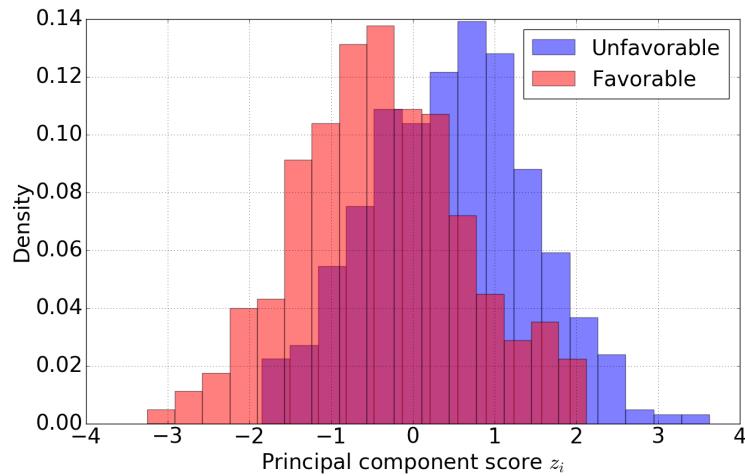


Figure 7: Distributions of favorable and unfavorable candidates when flow simulation parameters are projected along the direction of maximal separation in parameter space. The top and bottom 25 percentiles of favorable and unfavorable candidates are displayed.

In the above figure, the fracture, reservoir and operational parameters in Table 1 and Table 2 have been projected along the direction of maximum separation between favorable and unfavorable restimulation candidates in parameter space. The blue and red histograms represent the distribution of principal component scores for unfavorable and favorable candidates respectively. The figure shows the wells in the top and bottom 25 percentiles of cumulative production increase (total of 1250 wells), based on cumulative production enhancement. The principal component score z_i for well i is given by,

$$z_i = \sum_{k} l_k \psi_{ki} \quad (2)$$

In Equation 2, ψ_{ki} represents the flow simulation parameters, which have been standardized in order to eliminate the influence of the relative magnitude of each individual parameter (i.e. scaled to yield zero mean). In the summation, l_k is the principal component vector loading coefficient associated with each flow simulation parameter k . The “loadings” for the flow simulation parameters used in generating the training data are shown in Table 3 below. In multivariate flow parameter space, the loadings shown in Table 3 are the coordinates of the vector along which we observe the largest separation between favorable and unfavorable restimulation candidates, as depicted in Figure 7 above. These loadings are sorted in descending order of absolute magnitude. Under the current framework, the significance of the Table 3 is that summarizes the relative influence of the fracture, reservoir and operational parameters in separating favorable and unfavorable restimulation candidates. At the top of the list, we see that the rate of pressure-dependent hydraulic fracture closure exerts the highest influence in distinguishing between both categories, in addition to initial gas saturation, hydraulic fracture tip conductivity and spacing. On the other end, the loading vector coefficients suggest that the restimulation outcome is largely uninfluenced by factors such as reservoir depth, shale matrix permeability, pressure-dependent closure of complex fractures and number of hydraulic fracture stages.

Parameter, k	Loading, l_k	Parameter, k	Loading, l_k
Hydraulic fracture closure rate	0.509	Langmuir pressure	-0.132
Initial gas saturation	0.432	Initial reservoir temperature	-0.098
Hydraulic fracture tip conductivity	-0.368	Complex fracture conductivity	0.094
Hydraulic fracture spacing	-0.286	Complex fracture spacing	0.078
Langmuir volume	0.213	Initial reservoir pressure	-0.069
Diffusivity	-0.204	Bulk density	0.069
Shale porosity	-0.204	Natural fracture conductivity	0.047
Reservoir thickness	0.201	Natural fracture spacing	-0.042
Flowing bottomhole pressure	-0.179	Reservoir depth	0.016
Hydraulic fracture half length	-0.174	Shale permeability	-0.008
Hydraulic fracture conductivity	-0.169	Complex fracture closure rate	0.003
Shale compressibility	0.146	Number of hydraulic fracture stages	-0.001

Table 3: Principal component loadings for training data. Loadings are sorted in descending order of absolute magnitude.

In Figure 7, we can observe that favorable candidates tend to have more negative principal component scores, while unfavorable candidates tend to have more positive scores. Although caution must be taken in interpreting the signs of the loadings, this suggests that on aggregate, larger values of parameters with negative loading coefficients in Table 3 would lead to increased chance of restimulation success. However, there are multiple combinations of properties that can yield similar principal component scores, when we compute the sum of products shown in Equation 2. This illustrates one potential pitfall of production data analysis models and techniques which take into account only a subset of these parameters. Additionally, the majority of these input parameters are impossible or expensive to infer accurately in practice. However, given that restimulation success can be quantified based on rate response, it is clear the rate data directly holds information about the aggregate effect of different combinations of parameters and how their influence is manifested in the outcome of restimulation. For this reason, in the next section, we propose an alternative approach which decomposes production data into a collection of subtle rate decline characteristics, to serve as input to a classification model.

Finally, in order to explore the sensitivity of the classification performance to different restimulation candidate selection criteria, we have considered the following two additional metrics:

- Restimulation candidate selection based on instantaneous percent increase in flow rate, post-refrac.
- Change in decline rate after restimulation. This is determined by fitting the Arps' hyperbolic decline model (Arps, 1945) to the data prior to restimulation and after restimulation, and defining restimulation success based on the percent increase in Arps' hyperbolic decline constant.

Table 4 below summarizes the criteria used in assigning favorable and unfavorable candidate labels to the training production data. As shown, we have also considered two different options for restimulation timing – two and three years of production prior to treatment.

Timing of refrac	2 years			3 years		
	Min	Max	Threshold (P50)	Min	Max	Threshold (P50)
Change in one-year cumulative production	0.0724	0.565	0.285	0.0390	0.634	0.287
Change in instantaneous rate	0.595	10.135	1.527	0.601	13.055	1.602
Change in decline rate	-0.429	2.767	0.637	-0.133	2.886	0.944

Table 4: Distribution of restimulation candidate selection criteria computed from training data. All values are expressed as proportions (unitless).

Production data classification framework

In this section, we describe the proposed pattern recognition framework for analyzing well production data, motivated by techniques in face detection (Viola and Jones, 2001). The overall aim is to develop a fast and robust probabilistic binary classification model which utilizes input features obtained directly from the rate data. The advantage of the binary classification approach is that it allows us to develop a flexible tool which can be used to address a wide array of problems related to well performance. In other words, if we can define two sets of well production data which conform to any binary criteria of economic interest, such as “favorable” or “unfavorable” restimulation candidates, then we train a model to distinguish and accurately identify a new candidate well. In addition to these categorical predictions, we can extract probability estimates, which provide a means to rank candidate wells. In this section we discuss the proposed well data representation, after which we show how discriminative features can be extracted as inputs to the classification algorithm. Thirdly, we present the statistical framework for distinguishing between binary categories of well data.

Well data representation

How exactly do we apply face detection to analyze well data? This is achieved by representing well production data as 1-D vectorized images with pixel values indicating flow rate magnitudes. This approach is depicted in Figure 8 below.

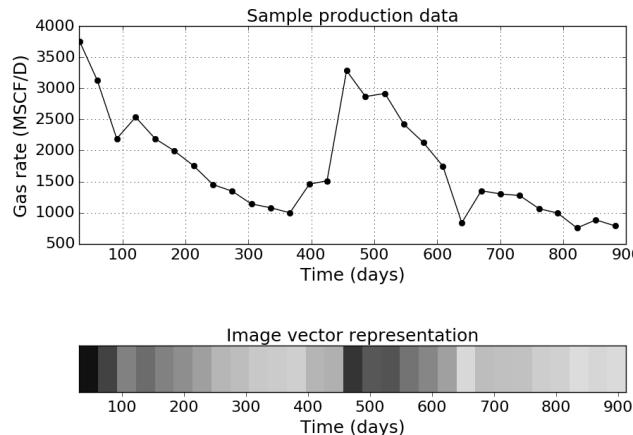


Figure 8: Gas production rate from a restimulated well (top), and the corresponding 1-D image vector representation (bottom). The color (pixel intensity) is proportional to the gas rate.

In the above figure, the gas production rate from a restimulated shale gas well is shown, with the corresponding image vector representation depicted underneath. In other words, the volume of gas produced during each time interval is proportional to the “pixel intensity” of each element in the image vector. Intuitively, this can be thought of as a “strip” of an image.

Let us now define the notation used in representing the well data. We begin with a well production dataset, where the gas volume from an arbitrary well i at time t is denoted as $q_i(t)$, for all $i = 1, \dots, N$ and $t = 1, \dots, t_f$. This implies a structured dataset containing N wells, with each well profile containing rate data sampled at equal and constant intervals for t_f time units. Uniform time sampling is generally a valid assumption for most field-collected well data. However, in cases with missing data, decline curve methods may be used to estimate unavailable production information. Also, in a dataset containing wells with dissimilar sampling intervals, all wells would need to be upscaled to conform to the coarsest time step available.

As a preprocessing step, in order to provide a uniform basis for comparison between data, we have normalized all samples by their initial production rate, i.e.

$$\tilde{q}_i(t) = \frac{q_i(t)}{q_i(t=1)} \quad (3)$$

Input feature extraction

Given a dataset with normalized rates, the first step in the pattern recognition workflow is to define a set of attributes for each well profile, to serve as a basis for distinguishing between wells that exhibit different characteristics. As earlier mentioned, one common approach in existing data-driven models is to specify a set of uncertain reservoir and completion input parameters into a pattern recognition scheme such as ANN, which are largely selected based on availability. In addition to requiring detailed and potentially expensive data, this modeling approach implicitly assumes that well productivity is only related to the finite set of selected input parameters, and thus neglects the influence of unmeasured reservoir properties. This has the potential to yield non-intuitive model coefficients, and also limits the applicability of the model to wells where exhaustive data is available.

As an alternative, by recognizing that the rate response from a producing well holds the most reliable truth regarding the multivariate relationship between subsurface characteristics, we instead utilize numerical attributes that are extracted directly from rate production data. In this way, the pattern recognition scheme is able to relate well performance to the subtle rate decline characteristics observed at different time scales and time periods in large well production datasets, which reflect the aggregate effect of competing reservoir influences.

With this objective in mind, we have characterized each production profile by computing 1-D Haar-like features, in a manner similar to the Viola-Jones face detection approach. As the name implies, these features are analogous to Haar wavelets used in image processing applications. The feature geometries we have used are shown in Table 5 below.

Feature template, z	Geometry
1	
2	

Table 5: Features used for production data classification. In order to compute a feature score, the sum of rates in the white region is subtracted from the sum of rates in the white region, for each geometry. These templates are shifted and scaled throughout the well profile, and a feature score is computed in each case.

For each well profile, we generate a list of feature scores $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$ by shifting and scaling each template shown in Table 5 throughout the producing life, and computing a real-valued feature score in each case. These feature scores are computed by simply subtracting the sum of normalized rates in the white region from the sum of normalized rates in the grey region. More formally, for each feature geometry $z \in \{1, 2\}$, given a feature scale Δt and starting time index $t = 1, \dots, t_f$, we compute each feature score $x_{ij}(\tilde{q}_i; t, \Delta t, z)$ as,

$$x_{ij}(\tilde{q}_i; t, \Delta t, z) = \begin{cases} \sum_{t'=t}^{t+\frac{\Delta t}{2}-1} \tilde{q}_i(t') - \sum_{t'=t+\frac{\Delta t}{2}}^{t+\Delta t-1} \tilde{q}_i(t') & (z = 1) \\ \sum_{t'=t}^{t+\frac{\Delta t}{3}-1} \tilde{q}_i(t') - \sum_{t'=t+\frac{2\Delta t}{3}}^{t+\frac{2\Delta t}{3}-1} \tilde{q}_i(t') + \sum_{t'=t+\frac{2\Delta t}{3}}^{t+\Delta t-1} \tilde{q}_i(t') & (z = 2) \end{cases} \quad (4a) \quad (4b)$$

Where,

$$\Delta t = \begin{cases} 2, 4, 6, \dots, t_f & (z = 1) \\ 3, 6, 9, \dots, t_f & (z = 2) \end{cases} \quad (5)$$

In the above expressions, the index $j = 1, \dots, p$ represents a unique combination of feature geometry, time and scale. For example, we can extract a total of $p = 236$ feature combinations from a two-year well profile with monthly-sampled data (i.e. $t_f = 24$).

Intuitively, these texture-based features adapted from image objection detection applications can be considered analogous to rate derivatives, which have traditionally been used in rate transient analysis to infer flow regime information from well data. In other words, using basic forward-difference approximations,

$$\frac{d\tilde{q}_i}{dt'} \Big|_{t'} \cong \frac{\tilde{q}_i(t') - \tilde{q}_i(t' + \Delta t')}{\Delta t'} \quad (6)$$

$$\frac{d^2\tilde{q}_i}{(dt')^2} \Big|_{t'} \cong \frac{\tilde{q}_i(t') - 2\tilde{q}_i(t' + \Delta t') + \tilde{q}_i(t' + 2\Delta t')}{(\Delta t')^2} \quad (7)$$

Comparing Equation 6 with Equation 4a, we can see that features computed using two-rectangle template ($z = 1$) are related to the first derivative of normalized flow rate. Furthermore, if we consider the summation terms in Equation 4a as upscaling operations which coarsen the sampling interval from one unit of t to a larger time step of $\frac{\Delta t}{2}$, then Equation 6 collapses to Equation 4a, since $\Delta t' = 1$ at this coarser scale. This means that when we compute multiple combinations of these features for a given well profile at different t and Δt , we are effectively decomposing the signal into a series of rate derivatives at multiple time periods and scales. Importantly, this implies that the two-rectangle features carry information about flow regimes experienced over the productive life of the well. Similarly, if we compare Equation 7 with Equation 4b, we see that the three-rectangle features ($z = 1$) are analogous to the second derivative of normalized rate. This attribute on the other hand captures the curvature characteristics at different periods and scale within the well data. One advantage of considering each of these features at multiple scales is that it helps to mitigate the impact of instantaneous fluctuations in rate or measurement error, making the method robust to noise.

In Big Data applications involving a significant amount of finely-sampled well data, it becomes more important to minimize the computational cost incurred in calculating these feature scores. In order to achieve this, we can take advantage of the simple geometries of these feature templates. In other words, we can pre-compute the cumulative normalized rate for each well at each time instance using the Equation 8 below:

$$\tilde{Q}_i(t) = \sum_{t'=0}^t \tilde{q}_i(t') \quad (8)$$

The summation $\tilde{Q}_i(t)$ is calculated for all $t = 1, \dots, t_f$. After pre-computing $\tilde{Q}_i(t)$ for each well, we can then easily determine all feature scores at each time t and scale Δt using Equations 9 as follows:

$$x_{ij}(t, \Delta t, z) = \begin{cases} 2\tilde{Q}_i\left(t + \frac{\Delta t}{2} - 1\right) - \tilde{Q}_i(t - 1) - \tilde{Q}_i(t + \Delta t - 1) & (z = 1) \\ 2\tilde{Q}_i\left(t + \frac{\Delta t}{3} - 1\right) - \tilde{Q}_i(t - 1) - 2\tilde{Q}_i\left(t + \frac{2\Delta t}{3} - 1\right) + \tilde{Q}_i(t + \Delta t - 1) & (z = 2) \end{cases} \quad (9)$$

In other words, the summation operations in Equations 4a and 4b have been represented as a set of basic arithmetic operations in Equation 9, thus minimizing computational expense incurred through unnecessary iterations. Note that $\tilde{q}_i(0) = \tilde{Q}_i(0) = 0$. By translating each feature template in time at different scales, the final output from this step is an exhaustive set of feature scores $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$ for each well $i = 1, \dots, N$, where p is the total number of feature combinations available within each production profile.

Statistical framework

Given a dataset made up of p Haar-like features scores extracted from N wells, the next step is to define a statistical basis for identifying restimulation candidates. In a manner similar to Viola and Jones (2001), we have framed this problem using a binary classification approach, and extended the methodology to extract probability estimates. In other words, instead of explicitly forecasting a real-valued flow rate after restimulation, the predictive model outputs a categorical response – favorable or unfavorable restimulation candidate – along with an estimated probability or refrac success.

The engine of the binary classification framework used in the Viola-Jones procedure is a statistical learning method, known as AdaBoost. The algorithm takes in a list of feature scores $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$ and a categorical outcome vector $y_i \in \{-1, 1\}$ for all training well production data $i = 1, \dots, N$ and features $j = 1, \dots, p$. In the vector y_i , -1 and 1 are categorical labels which denote unfavorable and favorable candidates respectively. For each unique feature j , we can collate the feature scores computed from all training data and use y_i to develop a simple binary classification model based on a single threshold b_j . This is known as a single-node binary decision tree (decision stump), as depicted in Figure 9 below. The threshold b_j is selected by minimizing an impurity function known as the Gini index (Breiman et al, 1984), which serves to minimize the number of misclassifications with respect to the training data. This fitting procedure can be performed using common statistical software packages.

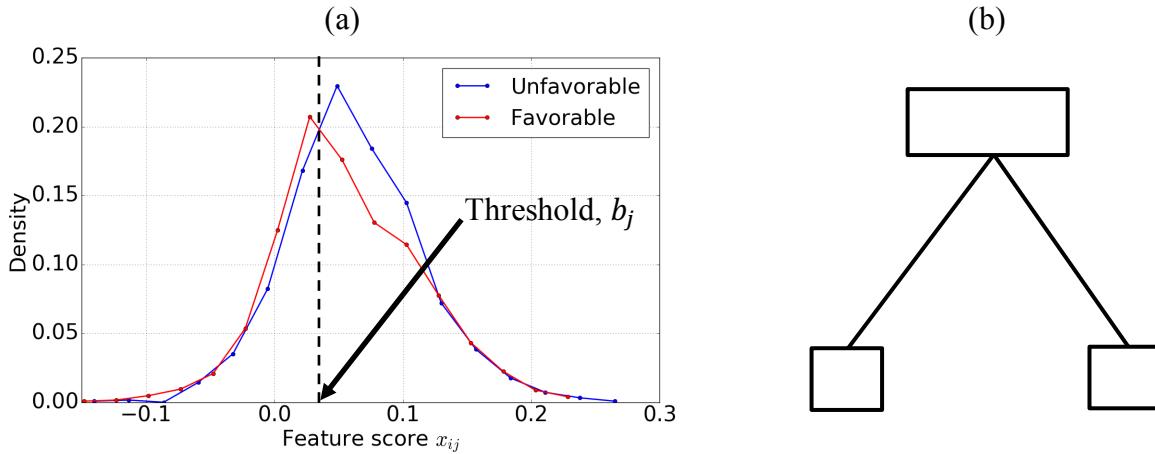


Figure 9: Left: Distribution of scores for feature $j = 1$ computed with respect to all wells in training set. b_j is the optimal threshold which minimizes overlap between favorable and unfavorable categories. If x_{ij} (the feature score computed with respect to well i and feature index j) is greater than threshold b_j , the well is assigned a categorical label of -1 (unfavorable candidate), and vice versa.

Right: Depiction of decision stump classification model.

In the above figure, we denote the decision stump trained using feature j as $f(x_{ij}; b_j) \in \{-1, 1\}$. All data with feature scores greater than b_j are predicted as unfavorable candidates, and vice versa. Note that the inequality can be reversed depending on the distribution of feature scores. For each candidate feature $j = 1, \dots, p$, we can quantify the performance of each decision stump by computing the weighted classification error ϵ_j with respect to the training data using Equation 10 below (Freidman et al., 2009):

$$\epsilon_j = \frac{\sum_{i=1}^N w_i I(f(x_{ij}) \neq y_i)}{\sum_{i=1}^N w_i} \quad (10)$$

In the above expression, ϵ_j lies between 0 and 1 and w_i represents the weights assigned to each well production profile, which are equal by default. $I(f(x_{ij}) \neq y_i)$ is an indicator function which returns 0 if the prediction matches the true categorical label and returns 1 in the case of a mismatch. As is visible in Figure 9a, there is a significant amount of overlap between the distributions of favorable and unfavorable candidates, which would lead to a large number of misclassifications. Due to the simplistic nature of these

single-threshold classification rules, they are in practice expected to yield only slightly better predictions than a random guess (Friedman et al., 2000). For this reason, they are referred to as “weak classifiers”. The AdaBoost algorithm however utilizes a sequential fitting procedure to combine or “boost” the poor predictions from these individual decision stumps into a more accurate classifier. The AdaBoost training procedure is summarized in Table A-1.

As shown in Table A-1, for a given number of iterations $m = 1, \dots, M$, the overall strategy is to train one individual decision stump $f(x_{ij})$ for each of p available features, compute the associated weighted error ϵ_j using Equation 10, and then select the weak classifier f_m which minimizes the weighted classification error ϵ_m at the current iteration m . The weighted error is used to compute a “voting weight” α_m , which is indirectly proportional to the error. This weak classifier is then reapplied to the training data and misclassified examples are assigned higher weights for the next iteration, $w_i^{(m+1)}$. The final outcome of the training procedure is a list of the weak classifiers f_m selected after each iteration, and their corresponding “voting weights” α_m for $m = 1, \dots, M$, which are linearly combined to return a categorical prediction as follows:

$$\hat{y} = \text{sign}[G(\mathbf{x}_i)] = \sum_{m=1}^M (\alpha_m f_m) - \theta \quad (11)$$

In Equation 11, $G(\mathbf{x}_i)$ returns a real number between $-\infty$ and ∞ . As indicated, the sign of the result gives the predicted category $\hat{y} \in \{-1, 1\}$, where 1 and -1 denote favorable and unfavorable candidates respectively. In other words, if we evaluate $G(\mathbf{x}_i)$ for any arbitrary well production profile, a negative value will be assigned to unfavorable candidates, and a positive value would be assigned to favorable candidates. The variable θ is a tuning parameter which is zero by default, and will be discussed shortly. The intuitive explanation of Equation 11 is that each weak classifier f_m has received a weight α_m which is proportional to its individual accuracy in distinguishing between favorable and unfavorable candidates. In this way, the predictions from M weak classifiers are combined via a “committee vote” (Freund and Schapire, 1996). The AdaBoost fitting procedure shown in Table A-1 can be performed using common open source statistical software (such as Scikit-Learn and R).

There are a number of properties which make AdaBoost a model for our problem. Firstly, the training algorithm implements a weighting procedure (step 2d in Table A-1) in which misclassified examples are assigned higher importance for the next iteration. It can be shown that these weights serve to minimize an exponential loss function, which is monotone and smooth. For this reason, AdaBoost is empirically considered to be robust to overfitting (Friedman et al., 2000), in contrast to methods such as ANN and support vector machines (SVM). Secondly, we can typically build a strong AdaBoost classifier using $M < p$ iterations. In other words, it functions as a natural feature selection process, in which the more discriminative features receive higher voting weights α_m . This greatly mitigates the computational cost associated with calculating all p features during training and also while assessing unseen candidate wells, which becomes more significant with larger datasets. Thirdly, the additive model $G(\mathbf{x}_i)$ can be probabilistically interpreted. Following a log odds-ratio (logit) model or its inverse, the anti-logit probabilistic model, the probability of an arbitrary data sample belonging to either category (1 or -1), can be expressed as (Friedman et al., 2000):

$$P(y = 1 | \mathbf{x}_i) = \frac{e^{G(\mathbf{x}_i)}}{e^{G(\mathbf{x}_i)} + e^{-G(\mathbf{x}_i)}} \quad (12)$$

$$P(y = -1 | \mathbf{x}_i) = \frac{e^{-G(\mathbf{x}_i)}}{e^{-G(\mathbf{x}_i)} + e^{G(\mathbf{x}_i)}} \quad (13)$$

Using this information, we can therefore scan an ensemble of production profiles and estimate the probability that each well is either a favorable or unfavorable candidate. This provides a means to quantify the uncertainty in predicting each well to be a favorable or unfavorable restimulation candidate. If the

probabilities are not very discriminatory, the procedure would suggest that the observed production information is inadequate for making reservoir development decisions and other sources of information need to be investigated in order to lead to more robust decisions. Additionally, using these probability estimates, we have a basis for ranking candidate wells.

The final step in the overall production data classifier training framework is the cascade training procedure (Viola and Jones, 2001). The overall idea is to train a series of AdaBoost models which are increasingly more discriminative. In other words, in each successive stage, more iterations are dedicated to improving the overall classification accuracy. In this way, wells which display strongly unfavorable characteristics are weeded out early in the process, which means that we only need to compute a much smaller subset of Haar-like features. This has the potential to lead to significant time savings in applications involving massive datasets. The cascaded classifier structure is depicted in Figure 10 below.

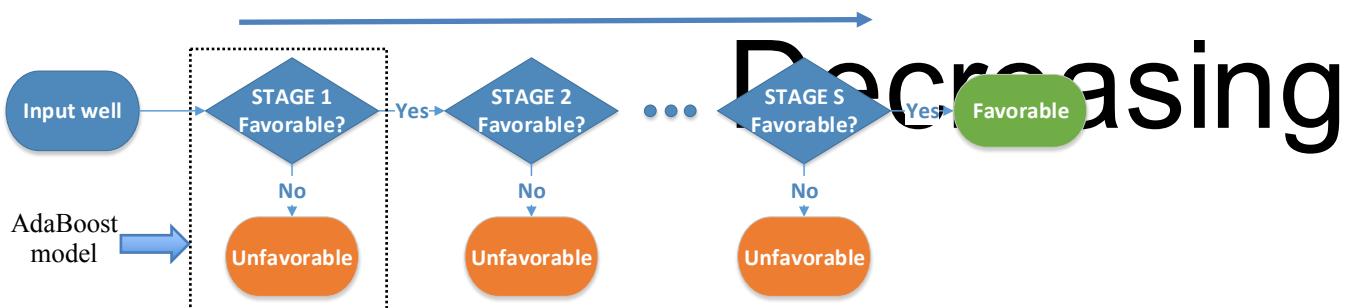


Figure 10: Cascaded classifier prediction protocol (modified from Viola and Jones, 2001). Each cascade stage contains one AdaBoost model. Each successive stage is increasingly more discriminative (i.e. lower false positive error. If at any stage an unfavorable candidate is predicted, the process is terminated. The probability of restimulation success can then be estimated at the stage of termination, using Equation 12.

In the above figure, during the cascade prediction protocol for any arbitrary well profile, we run the AdaBoost model in each successive stage by evaluating Equation 11. At each stage, we only proceed to the next stage if a value of 1 is predicted. The probability of restimulation success can be estimated at the stage of termination, using Equation 12.

The cascade model shown in Figure 10 is trained by tuning the true positive and false positive classification ratios in each successive AdaBoost model to impart a desired level of accuracy in predicting favorable and unfavorable candidates. In binary classification literature, the true positive ratio represents the proportion of data which belong to the favorable category (1) and are correctly predicted. On the other hand, the false positive ratio refers to the proportion of data which belong to the unfavorable category (-1) and are erroneously predicted as members of the favorable category. In Equation 11 above, when we take the sign of $G(\mathbf{x}_i)$ as a basis for assigning data to classes -1 or 1, an AdaBoost binary classification threshold of zero is implied. The θ parameter, which is zero by default, can therefore be decreased in order to ensure more training data are assigned to the favorable class (i.e. increase true positive ratio) and vice versa. In the limit, if we set $\theta = -\infty$, all data will be assigned to the favorable class, with true positive and false positive ratios both equaling 1. In the cascade training process, we take advantage of θ in tuning each successive AdaBoost model to accept increasingly fewer false positive classifications, while simultaneously maximizing true-positive classification accuracy.

The pseudocode for the cascaded classifier training algorithm is presented in Table A-2. In the table, F_{ps} and D_{ps} are user-defined multiplicative factors used to control the false negative and false positive ratios per cascade stage, while the optional parameter n is used to control the complexity of the AdaBoost model trained in each stage. In addition to improved computational efficiency, this approach has the potential to serve as a flexible training procedure which can be modified to prioritize classification accuracy of favorable or unfavorable wells, based on economic objectives.

Results and discussion

In this section, we present the results of model training and evaluation for the proposed production data classification algorithm. Next we compare to the performance of a type curve approach in identifying restimulation candidates. Finally we show the results of a field validation study, using the proposed pattern recognition approach.

Classifier training and evaluation

Using the generated training production data along with categorical labels denoting favorable (1) and unfavorable (-1) restimulation candidates, we have applied the proposed Viola-Jones framework to train a production data classification model. The training and evaluation procedure involves the following steps:

1. Given a collection of training data, compute feature scores for all combinations of Haar-like feature geometry, scale and starting time within the individual well profiles.
2. Train AdaBoost cascade model with a random sample of N well profiles, using feature scores computed with respect to the data before restimulation.
3. Assess the model predictive performance by scanning V wells and comparing the predicted restimulation outcome to the known outcome.

Table 6 below displays the cascade training parameters we have implemented.

Number of well production profiles used for training, N	400
Number of well production profiles used for testing, V	100
True positive ratio per stage, D_{ps}	0.97
False positive ratio per stage, F_{ps}	0.90
Features per stage multiplier, n	10

Table 6: Cascaded classifier training parameters.

As earlier discussed, in defining restimulation success, we have considered three separate candidate selection criteria. In each case, in order to capture the impact of timing of refracturing, we have utilized either two or three years of production data prior to the restimulation event. After the training process, the predictive performance of each of these classifiers has been tested by assessing a synthetic test set containing $V = 100$ well production profiles, which were not available to the training algorithm. Table 7 below summarizes the results of the production data classifier training exercise, and their predictive performance with respect to synthetic data.

Criteria	Timing			2 years			3 years		
	Accuracy (%)	Cascade stages	Training time (s)	Accuracy (%)	Cascade stages	Training time (s)			
Change in one-year cumulative production	83	13	118	81	11	215			
Change in instantaneous rate	78	13	375	76	14	831			
Change in decline rate	84	10	115	82	13	422			

Table 7: Production data classifier training results and performance with respect to synthetic test data.

The above results show that regardless of the restimulation candidate selection criteria, we are able to distinguish between favorable and unfavorable candidates with accuracy of 76-84%. By comparison, a monolithic (single-stage) AdaBoost classifier yields an overall classification accuracy of 76% (for 200

iterations) in the case involving two years of production, based on the cumulative production improvement criteria. The results in Table 7 suggest that improvement in instantaneous rate after restimulation is the least reliable indicator of restimulation success, as highlighted by the lower classification accuracy for both 2-year and 3-year cases.

In Table 7, we see that the trained classifiers have 10-14 cascade stages. We can observe that the instantaneous rate change criterion requires more training time for both 2-year and 3-year cases. This can be attributed to the fact that more AdaBoost iterations are required in each cascade stage in order to meet the convergence criteria shown in step 3b of Table A-2. This is further evidence that instantaneous increase in flow rate after restimulation is the least discriminative metric for restimulation success. Nevertheless in all cases, while training may potentially be computationally intensive, the final trained classifier is able to provide predictions on the test set in near real-time. For example, the classifier based on the cumulative production improvement criterion can make predictions on 100 samples of two-year well production in 0.35 seconds. All reported runtimes are based on a 64-bit system with 3.4 GHz processor and 32 GB RAM.

As earlier discussed, only a subset of all Haar-like feature combinations within a well production profile are utilized in the AdaBoost classification model within each cascade stage. In other words, the sequential fitting procedure described in Table A-1 is aimed at selecting the most discriminative features which best enable us to distinguish between restimulation candidates. Therefore, by looking at the features selected by the AdaBoost model at each stage, we can gain additional insight into the mode of operation of the proposed classification scheme. Appendix B summarizes the features selected by the trained model which utilizes two years of data prior to the restimulation event in classifying wells on the basis of cumulative production improvement.

Firstly, from Table B-1, we see that within all cascade stages, the majority of features selected are two-rectangle templates ($z = 1$), with relatively fewer occurrences of three-rectangle templates ($z = 2$). This suggests that overall, features which relate restimulation success to first order rate-derivatives at different scales and time periods, and help to highlight changes in flow regimes, are more discriminative than those which consider curvature-related information. Secondly, from Table B-2, we see that across all cascade stages, features evaluated within the first 6 months of production have been selected more frequently than those evaluated afterward. This suggests that the most informative rate decline characteristics useful in identifying restimulation candidates are available at early periods in the life of the well. Thirdly, as shown in Table B-3, features with maximum scales of 6 months offer the most discriminative information across different cascade stages. Overall, larger features evaluated at later periods tend to offer limited useful information in discerning between favorable and unfavorable restimulation candidates.

Comparison with type curve approach

In order to compare the performance of our classification model to traditional approaches, we have analyzed our synthetic well data using the trilinear flow solution proposed by Brown et al (2009). We have selected the trilinear model as a representative analytical solution among a number of existing type curve methods developed specifically for multistage hydraulically fractured shale wells. In Figure 11a below, we have plotted the pre-refrac production data on dimensionless axes based on the trilinear analytical solution. The characteristic dimensionless quantity P_{WD} is proportional to the rate-normalized pseudo-pressure drop, and takes into consideration several well, reservoir, fluid, and dual permeability fracture parameters.

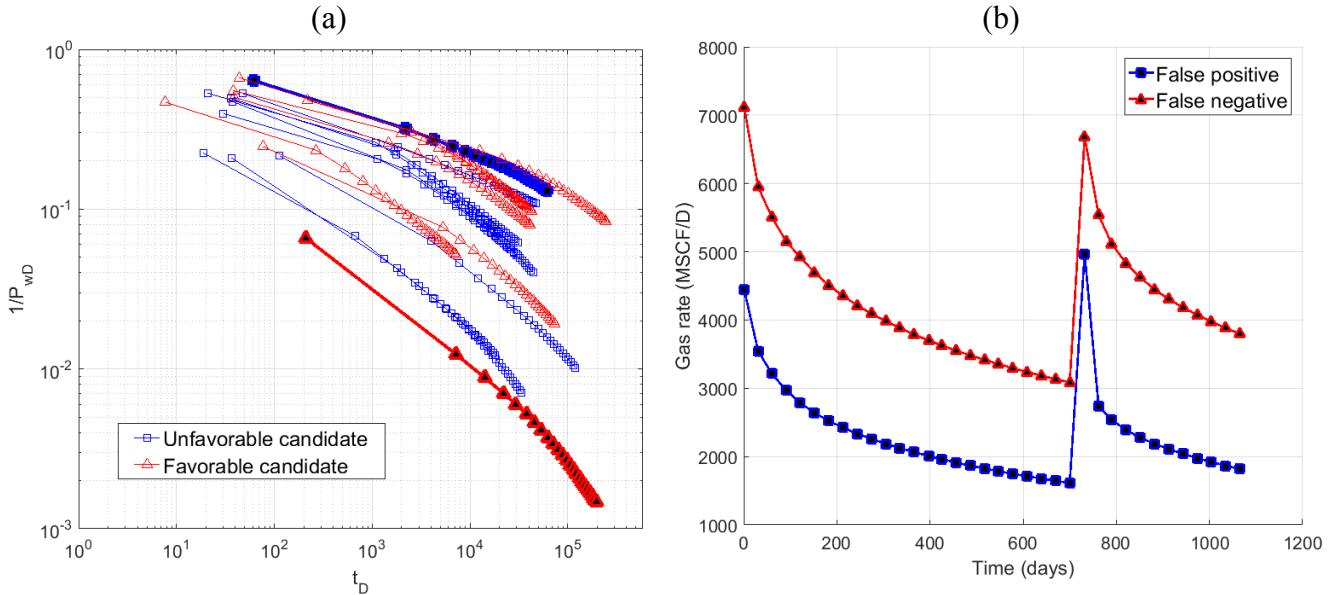


Figure 11: Restimulation candidate selection using type curve analysis (trilinear flow solution, Brown et al., 2011). Left: Well data plotted on dimensionless axes. Right: Examples of wells that would be misclassified if restimulation candidates were selected based on qualitative criteria observable on dimensionless axes, such as cumulative production.

In Figure 11a, we have plotted the rate data on $1/P_{WD}$ versus t_D . We take the inverse of P_{WD} in order to put the data in terms of pressure-normalized rate. Also, t_D has been computed using material balance pseudo-time. These steps, as are standard in rate-transient analysis, serve to account for non-constant rate and non-linear gas properties as a function of pressure. In order to plot the well data, information on hydraulic fracture properties and geometry, fluid and reservoir properties as well as pressure information is required. Additionally, given that the solution is developed in Laplace space, a numerical technique such as the Stehfest (1970) algorithm is required to represent the results in the time domain.

One potential workflow using this type curve approach would be to select restimulation candidates based on qualitative rate-related criteria inferable from the dimensionless axes, such as cumulative production. In other words, the wells with similar cumulative production profiles observed on the dimensionless axes would be grouped together and classified as favorable or unfavorable candidates. For example, on this basis, curves clustered toward the top right may be considered favorable candidates, while those located toward the bottom left may be considered unfavorable candidates. However, this approach can potentially lead to misclassifications and consequently missed opportunities, as evidenced by the significant overlap between unfavorable and favorable candidates in Figure 11a. To illustrate this, we consider the two highlighted wells in Figure 11a, and observe their post-refrac time-domain response in Figure 11b. Judging from the steeper production decline shown by the lower curve (blue) in Figure 11b after refrac, we see that the higher curve (red) has been missed as a favorable restimulation candidate.

We can investigate the combination of fracture, reservoir and operational parameters that have led to these misclassifications, by revisiting the previously discussed principal component analysis results. The flow simulation parameters corresponding to both wells and the resulting principal component data are displayed in Appendix C. Overall, the properties used in simulating the false positive well in Figure 11b (blue) yield a principal component score of -0.437 while those used in simulating the false negative well (red) yield a score of 1.398. Looking at Figure 7 we see that these scores both fall within the overlapping region between favorable and unfavorable candidates, which is characterized by high uncertainty in determining the outcome of restimulation. This can be attributed to the counterbalancing effect of competing reservoir influences. Specifically, if we consider the flow model parameters used in simulating the false positive case, we see that it is characterized by relatively tight hydraulic fracture spacing and low reservoir pressure, based on the uncertainty ranges displayed in Table 1 and Table 2. However, these are mitigated by factors such as relatively gentle rate of pressure-dependent hydraulic fracture closure, high hydraulic fracture conductivity and half length. As shown in Table C-1, when we evaluate the net weighted

sum of these competing reservoir influences (using Equation 2) along the vector of maximal separation between favorable and unfavorable candidates in multivariate parameter space, the aggregate effect approaches net zero (i.e. $z_i = -0.437$). Similarly, in the false positive case, the influence of high reservoir pressure and half length, are counteracted by poor completion evidenced by poor hydraulic fracture spacing and more drastic rate of pressure-dependent hydraulic fracture closure.

Overall, these results demonstrate that the combination of reservoir and well parameters and their influence on the rate response can render it difficult to gauge good restimulation candidates based solely on traditional data analysis methods such type curve approaches. In other words, we are not able to resolve influence of competing reservoir effects on phenomena such a restimulation success by visualizing rate data on dimensionless axes, such as those which take into account a limited set of fracture, fluid, and reservoir properties. On the other hand, by decomposing the raw production data into a set of attributes which are related to rate derivatives at multiple time periods and scales within the well life, we are able to train a probabilistic classification model to better capture the distinction between favorable and unfavorable candidates. More detailed modeling to incorporate geomechanics and a wider array of well completion strategies is required to completely understand the full range of reservoir parameters which affect restimulation candidate selection. However, these results depict the manner in which competing reservoir influences can be masked, and show that we are able to better capture the multivariate separation between restimulation candidates using flow rate-derived features.

Validation with field data

Finally, we have validated the proposed approach using publicly available field production data. In order to achieve this, we have considered 17 hydraulically fractured gas wells from Freestone county, within the Haynesville/Bossier Shale Play in East Texas. The sample size is limited by the rigorous screening criteria which we have applied to ensure that the field data match the simulated conditions as closely as possible. Specifically, each selected case is a hydraulically fractured horizontal shale gas wells which has undergone restimulation approximately within the first 2-3 years of production. Each restimulation event has been corroborated by completion reports which show evidence of re-perforation within the same formation. Also, in order to limit the effect of variable completion procedures, the selected wells have been constrained to the same operator within the same county. The production data from the wells used in this study are displayed in Appendix E and Appendix F.

In order to quantitatively define whether or not each of these wells underwent successful restimulation, we have extrapolated the flow rate from the primary production period using Arps' hyperbolic model and estimated the percent increase in cumulative production one year after restimulation. On this basis, the field production profiles shown in Appendix E have been defined as favorable restimulation candidates, while those in Appendix F have been defined as unfavorable candidates for restimulation. These class assignments have been determined using the thresholds specified in Table 4.

Next, we have attempted to predict suitability for restimulation treatment in each of these wells by making predictions solely based on the gas rate data prior to the refracturing events. In order to achieve this, we have assessed each of these wells using the cascaded classifiers which were trained based on cumulative production improvement one year after restimulation. As earlier mentioned, the selected wells were each refractured around the second or third year of production. Wells which were refractured close to the second year of production have been assessed using the trained two-year classifier, while those refractured around the third year have been assessed using the trained three-year classifier. Due to limited availability of field examples, missing production data has been estimated using the Arps' empirical relationship where necessary. For example, since Well 7 in Appendix E has 21 months of available data, we have extrapolated the final 3 months of production order to provide two complete years of production to the algorithm, so as to maintain consistency with the trained two-year classifier. Additionally, given that the field data is reported in terms of monthly production volumes, we have used the Arps empirical relationship to estimate the instantaneous daily production at the start of each month, in order to maintain

uniformity with the output of the flow simulator. To account for time-varying hyperbolic decline constant, we have achieved this by fitting the field data in intervals of six months.

Table F-1 displays results of evaluating the field cases using the trained classifiers. The assumed timing of restimulation for each field case, the class labels we have assigned based on increase in cumulative production after refrac, the predictions made by the cascaded classifier, as well as the estimated probability of restimulation success are shown. From these results, we see that 12 out of 17 of the field production profiles have been correctly classified, using the defined criteria. In other words, using only gas rate data prior to restimulation, we are able to correctly predict the outcome of restimulation with approximately 71% accuracy. This is a statistically significant result, based on single proportions t-test (p -value = 0.02). Overall, these results demonstrate the merit of the proposed classification approach in analyzing production data from hydraulically fractured gas wells to assist reservoir management and decision-making. In contrast to existing approaches which aim to directly predict gas production after restimulation, this classification method returns a categorical outcome and an accompanying probability of restimulation success or failure. The computation time required to assess all 17 field case was instantaneous. The estimated probabilities of fracture success reported in the table are based on the multivariate distribution of feature scores derived from the synthetic training dataset. In other words, they are influenced by the similarity between the rate decline characteristics exhibited by the field data and the synthetic examples, leading up the restimulation event. The predictive performance of the trained models therefore suggest overall agreement between the synthetic rate profiles used in training and the shale gas production observed in these field cases.

From the results in Table F-1, we can observe that when we run our classification model on field data, 7 out of 10 wells restimulated around the third year have been accurately predicted, while 5 out of 7 of the wells restimulated after two years have been accurately predicted. This is consisted with the assessment of synthetic data in Table 7, where we see slightly better performance in predicting two-year cases than three-year cases. Table 8 below displays two confusion matrices which describe the classification performance of the trained cascades in assessing both synthetic and field data, based on the criteria of change in cumulative production one year after restimulation.

		(a)				(b)	
		Unfavorable	Favorable			Unfavorable	Favorable
Predicted	Actual			Predicted	Actual		
Unfavorable	Unfavorable	83	17	Unfavorable	3	2	
Favorable	Favorable	19	81	Favorable	3	9	

Table 8: Confusion matrices summarizing classifier error in predicting favorable and unfavorable candidates based on (a) synthetic test data, and (b) field data.

Table 8a summarizes the model predictions based on the synthetic test set, for both two-year and three-year cases. We see that in the cases of misclassification, there is a slightly greater tendency to predict favorable restimulation candidates as unfavorable, than unfavorable candidates as favorable (i.e. the number of misclassifications in the lower left corner is slightly greater than those in the upper right). This implies that the cascade is inherently risk-averse, in that it is more inclined to allow a practitioner to bypass reserves (false negative or Type II error) than to lose money on poor performance after refracturing treatment (false positive or Type I error). This behavior is a consequence of the continuation criterion utilized in the cascade training procedure (Table A-2), which is given by:

$$F_s > 1 - D_s \quad (14)$$

The inequality in Equation 14 means that cascade training is continued until the false positive classification ratio F_s is less than or equal to one minus the true positive ratio D_s . This criterion is designed

to provide balanced allocation of error between both favorable and unfavorable categories. However, with fewer number of training and testing data samples, there is a greater tendency toward convergence at a value of F_s slightly lower than $1 - D_s$. This factor can be mitigated by using more data for cascade training and testing. Additionally, we can adjust Equation 14 as follows:

$$F_s > 1 - D_s + \beta \quad (15)$$

In the above expression, β is a positive or negative fraction. A positive value would lead to increased emphasis on correctly classifying favorable candidates at the expense of error in predicting unfavorable candidates, while a negative value would have the opposite effect. This can serve as a flexible tuning parameter to train the cascaded classifier based on various economic objectives.

We can observe some disparities in predictive performance between the synthetic and field cases. For example, in Table 8b, we see that 2 out of 5 unfavorable wells have been misclassified (false negative error of 40%) compared to a false positive ratio of 25%. This contrasts the results we observe in Table 8a, where the false positive error is slightly higher. This can likely be attributed to the small sample size of unfavorable candidates in the field data pool (5 wells). In other words, if we consider additional field data, the overall classification accuracy is likely to be more similar to those reported with the synthetic cases. Also, looking more closely at the estimated probabilities of restimulation success in Table F-1, we see that while correctly classified restimulation candidates have high probability, the probability estimates for misclassified wells are comparatively high. One potential source of error is the estimation of multiple months of production using simple empirical relationships, which may smooth out important decline characteristics and flow regime changes that are useful in distinguishing between favorable and unfavorable candidates. Overall, the results of this field study are likely to improve with more complete knowledge of reservoir and completion information, as well as more detailed flow modeling which is calibrated to the specific field.

Conclusions

Overall, in this paper we have proposed a novel probabilistic classification approach for analyzing massive hydraulically fractured shale gas well production datasets in order to yield information on the impact of near-wellbore fracture characteristics on well productivity. We have demonstrated the viability of this classification approach for a use-case in restimulation candidate selection. The results show that we are able to distinguish refrac candidates in the presence of erratic rate fluctuations, using a cascaded AdaBoost classifier algorithm which provides both categorical predictions on restimulation success and probability estimates. This serves to quantify the uncertainty associated with the classifier predictions.

Adapted from face detection methods, we use a series of texture-based attributes called Haar-like features to characterize well data as input to the classification algorithm. This data-driven approach is based on the understanding that the rate decline characteristics invariably reflect the multivariate combination of fracture, fluid, reservoir properties which contribute to restimulation success. We have shown that these Haar-like features are analogous to first and second derivatives of flow rate. Therefore, by using these features as input to our cascaded AdaBoost classification model, we have developed a fast and robust statistical model which is sensitive to flow regime changes as well as curvature characteristics at multiple scales and time periods within the life of the well.

To serve as training examples for the proposed classification algorithm, we have generated production data from restimulated wells using a forward flow simulation approach, and quantified restimulation success using numerical metrics such as cumulative production improvement one year after restimulation. By considering the multivariate distribution of input flow simulation parameters, we have estimated the relative importance of physical parameters that influence restimulation success. Through analysis of the principal component loadings, we have shown that multiple combinations of physical parameters have the potential to counteract one another, thereby making it difficult to distinguish between restimulation

candidates using the type curve approach considered. However, by capturing a series of subtle rate fluctuations embedded in Haar-like features, which are related to rate derivative and curvature attributes at different time periods and scales, the proposed approach is able to distinguish between restimulation candidates to a high degree of accuracy for practical purposes.

It is important to emphasize that the use of simulated data in this study is due to a lack of available field data with known reservoir properties and completion information. However, with access to field production information, the proposed method can easily also be applied to train a classification model capable of distinguishing between wells that exhibit any two different criteria of interest, and to develop fast and robust models to analyze future wells. One potential limitation of this approach for use with field data is the dependence of the Haar-like features on the duration and sampling interval of the production data, since they are parameterized by time and scale. This confines the application of our methodology to regularly-sampled production data. The method proposed in this work is not intended to serve as a replacement for traditional simulator-based analyses. On the contrary, it can serve as a complement to stochastic history matching procedures, in order to screen candidate reservoir models that match specific physical criteria.

Finally, the method proposed in this research addresses challenges related to Big Data from hydraulically fracture wells within fractured shale reservoirs by developing a flexible probabilistic screening tool which is specifically geared towards statistical robustness, computational efficiency, and scalability required for applications involving massive datasets. In other words, while we have demonstrated this approach within the context of restimulation candidate selection, it can be easily modified to address other reservoir management/development decisions by harnessing the rate-related information embedded within massive collections of production data. Although the training procedure may potentially be relatively computationally intensive, assessment of production data using the trained model approaches real-time in all cases we have considered. This renders the method amenable to applications requiring real-time analysis of well performance, which has the potential to impact the management of massive datasets obtained from shale gas fields. This helps to facilitate a transition toward workflows involving continuous analysis of well data, in order to bolster confidence and mitigate uncertainty in reservoir management and decision-making.

Acknowledgements

Support from the National Science Foundation (NSF Grant 1546553) is gratefully acknowledged.

Nomenclature

b	Arps' hyperbolic decline constant
b_j	Threshold for binary classification in weak classifier trained on feature index j
D_{ps}	Minimum true positive ratio per stage in cascade training
d_f	Pressure-dependent fracture conductivity decline constant (1/psi)
F_{ps}	Maximum false positive ratio per stage in cascade training
x_{ij}	Haar-like feature score computed using production data from well i based on feature j
$f(x_{ij})$	Binary decision stump corresponding to feature j
G	AdaBoost classifier model
k_f	Fracture permeability (md)
N	Number of production data profiles used in training
n	Number of features per stage multiplier in cascaded training
t	Time (days)
p	Total number of feature combinations
q_i	Gas flow rate for well i (MSCF/D)
\tilde{q}_i	Standardized gas flow rate for well i

\tilde{Q}_i	Standardized cumulative gas flow rate for well i
t_f	Total producing time (days)
V	Number of production data profiles used in testing
w_i	Weight assigned to data from well i during AdaBoost training iteration
y_i	True category of well i
z	Feature template indicator

Greek

α_m	Expansion coefficient in AdaBoost model at iteration m
β	Tuning parameter for cascade training termination criteria
ϵ_m	Classification error associated with feature j at stage m
θ	AdaBoost model classification threshold

References

- Anderson, D. M., Nobakht, M., Moghadam, S., & Mattar, L. (2010). Analysis of Production Data from Fractured Shale Gas Wells. *SPE Unconventional Gas Conference, SPE-131787-MS*. Pittsburgh: Society of Petroleum Engineers.
- Arps, J. J. (1945). Analysis of decline curves. *Transactions of the AIME*, 228-247.
- Aulia, A., Rahman, A., & Velasco, J. Q. (2014). Strategic Well Test Planning Using Random Forest. *SPE Intelligent Energy Conference & Exhibition*. Utrecht: Society of Petroleum Engineers, SPE-167827-MS.
- Bello, R. O., & Wattenbarger, R. A. (2010). Multi-stage hydraulically fractured horizontal shale gas well rate transient analysis. *North Africa technical conference and exhibition*. Society of Petroleum Engineers.
- Brieman, L., Friedman, J., Stones C, J., & Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.
- Brown, M., Ozkan, E., Raghavan, R., & Kazemi, H. (2011). Practical solutions for pressure-transient responses of fractured horizontal wells in unconventional shale reservoirs. *SPE Reservoir Evaluation & Engineering*.
- Chin, L. Y., & Raghavan, R. (2002). Fully Coupled Geomechanics and Fluid-Flow Analysis of Wells With Stress-Dependent Permeability. *SPE Journal, SPE-58968-PA*, 32-45.
- Cipolla, C., & Wallace, J. (2014). Stimulated Reservoir Volume: A Misapplied Concept? . *SPE Hydraulic Fracturing Technology Conference, SPE-168596-MS*. The Woodlands: Society of Petroleum Engineers.
- Clarkson, C. R. (2013, April 1). Production data analysis of unconventional gas wells: Review of theory and best practices. *International Journal of Coal Geology*, pp. 101-146.
- Feblowitz, J. (2013). Analytics in Oil and Gas: The Big Deal About Big Data. *SPE Digital Energy Conference*. The Woodlands: Society of Petroleum Engineers, 163717-MS.
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *icml*, 96, 148-156.
- Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The annals of statistics*, 28(2), 337-407.
- Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics.
- Gilman, J., & Kazemi, H. (1988). Improved calculations for viscous and gravity displacement in matrix blocks in dual-porosity simulators. *Journal of petroleum technology*, 60-70.
- Jayakumar, R., Rai, R., Boulis , A., & Araque-Martinez, A. (2013). A Systematic Study for Refracturing Modeling Under Different Scenarios In Shale Reservoirs. *SPE Eastern Regional Meeting*. Society of Petroleum Engineers.
- Kulga, I. B. (2014). *Analysis of the efficacy of carbon dioxide sequestration in depleted shale gas reservoirs*. UNiversity Park: The Pennsylvania State College Department of Energy and Mineral Engineering.
- Mayerhofer, M. J., Lolon, E. P., Youngblood, J. E., & Heinze, J. R. (2006). Integration of microseismic-fracture-mapping results with numerical fracture network production modeling in the Barnett Shale. *SPE annual technical conference and exhibition*. SPE.
- Mohaghegh, S., Reeves, S., & Hill, D. (2000). Development of an intelligent systems approach for restimulation candidate selection. *SPE/CERI Gas Technology Symposium*. Society of Petroleum Engineers.
- Moore, L. P., & Ramakrishnan, H. (2006). Restimulation: candidate selection methodologies and treatment optimization. *SPE Annual Technical Conference and Exhibition*. San Antonio: Society of Petroleum Engineers.
- Ozkan, E., Brown, M. L., Raghavan, R. S., & Kazemi, H. (2009). Comparison of Fractured Horizontal-Well Performance in Conventional and Unconventional Reservoirs . *SPE Western Regional Meeting*. San Jose: Society of Petroleum Engineers, SPE-121290-MS.
- Patri, O. P., Panangadan, A. V., Chelmis, C., McGee, R., & Prasanna, V. (2014). Predicting Failures from Oilfield Sensor Data using Time Series Shapelets. *SPE Annual Technical Conference and Exhibition*. Amsterdam: Society of Petroleum Engineers, SPE-170680-MS.
- Roussel, N. P., & Sharma, M. M. (n.d.). Selecting Candidate Wells for Refracturing Using Production Data. *SPE Annual Technical Conference and Exhibition*. 2011: Society of Petroleum Engineers.
- Rubin, B. (2010). Accurate Simulation of Non Darcy Flow in Stimulated Fractured Shale Reservoirs. *SPE Western Regional Meeting*. Anaheim: Society of Petroleum Engineers, SPE-132093-MS.
- Seemann, D., Williamson, M., & Hasan, S. (2013). Improving Reservoir Management through Big Data Technologies. *SPE Middle East Intelligent Energy Conference and Exhibition*. Manama: Society of Petroleum Engineers, SPE-167482-MS.

- Shelley, R. F. (1999). Artificial Neural Networks Identify Restimulation Candidates in the Red Oak Field. *SPE Mid-Continent Operations Symposium*. Society of Petroleum Engineers.
- Stehfest, H. (1970). Algorithm 368: Numerical inversion of Laplace transforms. *Communications of the ACM*, 47-49.
- Tavassoli, S., Yu, W., Javadpour, F., & Sepehrnoori, K. (2013). Selection of Candidate Horizontal Wells and Determination of the Optimal Time of Refracturing in Barnett Shale (Johnson County). *SPE Unconventional Resources Conference Canada*. Society of Petroleum Engineers.
- Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Volume 1)* (pp. I-511 - I-518). IEEE.

Appendix A

Classifier training algorithms

Input: N training data

- $N/2$ examples each of favorable and unfavorable cases
- Feature scores, $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}] \forall i = 1, \dots, N$
- Categorical labels, $y_i = \{-1, 1\}$ i.e. unfavorable and favorable cases respectively

Procedure:

1. Begin with equal weights for first iteration, $w_i^{(m)} = 1/N$
2. For each iteration m
 - a. Normalize weights $w_i^{(m)} = \frac{w_i^{(m)}}{\sum_{i=1}^N w_i^{(m)}}$
 - b. Determine weak classifier f_m which minimizes error, $\epsilon_m = \frac{\sum_{i=1}^N w_i^{(m)} I(f_m \neq y_i)}{\sum_{i=1}^N w_i^{(m)}}$
 - c. Compute expansion coefficient for current iteration, $\alpha_m = \ln \left[\frac{1 - \epsilon_m}{\epsilon_m} \right]$
 - d. Up-weight misclassified data for next round, $w_i^{(m+1)} = w_i^{(m)} \exp[\alpha_m I(f_m \neq y_i)]$
3. Final classifier prediction, $\hat{y} = \text{sign}[\sum_{m=1}^M (\alpha_m f_m) - \theta]$

Table A-1: AdaBoost classification algorithm modified from Friedman et al. (2009).

Input:

- Pool of K training data with class labels, where $K \gg N$
 - Training-testing data split ratio, $N:V$, where $N + V = 1$
- User-defined objectives
 - Maximum false positive ratio per stage, F_{ps}
 - Minimum true positive ratio per stage, D_{ps}
 - Number of features per stage multiplier, n (optional)

Procedure:

1. Randomly select $N/2$ favorable training cases and V test samples from pool
2. Initialize $F_0 = 1; D_0 = 1; s = 0$
3. While $F_s > 1 - D_s$:
 - a. Increment stage $s = s + 1$; Initialize number of iterations, $m = 0$
 - b. While $F_s > F_{ps} * F_{s-1}$ or $m \leq n * s$
 - i. Begin new AdaBoost iteration, $m = m + 1$
 - ii. Train AdaBoost classifier with selected training samples
 - iii. While $D_s \leq D_{ps} * D_{s-1}$,
 - Decrease the threshold θ of the classifier at current stage s
 - c. Evaluate new cascade classifier on negative examples in the training pool and select $N/2$ false positives as the new unfavorable training cases

Table A-2: Cascade training procedure adapted from Viola and Jones (2001).

Appendix B

Summary of features selected during cascaded classifier training

Stage	Ratio of two-rectangle features (z=1)	Ratio of three-rectangle features (z=2)
1	1.00	0.00
2	0.67	0.33
3	0.75	0.25
4	0.88	0.13
5	0.86	0.14
6	0.88	0.13
7	0.94	0.06
8	0.90	0.10
9	0.88	0.13
10	0.94	0.06
11	0.67	0.33
12	0.75	0.25
13	1.00	0.00
Average	0.85	0.15

Table B-1: Proportions of features selected in each cascade stage, for based on feature template.

Stage	Ratio of features evaluated at t=[1,6] months	Ratio of features evaluated at t=[7,12] months	Ratio of features evaluated at t=[13,18] months	Ratio of features evaluated at t=[19,24] months
1	0.44	0.11	0.22	0.22
2	0.67	0.00	0.00	0.33
3	0.25	0.38	0.38	0.00
4	0.40	0.23	0.20	0.18
5	0.43	0.14	0.29	0.14
6	0.50	0.38	0.13	0.00
7	0.29	0.26	0.35	0.10
8	0.47	0.20	0.27	0.07
9	0.63	0.00	0.38	0.00
10	0.35	0.35	0.24	0.06
11	1.00	0.00	0.00	0.00
12	0.00	0.75	0.00	0.25
13	0.25	0.00	0.50	0.25
Average	0.44	0.21	0.23	0.12

Table B-2: Proportions of features selected in each cascade stage, based on time instance t at which the features are evaluated.

Stage	Ratio of features with scale $\Delta t = [1,6]$ months	Ratio of features with scale $\Delta t = [7,12]$ months	Ratio of features with scale $\Delta t = [13,18]$ months	Ratio of features with scale $\Delta t = [19,24]$ months
1	0.44	0.11	0.22	0.22
2	0.67	0.00	0.00	0.33
3	0.25	0.38	0.38	0.00
4	0.40	0.23	0.20	0.18
5	0.43	0.14	0.29	0.14
6	0.50	0.38	0.13	0.00
7	0.29	0.26	0.35	0.10
8	0.47	0.20	0.27	0.07
9	0.63	0.00	0.38	0.00
10	0.35	0.35	0.24	0.06
11	1.00	0.00	0.00	0.00
12	0.00	0.75	0.00	0.25
13	0.25	0.00	0.50	0.25
Average	0.44	0.21	0.23	0.12

Table B-3: Proportions of features selected in each cascade stage, based on feature scale Δt .

Appendix C

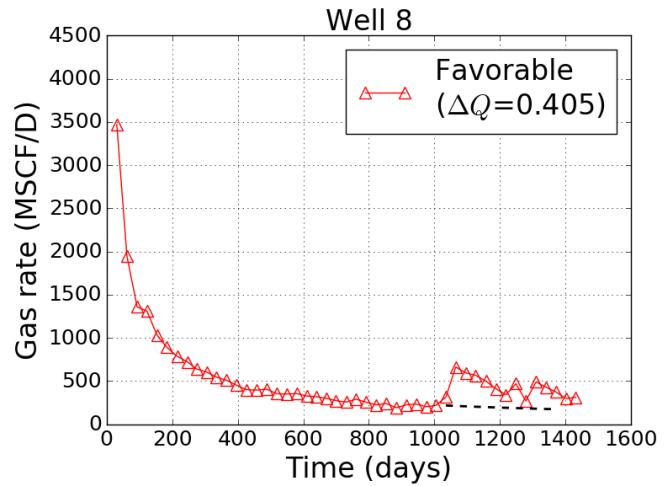
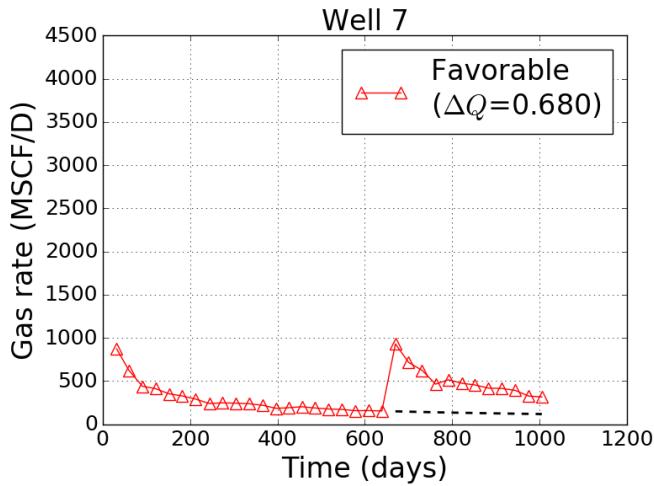
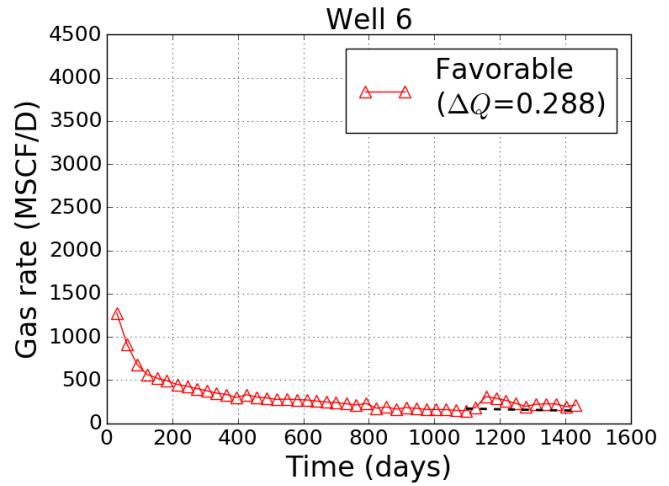
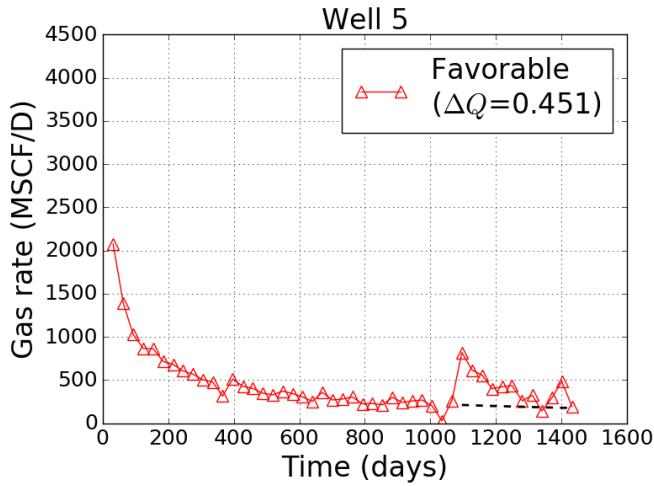
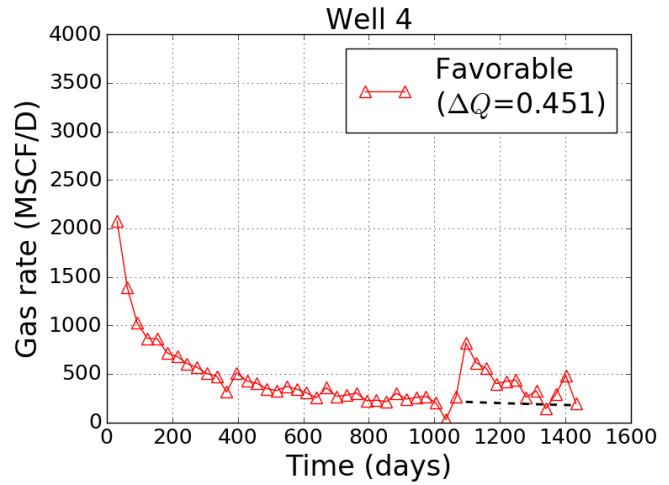
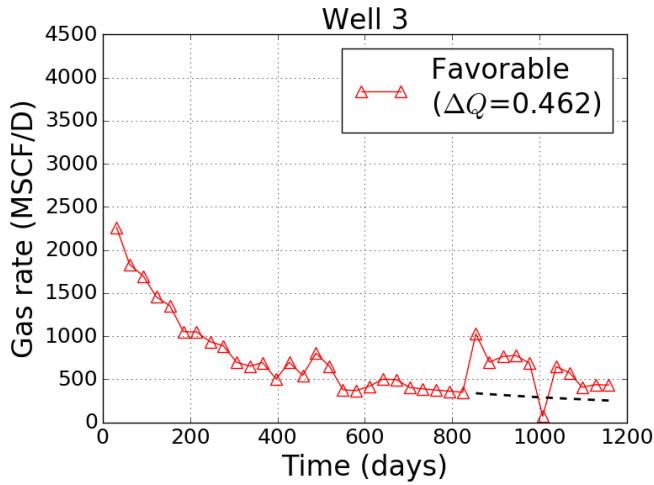
Principal component analysis data

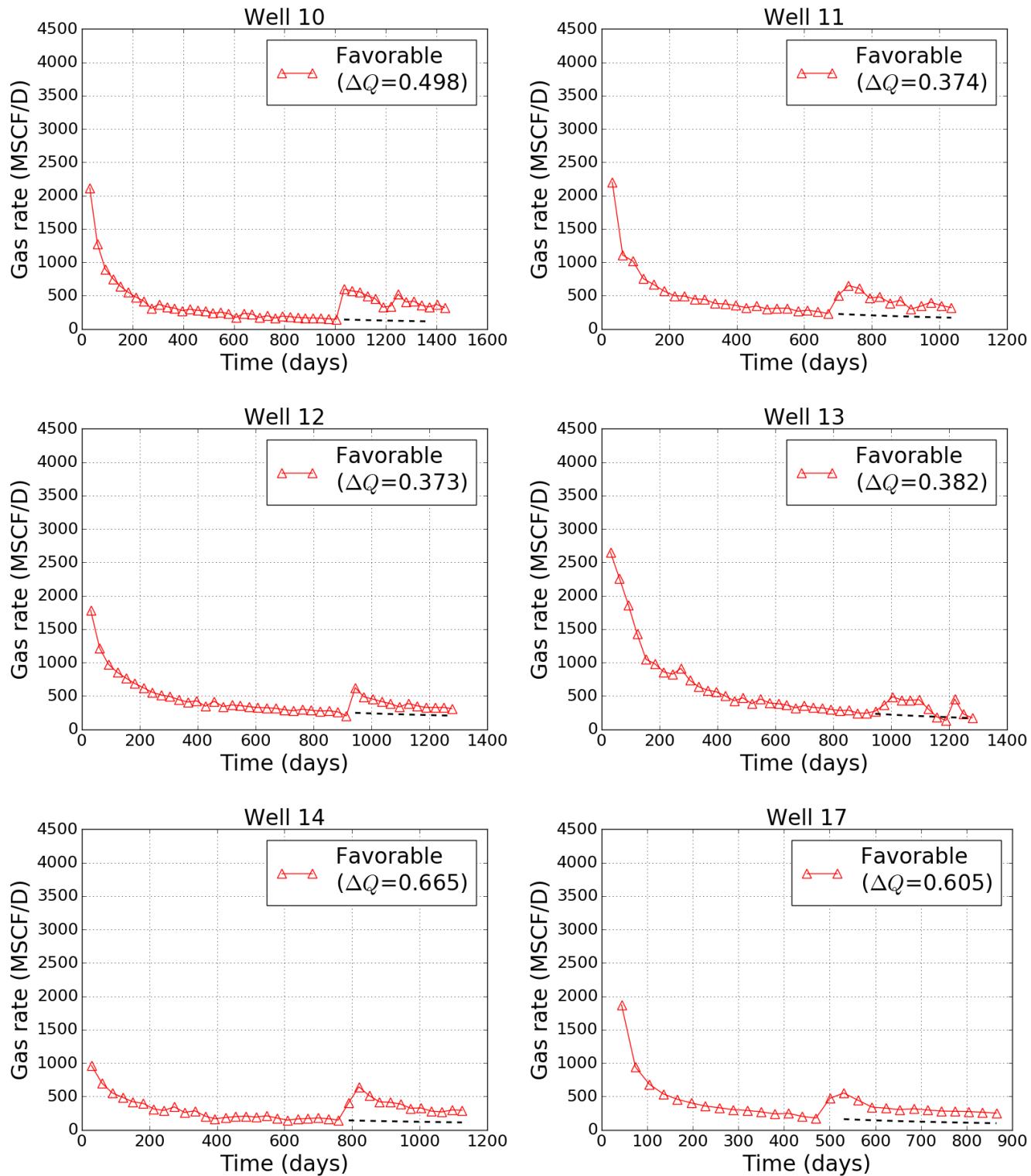
		False positive case		False negative case
Parameters, k	Value	$l_k \psi_{ki}$	Value	$l_k \psi_{ki}$
Hydraulic fracture closure rate	3.37	-0.548	4.19	0.169
Initial gas saturation	0.797	0.343	0.801	0.377
Hydraulic fracture tip conductivity	19.38	-0.518	5.56	-0.091
Hydraulic fracture spacing	120	0.378	120	0.378
Langmuir volume	86.04	-0.232	54.56	-0.348
Diffusivity	5.27E-05	-0.024	6.26E-05	-0.142
Shale porosity	8.89E-02	-0.065	4.26E-02	0.264
Reservoir thickness	253	0.187	276	0.268
Flowing bottomhole pressure	490.34	0.107	451.83	0.139
Hydraulic fracture half length	570	-0.253	510	-0.151
Hydraulic fracture conductivity	28.24	-0.239	11.90	0.239
Shale compressibility	2.57E-06	0.145	2.25E-06	0.062
Langmuir pressure	580.28	0.186	586.41	0.183
Initial reservoir temperature	325.02	-0.136	149.35	0.102
Complex fracture conductivity	11.05	0.148	3.27	-0.105
Complex fracture spacing	0.519	-0.110	1.561	0.048
Initial reservoir pressure	2455.30	0.102	6462.52	-0.060
Bulk density	2.955	0.106	2.615	0.014
Natural fracture conductivity	4.43E-03	-0.065	3.02E-02	0.019
Natural fracture spacing	2.948	0.031	2.184	0.042
Reservoir depth	8825	0.014	5840	-0.018
Shale permeability	3.007E-04	0.008	1.856E-04	0.011
Complex fracture closure rate	4.26	0.001	4.07	0.000
Number of hydraulic fracture stages	7	0.000	9	-0.001
Principal component score, z_i		-0.437		1.398

Table C-1: Principal component analysis data corresponding to false positive and false negative cases identified in type curve analysis. $l_k \psi_{ki}$ represents the product of the standardized flow parameter and the corresponding principal component loading (see Equation 2).

Appendix D

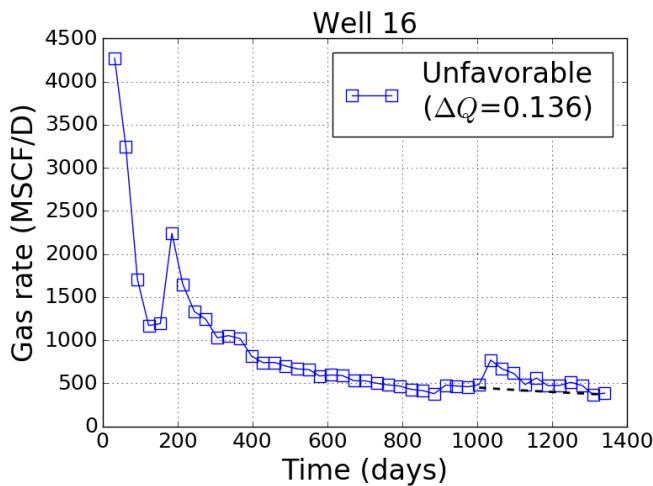
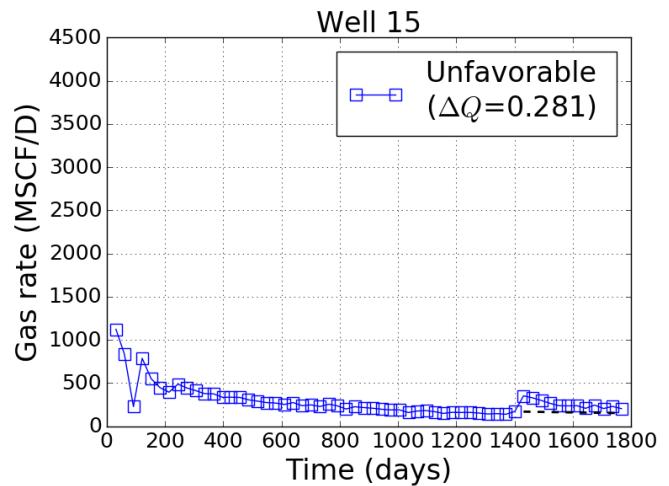
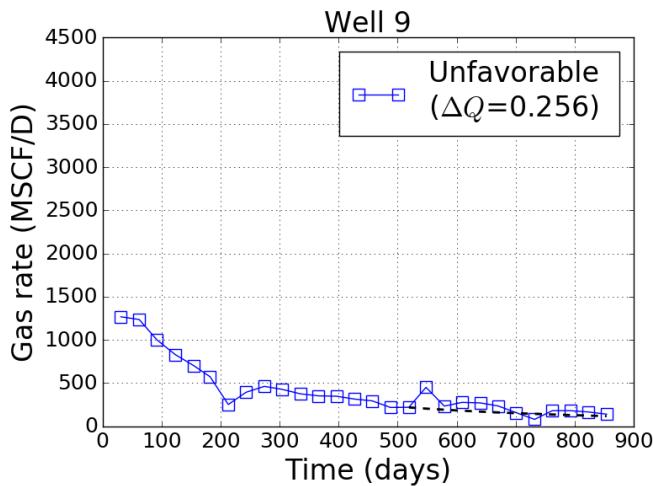
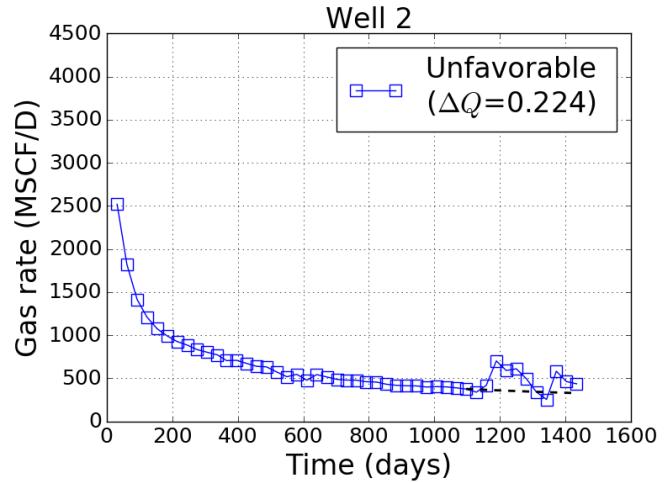
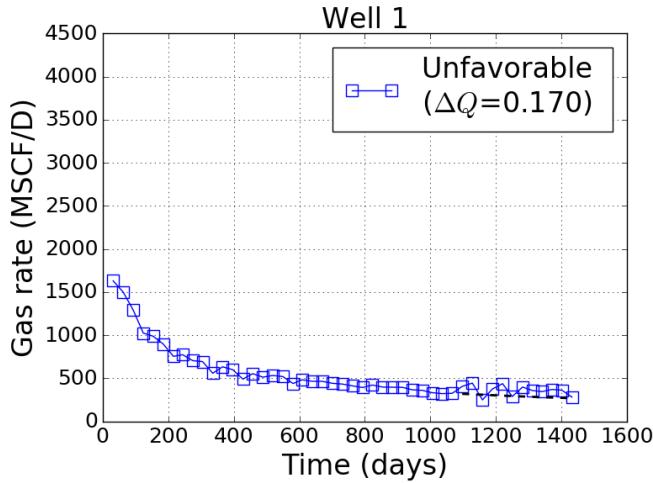
Favorable candidate wells used in field validation study





Appendix E

Unfavorable candidate wells used in field validation study



Appendix F

Field validation study results

Well	Assumed timing of restimulation (years)	Change in one-year cumulative production	Assigned category	Predicted category	Estimated probability of restimulation success
1	3	0.170	Unfavorable	Unfavorable	0.32
2	3	0.224	Unfavorable	Favorable	0.92
3	2	0.462	Favorable	Unfavorable	0.31
4	3	0.451	Favorable	Favorable	0.92
5	2	0.466	Favorable	Favorable	0.97
6	3	0.288	Favorable	Favorable	0.98
7	2	0.680	Favorable	Favorable	0.97
8	3	0.405	Favorable	Favorable	0.92
9	2	0.256	Unfavorable	Unfavorable	0.31
10	3	0.498	Favorable	Unfavorable	0.36
11	2	0.374	Favorable	Favorable	0.97
12	3	0.373	Favorable	Favorable	0.93
13	3	0.382	Favorable	Favorable	0.78
14	2	0.665	Favorable	Unfavorable	0.31
15	3	0.281	Unfavorable	Unfavorable	0.36
16	3	0.136	Unfavorable	Favorable	0.98
17	2	0.605	Favorable	Favorable	0.98

Table F-1: Results of field study to validate classification approach.