PRODUCTION DATA ANALYSIS BY MACHINE LEARNING

A Dissertation

by

PENG ZHOU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	John Lee
Co-Chair of Committee,	Huiyan Sang
Committee Members,	Duane McVay
	Peter Valko
Head of Department,	Jeff Spath

May 2019

Major Subject: Petroleum Engineering

ABSTRACT

In this dissertation, I will present my research work on two different topics. The first topic is production data analysis of low-permeability well. The second topic is a quantitative evaluation of key completion controls on shale oil production.

In Topic 1, I propose and investigate two novel methodologies that can be applied to improve the results of low-permeability well decline curve analysis. Specifically, I first proposed an iterative two-stage optimization algorithm for decline curve parameter estimation on the basis of two-segment hyperbolic model. This algorithm can be applied to find optimal parameter results from the production history data. By making use of a useful relation that exits between material balance time (MBT) and the original production profile, we propose a three-step diagnostic approach for the preliminary analysis of production history data, which can effectively assist us in identifying fluid flow regimes and increase our confidence in the estimation of decline curve parameters. The second approach is a data-driven method for primary phase production forecasting. Functional principal component analysis (fPCA) is applied to extract key features of production decline patterns on basis of multiple wells with sufficiently long production histories. A predictive model is then built using principal component functions obtained from the training production data set. Finally, we make predictions for the test wells to assess the quality of prediction with reference to true production data. Both methods are validated using field data and the accuracy of production forecasts gives us confidence in the new approaches.

In Topic 2, generalized additive model (GAM) is applied to investigate possibly nonlinear associations between production and key completion parameters (e.g., completed lateral length, proppant volume per stage, fluid volume per stage) while accounting for the influence of different geological environments on hydrocarbon production. The geological cofounding effect is treated as a random clustered effect and incorporated in the GAM model by means of a state-of-the-art statistical machine learning method graphic fused LASSO. We provide several key findings on the relation between completion parameters and hydrocarbon production, which provide guidance in the development of efficient completion practices.

DEDICATION

To my wife Chen Lin, my son Samuel Zhou and my daughter Elim Zhou

for their endless support, love and encouragement

ACKNOWLEDGEMENTS

I would like to thank my committee chair and co-chair, Dr. John Lee and Dr. Huiyan Sang, who provides me the opportunity, financial support and tremendous guidance so that I can complete my Ph.D. program. Sincere thanks to my other committee members, Dr. McVay, Dr. Peter Valko and Dr Laprea Bigott Marcelo for their guidance during the course of this research.

Special thanks to my mentor Ligang Lu for his guidance and support during my time at Shell in summer 2017. Many thanks to the Advanced Analytics team for making my summer in 2018 at Maxim Integrated in California a great experience and unforgettable memory.

In the end, I would not finish my Ph.D. program without the support from my wife, my lovely son and my lovely daughter. Their support and encouragement are always important no matter where I am.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor John Lee (chair), Duane McVay and Peter Valko of Department of Petroleum Engineering and Professor Huiyan Sang (co-chair) of Department of Statistics.

The data analyzed for Chapter II and III is from Drilling Info. The data analyzed for Chapter IV was provided by Shell Technology Center in Houston.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

All the funding sources I received during the Ph.D. period are from National Science Foundation, Department of Petroleum Engineering at Texas A&M University and Shell International Exploration and Production Company, respectively.

NOMENCLATURE

b_i	Arps b parameter in transient flow regime
b_f	Arps b parameter in boundary dominant flow regime
D_i	Initial decline rate, $1/\mathrm{D}$
$D_{\min}$	Minimum decline rate, $1/\mathrm{D}$
D_∞	Power law decline at "infinite time" constant, $1/\mathrm{D}$
G_p	Oil cumulative production, bbl
q_i	Initial production rate, $\mathrm{bbl/D}$
q_1	Production rate at Day 1, $\mathrm{bbl/D}$
q_∞	Production rate at "infinite time", $\mathrm{bbl/D}$
t_c	Switching time from transient flow to boundary dominant flow
τ	Characteristic time in stretch exponential model, $1/\mathrm{D}$
β_l	Constant for late-time period in extended exponential model
β_e	Constant for early-time period in extended exponential model
ACE	Alternating conditional expectation
BDF	Boundary dominant flow
CI	Confidence interval
DCA	Decline curve analysis
DEN	Rock density
EEDCA	Empirical extended exponential model
EUR	Estimated ultimate recovery

fPCA	Functional principal component analysis
GAM	Generalized additive model
GR	Gamma ray
GOR	Gas Oil Ratio, scf/STB
LOO	Leave one out
LASSO	Least absolute shrinkage and selection operator
MBT	Material balance time
MCMC	Markov Chain Monte Carlo
MSE	Mean square error
NEU_LIM	Neutron porosity
PCA	Principal component analysis
RESDEP	Deep resistivity
SSE	Sum of square error
WOR	Water Oil Ratio, bbl/STB

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

The US shale oil boom has reshaped the global energy landscape (Holditch 2010) and according to preliminary estimates published recently by the Energy Department, the United States is the world's largest producer of crude oil for the first time since 1973 (**Figure** 1). This great economic success in developing these resources have been largely driven by the advances in drilling technologies such as multistage horizontal well drilling and multistage fracturing. However, the lack of sufficient knowledge in physical properties and the physics controlling production from shale formation limits our ability to model and forecast with confidence production and reserves from these important plays. Advances in the industry's ability to forecast future production more accurately impacts financial forecasts, perceived asset values and accuracy of reserves disclosed to the public.

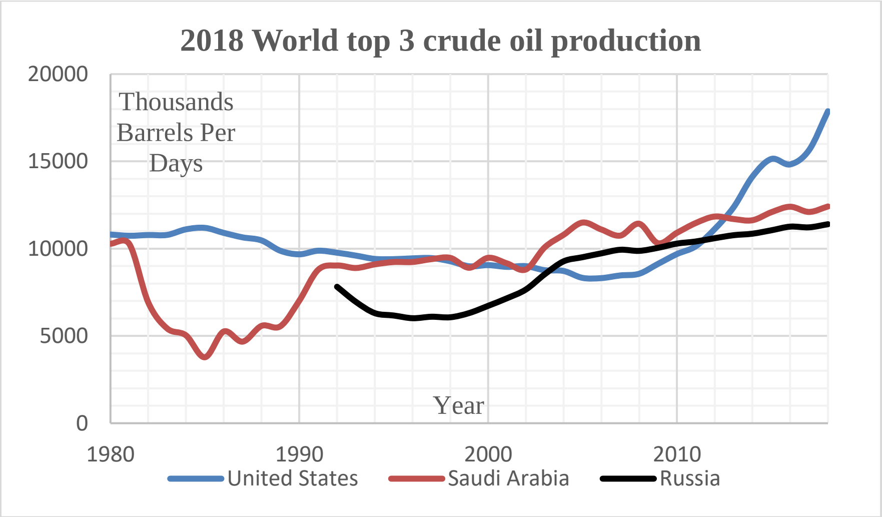


Figure 1: Monthly crude oil production

A variety of tools have been developed for reserve evaluation and production forecasting in the petroleum industry. One approach most frequently used in industry is history matching through numerical simulation (Aziz, K. and Settari, A. 1979). This method incorporates the observed production data into the geological model by calibration. Once unknown reservoir and well properties are determined through history matching, the improved numerical model can be used to forecast future production and remaining reserves. The advantage of this method is that it accounts for complex reservoir heterogeneity, rock compaction, pressure dependence of reservoir fluid properties and many other important features of shale reservoirs. The disadvantage is that this method requires much work in reservoir characterization and numerical simulation, which is typically time consuming. For unconventional reservoirs, fracture geometry, distribution and connections are difficult to characterize and therefore, construction of a robust detailed numerical model can be difficult, and requires extensive manpower, cost, time and data collection.

In practice, decline curve analysis (DCA) is a quick and efficient method for engineers to make prediction on well future production. However, it turns out that the traditional Arps equation (Arps 1944) is not applicable to ultra-low permeability formations because production data from these shale wells differ significantly from conventional production data. As a result, a large amount of work has been done for unconventional well decline curve analysis and a review of different decline curve models will be given in next section.

1.1 Decline curve literature review

In the study of production data from unconventional reservoirs, engineers always have the difficulties of simultaneously matching the high initial production rate, extremely sharp decline rate in the transient period, and shallow declines in the boundary-dominated flow (BDF) period. In the early life of the shale producer, the decline rate can only be captured with a b factor greater than 1, which is out of the bounds of traditional Arps equation. Thus, researchers have developed many different empirical models with the goal of finding models more suitable for low permeability reservoirs. In this section, the current existing decline curve model are reviewed.

Ilk *et al.* (2008) proposed a power-law model for the production data from shale reservoirs. This model is a rate equation:

$$q(t) = q_i \exp[-D_\infty t - D_i t^n] \tag{1.1}$$

where q_i is the initial flow rate; D_∞ is a decline constant at “infinite time”; D_i is another decline constant that is introduced for the fitting on the production data at early transient flow period; n is an exponential index that is less than 1, but greater than 0. According to the definition of decline rate, $D(t)$ can be written as

$$D(t) = D_\infty + nD_i t^{n-1} \tag{1.2}$$

In this model, there are four parameters to be determined.

Valko *et al*. (2010) proposed another empirical stretched exponential model for unconventional production data analysis. The rate equation is given as

$$q(t) = q_i \exp\left[-\left(\frac{t}{\tau}\right)^n\right] \tag{1.3}$$

where q_i is the initial flow rate; τ is a characteristic time parameter; n is an exponential index with the value being in the domain from 0 to 1. The decline rate of this model is given as

$$D(t) = \frac{n}{\tau}\left(\frac{\tau}{t}\right)^{1-n} \tag{1.4}$$

In this model, there are three parameters to be determined.

It turns out that these two models can be derived based on the following two conditions:

(1) The decline rate is always positive

(2) The production decline at late time is assumed to be exponential decline.

Based on these two conditions, the form of rate equation should be given as follows:

$$q(t) = q_i \exp[-(\beta_l + f(t))t] \tag{1.5}$$

where β_l is a nonnegative constant to account for the late-life flow rate performance and $f(t)$ is a function we can select so that $q(t)$ will satisfy the above two conditions. The selection will follow some reasonable principles. The first principle for $f(t)$ is that the limit value of $f(t)$ should be zero at t goes to infinity and $f(t)$ is a monotonically decreasing function (the first derivative of $f(t)$ is always negative). The expression for the first principle is, therefore, given as

$$f(t) \rightarrow 0 \text{ as } t \rightarrow \infty \text{ and } f'(t) < 0 \text{ for all } t \tag{1.6}$$

The first principle guarantees the condition 2 is automatically satisfied.

The second principle is associated with the decline rate. The equation of decline rate is

$$D(t) = \beta_l + f(t) + tf'(t) \tag{1.7}$$

To guarantee that $D(t)$ is always positive definite, a sufficient condition is that

$$f(t) + tf'(t) > 0 \tag{1.8}$$

Since $f'(t)$ is required to be always negative according to (1.6), then we can write the condition (1.8) as

$$\frac{1}{t} > -\frac{f'(t)}{f(t)} > 0 \tag{1.9}$$

One solution that satisfies condition (1.9) can be easily seen

$$-\frac{f'(t)}{f(t)} = \frac{n}{t}, 0 < n < 1 \tag{1.10}$$

Then, solve equation (1.10) and we have $f(t)$ as

$$f(t) = C_1 t^{-n} + C_2 \tag{1.11}$$

where C_1 and C_2 are two constants. To make sure that $f(t)$ goes to zero as t approaches infinity, C_2 is set to be zero. Then $f(t)$ is $C_1 t^{-n}$ with $0 < n < 1$. This solution is just the Power law rate equation. If β_l is set to be zero, then the rate equation is reduced to stretched exponential model. Therefore, power law model and stretched exponential model both have the same origin and their difference is only whether the production decline at late BDF is assumed to be exponential decline or not. The stretched exponential model has problems in switching from linear to BDF flow regimes (Freeborn and Russel, 2014).

The Duong model (Duong, 2010) was developed specifically for low-permeability reservoirs. The equation of Duong model is

$$G_p = \frac{q_1}{a} e^{\frac{a}{1-m}(t^{1-m}-1)} \tag{1.12}$$

where q_1 is rate at day 1, a and m are constant parameters to be determined from $\log \frac{q(t)}{G_p}$ vs $\log t$ and t is in days. The estimation of a and m can be derived from the intercept on y axis and the slope, respectively. This method has been applied to some shale gas wells and it is found that this method is unable to work with mixed flow regimes, like switching from transient to BDF regimes (Freeborn and Russel, 2014). As a result, modified Duong's model (Joshi and Lee, 2013) is suggested where the regression line is forced through the origin ($q_\infty = 0$) and a change from transient flow to boundary dominated flow is modeled using a hyperbolic decline of 5%.

Zhang *et al.* (2015) proposed an empirical extended exponential decline model for shale reservoir. The equation of extended exponential decline model is

$$q(t) = q_i e^{-at} \tag{1.13}$$

where $a = \beta_l + \beta_e e^{-t^n}$; β_l is a constant to account for the late-life period; β_e is a constant to account for the early transient period; n is an empirical exponent and t is the time in months. The decline rate derived from extended exponential decline model is

$$D(t) = \beta_l + \beta_e e^{-t^n}(1 - nt^n) \tag{1.14}$$

The decline rate (1.14) is not always positive definite. The second term of (1.14) is not a monotone function. The decline rate will decrease as t increases at early time and then at some point t_{crit}, the first derivative of $D(t)$ will be equal to zero. After that critical point, the decline rate starts to increase and the limit value of $D(t)$ is β_l as t goes to infinity. In other words, the value of $D(t)$ at the critical point t_{crit} is a minimum value. However, the minimum value of $D(t)$ is not always positive. The equation of minimum decline rate is

$$\min D(t) = \beta_l - \beta_e e^{-\frac{n+1}{n}} \tag{1.15}$$

If $\beta_l > \beta_e e^{-\frac{n+1}{n}}$, $D(t)$ is always positive; otherwise, $D(t)$ can be negative, which means the production rate does not always decline and may increase in some period of time. As a result, we should be cautious to use extended exponential decline model in production data analysis for shale wells.

Modified hyperbolic model is the most popular decline model used in the industry due to its simplicity. The early transient flow is modeled by hyperbolic decline with Arp's b parameter greater than 1 and the late BDF is modeled by exponential decline, which will yield finite estimates of ultimate recovery (EUR). The switching time from transient flow to BDF is assumed to be at the time when the decline rate is approximately 5% according to engineering experience. Modified hyperbolic model is always used as a benchmark model to assess the quality of other new models.

1.2 Machine learning in petroleum engineering

Machine learning has proved to be a very powerful technology in IT industries. In machine learning, we use statistical and optimization knowledge to analyze data with the goal of constructing reliable models for forecasting and classification. This technology is appealing to the petroleum industry since it might bring some breakthroughs on many problems that require data analysis. There is a massive amount of studies being done using machine learning in petroleum engineering. In this section, a review is given on several popular topics that researchers have tried to solve by machine learning.

The first topic is about sweet spots exploration. A shale reservoir may be thought of a sweet spot if it has high total organic carbon (TOC) and high fracability. TOC is one of the main factors for identifying the reservoir with higher potential gas production; the fracability affects the flow of hydrocarbons in a shale reservoir and future fracking in it (Tahmasebi *et al*., 2017), and brittle shale is more likely to be naturally fractured and get good response to it. The related sub-topics on sweet spots exploration include TOC estimation, permeability and porosity prediction, lithofacies classification, etc, as is shown in **Table** 1.

Topic	Inputs	Outputs	Techniques
TOC estimation	Log data Limited core data (to train algorithm)	TOC values	Support vector regression/machine Neural network Fuzzy logic Others: multi-linear regression, extreme learning machine
Permeability and Porosity Prediction	Log data volume of shale the sonic and density logs the porosity log core permeability measurements	Permeability values	Discriminant analysis Fuzzy/Neural Inference Principal component analysis Others: Hybrid Genetic Programming
Lithofacies analysis	Core analysis data log data seismic data	Lithofacies class	Neural network Support vector machine
Quantitative seismic interpretation	Seismic Volume well log	frackability lithofacies	Neural network Support vector machine Others: self-organizing map

Table 1: Summary for sweet spots exploration

The second topic is about well operations. The related sub-topics include completion design analysis, well design, hydraulic fracture performance, stage completion performance, fracture spacing analysis, etc. A summary for well operations is given in **Table** 2 below.

Topic	Input	Output	Techniques
Production optimization	Reservoir quality proxies Well architecture Well completion	Oil Production	Gradient Boosting Models Random Forests Support Vector Machine Kriging Method
Completion design analysis	Reservoir quality proxies Well architecture Well completion	Oil production	Gradient Boosting Machine/Boosted Tree Regression
Well design	Geological, geophysical, geomechanical, stimulation, petrophysical, reservoir, production, etc.	Recommendations for well stimulation, well location, orientation, design and operation	Others: a self-teaching expert system
Hydraulic fracture performance	Reservoir quality proxies Well architecture Well completion Stimulation data	Production	Boosted tree method Random forest Others: learning machine
Stage completion performance	near-wellbore geologic parameters Stage-level engineering parameters Seismic-derived inversion reservoir properties	pressure response	Random Forests
Fracture spacing analysis	bed thickness Structural position	fracture spacing	Neural network

Table 2: Summary for well operations

Specifically, Baker Hughes implemented random forest and gradient boosting models to well production optimization via completion design (Lafollette *et al*., 2014; Zhong *et al.*, 2015; Schuetter *et al*., 2015; Vera *et al*., 2015; Lafollette *et al*., 2012). Ruths *et al.* (2017) applied random forests to merge seismic, geologic, and engineering data to predict completion performance, with respect to well design, hydraulic fracture performance, pressure response, stage completion performance, etc. Moreover, Lolon *et al*. (2016) from

Liberty Oilfield Services analyzed completion design and fracture treatment by gradient boosting models and random forests to determine the important predictors.

Additionally, there are several studies using MCMC for production analysis. For example, Fulford *et al*. (2015) used MCMC and transient hyperbolic model to analyze the rate time decline behavior. Khanal *et al*. (2017) proposed a new forecasting method combining principal component analysis and MCMC. McLane and Gouveia (2015) applied Monte-Carlo for validating analog production type curves. More Bayesian framework methods are implemented into history matching studies.

Other than the topics mentioned above, there are several studies for fracture spacing analysis. Kaviani *et al*. (2006) applied radial basis function (RBF) into artificial neural network for small datasets. Inverse fluid modeling facilities designing drilling, spacer, cement or fracturing fluids is studies by Tarrahi and Shadravan (2016) through Gaussian Process Regression.

The third topic is about asset evaluation. Asset evaluation includes studies related with production forecasting, history matching, reservoir evaluation (i.e., PVT analysis) and asset management. A summary for asset evaluation is given in **Table** 3 below:

Topic	Inputs	Outputs	Techniques
Production prediction	Historical production data Tubing head pressure Production rate Geological map	Oil production	Decision Tree Time series analysis Clustering analysis Neural Network Others: learning machine systems, Bayesian machine learning methods, self-organizing map Principal Component Analysis
History matching	Production data	Facies model; Reservoir model	PCA Others: Ensemble Kalman filter, Bayesian analysis
Bubble point pressure	data from a variety of crude oil of various composition ranges and from various geographical locations	Bubble point pressure	Support vector machine Principal component analysis Neural network
Crude oil viscosity	experimental PVT data	Oil viscosity	Support vector machine
Condensate-to gas ratio	experimental data and some PVT data available in the literature	condensate-to-gas ratio	Neural network
Asset management/quality assessment	High-frequency data, i.e., water rate, oil rate, fluid temperature and water content in oil.	Expert decision, i.e., healthy or faulty	Others: active learning, semi-supervised learning

Table 3: Summary of asset evaluation

Several learning systems have been created for production prediction in recent years. AKW Analytics Inc developed a generalized Petroleum Analytics Learning Machine (PALM) system (Anderson *et al.*, 2016a; 2016b). This learning machine includes many machine learning algorithms, Systems Integration Database (SID) and cloud-based file system. This learning system enables users to model oil and gas production based on hundreds of geological, geophysical, and petroleum engineering attributes, not only measured in the field, but also computed form models such as reservoir simulations. Different from other study systems, this learning machine collects not only statistical machine learning methods, but also other big data analytics methods, such as information extraction, noisy test processing, knowledge discovery, etc. Those methods make the system more powerful in unstructured data analysis. PALM can be used in many ways. Anderson *et al*. (2016a, 2016b) provided several examples, such as forecasting production, classifying hydraulic fractures and discovering the correlations between machine recognized Frac-Classes, completion improvements and production performance.

History matching is simulation modelling-matching historical data through tuning the parameters in the simulation model. It has been analyzed for a long time, and different stochastic approaches have been proposed (Landa *et al*., 2005; Gao *et al*., 2004; Oliver *et al.*, 2008; Oliver and Chen, 2011; Rwechungura *et al.*, 2011; Emerick and Reynolds, 2012; Tavakoli *et al*., 2014; Vink *et al*, 2015; Chen *et al*., 2012, 2014, 2016; Gao *et al*., 2016). Chen *et al*. (2014, 2016) implemented a pluri-principal-component-analysis (PCA), which combines PCA with pluri-Gaussian simulation, to extract the major geological features and generate real-valued facies. The authors integrated pluri-PCA with a derivative-free optimization algorithm and built the rock-type rules automatically according to the proposed method. The proposed model allows gradually changing facies distribution to match production data, and it is geologically and physically consistent (Chen *et al*., 2016). Based on their pluri-PCA model, Honorio *et al*. (2015) extended the model with PRaD method (Piecewise Reconstruction from a Dictionary). Compared with the pluri-PCA method, integration of PRaD with pluri-PCA may further reduce the misclassified facies by 80% when 200 or more PCs are used. Ensemble Kalman Filter (EnKF), introduced by Evensen (2003), is also a popular method in petroleum engineering. Tavakoli *et al*. (2014), Emerick and Reynolds (2013), Oliver and Chen (2011) applied EnKF as a useful history-matching tool.

As for reservoir evaluation, there is a group of research related with PVT analysis which describes the reservoir characteristics, i.e., bubble point pressure, crude oil viscosity, saturation pressure, etc; other studies focus on breakthrough time of water coning, condensate-to-gas ratio. Support vector machine and neural network are widely used in this area (Ramirez *et al*., 2017; Farasat *et al*., 2013; Hemmati Sarapardeh *et al.*, 2014; Zendehboudi *et al.*, 2012; Rafiee-Taghanaki *et al*., 2014). Most of the models show stable performances and have good agreement with experimental or field data.

Last but not the least, asset management is especially challenging due to its multi-disciplinary, cross-functional, and human expertise-intensive (El-Bakry *et al*., 2012). Meanwhile, the large amount of data and the complexity of the components and processes also create difficulties for real-time decision making. The related problems include data acquisition and storage, asset health monitoring (Shyeh *et al*., 2008; El-Bakry *et al*., 2012; Subrahmanya *et al*., 2014). ExxonMobil is a leader in the asset management area.

Overall, it is worth more research in petroleum engineering by machine learning.

1.3 Motivation and scope of the work

As we mentioned in section 1.1, the switching time required in the modified hyperbolic model always depends on an evaluator's experience. In addition, in the modified hyperbolic model we assume that the decline during late BDF is exponential. As a result, based on a two-segment hyperbolic model, in Chapter 2 we propose an optimization algorithm with the objective of removing the two assumptions required by modified hyperbolic model. Next, in Chapter 3 a data-driven method is proposed for primary phase production forecasting and this method is based on the well-established statistical machine learning technique, functional principal component analysis (fPCA). Lastly, in Chapter 4, we develop a statistical machine learning method to investigate the association between oil production and key completion parameters while accounting for the confounding geological effect. The objective of this research is to use an efficient way to find optimal completion design that maximize well productivity. Chapter 3 and Chapter 4 are associated with machine learning.

CHAPTER II

ALGORITHM FOR DECLINE CURVE PARAMETER INFERENCE[1]

2.1 Overview

In this chapter, I first present important criteria that have been overlooked in existing empirical models but that are necessary for a robust and accurate production decline model. It turns out that the modified hyperbolic model does not violate these criteria. As a result, on the basis of a two-segment hyperbolic model, an optimization algorithm is proposed to estimate five important production parameters for a well that is in the boundary dominant flow (BDF) regime (1) initial production rate q_i, (2) initial decline rate D_i, (3) initial Arps b parameter b_i for early transient flow, (4) final Arps b parameter b_f for late BDF, and (5) optimal switching time t_c from early transient flow to late BDF. In addition, we developed a three-step diagnostic approach for the analysis of flow regime that a well has gone through. The proposed diagnostic approach is effective in reducing the uncertainty in well flow regime and the estimation of production parameters. The merits of this new algorithm are demonstrated with the application to analysis of real production data in Eagle Ford and Bakken reservoirs.

The methods proposed in this chapter are the cornerstone to predict well estimated ultimate recovery (EUR). In addition, this work has significant impact on many other

[1] The following URTeC paper, Criteria for Proper Production Decline Models and Algorithm for Decline Curve Parameter Inference, is reprinted with permission from the Unconventional Resources Technology Conference, whose permission is required for further use.

related projects such as the construction of type well production profiles, optimal completion design and probabilistic decline curve analysis since there all depend on the five production parameters.

2.2 Formulation and algorithm

In the traditional Arps hyperbolic decline model, two hyperbolic parameters are important:

$$a(t) = \frac{1}{D(t)} = -\frac{q(t)}{dq(t)/dt}; \qquad \text{(Loss-Ratio)} \qquad (2.1)$$

$$b(t) = \frac{da(t)}{dt} = \frac{d}{dt}\left[\frac{1}{D(t)}\right] = -\frac{d}{dt}\left[\frac{q(t)}{dq(t)/dt}\right]; \qquad \text{(Derivative of Loss Ratio)} \qquad (2.2)$$

where $q(t)$ denotes the oil/gas production rate at time t. Several important criteria are necessary for a robust and accurate production decline model. The first condition is that the decline rate $D(t)$ must be finite and positive. In addition, consistent with field experience, the Arps b parameter must be nonnegative. The last condition is that EUR must be finite at long time, which implies that the Arps b parameter must be less than 1 during the late BDF regime. From (2.1) and (2.2) we can write the production rate as

$$\ln \hat{q}(t) = \ln \hat{q}_i - D_i \int_0^t d\tau \frac{1}{1 + D_i \int_0^\tau b(s)ds} \qquad (2.3)$$

where $\hat{q}_i$ and D_i are initial production rate and decline rate, respectively; the Arps b parameter is a nonnegative function of time. The equation of decline rate is obtained from (2.3) as follows:

$$D(t) = \frac{D_i}{1+D_i \int_0^t b(s)ds} \tag{2.4}$$

From (2.4) we can see that the decline rate $D(t)$ is positive and finite provided that initial decline rate D_i is positive.

Now we see that the first two conditions are satisfied with the constraints on $b(t)$ and D_i. Next, suppose that we have a series of production data points $q(t_i)(i = 1,2,3, \cdots, N)$. Then our goal is to find a smoothing decline curve by the model $\ln q_\alpha = \ln \hat{q}_\alpha + \epsilon_\alpha$ where $\hat{q}_\alpha$ is the estimator of production rate at time t_α and the values ϵ_α are assumed independent and identically distributed error with mean 0 and variance σ^2. Then the objective optimization problem considered here is

$$F(\mathbf{q}|w) = N^{-1} \sum_\alpha \{\ln q_\alpha - \ln \hat{q}_i + D_i m(t_\alpha)\}^2 \tag{2.5}$$

where

$$m(t) = \int_0^t d\tau \frac{1}{1+D_i \int_0^\tau b(s)ds} \tag{2.6}$$

Eq. (2.5) represents the sum of square residual error and could be generalized to include variable weights for the observations. An estimate of σ^2 is

$$\hat{\sigma}^2 = (N-2)^{-1}N^{-1}\sum_{\alpha}\{\ln q_{\alpha} - \ln \hat{q}_i + D_i m(t_{\alpha})\}^2 \tag{2.7}$$

The confidence interval (CI) can be estimated using bootstrapping (Casella and Berger, 2002).

The transient Arps b parameter is hard to estimate since it is correlated with the second derivative of production rate $\hat{q}(t)$. A simplification is necessary for Arps b parameter. Here we consider the Arps b parameter is modeled by a two-segment piecewise constant function

$$b(t) = \begin{cases} b_i; 0 < t < t_c \\ b_f; \qquad t > t_c \end{cases} \tag{2.8}$$

where t_c denotes the switching time from early transient flow to late BDF flow. b_i and b_f are the characteristic b parameter for early transient flow and late BDF flow. The final Arps b parameter b_f must be less than 1 to make sure that EUR is finite at long time. The corresponding production rate $\hat{q}(t)$ of two-segment hyperbolic model is

$$\hat{q}(t) = \begin{cases} \frac{q_i}{(1+D_i b_i t)^{\frac{1}{b_i}}}; & t \leq t_c \\ \frac{q_i}{(1+D_i b_i t_c)^{\frac{1}{b_i}}} \left[\frac{1+D_i b_i t_c}{1+D_i b_i t_c - D_i b_f t_c + D_i b_f t}\right]^{\frac{1}{b_f}}; & t > t_c \end{cases} \tag{2.9}$$

Next, we will consider how to estimate the five important parameters q_i, D_i, b_i, b_f, t_c. On the basis of two-segment modified hyperbolic model, $m(t)$ in (2.6) can be written as

$$m(t) = \begin{cases} \frac{1}{D_i b_i} \ln(1 + D_i b_i t)\,; & t \leq t_c \\ \frac{1}{D_i b_i} \ln(1 + D_i b_i t_c) + \frac{1}{D_i b_f} \ln \frac{1+D_i(b_i - b_f)t_c + D_i b_f t}{1+D_i b_i t_c}; & t > t_c \end{cases} \tag{2.10}$$

Let $(b_D)_i \equiv D_i b_i$, $(b_D)_f \equiv D_i b_f$ and we can see that in Eq. (2.10) the variables b_i and b_f are both bounded with initial decline rate D_i. This observation motivates us to propose the following method for optimal solution of q_i, D_i, b_i, b_f, t_c.

Since the final Arps b parameter b_f must be bounded between 0 and 1 and it is sufficient that the estimation for b_f is numerically precise up to the order of 0.1, we can use the parameter sweeping method to find optimal b_f by searching exhaustively in the interval from 0 to 1 at step size of 0.05 (candidate array). The number of search required for optimal b_f is equal to the size of candidate array and in the ideal case the optimal b_f should be the element that yields the minimum least square error. The ideal case refers to the situation that the residual error is identical and independent distributed. However, in real cases this condition may be violated and as a result, we use a relatively weak criterion to determine b_f (step 12 in the algorithm below). For each element in the candidate array of b_f, we apply an iterated two-stage algorithm to find the optimal solution of q_i, D_i, b_i, t_c: first, begin with an initial nonzero estimate $b_i^{(0)}$, $t_c^{(0)}$ and estimate initial production rate $q_i^{(0)}$ and initial decline rate $D_i^{(0)}$ by linear regression. $m(t)$ is considered as a predictor for linear regression. However, since D_i is also involved in the equation for $m(t)$, the estimate for $D_i^{(0)}$ and $q_i^{(0)}$ requires a secondary iteration for a consistent convergent estimate of $D_i^{(0)}$ and $q_i^{(0)}$. Then on any main iteration $\nu > 0$ for which $q_i^{(\nu-1)}, D_i^{(\nu-1)}, b_i^{(\nu-1)}, t_c^{(\nu-1)}$ are estimates on the previous iteration, we update b_i and t_c by the L-BFGS-B method to obtain $b_i^{(\nu)}$ and $t_c^{(\nu)}$, and then compute $q_i^{(\nu)}$ and $D_i^{(\nu)}$ by linear regression. The algorithm pseudocode and secondary iteration detail are given below:

Decline Curve Analysis (DCA) Algorithm for Unconventional Reservoir

```
Input:   a series of production data points ($q(t_i), i = 1,2,\dots,N$), a weight vector $\boldsymbol{w}$,
         maximum number of iterations $\mathrm{MAX.MAIN.ITER}$ and error tolerance $\delta_1, \epsilon$
Output:  $q_i, D_i, b_i, b_f, t_c$
1.  Parameter sweeping: set up a candidate array $S$ of length $L$ for $b_f$ by equal space
    discretization of the interval [0,1].
2.  for $i = 1$ to $L$ do
3.      Start with $b_f = S[i]$ and an initial estimate of $b_i^{(0)}, t_c^{(0)}$
4.          Estimate initial production rate $q_i^{(0)}$ and initial decline rate $D_i^{(0)}$ by linear
            regression. An iteration (secondary iteration) is required here to find consistent
            convergent estimate of $D_i^{(0)}$ and $q_i^{(0)}$. Then compute the residual error $f^{(0)}$.
5.      for $\nu = 1$ to $\mathrm{MAX.MAIN.ITER} + 1$ do
6.          given $q_i^{(\nu-1)}, D_i^{(\nu-1)}, b_i^{(\nu-1)}, t_c^{(\nu-1)}$, update $b_i$ and $t_c$ by L-BFGS-B method to
            obtain $b_i^{(\nu)}$ and $t_c^{(\nu)}$.
```

```
7.          given $b_i^{(\nu)}$ and $t_c^{(\nu)}$, compute $q_i^{(\nu)}$ and $D_i^{(\nu)}$ by linear regression. An iteration
            (secondary iteration) is required here to find consistent convergent estimate of
            $D_i^{(\nu)}$ and $q_i^{(\nu)}$. Then compute the residual error $f^{(\nu)}$.
8.        if $\left(f^{(\nu-1)} - f^{(\nu)}\right)/f^{(\nu)} < \delta_1$ then
                break
          else
              if $\nu >$ MAX. MAIN. ITER then
                  throw exception and print out the message on non-convergence.
              end if
          end if
9.      end for
10.     $\text{res. cur} = f^{(\nu)}$
11.     if $i = 1$ then
            $q_i = q_i^{(\nu)}, D_i = D_i^{(\nu)}, b_i = b_i^{(\nu)}, b_f = b_f^{(\nu)}, t_c = t_c^{(\nu)}, \text{res. prev} = f^{(\nu)}$
        end if
12.     if $(\text{res. prev} - \text{res. cur}) > \epsilon$ then
            update $q_i, D_i, b_i, b_f, t_c$ and res. prev
        end if
13.   end for
```

Secondary Iteration for Step 5 and 8 in DCA algorithm

```
1. Begin with an initial estimate of $D_i^{(p)}$ ("$p$" stands for old value), maximum number of
   iterations MAX. SECONDARY. ITER and error tolerance $\delta_2$
2. Compute $m(t)$ at different time $(t_1, t_2, \dots, t_N)$ using $D_i^{(p)}$
3. Compute least square solution of $(\ln q_i)^{(c)}$ and $D_i^{(c)}$ from linear regression ("$c$" stands
   for current value)
4. if $\left|D_i^{(p)} - D_i^{(c)}\right| / D_i^{(c)} < \delta_2$ then
       return $D_i^{(c)}$ and $(\ln q_i)^{(c)}$
   else
       $D_i^{(p)} = D_i^{(c)}$
   end if
5. Repeat steps from 2 to 4 until the number of iterations is equal to
   MAX. SECONDARY. ITER. If $D_i$ fails to converge after MAX. SECONDARY. ITER
   iterations, stop and print out the message on non-convergence.
```

In the first part of the algorithm above, the criterion used in step 12, which needs to be consistent with the tolerance error in step 8, is necessary to prevent erroneous estimation of b_f due to some unknown interference in the production history that distorts the production decline pattern. In addition, post-analysis of production parameter results is important to make sure the resulting predictive model is correct and reliable for EUR determination. Given the amount of well production history data and the number of parameters for estimation, sweeping from 0 to 1 for the optimal b_f is not a computational issue. This sweeping method is similar to type curve matching where we compare the well production history data with multiple dimensionless type curves and find a type curve that fits the production data best. For each element in the candidate array of b_f, the algorithm converges very fast. The most difficult step in the algorithm is initial estimate of t_c since updating b_i and t_c is a non-convex problem and the final result of b_i and t_c depends on the initial starting point of t_c. Therefore, it is important to conduct a preliminary analysis of production data using log-log diagnostic plot to estimate the possible value of t_c. With this step we will know whether the well is in the boundary dominated flow regime. If boundary dominated flow is hard to identify from diagnostic plot or the production data is too noisy, we recommend fitting the data using the hyperbolic model with a single Arps b parameter. We can make simple modifications to the DCA algorithm so that it is applicable for estimating parameters for the hyperbolic model with a single b parameter. If the hyperbolic model has only one b parameter, using the DCA algorithm we can find the parameters that minimize eq. (5) globally. If production data is too noisy, the assumption on the residual error is invalid and we should be cautious with the model from least square regression. The implementation of the DCA algorithm is easy since many robust free package are available online for the step of updating b_i and t_c with the L-BFGS-B method. If we assume the final Arps b parameter b_f is a fixed value ($1/3$ for oil well or 0.4 for gas well according to Fetkovich and Spivey *et al.*), then we will have only four parameters to estimate and the computation time will be reduced.

2.3 Flow regime diagnosis

As we mentioned earlier, the preliminary analysis of well production data is important to make sure that the production predictive model is correct and reliable for EUR estimation. The chief goal of this step is to determine the flow regime well production has gone through using a log-log diagnostic plot. There are two different diagnostic plots frequently used for diagnostic analysis. One is the plot based on production rate $q(t)$ and production time t. The other one is plotted with respect to material balance time (MBT), $t_{MBT}(t)$, instead of real production time. If we use a two-segment hyperbolic model for shale well production decline, the slope k_{PT} of $\ln q(t)$ vs. $\ln t$ is

$$k_{PT} = \left|\frac{d\ln q(t)}{d\ln t}\right| = \begin{cases} \frac{D_i t}{1+D_i b_i t}; t \leq t_c \\ \frac{D_{t_c} t}{1+D_{t_c} b_f (t-t_c)}; t > t_c \end{cases} \tag{2.11}$$

where $D_{t_c} = \frac{D_i}{1+D_i b_i t_c}$ denotes the decline rate at the switching time t_c. The slope k_{MBT} of $\ln q(t)$ vs. $\ln t_{MBT}(t)$ (appendix A) is

$$k_{MBT} = \left|\frac{d \ln q(t)}{d \ln t_{MBT}(t)}\right| =$$

$$\begin{cases} \dfrac{1-(1+D_i b_i t)^{\frac{1}{b_i}-1}}{b_i-(1+D_i b_i t)^{\frac{1}{b_i}-1}} \ (b_i > 1); & t \le t_c \\ \dfrac{\frac{b_f-1}{b_i-1}\left[1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}\right]-1+ \quad \left[1+D_{t_c} b_f (t-t_c)\right]^{-\frac{1}{b_f}+1}}{\frac{b_f-1}{b_i-1}\left[1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}\right]-1+b_f\left[1+D_{t_c} b_f (t-t_c)\right]^{-\frac{1}{b_f}+1}} \ \left(0 < b_f < 1\right); & t > t_c \end{cases} \tag{2.12}$$

The equation of k_{MBT} in the limit of $b_i = 1$ or $b_f = 0$ can be derived using three equations as shown below:

$$\lim_{b_i \to 1} \frac{1-(1+D_i b_i t)^{\frac{1}{b_i}-1}}{b_i-(1+D_i b_i t)^{\frac{1}{b_i}-1}} = \frac{\ln(1+D_i t)}{1+\ln(1+D_i t)} \tag{2.13.1}$$

$$\lim_{b_i \to 1} \frac{1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}}{b_i-1} = \ln(1+D_i t) \tag{2.13.2}$$

$$\lim_{b_f \to 0} \left[1 + D_{t_c} b_f (t - t_c)\right]^{-\frac{1}{b_f}+1} = \exp[-D_{t_c}(t - t_c)] \tag{2.13.3}$$

Eqs. (2.11) and (2.12) are both monotonically increasing functions with respect to production time t. If $t \le t_c$, we have

$$k_{PT} \le \frac{1}{\frac{1}{D_i t_c}+b_i}; k_{MBT} \le \frac{1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}}{b_i-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}} \tag{2.14}$$

In addition,

$$k_{PT} \rightarrow \frac{1}{\frac{1}{D_i t_c}+b_i};\, k_{MBT} \rightarrow \frac{1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}}{b_i-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}}, \text{as } t \rightarrow t_c \tag{2.15}$$

If the switching time t_c has a very large value ($D_i t_c \gg 1$), $k_{PT} \rightarrow \frac{1}{b_i}$ and $k_{MBT} \rightarrow \frac{1}{b_i}$ as $t \rightarrow t_c$. This implies that when the value of t_c is large enough, in the vicinity of t_c the rate curve based on production time (PT curve) is almost parallel to the curve based on material balance time (MBT curve). If $t > t_c$, we have $k_{PT} \rightarrow \frac{1}{b_f} > 1$ and $k_{MBT} \rightarrow 1$ as $t \rightarrow \infty$. Since $k_{PT} \geq k_{MBT}$ at any time t (appendix B), the following method is proposed to determine the flow regime more accurately.

(1) Plot the PT curve and MBT curve on the same log-log plot.

(2) Draw a line L tangent to the end of MBT curve. If the slope of line L is very close to unit slope, then the well should be in BDF regime. However, in most cases the slope of line L is less than 1 and we need step 3 to determine whether the well is in BDF regime or not.

(3) Move line L onto the PT curve at the point P where L is approximately tangent to PT curve. Since $k_{PT} \geq k_{MBT}$ at any time t, we expect that, in the interval between point P and the end of well production, there should be some data points the tangent slope of which is greater than the slope of line L. If we can draw another line for the data

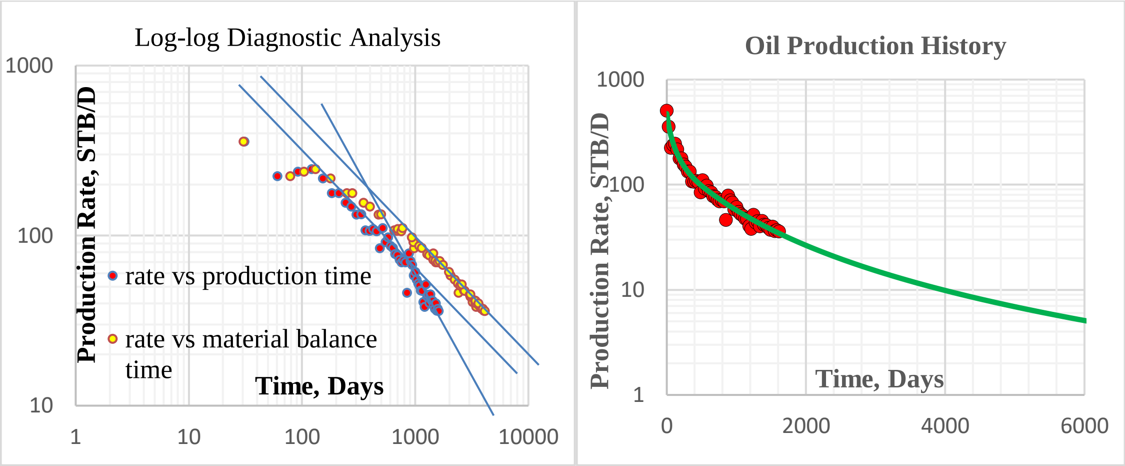
after point P and the slope of second line is greater than or equal to 1, we have high confidence that the well is in BDF regime. If there are very few data points after point P and it is hard to draw a line with slope greater than that of line L, then there is a high chance that the well is still in the transient flow regime. In addition, if we denote the slope of line L as k, the initial Arps b parameter b_i must be less than or equal to $\frac{1}{k}$.

As we will see in the examples shown later, the above three steps will be utilized in the preliminary diagnostic analysis of well production data. Using these three steps makes it easy for us to identify flow regime and effectively reduce the uncertainty on the results of production parameters.

2.4 Method validation

We applied the proposed algorithm to analyze eight wells from Bakken and Eagle Ford – 4 oil wells and 4 gas wells. Data cleaning is required to remove the initial “ramp-up” part in production history. From the results of the following eight wells, we can see the power of the new diagnostic approach and new optimization algorithm. If the well production has switched from transient flow regime to BDF regime, we can determine the optimal switching point. We can also determine the decline rate at the switching point, which is an advantage over the traditional modified hyperbolic model that requires the evaluator to provide a minimum decline rate before knowing the switching time point.

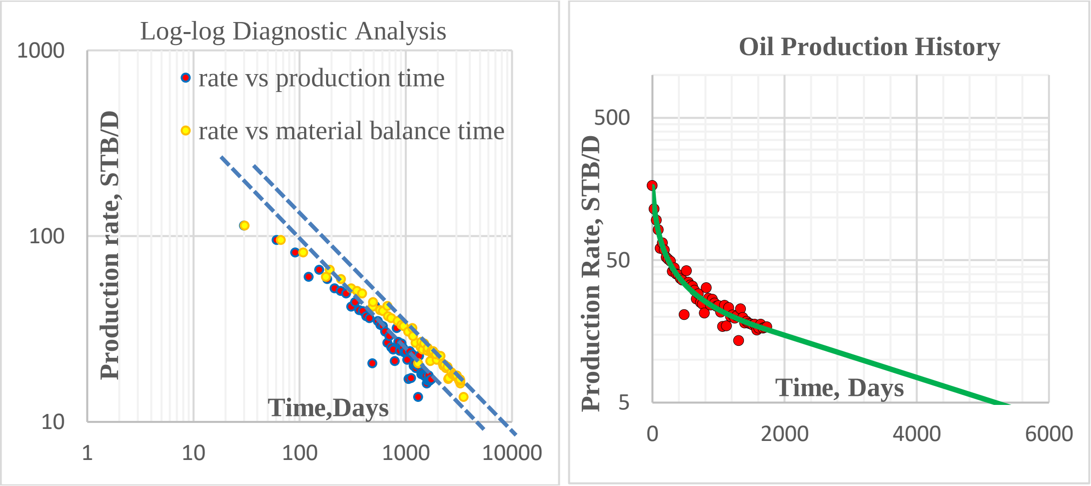
Example #1: a shale oil well located in Billings County, North Dakota, producing from Bakken reservoir - API number $33 - 007 - 01707$. From the log-log diagnostic plot, if we draw two parallel straight lines, one tangent to the data at the end of material balance time (yellow points), we can see the data at the end of production time (red points) lies in a line with slope greater than 1, indicating that the well should be in the BDF regime. The estimation results of production parameters are given in **Figure** 2 below. From the results of production parameters, the estimated switching time is at 538 days.



Production Parameters		
Initial production rate q_i	474	STB/D
Initial decline rate D_i	0.0115	1/D
Initial Arps b parameter b_i	1.39	
Final Arps b parameter b_f	0.5	
Decline rate at switching D_{sw}	0.0012	1/D
Switching time t_c	538	Days

Figure 2: Well #1 decline curve analysis result (Zhou *et al.* 2018)

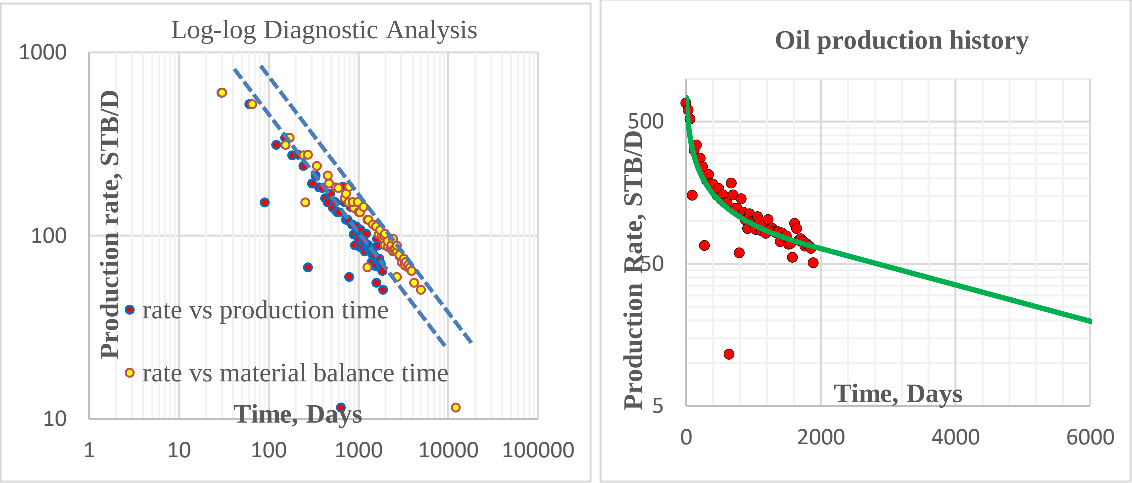
Example #2: a shale oil well located in Billings County, North Dakota, producing from Bakken reservoir - API number $33 - 007 - 01656$. From the diagnostic plot, we have high confidence that well #2 is in the transient flow regime. The analysis result is given in **Figure** 3 below. The estimation results of production parameters are consistent with diagnostic plot. In order to predict EUR, we need to know switching time t_c from transient flow to BDF and the final Arps b parameter b_f. If we can find an analogous well which is already in the BDF regime, located in similar geological environment and has similar completion parameters, we may use the values of t_c and b_f from the analogous well to predict EUR of well #2.



Production Parameters		
Initial production rate q_i	164	STB/D
Initial decline rate D_i	0.0155	1/D
Initial Arps b parameter b_i	1.65	
Final Arps b parameter b_f	NA	
Decline rate at switching D_{sw}	NA	1/D
Switching time t_c	NA	Days

Figure 3: Well #2 decline curve analysis result (Zhou *et al.* 2018)

Example #3: a shale oil well located in McKenzie County, North Dakota, producing from Bakken reservoir - API number $33 - 053 - 03247$. Similar to well #2, from the diagnostic plot we have high confidence that well #3 is still in the transient flow regime. The analysis result is given in **Figure** 4 below and it is consistent with the analysis from diagnostic plot.

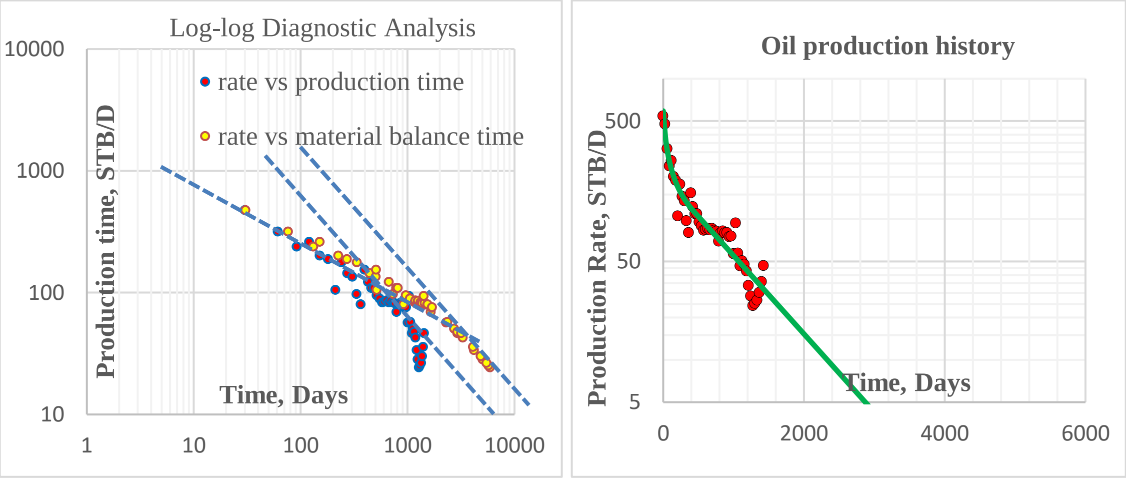


Production Parameters		
Initial production rate q_i	738	STB/D
Initial decline rate D_i	0.0209	1/D
Initial Arps b parameter b_i	1.76	
Final Arps b parameter b_f	NA	
Decline rate at switching D_{sw}	NA	1/D
Switching time t_c	NA	Days

Figure 4: Well #3 decline curve analysis result (Zhou *et al.* 2018)

Example #4: a shale oil well located in McKenzie County, North Dakota, producing from Bakken reservoir - API number $33 - 053 - 03318$. The analysis result is given in **Figure**

5 below. According to the results of production parameters, we can see that well #4 follows exponential decline after 436 days from the first production day.

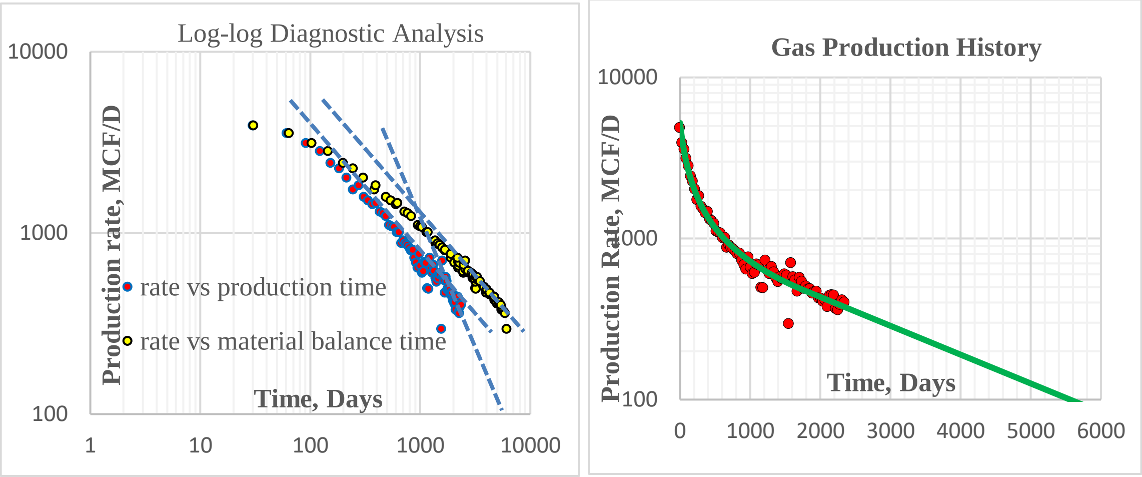


Production Parameters		
Initial production rate q_i	589	STB/D
Initial decline rate D_i	0.0202	1/D
Initial Arps b parameter b_i	1.68	
Final Arps b parameter b_f	0	
Decline rate at switching D_{sw}	0.00128	1/D
Switching time t_c	436	Days

Figure 5: Well #4 decline curve analysis result (Zhou *et al.* 2018)

Example 5#: a shale gas well in the La Salle County, Texas, producing from Eagle Ford - API number $42 - 283 - 32348$. From the diagnostic plot, we can see that well #5 has reached the BDF regime since the production data at the end of production time (red points) lies on a line with slope greater than 1. The production parameter result is given in

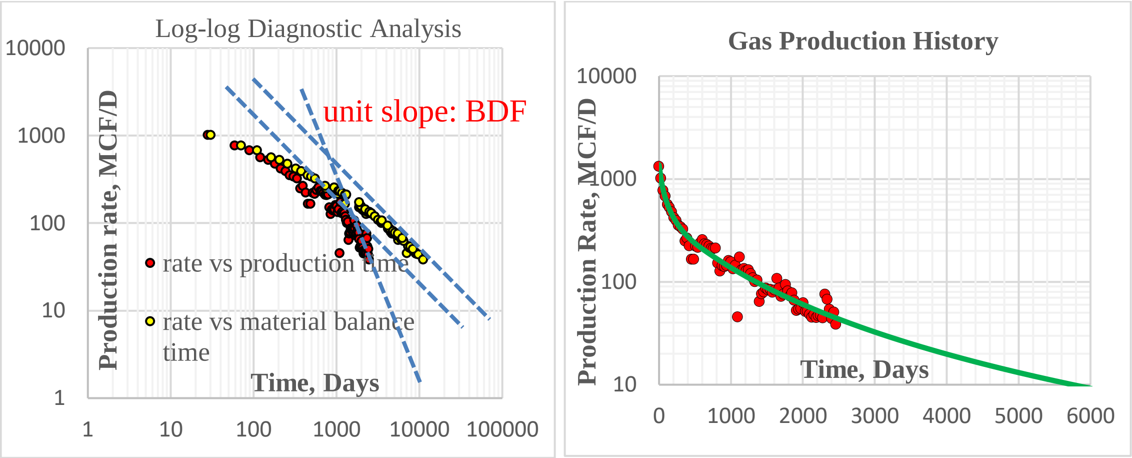
Figure 6 below. The production of well #5 switched to exponential decline after 1797 days from the first production day and it is consistent with our analysis from diagnostic plot.



Production Parameters		
Initial production rate q_i	5218	MCF/D
Initial decline rate D_i	0.00917	1/D
Initial Arps b parameter b_i	1.29	
Final Arps b parameter b_f	0	
Decline rate at switching D_{sw}	0.00041	1/D
Switching time t_c	1797	Days

Figure 6: Well #5 decline curve analysis result (Zhou *et al.* 2018)

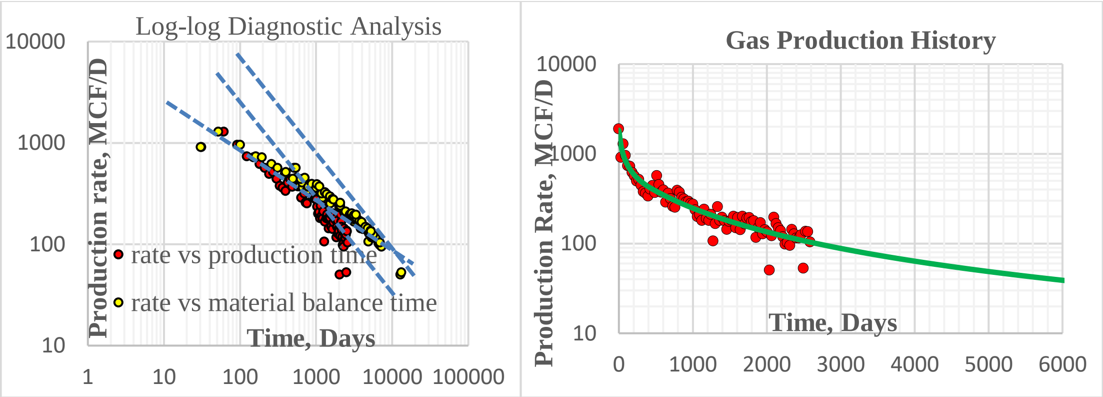
Example #6: a shale gas well in the county McMullen, Texas, producing from Eagle Ford - API number $42 - 311 - 34139$. From the diagnostic plot, we can see unit slope at the end of material balance time (yellow points) indicating BDF. If we draw a unit slope line that is tangent to production data (red points), we can see production data at the end of production time lies on a line with slope greater than 1. The production parameter results are given in **Figure** 7 below. The production of well #6 switched to BDF after 579 days from the first production day. The optimal final Arps b parameter b_f is 0.4 which is consistent with our observation in diagnostic plot.



Production Parameters		
Initial production rate q_i	1373	MCF/D
Initial decline rate D_i	0.0147	1/D
Initial Arps b parameter b_i	1.37	
Final Arps b parameter b_f	0.4	
Decline rate at switching D_{sw}	0.0012	1/D
Switching time t_c	579	Days

Figure 7: Well #6 decline curve analysis result (Zhou *et al.* 2018)

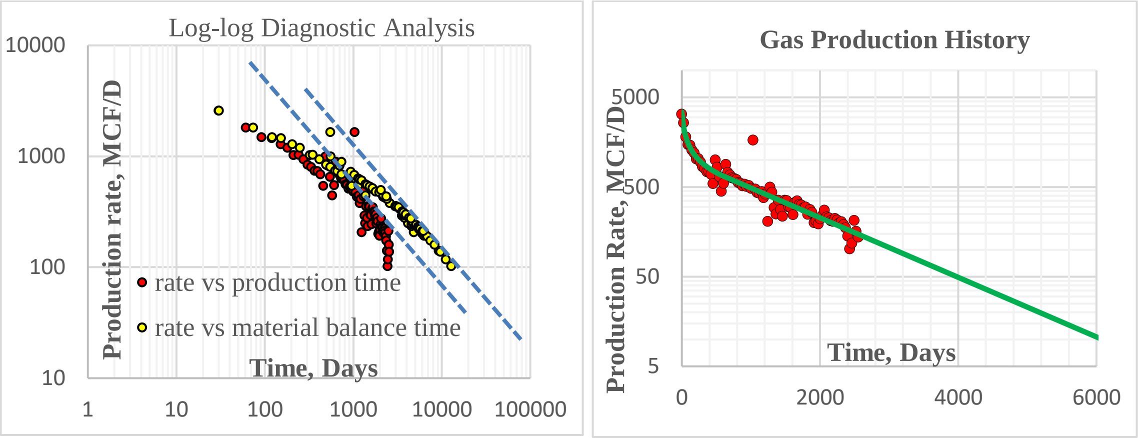
Example #7: a shale gas well in McMullen County, Texas, producing from Eagle Ford - API number $42-311-34131$. From the diagnostic plot, we can see unit slope at the end of material balance time (yellow points) indicating BDF. The production parameter results are given in **Figure** 8 below. The production of well #7 switched to BDF after 449 days from the first production day. In addition, the value of initial Arps *b* parameter is 1.97 indicating linear transient flow and the optimal final Arps *b* parameter for BDF is 0.7.



Production Parameters		
Initial production rate q_i	1832	MCF/D
Initial decline rate D_i	0.0210	1/D
Initial Arps b parameter b_i	1.97	
Final Arps b parameter b_f	0.7	
Decline rate at switching D_{sw}	0.00107	1/D
Switching time t_c	449	Days

Figure 8: Well #7 decline curve analysis result (Zhou *et al.* 2018)

Example #8: a shale gas well in McMullen County, Texas, producing from Eagle Ford - API number $42 - 311 - 34142$. From the diagnostic plot, we can see unit slope at the end of material balance time (yellow points) indicating BDF. The production parameter results are given in **Figure** 9 below. Well #8 starts with linear transient flow (initial Arps b parameter is 2.03) and then switched to exponential decline after 623 days from the first production day.



Production Parameters		
Initial production rate q_i	3422	MCF/D
Initial decline rate D_i	0.0221	1/D
Initial Arps b parameter b_i	2.03	
Final Arps b parameter b_f	0.0	
Decline rate at switching D_{sw}	0.00076	1/D
Switching time t_c	623	Days

Figure 9: Well #8 decline curve analysis result (Zhou *et al.* 2018)

2.5 Discussion

In the modified hyperbolic model, traditionally engineers specify a minimum annual decline rate D_{min} (e.g., 5%) as the criterion for the switching from transient flow to BDF.

Now let us review Example 1 in the previous section to see what problem we may encounter if we use minimum annual decline rate as the standard for the change of flow regime. If the minimum annual decline rate D_{min} is set to 5%, then the corresponding minimum daily decline rate will be approximately 1.37×10^{-4} $1/\mathrm{D}$. This means only when the daily decline rate is less than 1.37×10^{-4} $1/\mathrm{D}$, will the production decline reach exponential decline. According to our diagnostic plot shown in Example 1, we know that at 1000 days the well is already in the boundary dominated flow regime. Now we need to know at $t = 1000$ days if it is possible that the daily decline rate is below 1.37×10^{-4} $1/\mathrm{D}$. Before the switching time, the equation of decline rate is written as follows:

$$D(t) = \frac{1}{\frac{1}{D_i} + b_i t} \; ; t \leq t_c \tag{2.16}$$

According to our analysis we know the initial decline rate should be of the order of 10^{-2} and the value of initial Arps b parameter b_i should be less than 2, then we can see that at $t = 1000$ days, $D(t)$ will be at least $\frac{1}{2000}$ $1/\mathrm{D}$ (here we neglect $\frac{1}{D_i}$ since $\frac{1}{D_i} \ll b_i t$ at $t = 1000$ days). Therefore, $D(t)$ at $t = 1000$ days will be at least 5×10^{-4} $1/\mathrm{D}$ which is approximately 3.6 times of the value 1.37×10^{-4} $1/\mathrm{D}$. As a result, if we use D_{min} as the standard for the switching from hyperbolic decline to exponential decline, we will have the result that the well in Example 1 is still in the transient flow regime, which is not true according to our diagnostic analysis. Therefore, from this analysis we can see the drawback of using D_{min} for modified hyperbolic model.

CHAPTER III
APPLICATION OF STATISTICAL METHODS TO PREDICT PRODUCTION FROM LIQUID-RICH SHALE RESERVOIR [2]

3.1 Overview

In this chapter, we propose a data-driven method for primary phase production forecasting. This method is based on a well-established statistical machine learning technique, functional principal component analysis (fPCA) – an extension to principal component analysis (PCA) (Ramsay and Silverman, 2005). The PCA method is a well-established approach for data analysis in statistics and has been used for production data analysis by some researchers (Makinde and Lee, 2016; Bhattacharya and Nikolaou, 2013). PCA has several limitations in production analysis. One limitation is that from PCA we cannot determine EUR of a well and the prediction can be extended only to the maximum production time used in the training set. In this work, we propose to use fPCA for production data analysis and forecasting given the fact that production data collection from a well are a functional time series with strong temporal correlation. For the sake of convenience, we call this approach the fPCA method. Compared to the analytical approach and reservoir numerical simulation, the fPCA method requires neither an analytical production model nor a high-resolution geological model necessary for numerical simulation. The prediction is made based on the study of production histories from many different low permeability wells. This method is efficient in that the production analysis

[2] The following URTeC paper, Application of Statistical Methods to Predict Production From Liquid-Rich Shale Reservoirs, is reprinted with permission from the Unconventional Resources Technology Conference, whose permission is required for further use.

for a large number of wells can be conducted at the same time. Also, this method is very easy to implement with the use of the free R-package “fda” (Ramsay *et al*., 2009).

3.2 Data description and preprocessing

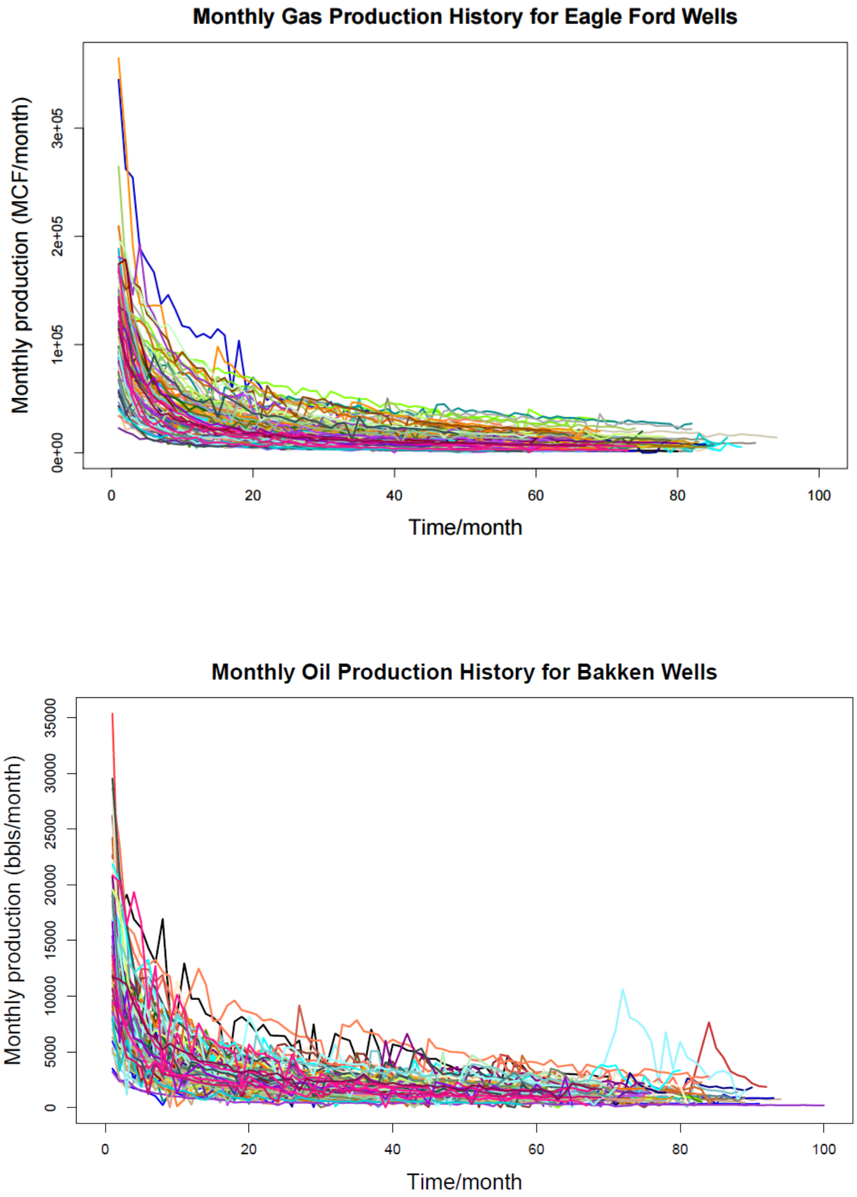


Figure 10: Unconventional well production data (a) Eagle Ford Reservoir (b) Bakken Reservoir (Zhou *et al.* 2017)

In this study, we selected 100 gas wells from the Eagle Ford and 100 oil wells from the Bakken. Both data sets are for wells that undergo multiphase flow eventually. These kinds of wells are more difficult to forecast that wells with single phase flow. All the wells are currently still active with production histories no longer than 9 years. The data available for analysis was monthly production with no other geological/well completion information available. In the preprocessing step, we discarded the initial “ramp-up” production data and focused on data after decline began. In addition, we ignored the exact physical time associated with the production data and aligned the peak production data in a line with the same time coordinate. The data resulting after these two pre-processing steps are shown in **Figure** 10**.**

In **Figure** 10, we see that the production data are still quite noisy, so one more preprocessing step was required to clean the data prior to the analysis. Since the cumulative production-time plot is much smoother than the rate-time plot, we used (and recommend) the cumulative production data with the aid of the Bourdet derivative algorithm to calculate monthly production rate. **Figure** 11 shows the result of the Bourdet production rate (red dots) and we see that it has much less noise than the original average production rate (purple dots).

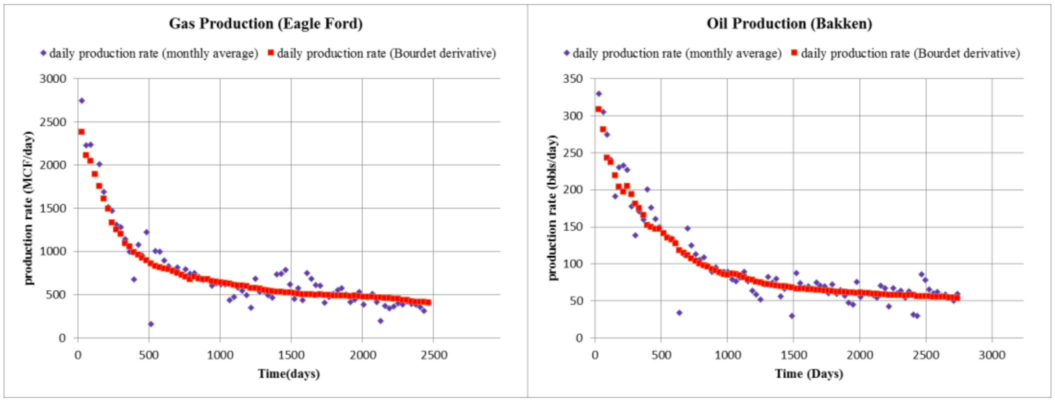


Figure 11: Monthly production rate (average vs. Bourdet derivative) (Zhou *et al.* 2017)

In addition, experience tells us that we be much more confident in our predictions if boundary-dominated flow (BDF) can be identified in the production data. As a result, log-log diagnostic plot analysis was conducted to determine how many wells in our data sets reached the BDF regime. We found that 14 of the 100 oil wells in the Bakken formation reached the BDF regime. In the Eagle Ford formation, 47 out of 100 gas wells reached BDF. **Figure** 12 provides examples of flow regime diagnostic plots.

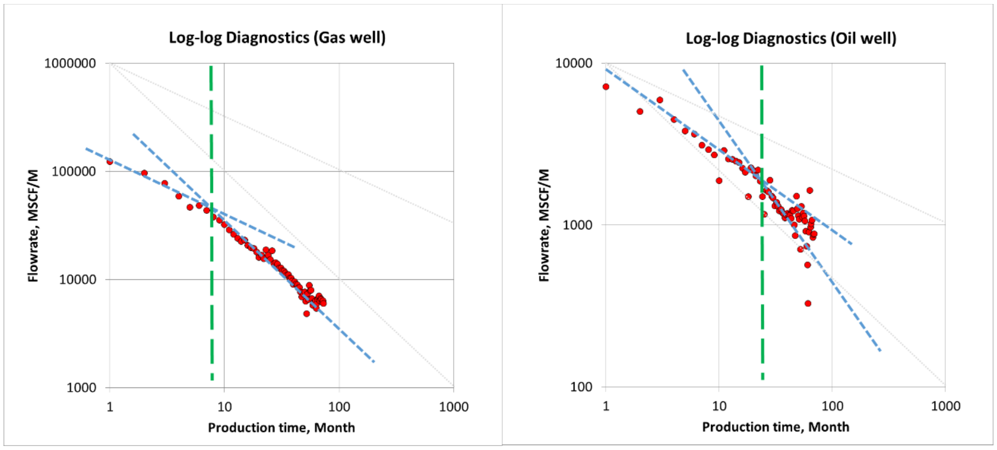


Figure 12: Log-log plot diagnostic analysis (Zhou *et al.* 2017)

3.3 Methodology

Production prediction starts with a study of the shale well production features. As mentioned above, production data from ultra-low permeability reservoirs has quite different features than data from conventional resources. In order to identify those features, we use the functional principal component analysis (fPCA) technique which accounts for the strong temporal correlation within a series of production data when learning decline curves. Specifically, we collect the production data from N wells at a set of locations S . These wells have long production histories and hence are used as training data sets. For each well, the production data after pre-processing will be scaled by dividing the production from each well by its initial maximum production rate. Next, the processed production data of each well are treated as observations of a functional time series and modeled by a smooth function by basis function representations. Here we used B-spline functions as the basis functions for interpolation since the production function is a non-periodic function. On the basis of a set of scaled production curves $q_i(t)(i = 1,2, \ldots, N)$, we can find a set of K principal component (PC) functions $\xi_j(t)(j = 1,2,3, \ldots, K)$ which satisfy the following equation

$$\int \nu(s,t)\xi(s)ds = \rho\xi(t) \qquad (3.1)$$

where ρ is some appropriate eigenvalue associated with the covariance function $\nu(s,t)$, defined as

$$\nu(s,t) = N^{-1}\sum_{i=1}^{N}[q_i(s) - \bar{q}(s)][q_i(t) - \bar{q}(t)] \tag{3.2}$$

In the covariance function $\nu(s,t)$, $\bar{q}(t)$ represents the scaled production mean value at time t. To obtain the PC functions, we can use the free R-package “fda”. The PC functions ξ_j are orthonormal to each other and can be treated as basis functions so that the expansion of each production curve in the training set in terms of these basis functions approximates the true production curve as closely as possible. The number of PC functions needed for expansion depends on the magnitude of the eigenvalue ρ. We retain only the PC functions that capture the most significant variation in the production function. Once the PC functions ξ_j are extracted from the training production curves, we use them as the interpolation basis for the production data of other testing wells since these PC functions are the characteristic functions of the well production in location S. Mathematically, the functional form of the testing well production is

$$\hat{q}_{test}(t) = \bar{q}(t) + \sum_{l=1}^{K} f_l \xi_l(t) \tag{3.3}$$

where $\hat{q}_{test}$ is a functional production estimator and f_l is the weighting coefficient which can be determined by the simple multiple regression. Assuming that a test well has p observed production data points, we determine the coefficients $f_l (l = 1,2, \dots, K)$ by minimizing the sum of squared errors defined as follows:

$$SSE = \sum_{j=1}^{p} \left[q_{test}^{obs}(t_j) - \hat{q}_{test}(t_j) \right]^2 \tag{3.4}$$

where $q_{test}^{obs}(t_j)$ denotes the observed production data at time t_j after normalization. Since the number of production data points p is always much larger than the number of PC expansion basis functions K, this minimization problem always has a unique solution for f_l. After determining the coefficients f_l, we can use the resulting model $\hat{q}_{test}(t)$ for production prediction. **Figure** 13 below summarizes the entire procedure for production forecasting.

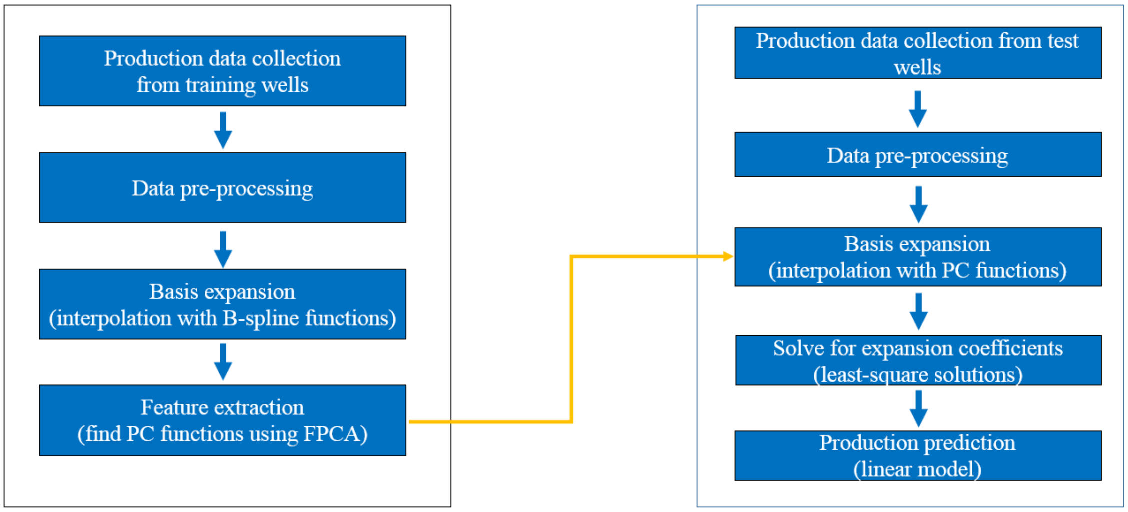

Production data collection from training wells
Data pre-processing
Basis expansion (interpolation with B-spline functions)
Feature extraction (find PC functions using FPCA)
Production data collection from test wells
Data pre-processing
Basis expansion (interpolation with PC functions)
Solve for expansion coefficients (least-square solutions)
Production prediction (linear model)

Figure 13: Workflow of functional principal component analysis for well production data (Zhou *et al.* 2017)

3.4 Results

3.4.1 Method validation

In order to assess the performance of fPCA in production prediction, the 100 wells in Eagle Ford/Bakken were both divided into two groups. 50 wells were used for training and the other 50 wells were test wells. The training wells have longer production histories. From the training well data, we computed the first ten leading eigenvalues and the corresponding PC functions. In this way we could decide how many PC functions should be kept for the linear production model. The eigenvalue results are shown in **Figure** 14. In **Figure** 14 we see in both cases the first eigenvalue is dominant compared to other eigenvalues and the eigenvalues decline very rapidly. In addition, **Figure** 15 shows the first three principal component functions by displaying the mean curve along with +'s and –'s indicating the consequences of adding and subtracting a small amount of each principal component. This occurs because a principal component represents variation around the mean, and therefore is naturally plotted as such. In the Eagle Ford we see that the first three PC functions account for 99.2% of the variation while in Bakken the first PC function accounts for 97.9% of the variation. Therefore, it is accurate enough to retain only the first three PC functions as the expansion basis for the linear production model.

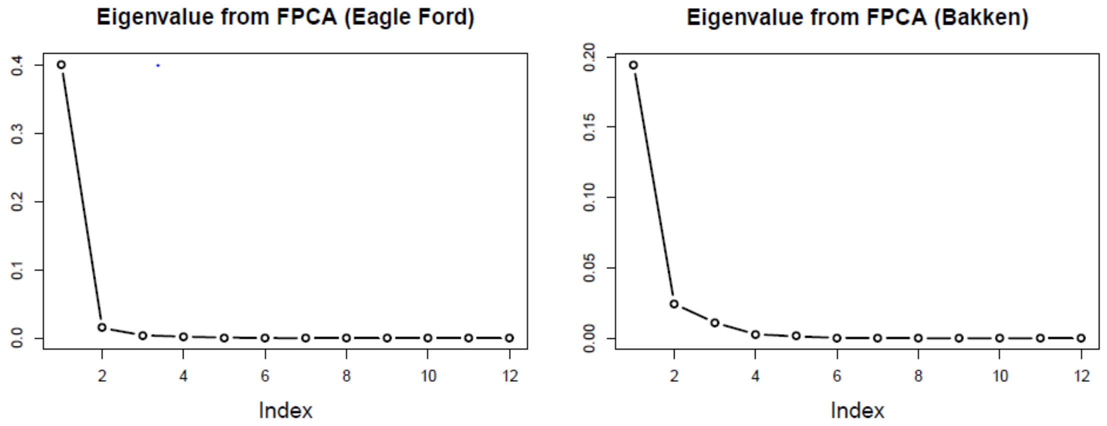
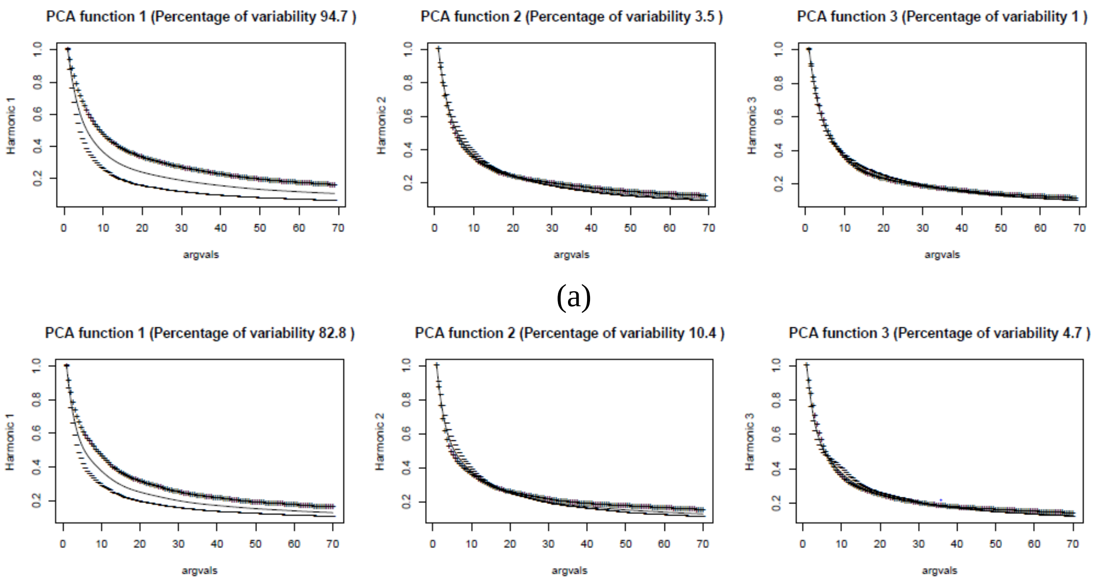


Figure 14: Eigenvalues from fPCA (Zhou *et al.* 2017)



(a)

(b)

Figure 15: Three principal component functions (a) Eagle Ford; (b) Bakken (Zhou *et al.* 2017)

The test wells were used for validation. For each test well, we constructed a linear production model with a portion of the observed production data as the response input.

With the resulting production function, we predicted the production rate at late times and compared the results with the observed rates to check the prediction quality. In this experiment, we predicted the final two years of production for test wells; **Figure** 16 (Eagle Ford) and **Figure** 17 (Bakken) show the results. In **Figures** 16 **and** 17, the red line on the right side of the blue bar is the predicted result and the black data points on the left side of the bar are input data. These figures show that the predictions obtained with the fPCA method are reasonably good.

Eagle Ford

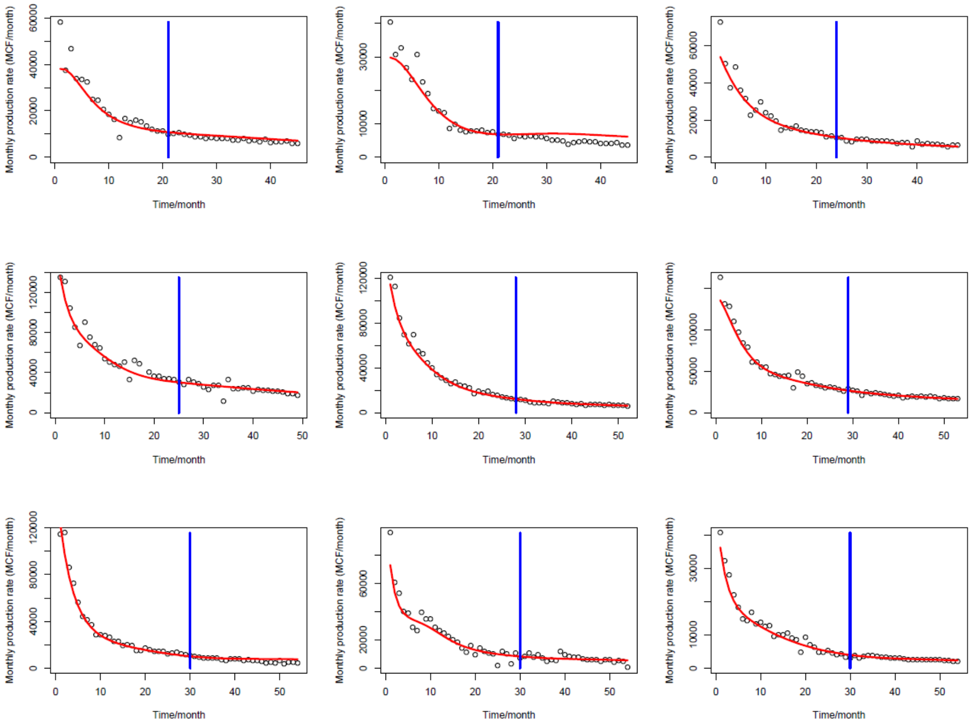


Figure 16: Test well production prediction verification (Eagle Ford: monthly rate vs. time) (Zhou *et al.* 2017)

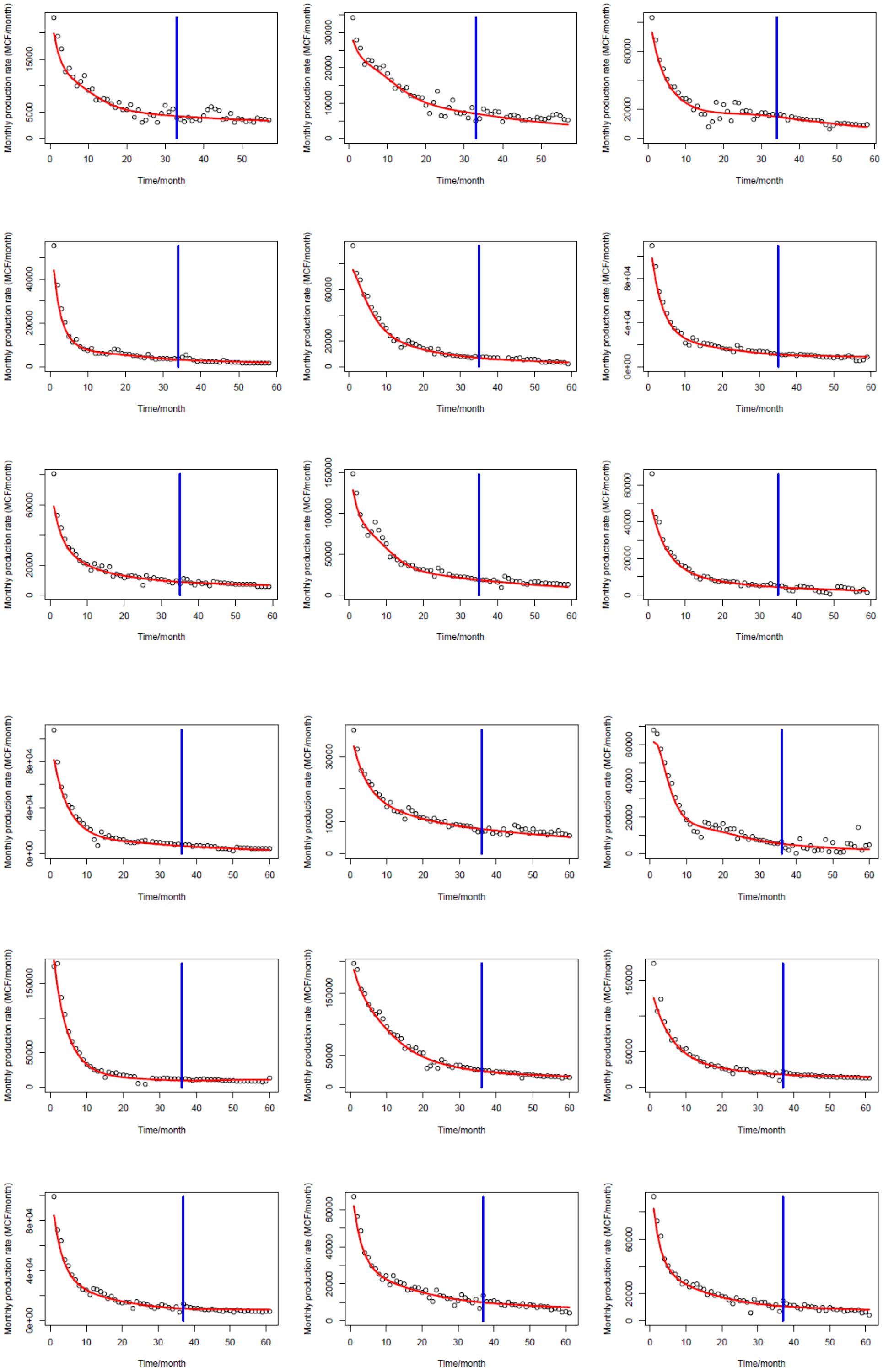


Figure 16: Continued

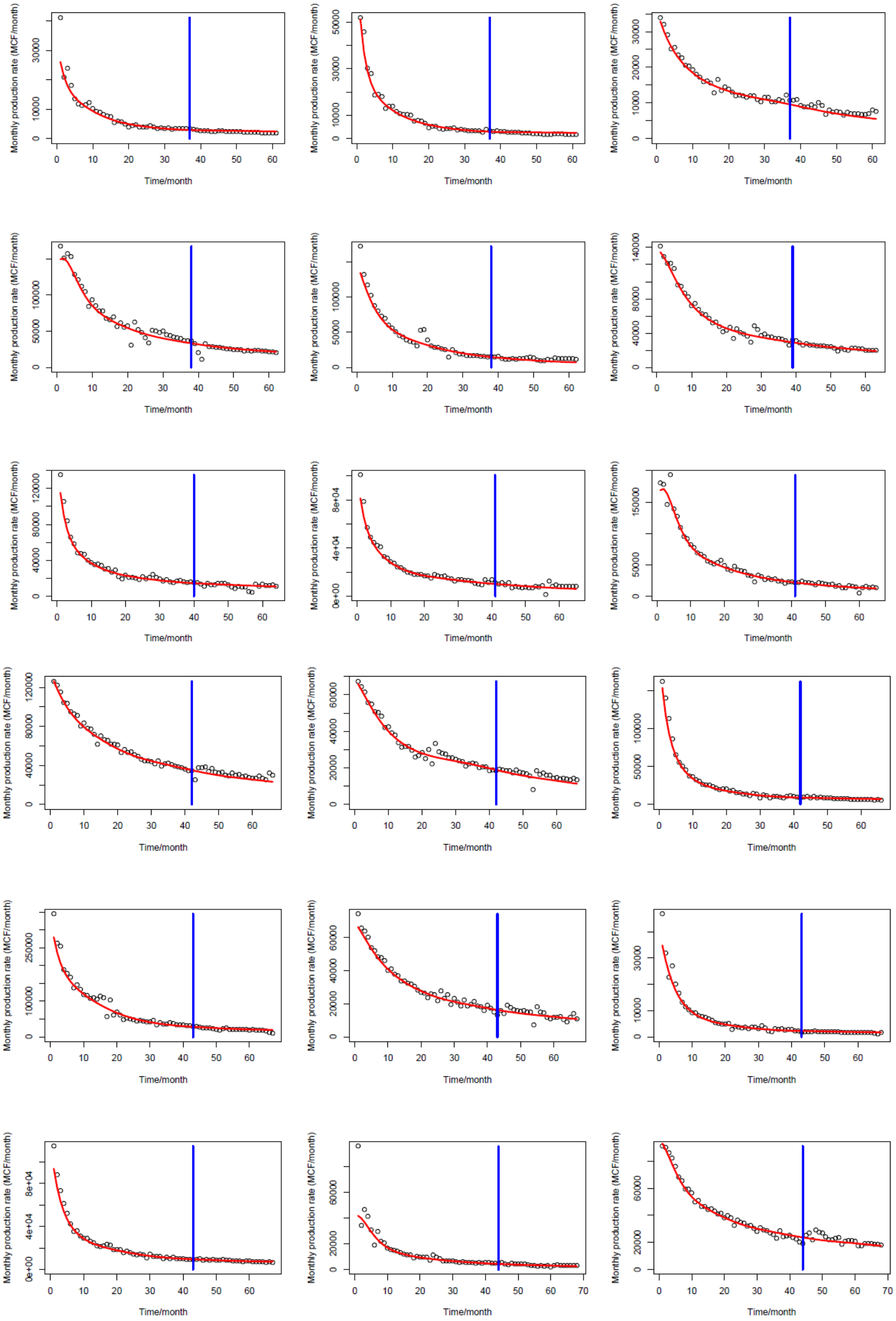


Figure 16: Continued

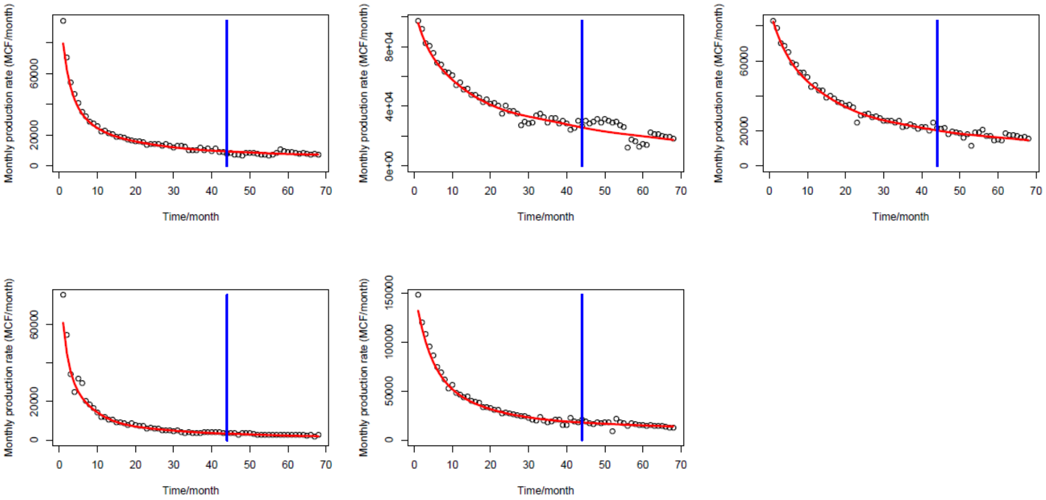


Figure 16: Continued

Bakken

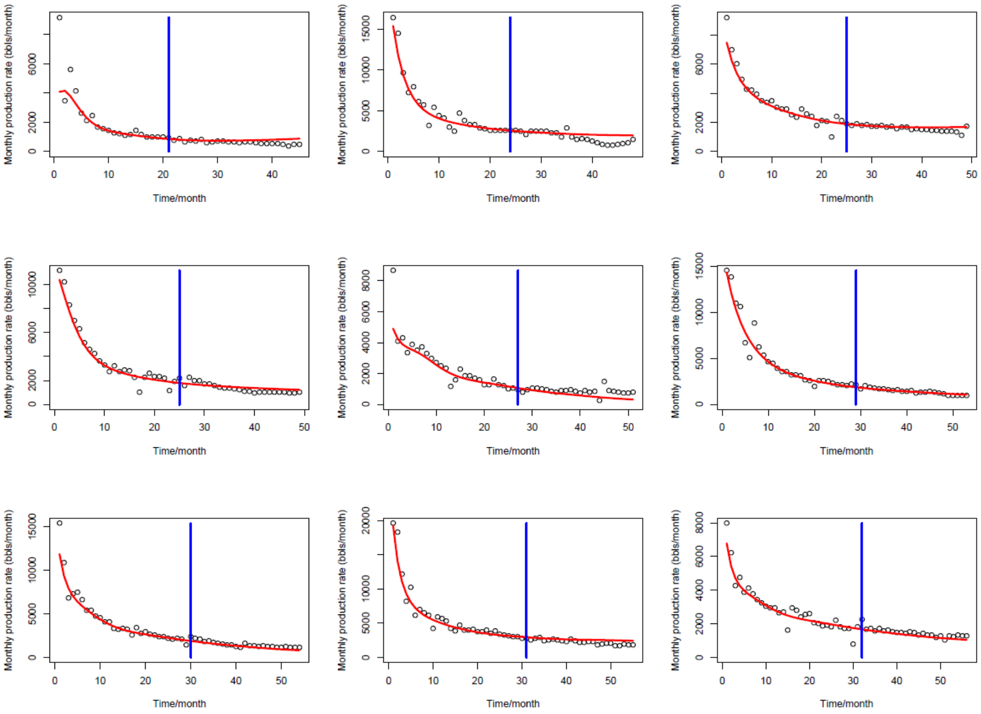


Figure 17: Test well production prediction verification (Bakken: monthly rate vs. time) (Zhou *et al.* 2017)

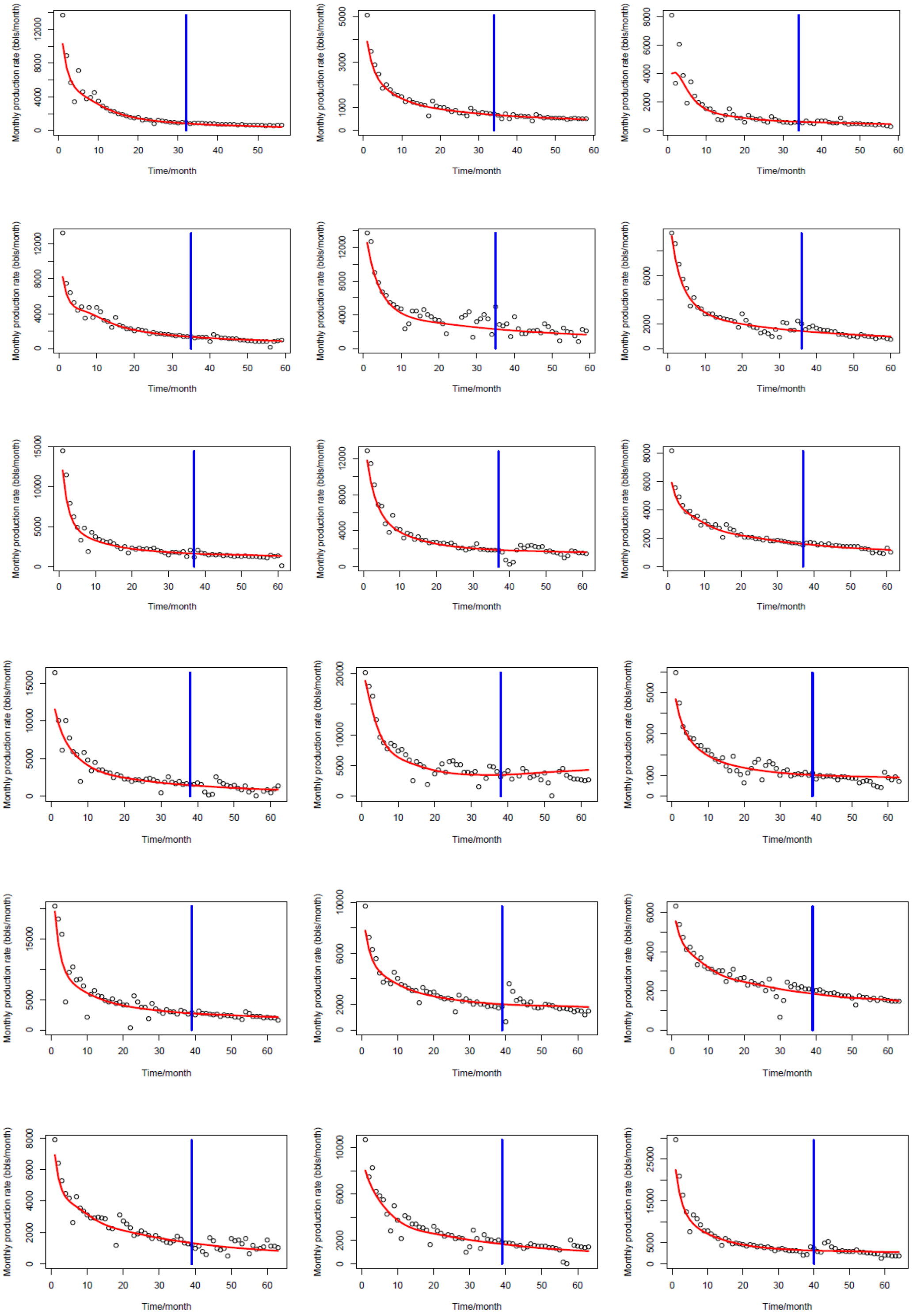


Figure 17: Continued

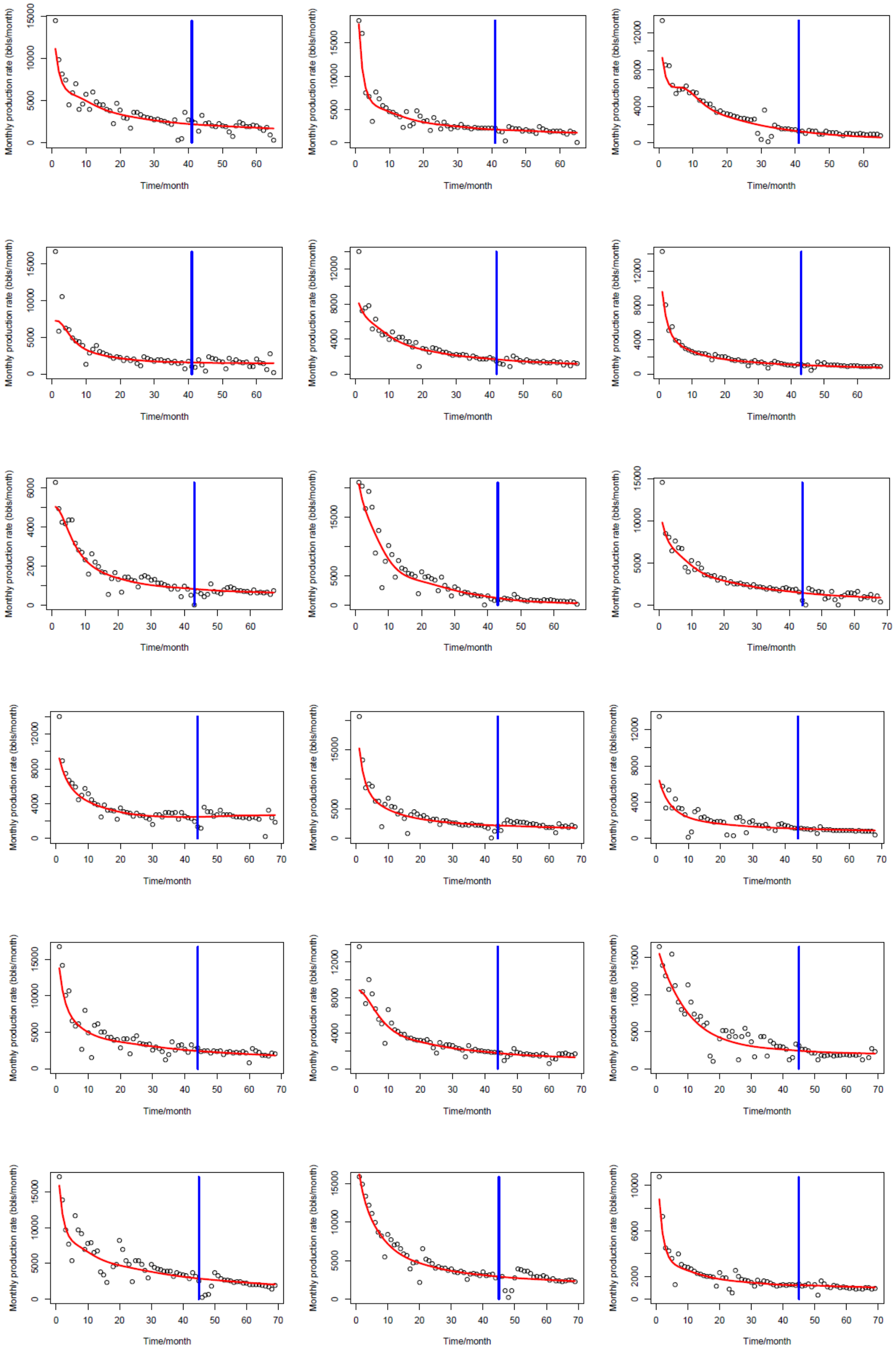


Figure 17: Continued

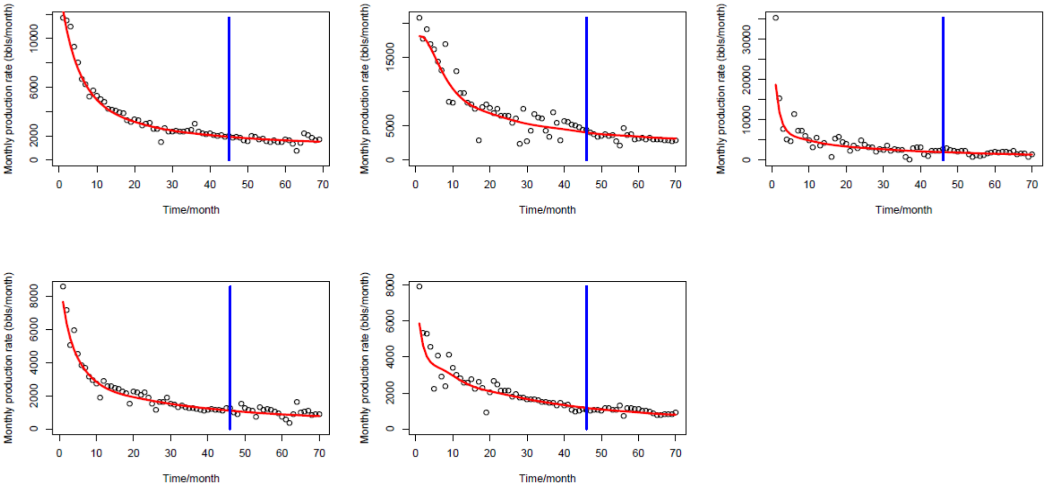
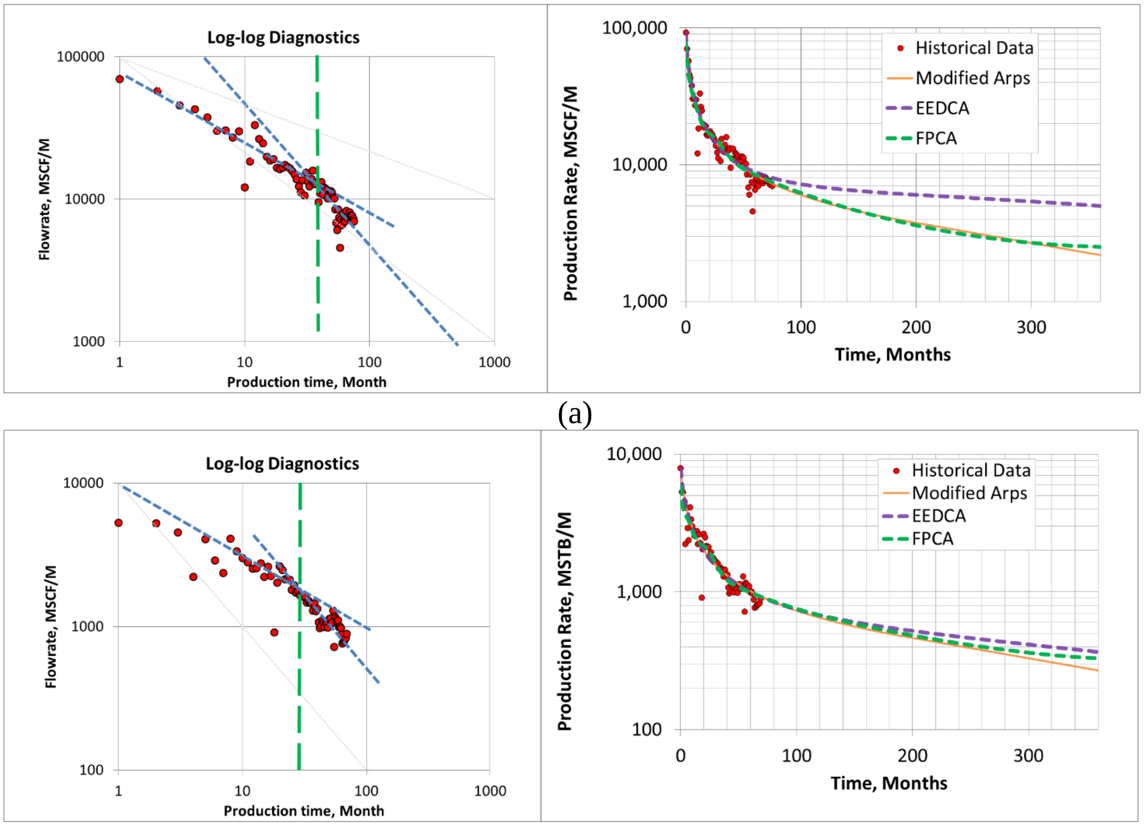


Figure 17: Continued

3.4.2 Comparison of fPCA method and empirical decline model forecasts

The production predictive model derived from fPCA has a functional form from which we can forecast future production and calculate the EUR of a well. We compared these forecasts to those obtained from the traditional modified Arps hyperbolic model and the empirical extended exponential model (EEDCA) (Zhang et al., 2015; Zhang et al., 2016) published recently. For the modified Arps hyperbolic model, we chose a minimum terminal decline rate of 4 %. 15 gas wells are selected from the Eagle Ford and 15 oil wells from Bakken. We make prediction on the primary phase production in the next 10-20 years. The detailed results for one gas well from Eagle Ford and one oil well from Bakken are shown in **Figure** 18. In **Figure** 18 we observe boundary-dominated flow at late times on both plots, and the prediction from fPCA is close to the estimate from the modified hyperbolic. The EEDCA result is higher than that from the other two methods. To be more comprehensive, we present results from analysis of 30 wells in Appendix C. The results presented in Appendix C indicate that fPCA produces predictions comparable to the modified hyperbolic model and extended exponential model.



(a)

(b)

Figure 18: DCA with fPCA, modified Arps and EEDCA (upper: Eagle Ford; lower: Bakken) (Zhou *et al.* 2017)

CHAPTER IV

QUANTITATIVE EVALUATION OF KEY COMPLETION CONTROLS ON SHALE OIL PRODUCTION

4.1 Overview

Since the oil price downturn in 2014, US oil industry has been focusing on reducing the operational expenditures and improving well productivity to stay profitable in the sub-$60 oil price environment (Curtis and Montalbano, 2017). Since well productivity is strongly influenced by the completion design, completion optimization is a key cost-saving method to increase oil production. However, this problem is challenging because well production is influenced simultaneously by a complex combination of many factors such as geological environment, reservoir fluid properties and well completion parameters. It remains largely elusive which features should be extracted from a rich pool of available measurements on reservoir properties to predict production. Even with the knowledge of feature extraction, the functional association between geological features and production can be complex in nature and challenging to characterize. This challenging partially stems from the spatial misalignment issue between horizontal producers and geological data from nearby vertical wells. Wells with complete measurements in the Permian database are quite limited. The database has both vertical deeper-well logs that penetrate Permian Shale but do not have Permian production, and Permian horizontal wells that do not penetrate the full formation but have production data. In addition, a common relationship may not hold across different sub-regions in a reservoir area.

In this chapter, based on production data, completion data, geological data and other relevant information such as water oil ratio (WOR) and gas oil ratio (GOR), we develop a generalized additive model (GAM) (Hastie and Tibshirani, 1986) to investigate possibly nonlinear functional associations between production and key completion parameters. The geological confounding effect can be incorporated in the model in two ways. In the first method, the categorical geological variable “internal zone” (internal layer where the horizontal lateral is located) is included in the GAM model. We can see the importance of “internal zone” according to the statistical significance p value and the leave one out (LOO) cross validation error. In case of the missing of the feature “internal zone”, we propose the second method where the geological effect extracted from well logging data is treated as a clustered random effect by extending a state-of-the-art statistical machine learning method via homogeneity pursuit regularizations that allows us to automatically find clusters without the need to specify the number of clusters in prior and estimate completion control effects simultaneously. The model performance is assessed in terms of prediction accuracy using LOO cross validation. To assess the merits of GAM model, a comparison is made with another additive model derived from alternating conditional expectation (ACE) algorithm (Breiman and Friedman, 1985). It turns out the GAM model outperforms ACE model in prediction accuracy. In addition, GAM model has advantages over ACE model in two folds: (1) we can directly see the relation of oil production with key completion parameters while in ACE model we need an additional inverse transform on both the response and predictors. (2) ACE model is strongly based on the assumption of the additivity of predictors and fails to account for the interaction between predictors while GAM model allows us to include interaction in the model. The learning results of such associations can provide guidance in the development of efficient completion practices.

4.2 Literature review

Using reservoir simulation, Malayalam *et.al* (2014) presented an approach to answer the questions such as how long the lateral to drill, how many stages to complete and how far apart to place the laterals. For the single well optimization study, they made the conclusion that (1) incremental recovery decreases as more stages are added and (2) for a set number of stages longer laterals yields more production due to less fracture interference. In addition, a case study by Zhong *et.al* (Zhong, *et al.*, 2015) using data mining approach reveals that completion lateral length and total proppant amount are important factors driving the first 12-month cumulative oil production. Recently, Pradhan and Xiong (Pradhan and Xiong, 2018) conducted a study aiming to determine optimal lateral lengths and trajectories in Permian Basin. A finding from this study is that long lateral lengths do not completely ensure proportionately more production and some other factors could influence recoveries per lateral length when drilling longer laterals. Another study by Yuan *et.al.* (Yuan *et al.,* 2017) also made a similar claim that through data analytical studies no distinctive advantage of drilling a well with higher lateral length was found. No clear correlation trend was observed between higher lateral length and better production performance in the Barnett.

Although the recent progress in the study of completion design has greatly increased our understanding of key completion controls on shale reservoir productivity, a quick and efficient method that clearly reveals the quantitative relation between production and key completion engineering parameters is still missing. In practice, engineers prefer a method which can quickly provide a clear solution on the questions such as how long the lateral to drill, what is the stage spacing and how much proppant should be used. The major objective of this study is to answer those questions using statistical machine learning method. Our method is a data driven method which is based on all the useful data information we can collect from the existing wells. We claim that our method can provide engineers useful guidance that maximizes well productivity and save operational cost.

4.3 Data and preliminary analysis

In this study, the raw data consists of 355 vertical well logging data, production histories and relevant completion data of 1532 horizontal producing wells and 2221 vertical producing wells from Permian Basin. The target variable is 1-year cumulative oil production and the relevant nongeological features include amount of proppant per stage (lb/stage), fluid per stage (gal/stage), stage spacing (feet), completed lateral length (feet), water oil ratio (WOR), cumulative gas oil ratio (cumGOR) and shut in days (Days). The geological feature is either categorical variable “internal zone” or well logging data. We assume that the logs of each horizontal lateral can be inferred from the logging data of its nearest vertical well at approximately the same depth. The logging data of horizontal wells are used for geological feature clustering and it consists of gamma ray (GR), rock density

(DEN), deep resistivity (RESDEP) and neutron porosity (NEU_LIM). GR is known to be associated with the rock type in the reservoir. High GR value indicates high shale volume and lower GR value indicates lower shale volume. RESDEP indicates the water saturation level at the well location. High deep resistivity indicates lower water saturation since the hydrocarbon has high resistivity compared to the fresh water in the formation. NEU_LIM and DEN are associated with the medium porosity at the well location. The density of pure sandstone is $2.65\ \mathrm{g/cm^3}$ and the density of pure limestone is $2.71\ \mathrm{g/cm^3}$, but the hydrocarbon and water in the pore will change the rock density. We can estimate the medium porosity according to NEU_LIM and DEN logs. In this study, our interest is focused on a Beta zone which has four sub-layers named as WC_SH_B, WC_SH_B1, WC_SH_B2 and WC_SH_B3. Finally, the number of wells with 1-year production history in Beta zone is 106. **Figure** 19 below shows the spatial distribution of vertical wells and our target horizontal wells. From **Figure** 19, vertical wells and horizontal wells are spatially close to each other and as a result, it is reasonable for us to estimate a geological parameter at the location of a horizontal well by using the corresponding nearest vertical log.

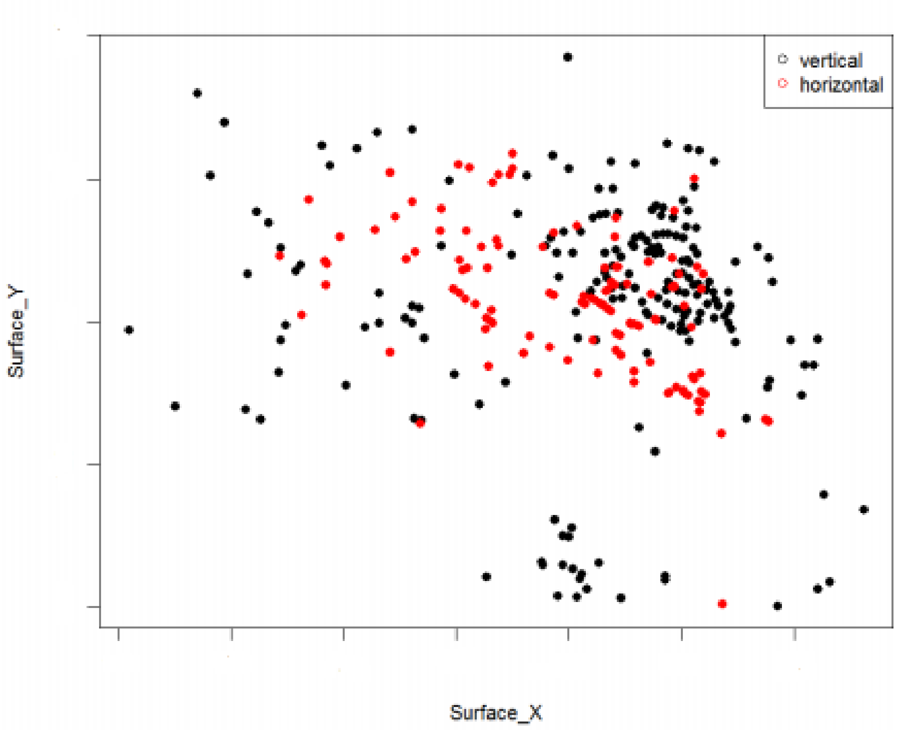


Figure 19: Spatial distribution of horizontal (red) and vertical (black) wells for completion design analysis

The histogram of the response and relevant features are shown in **Figure** 20 below. As we can see, the response variable 12-month cumulative oil production approximately follows the log normal distribution. As we will show later in the LOO cross validation result, in the modeling it is better to take log transformation on the response variable, which yields lower cross validation error compared to the model using the original response. In addition, from the histogram of completed lateral length, we can see there are very few wells with long lateral length from which we expect that the model will have high uncertainty in the relation between oil production and long completed lateral length.

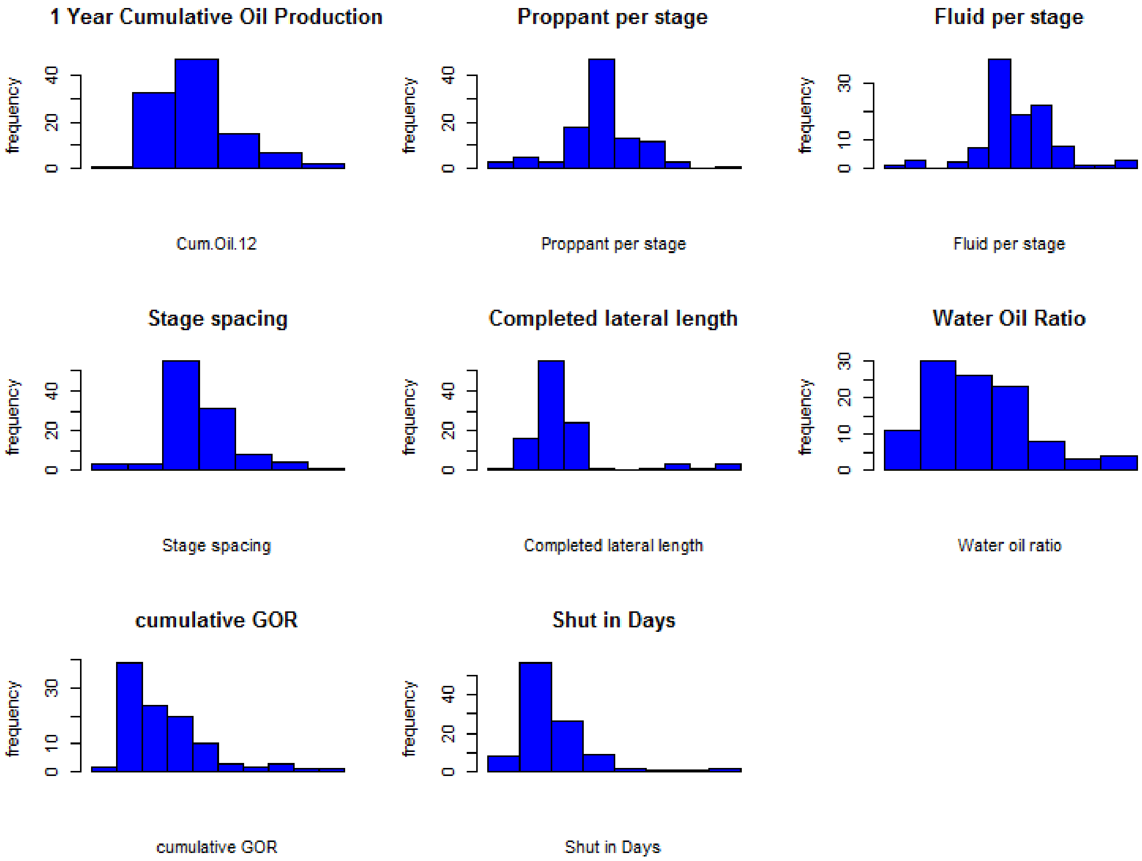


Figure 20: Histogram of response and predictors

The correlation of the response with predictors is shown in **Figure** 21 below:

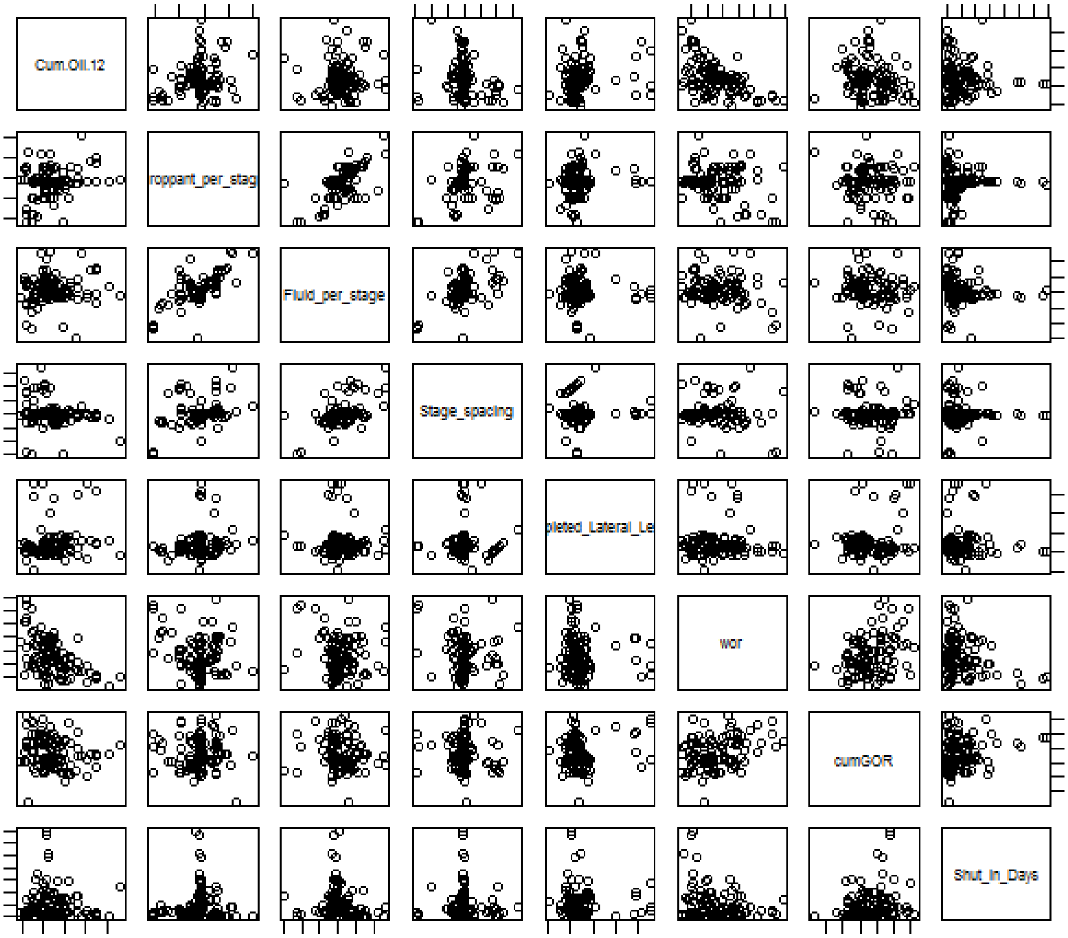


Figure 21: Correlation of cumulative oil production with relevant features

From **Figure** 21, we see that proppant per stage and fluid per stage have strong correlation. In addition, the cumulative oil production is negatively correlated with WOR. It is hard to see the rest of relation pattern from **Figure** 21 and as a result, the correlation matrix is given in **Table** 4 below:

	Cum. Oil.12	Proppant per stage	Fluid per stage	Stage spacing	Completed Lateral Length	WOR	cumGOR	Shut in Days
Cum. Oil.12	1.0	0.2428	0.083	-0.221	0.168	-0.531	-0.170	-0.082
Proppant per stage	0.2428	1.0	0.654	0.372	0.125	-0.233	-0.038	0.029
Fluid per stage	0.083	0.654	1.0	0.567	-0.042	-0.120	-0.085	-0.167
Stage spacing	-0.221	0.372	0.567	1.0	0.029	-0.111	-0.061	-0.028
Completed Lateral Length	0.168	0.125	-0.042	0.029	1.0	-0.037	0.151	-0.014
WOR	-0.531	-0.233	-0.120	-0.111	-0.037	1.0	0.319	-0.269
cumGOR	-0.170	-0.038	-0.085	-0.061	0.151	0.319	1.0	0.169
Shut in Days	-0.082	0.029	-0.167	-0.028	-0.014	-0.269	0.169	1.0

Table 4: Correlation matrix of cumulative oil production with relevant features

The table above conveys us the general information on the relation of oil cumulative production with each predictor and more investigation is required to characterize the influence of key completion parameters on oil production.

4.4 Modeling

In this section, we will have a discussion on the statistical model that characterizes the associations of production with geological and nongeological variables. Most of nongeological variables are well completion parameters. First, we introduce notations. We denote Y^i as the (transformed) production at well location $\mathbf{s}_i$. Let N denote the number of production wells. The geological feature at well location $\boldsymbol{s}_i$ is denoted as $\mathbf{X}_g^i$. This geological feature can be either the categorical variable "internal zone" or a group of variables derived from well logging data. As we will show later, different methods will be applied to the two different cases of geological features. The vector of nongeological variables at well location $\mathbf{s}_i$ is denoted by $\mathbf{X}_{ng}^i = (x_{ng,1}^i, x_{ng,2}^i, \dots, x_{ng,7}^i)$ where the variables from $x_{ng,1}^i$ to $x_{ng,7}^i$ stand for proppant per stage, completed lateral length, stage spacing, fluid per stage, WOR, cumulative GOR and shut in days at well location $\mathbf{s}_i$, respectively. Whatever geological feature we use, we consider oil production model as an additive model with the form as follows:

$$Y^i = f^{(ng)}\left(\mathbf{X}_{ng}^i\right) + f^{(g)}\left(\mathbf{X}_g^i\right) + \epsilon^i \tag{4.1}$$

where ϵ^i is the residual, and $f^{(ng)}$ and $f^{(g)}$ are nongeological effect and geological effect, carrying information on how geological parameters and nongeological parameters influence production, respectively.

The nongeological effect $f^{(ng)}$ is expected to be a relatively smooth function and the interpretability of this function is of high priority for practical engineering controls. We choose to model it as a generalized additive model (GAM) , which has been acknowledged as an appealing choice to model multivariate function models. It represents the relationships between the predictors and the dependent variable as a sum of unknown smooth functions, which flexibly allow both linear or nonlinear fits with relaxed assumptions on the actual relationship between response and predictor with interpretable results. Specifically, the GAM model for $f^{(ng)}$ takes the form

$$f^{(ng)}\left(\mathbf{X}_{ng}\right) = \sum_{i=1}^{7} f_i^{(ng)}\left(x_{ng,i}; \beta_{ng,i}\right) \tag{4.2}$$

in which each $f_i^{(ng)}$ is modeled as a smooth function with parameters $\boldsymbol{\beta}_i$ for $i = 1, \ldots, 7$. Here the equation of $f^{(ng)}$ does not consider the interaction between predictors, but it is easy to generalize this model to take account of the interaction if necessary. For the sake of convenience, the discussion below does not include the interaction. One popular choice of such functions is the cubic smoothing splines (Wood, 2006), where the natural spline basis functions are used with the knots placed at all the observed points to circumvent the problem of knot selection, and the coefficients of these basis functions are regularized to suppress overly wiggly components and to avoid overfitting.

For the k-th nongeological predictor, let $\phi_k^1, \ldots, \phi_k^N$ be the truncated power basis functions for natural cubic spline with knots at $x_{ng,k}^1, \ldots, x_{ng,k}^N$. Then each individual function $f_k^{(ng)}$ can be expressed as $f_k^{(ng)}\left(x_{ng,k}\right) = \sum_{j=1}^{N} \phi_k^j(x_{ng,k})\, \beta_k^j$, the smoothing penalty term is $\lambda_{ng}\boldsymbol{\beta}_k^T \Omega_k \boldsymbol{\beta}_k$, where Ω_k is the $n \times n$ smoothing penalty matrix whose (i,j)-th element is $\omega_k^{ij} = \int \phi_k''^i(t)\phi_k''^j(t)dt$. It plays an important role to control for overfitting by penalizing the wiggliness for each $f_k^{(ng)}$. The penalty term involves a so called smoothing parameter λ_{ng}, controlling the level of penalty; the larger the value of λ_{ng}, the smoother the function.

A remaining challenging is to determine the specification for the geological effect $f^{(g)}$. It is known that subsurface consists of complex and heterogeneous multiple layers and hence underground geophysical properties are often expected to change abruptly. Therefore, relationships between production and geological covariates are expected to exhibit spatially non-smoothly varying patterns. Detecting these clusters allows straightforward interpretations of local associations between response variables and covariates. There are two methods that can be applied to determine the well cluster membership. The first method is to directly use the geological categorical variable “internal zone” as well cluster label. Then the function $f^{(g)}$ is a piecewise constant function of “internal zone”. This method is easy and straightforward. In case of the missing of “internal zone”, we can resort to well logging data for clustering and in the next section, we will mainly discuss geological clustering by well logging data.

4.5 Geological clustering by logging data

In the second method, we consider a flexible regularization model for $f^{(g)}$ that extends the spatial fused lasso and the k-nearest-neighbor $(\mathrm{K}-\mathrm{NN})$ lasso for non-parametric regression. This method is performed in the following steps:

(1) Combine the geological covariates and the spatial coordinates, and define a distance metric between pairs of covariate vectors.

(2) Construct a k-nearest-neighbor $(\mathrm{K}-\mathrm{NN})$ graph, denoted it as $G=(V,E)$ where $V=\{v_1,\ldots,v_n\}$ is the vertex set with n vertices and E is the edge set.

(3) Use this $\mathrm{k}-\mathrm{NN}$ graph to construct the fused lasso penalty for $f^{(g)}$ as follows:

$$\lambda_g \sum_{(i,j)\in E} \left| f^{(g)}(\mathbf{X}_g^i) - f^{(g)}(\mathbf{X}_g^j) \right|$$

The regularization, referred to as the fused lasso penalty (Tibshirani and Taylor, 2011), is to encourage homogeneity between the geological effects at two locations if they are connected by an edge in E. We will discuss how to construct the edge set E later. λ_g is a regularization parameter determining the strength of fused lasso penalty. Since the solution of l_1 penalty results in exact fusion or separation between $f^{(g)}(\mathbf{s}_i)$ and $f^{(g)}(\mathbf{s}_j)$, this regularization automatically leads to a spatially clustered geological random effect.

The values of nearest neighbors k and clustering number are two tuning parameters that are required as input. The optimal number of nearest neighbors and the optimal number of clusters are determined by LOO cross validation. The advantages of using fused lasso for cluster detection are in three folds; it provides an integrated approach that allows to detect clusters and estimate model parameters simultaneously; the number of clusters is completely data driven and there is no strong restrictions on the shape of clusters; furthermore, although designed for piecewise constant coefficients, previous theoretical studies show that this penalty has strong local adaptivity in that it can also successfully capture piecewise Lipschitz continuous functions.

The edge set E is a key ingredient in the model since it reflects the prior assumption on the homogeneity structure of geological effects. Note the fact that similar geological conditions are likely to lead to similar effect on production. It is therefore desirable to construct E such that pairs of locations that have similar values of geological parameters are included to reflect homogeneity among them. As aforementioned, in this study, we choose to include all edges that connect a well location with each of its k nearest neighbors. Here, the neighbors are searched by using the distance metric defined on principle component scores of geological parameter values. By principle component analysis (PCA), all wells share one space coordinate system spanned by the principle components and for each well, the corresponding score vector can be interpreted as the projection of the original vector (defined by the spatial coordinates and geological parameters at that specific well location) onto each unit principle component coordinate.

As a result, we can take the principle component score vector as point coordinates that specify the location in the space spanned by the principle component vectors. Thus, we propose to use the principle component score metric to measure the geological similarity, which is analogous to the method we normally apply to define two-point distance in Euclidean space.

Using the above regularization models for $f^{(g)}$ and $f^{(ng)}$, we have an optimization problem as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\left\{Y^i - \sum_{k=1}^{7}\sum_{j=1}^{N}\phi_k^j\left(x_{ng,k}\right)\beta_k^j - f^{(g)}\left(\mathbf{X}_g^i\right)\right\}^2 + \sum_{k=1}^{7}\lambda_{ng}\boldsymbol{\beta}_k^T\Omega_k\boldsymbol{\beta}_k + \lambda_g\sum_{(i,j)\in E}\left|f^{(g)}\left(\mathbf{X}_g^i\right) - f^{(g)}(\mathbf{X}_g^j)\right| \qquad (4.3)$$

Our goal is to find the estimates of $\boldsymbol{\beta}_k$ and $f^{(g)}$ that maximizes the above objective function.

4.6 Estimation

In this section, we will show an iterative optimization algorithm for the estimation of the parameter $\boldsymbol{\beta}_k$ and $f^{(g)}(\mathbf{X}_g^i)$ in Eq. (4.3).

Given values of $f^{(g)}(\mathbf{X}_g^i)$, the vector $\boldsymbol{\beta}_k(k = 1,2,\ldots,7)$ are estimated via a quadratically penalized least square method, i.e., by minimizing

$$\frac{1}{N}\sum_{i=1}^{N}\left\{Y^i - \sum_{k=1}^{7}\sum_{j=1}^{N}\phi_k^j\left(x_{ng,k}^j\right)\beta_k^j - f^{(g)}\left(\mathbf{X}_g^i\right)\right\}^2 + \sum_{k=1}^{7}\lambda_{ng}\boldsymbol{\beta}_k^T\Omega_k\boldsymbol{\beta}_k \tag{4.4}$$

We use the function gam in the R package mgcv to solve this optimization. And we follow the Generalized Cross Validation (GCV) criterion to estimate the smoothing parameters λ_{ng}.

Given values of $\boldsymbol{\beta}_k(k = 1,2,\ldots,7)$, $f^{(g)}(\mathbf{X}_g^i)$ is obtained by solving a regularized convex optimization as below:

$$\frac{1}{N}\sum_{i=1}^{N}\left\{Y^i - \sum_{k=1}^{7}\sum_{j=1}^{N}\phi_k^j\left(x_{ng,k}^j\right)\beta_k^j - f^{(g)}\left(\mathbf{X}_g^i\right)\right\}^2 + +\lambda_g\sum_{(i,j)\in E}\left|f^{(g)}\left(\mathbf{X}_g^i\right) - f^{(g)}(\mathbf{X}_g^j)\right| \tag{4.5}$$

We first reformulate the above equation as a generalized Lasso problem as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\left\{Y^i - \sum_{k=1}^{7}\sum_{j=1}^{N}\phi_k^j\left(x_{ng,k}^j\right)\beta_k^j - f^{(g)}\left(\mathbf{X}_g^i\right)\right\}^2 + \lambda_g\left\|\mathbf{H}\mathbf{f}^{(g)}\right\|_1 \tag{4.6}$$

where $\mathbf{f}^{(g)} = (f^{(g)}\left(\mathbf{X}_g^1\right),\ldots,f^{(g)}(\mathbf{X}_g^N))$, $\mathbf{H}$ is a $m \times n$ matrix constructed from the edge set E with m edges. For an edge connecting two locations s_i and s_j, we represent the penalty term $\left|f^{(g)}\left(\mathbf{X}_g^i\right) - f^{(g)}(\mathbf{X}_g^j)\right|$ as $\left|\mathbf{H}_m\mathbf{f}^{(g)}\right|$ where $\mathbf{H}_m$ is a row vector of $\mathbf{H}$ only containing two non-zero elements, 1 at i-th element and -1 at j-th. The standard ADMM proceeds by first decoupling the likelihood term and the regularization term by introducing new equality constraints $\mathbf{Hf}^{(g)} - \boldsymbol{\gamma} = \mathbf{0}$, that is to

$$\text{minimize}\left\|\mathbf{Y} - \hat{\mathbf{f}}^{(ng)} - \mathbf{f}^{(g)}\right\|_2^2 + \lambda_g \|\boldsymbol{\gamma}\|_1, \text{ subject to } \mathbf{Hf}^{(g)} - \boldsymbol{\gamma} = \mathbf{0}$$

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

Step 1: $\mathbf{f}_{(t+1)}^{(g)} = (\mathbf{I} + \rho\mathbf{H}^T\mathbf{H})^{-1}(\mathbf{Y} - \hat{\mathbf{f}}^{(ng)} + \rho\mathbf{H}^T(\boldsymbol{\gamma}_{(t)} - \mathbf{u}_{(t)}))$ (4.7.1)

Step 2: $\boldsymbol{\gamma}_{(t+1)} = S_{\lambda_g/\rho}(\mathbf{Hf}_{(t+1)}^{(g)} + \mathbf{u}_{(t)})$ (4.7.2)

Step 3: $\mathbf{u}_{(t+1)} = \mathbf{u}_{(t)} + \mathbf{Hf}_{(t+1)}^{(g)} - \boldsymbol{\gamma}_{(t+1)}$ (4.7.3)

where $S_{\lambda_g/\rho}$ is the soft thresholding function. We use the admm.genlasso function in the R package penreg to solve it. The two optimizations are run iteratively until cluster memberships and model parameters converge.

LOO cross validation is needed for either model assessment or tuning parameter selection since the number of wells available for the study is limited. Specifically, given n horizontal wells, we choose $n - 1$ horizontal wells as training wells to build a production model that incorporates both geological and completion effects and then make prediction on production for the well that is left out. Each well in our study will serve as a test well once and finally, we will compare the true production values with the prediction obtained from cross validation. The equation of cross validation error is

$$\text{C.V. Error} = \sqrt{\frac{\sum_i \left(y_{\text{true},i} - y_{\text{pred},i}\right)^2}{\sum_i y_{\text{true},i}^2}} \tag{4.8}$$

where y_{true} and y_{pred} denote the true value and prediction value of 1-year cumulative oil production, respectively. We will select the model or the tuning parameters that yield the minimum cross validation error.

4.7 Real data results

We first consider the GAM model that includes the feature “internal zone”. Not all the nongeological features are important to GAM production model and p value is used to find out features that are statistically significant. LOO cross validation is used to support our argument on feature selection. According to fracture mechanics, the amount of proppant and fluid used in fracking are correlated. Therefore, an additional term is added in GAM to take account of the interaction between proppant and fluid. To show the importance of interaction term, we compare the results with and without the interaction in terms of the criterion LOO cross validation error. Also, we will compare the results with and without applying log transformation to the response variable. Then we will present the results including geological clustering. Finally, the result from ACE method will be used for model comparison.

Functional terms of features	p-value (case 1)	p-value (case 2)	p-value (case 3)	p-value (case 4)
ti(proppant per stage)	0.0458	0.0296	0.0183	0.0200
ti(stage spacing)	1.45e-5	4.62e-6	8.99e-6	2.1e-6
ti(completed lateral length)	0.0342	0.0758	NA	NA
ti(fluid per stage)	0.0446	0.0286	0.0566	0.0315
ti(proppant per stage, fluid per stage)	0.0322	0.0207	0.0132	NA
ti(cumGOR)	0.167	NA	NA	NA
ti(WOR)	9.78e-9	2.76e-9	3.62e-9	9.04e-9
ti(shut in days)	0.0026	0.00454	0.00214	0.0039
Internal Zone	0.0244	0.0161	0.0122	0.0168
Cross validation	0.2576	0.2512	0.2490	0.2643

Table 5: p value table of GAM model

The second column of **Table** 5 above shows the p value of each feature for the model containing all the features (case 1). The LOO cross validation error is 0.2576. As we can see, the feature cumulative GOR has very high p value, which indicates this feature is not statistically significant in oil production model. Then we run the second model without cumulative GOR (case 2) and the p value of each feature is shown in the third column of **Table** 5. The LOO cross validation error is 0.2512. Therefore, without cumulative GOR, the LOO cross validation error becomes smaller. Furthermore, in case 2 we see the p value of completed lateral length is 0.0758 and therefore, the third model (case 3) without completed lateral length is tested. The LOO cross validation error is 0.2490. Thus, the most important features for oil production are proppant per stage, fluid per stage, stage spacing, WOR, shut in days and internal zone. In addition, the interaction between proppant per stage and fluid per stage is very important, which can be seen from the comparison between case 3 and case 4. In case 4, the model without interaction yields the cross validation 0.2643 much larger than the result from case 3. Thus, the final predictive model for oil production is chosen to be case 3. Completed lateral length does not show the statistical importance in predicting oil cumulative production. This is because in our dataset most of wells have completed lateral length highly concentrated within one interval and then we do not see the importance of this feature. However, according to the domain knowledge from production engineering, we believe completed lateral length is an important feature and in addition, since one of our goals is to answer the question whether longer completed lateral length necessarily indicates higher cumulative production, we will include completed lateral length in the final model to see the effect of long complete lateral length on oil production. The relations of oil production with proppant per stage, fluid per stage, stage spacing, WOR, shut in days and completed lateral length are given in **Figure** 22 below:

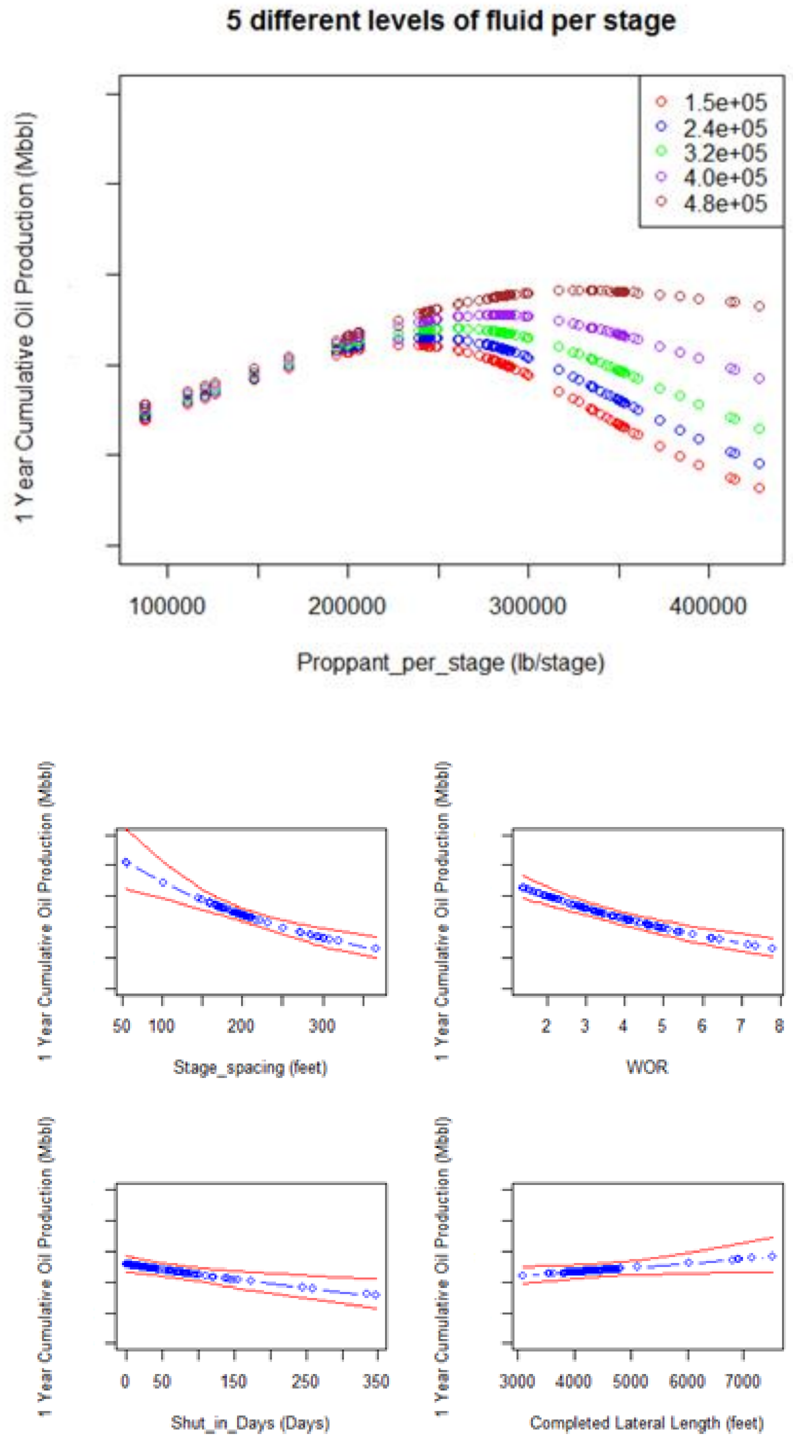


Figure 22: The relation of oil cumulative production with key nongeological features

In **Figure** 22, the first plot shows the functional association between oil cumulative production and proppant per stage at different levels of fluid per stage. Focus on the peak points of the 5 curves and we see at that point the amount of proppant and fluid are positively correlated. The more fluid we use during fracking, the more proppant we need. We can find physical explanation for this result. The fluid is injected mainly to stimulate fracture propagation and the proppant is injected into the fracture for supporting purpose which prevent the fracture from being closed again. The more fluid we inject, the larger the fracture volume will be and as a result, the more proppant is needed to keep the fracture open. In addition, when the amount of fluid per stage is fixed, injecting more proppant does not necessarily indicate the increase of oil production. At the average level of fluid per stage, the proppant efficiency starts to decline when proppant per stage is approximately over 3×10^5 $\mathrm{lb/stage}$. It is recommended to determine the optimal value of proppant per stage after we decide how much fluid we will use for fracking. Next, in the second plot the blue dot line represents the expected contribution to oil cumulative production from each of the other three features and the two red lines in each plot represent the uncertainty on the estimation. The feature stage spacing is negatively correlated with oil cumulative production. The smaller the stage spacing is, the higher the cumulative production will be. However, we also notice that the number of wells with stage spacing less than 150 ft is very few and it is recommended to have more wells with stage spacing around 100 ft to 150 ft and we expect higher cumulative production compared to other wells with stage spacing greater than 150 ft. As for WOR, we can see that the higher water oil ratio is, the lower the oil cumulative oil production will be. The same relation holds for the feature shut in days. The last plot is a relation between oil production and completed lateral length. From **Figure** 22, we see the oil production rises with the increase of completed lateral length. However, there are very few wells with completed lateral length higher than 5000 ft and the production uncertainty at long lateral length is large. Therefore, it is recommended that we should drill more wells with completed lateral length between 5000 ft and 7000 ft to help us understand more about production from long horizontal wells. This will reduce our uncertainty on the production from long horizontal wells. At least, from the current figure we have high confidence that the completed lateral length 5000 ft is better than completed lateral length 4000 ft in the aspect of maximizing well productivity. This observation should be helpful to engineers who need to decide how long the well should be drilled. In summary, according to **Figure** 22, we can determine the optimal value of stage spacing. The optimal value of proppant per stage depends on the amount of fluid we plan for each stage. The completed lateral length is recommended to be between 5000 ft and 7000 ft.

If we do not apply log transformation to the original oil cumulative production, the LOO cross validation error is 0.273. Therefore, we can see the model with log transformation yields lower cross validation error. This is one reason why the response variable is transformed in this study. Another reason for log transformation is that the GAM model with log transformation implies that the nonlinear function of true production is the product of a function for geological effect and another function for completion effect. If a reservoir is a sweet spot with good geological properties, then we expect the production will be high; if the geological condition is the same, then good completion design will also increase well productivity.

Now we will discuss the geological clustering. In geological clustering, the feature Internal zone is not used. The optimal number of neighbors and clusters are determined by LOO cross validation. The result is given in **Table** 6 below:

Clusters Neighbors	2	3	4	5	6	7	8	9	10
3	NA	NA	0.2584	0.2569	0.2534	0.2525	0.2523	0.2505	0.2534
4	0.2552	0.2552	0.2498	0.2498	0.2498	0.2512	0.2534	0.2557	0.2556
5	0.2552	0.2554	0.2519	0.2516	0.2518	0.2523	0.2531	0.2542	0.2547

Table 6: Leave one out cross validation error (geological clustering)

From **Table** 6 we can see that the optimal number of neighbors is 4 and the optimal number of clusters is 4. The LOO cross validation error is 0.2498 which is close to the error from the first model that has the component "internal zone". When the number of neighbors is

4, the minimum number of clusters is 4. Based on the optimal hyperparameter values, we have the geological clustering results in **Figure** 23 and **Figure** 24. From **Figure** 24, we see the relations of oil production with stage spacing, WOR, shut-in day and completed lateral length are the same as those from the GAM model based on “internal zone”. As for the relation between oil production and proppant per stage, we see a slight difference when the amount of fluid per stage is 4.8×10^5 $\mathrm{gal/stage}$ compared to the plot in **Figure** 22. This does not indicate that in the second approach the nonlinear functions of proppant per stage, fluid per stage and the interaction term are different from those in the first model. It turns out that we will see similar trend in the first model when we increase the amount of fluid in each stage. This difference arises from the difference of the two methods we proposed in handling the geological confounding effect. Geologists determine the internal zone according to their analysis on all the geological information available including well logging data, while in the second approach well logging data is only used as prior information and the final clustering is Bayesian result which incorporates both well logging and production information. In addition, we have a comparison on the estimated average production that is only due to the geology part by letting the completion effect on production be the same in the two models. It turns out that the two estimated mean values are very close to each other with a difference by only 3%. This indicates the similarity of the two models. According to the result from the second model (geological clustering by well logging), we recommend the amount of proppant per stage to be between 3.5×10^5 lb/stage and 4.0×10^5 lb/stage given the large uncertainty that exists at large proppant per stage. Fluid per stage is recommended to be no less than 4.0×10^5 gal/stage.

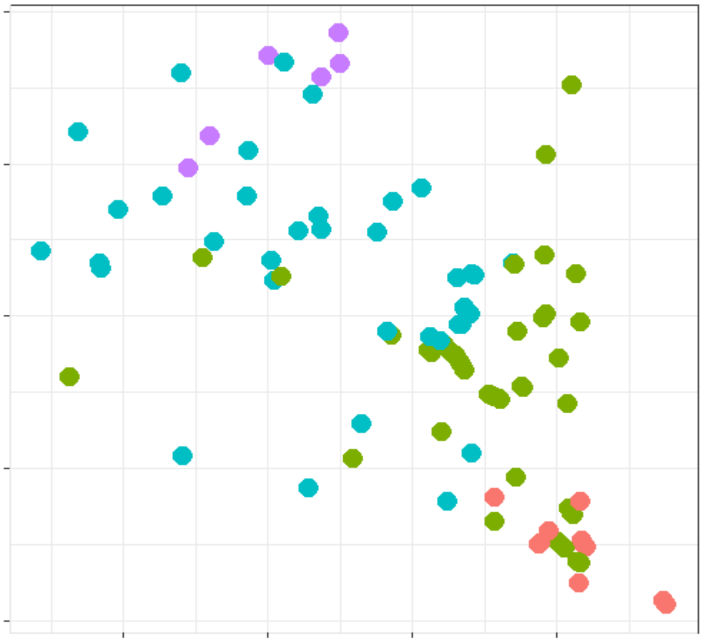


Figure 23: Geological clustering

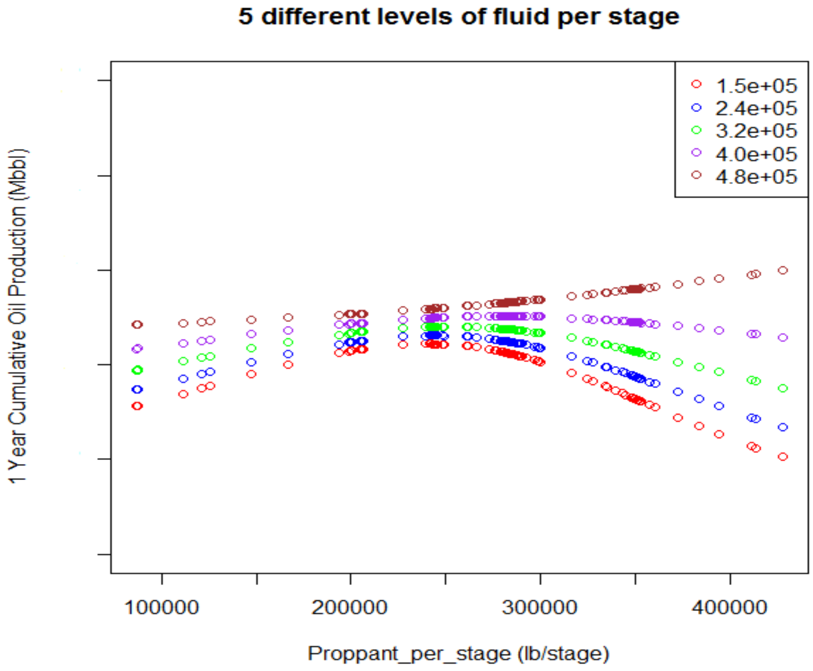


Figure 24: The relation of oil cumulative production with key nongeological features (geological clustering by logging data)

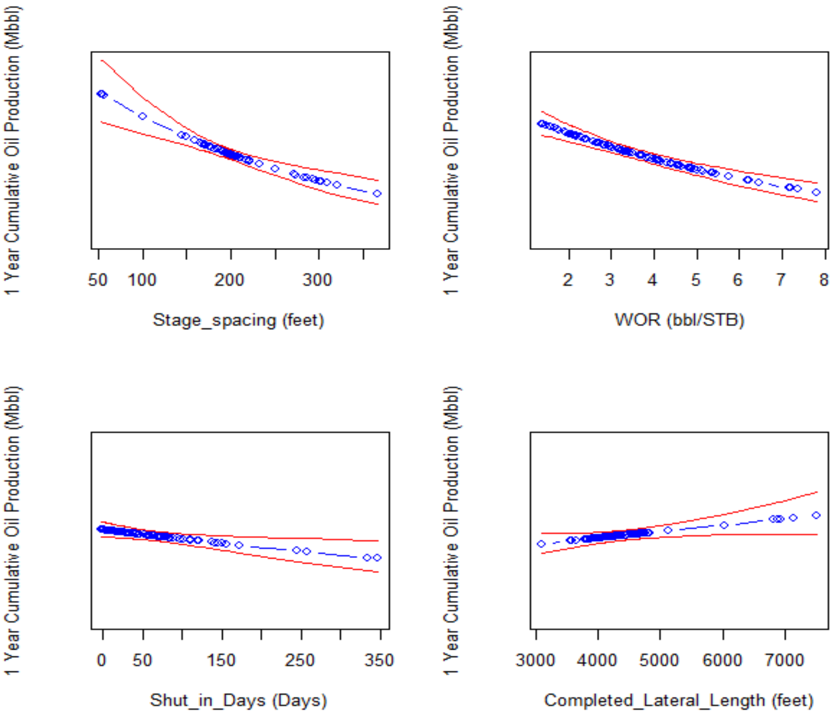


Figure 24: Continued

Lastly, we compare our method with the model derived from ACE algorithm. ACE builds transformations of the independent variables $X_i \to \phi(X_i)$ and the response variable $Y \to \theta(Y)$ which minimize the regression error variance in the transformed space:

$$e^2(\theta, \phi_1, \phi_2, \dots, \phi_n) = \frac{E\left\{\left[\theta(Y) - \sum_{i=1}^{n} \phi_i(X_i)\right]^2\right\}}{E[\theta^2(Y)]} \tag{4.9}$$

The features used in ACE model are the same as the features used in GAM model, but ACE model does not take account of the interaction between proppant and fluid. The result of ACE algorithm is given below:

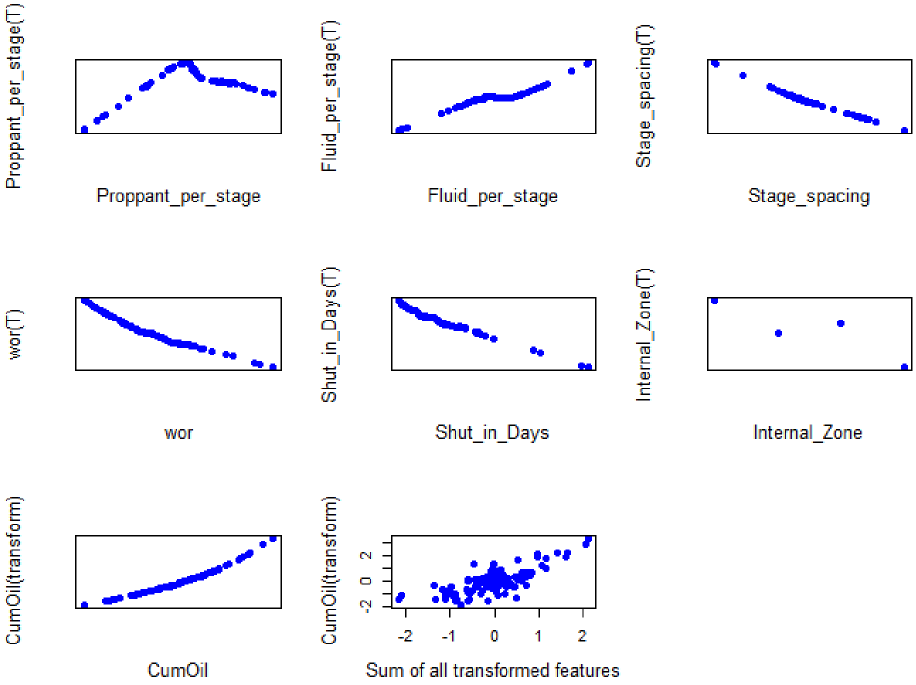


Figure 25: Nonlinear transform of features by ACE algorithm

In **Figure** 25, we can see the transformed functions of proppant per stage and fluid per stage are very similar to the functions we obtained in GAM model. Stage spacing, WOR and shut in days also hold similar relation. But the LOO cross validation error from ACE model is 0.29 which is 16% higher than the GAM model. As a result, we recommended the GAM model for the oil production model.

CHAPTER V

SUMMARY AND CONCLUSIONS

In chapter 2, we developed an optimization algorithm for shale well decline curve analysis. For a well that is already in the BDF regime, this algorithm can be combined with a three-step diagnostic method to find the optimal solution of five important production parameters q_i, b_i, D_i, b_f, t_c. Since there is a rigorous theoretical proof supporting the proposed three-step diagnostic method, we can use the new diagnostic method to effectively identify production flow regime, and this improvement can help us have more confidence on the estimated parameter values. In contrast, from the discussion in section 2.5, we see the drawback of using traditional modified hyperbolic model based minimum decline rate $D_{\min}$. If the wells are still in the transient flow regime, with simple modification of the algorithm we can uniquely determine the optimal value of q_i, D_i, b_i that yields global minimum residual errors. The foundation of the proposed algorithm is the assumption that the residual errors are identical and independent, and we can check its validity by computing the autocorrelation of the residual errors. We must check the value of decline curve parameters to see if the results are consistent with our physical domain knowledge of reservoir fluid flow. The proposed diagnostic plot is recommended for parameter verification. From diagnostic plot, we determine the flow regime and then we have a rough idea on the range of parameter values and then we check whether our result fits with our preliminary analysis. If not, we need further investigation on the well production data for the reason of mismatch (e.g., outlier). The value of switching time t_c can affect the value of initial Arps b parameter and as a result, the additional parameter verification is extremely important when the production data is severely discontinuous. Accurate estimation of q_i, b_i, D_i, b_f, t_c is important to EUR estimation. In probabilistic decline curve analysis (PDCA), we need to specify the distribution of production parameters which is the cornerstone to the MCMC step of PDCA. The distribution of production parameters is based on the analysis of all the existing well production data. If there are not enough wells for analysis, the additional assumption is made about the distribution function that each decline curve parameter may satisfy and then we can use the maximum likelihood function or method of moments to estimate the unknown parameters of each distribution function. Using the algorithm proposed in this chapter we can effectively improve the estimation accuracy of decline curve parameters for each single well. Therefore, we will also improve our estimation on the distributions of q_i, b_i, D_i, b_f, t_c. In addition, our work can be applied in the construction of type wells and EUR estimation for undrilled wells.

In chapter 3, we have developed a new method for forecasting primary phase production from liquid-rich shale reservoirs. Functional principal component analysis is the core of our approach. We conducted a hindcasting experiment, which demonstrated the ability of the fPCA method to predict production accurately according to the comparison between prediction and true observation. In the fPCA method we construct a linear production model with the principal component functions as the expansion basis. From the linear production model, we can construct a production decline curve which honors the observed production data and which also predicts the EUR of a well. The estimates of ultimate recovery (EUR) by fPCA method is close to the estimation based on modified hyperbolic model and in some cases the extended exponential decline model produces wrong estimation.

In contrast to decline-model techniques for low-permeability well production analysis, the fPCA approach is driven purely by the production data. In the decline-model based approach, our goal is to fit the observed production data with an empirical model by adjusting model parameters so that the discrepancy between the observed data and estimation is minimum. This is a nonlinear regression problem and we will always encounter the issue that the model parameter determination does not have a unique solution. Different parameter values may result in quite different predictions. In addition, for the modified hyperbolic model, we need to specify the switching point from a hyperbolic model to an exponential model. This switching point depends on a minimum decline rate which is selected by the analyst. This introduces uncertainty in the prediction results due to humans, who often introduce subjective biases into their analyses. The advantages of the FPCA approach is that it does not have the issue with non-uniqueness and it is also easy to implement with the aid of the free R-package “fda” available online. This approach is also efficient and can be used to analyze the production from a large number of producing wells at the same time.

One limitation of the fPCA approach is its requirement on the smoothness of production data. We developed an approach using a cumulative production curve coupled with the

Bourdet derivative algorithm to calculate rates which reduces the noise that exists in field data. For the 200 wells from Eagle Ford and Bakken formations, we see the smoothness of production data is improved after using the cumulative production data and Bourdet derivative algorithm. In addition, the new production profile matches well with the original profile. Another limitation of the fPCA approach is the need for the training wells with long production histories from which we can extract production features. Training well histories must include boundary-dominated flow; otherwise, the forecasts will be inaccurate. This limitation can be mitigated with the use of reservoir simulation when an insufficient number of wells are available for training. However, successful simulation requires a high-quality geoscience reservoir model.

Last, but not least, in this work we did not consider spatial correlation which might exist in the production data due to well interference. We recommend a study of this problem in future work. In addition, the important problem of predicting secondary phase production is important in liquid-rich low permeability reservoirs is worth investigating in the future.

In chapter 4, we consider the problem of completion design optimization. We propose a generalized additive model (GAM) to investigate possibly nonlinear associations between production and key completion parameters (e.g. completion lateral length, total proppant, number of hydraulic fracturing stages). The geological confounding effect is incorporated in the model in two ways. In the first method, we add the categorical feature "internal zone" (internal layer where the horizontal lateral is located) to the GAM model. In case of the missing of the feature “internal zone”, we propose the second method where the geological effect extracted from well logging data is treated as a clustered random effect by means of graphic fused LASSO. Feature selection is carried out based on the p value of each term in the generalized additive model. Leave-one out (LOO) cross validation error is used to assess the goodness of feature selection. The results show that the following features are important to the prediction of oil cumulative production (1) proppant per stage (2) fluid per stage (3) stage spacing (4) water oil ratio (WOR) (5) shut in days and (6) internal zone. Based on the Permian dataset used in this study, completed lateral length is not statistically significant in the predictive model of oil production. This is because the completed lateral length in our dataset has low variation and is highly concentrated within one region. According to our domain knowledge on well production, we believe completed lateral length is an important feature and as a result, we have completed lateral length included in our final model. In addition, it is recommended to collect more data from wells with long completed lateral length. The model using the feature “internal zone” has an advantage over the second model in computational efficiency. It requires no iteration and no hyperparameter tuning by cross validation. Hyperparameter tuning by cross validation is the most time-consuming part in the second model. The second model has the advantage over the first model in the aspect that the clustering result is a result that incorporates both geological and production information.

Furthermore, we have the following recommendation for well completion in practice after the analysis of the 106 horizontal wells in Permian basin: (1) the optimal amount of proppant at each stage depends on the amount of fluid to be injected for hydraulic fracturing. The more fluid we use for multistage fracturing, the more proppant is needed to keep the fracture open. Based on our dataset, when the amount of fluid in each stage is less than 4.0×10^5 gal/stage, we see that there is an optimal value for the average amount of proppant to be injected in each stage and the production will no longer increase when proppant per stage is more than that optimal value. On the other hand, when fluid per stage exceeds 4.0×10^5 gal/stage, from our dataset we do not see the decline of production due to the increase of proppant in each stage. We recommend that proppant per stage not be less than $3.5 \times 10^5 \text{ lb/stage}$ and the fluid per stage not be less than 4.0×10^5 gal/stage. (2) We recommend a stage spacing of 150 ft (3) We recommend a completed lateral length between 5000 ft and 6000 ft. To address the issue of economics, we need more data such as estimated ultimate recovery (EUR) and completion cost. This topic is, however, beyond the scope of this study.

REFERENCES

Anderson, R. N., Xie, B. Y., Wu, L., Kressner, A. A., Frantz Jr., J. H., Ockree, M. A. and Brown, K. G. 2016a. Petroleum Analytics Learning Machine to Forecast Production in the Wet Gas Marcellus Shale. *SPE/AAPG/SEG Unconventional Resources Technology Conference, San Antonio, Texas, USA*, DOI: https://doi.org/10.15530/URTEC-2016-2426612

Anderson, R. N., Xie, B. Y., Wu. L., Kressner, A. A., Frantz Jr., J. H., Ockree, M. A., Brown, K. G., Carragher, P. and McLane, M. A. 2016b. Using Machine Learning to Identify the Highest Wet Gas Producing Mix of Hydraulic Fracture Classes and Technology Improvements in the Marcellus Shale. *SPE/AAPG/SEG Unconventional Resources Technology Conference, San Antonio, Texas, USA*, DOI: https://doi.org/10.15530/URTEC-2016-2430481

Arps, J. J. 1944. Analysis of Decline Curves. SPE-945228-G. Trans., AIME, 160: 228-247.

Aziz, K. and Settari, A. 1979. Petroleum Reservoir Simulation, Applied Science Publishers.

Bhattacharya, S. and Nikolaou, M. 2013. Analysis of Production History for Unconventional Gas Reservoir with Statistical Methods. *SPE Journal*, 18(05), 878-896. SPE-147658-PA.

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

Chen, C. H., Ramirez, B. A., Vink, J. C. and Girardi, A. M. 2016. Assisted History Matching of Channelized Models by Use of Pluri-Principal-Component Analysis. *SPE Journal*, 21(05), 1793-1812. DOI: https://doi.org/10.2118/173192-PA

Chen, C. H., Gao, G. H., Honorio, J., Gelderblom, P., Jimenez, E. and Jaakkola, T. 2014. Integration of Principal-Component-Analysis and Streamline Information for the History Matching of Channelized Reservoirs. Paper SPE 170636 presented at

SPE Annual Technical Conference and Exhibition, Amsterdam, The Netherlands, 27-29 October, DOI: https://doi.org/10.2118/170636-MS

Chen, C. H., Jin, L., Gao, G. H., Weber, D., Vink, J. C., Hohl, D., Alpak, F. O. and Pirmez, C. 2012 Assisted History Matching Using Three Derivative-Free Optimization Algorithms. Paper SPE 154112 presented at SPE Europec/EAGE Annual Conference, Copenhagen, Denmark, 4-7 June, DOI: https://doi.org/10.2118/154112-MS

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

Doung, A. N. 2010. An Unconventional Rate Decline Approach for Tight and Fracture-Dominated Gas Wells. Paper SPE 137748 presented at Canadian Unconventional Resources and International Petroleum Conference, Calgary, Alberta, Canada, 19-21 October 2010. DOI: http://dx.doi.org/10.2118/137748-MS.

EI-bakry, A., Romer, M. C., Xu, P., Sundaram, A., Usadi, A. K., Morehead, H. L., Crawford, M. L., Holloway, B. and Knight, C. 2012. Decision Support and Workflow Automation for the Development and Management of Hydrocarbon Assets Using Multi-Agent Systems. Paper SPE 150285 presented at SPE Intelligent Energy International, Utrecht, The Netherlands, 27-29 March. DOI: https://doi.org/10.2118/150285-MS

Emerick, A. A. and Reynolds, A. C. 2013. Ensemble Smoother with Multiple Data Assimilation. *Computers & Geosciences,* 55, 3-15, DOI: https://doi.org/10.1016/j.cageo.2012.03.011

Evensen, G. 2003. The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation. *Ocean Dynamics*, 53(4), 343-367

Farasat, A., Shokrollahi, A., Arabloo, M., Gharagheizi, F. and Mohammadi, A. H. 2013. Toward an Intelligent Approach for Determination of Saturation Pressure of Crude Oil. *Fuel Processing Technology*, 115, 201-214.

Fetkovich, M. J. 1980. Decline Curve Analysis Using Type Curves. *J Pet Technol*, 32(6), 1065-1077. SPE-4629-PA.

Fetkovich, M. J., Vienot, M. E., Bradley, M. D. and Kiesow, U. G. 1987. Decline Curve Analysis Using Type Curves: Case Histories. *SPE Form Eval* 2(4): 637-656, Trans., AIME, 283, SPE-13169-PA.

Freeborn. R. and Russell, B. 2014. How to Apply Stretched Exponential Equations to Reserve Evaluation. Paper SPE 162631 presented at SPE Hydrocarbon Economics and Evaluation Symposium, Calgary, Alberta, Canada, 24-25 September, 2012.

Fulford, D. S., Bowie, B., Berry, M. E., Bowen, B. and Turk, D. W. 2015. Machine Learning as a Reliable Technology for Evaluating Time-Rate Performance of Unconventional Wells. Paper SPE 174784 presented at SPE Annual Technical Conference and Exhibition, Houston, Texas, USA, 28-30 September, DOI: https://doi.org/10.2118/174784-MS

Gao, G. H. and Reynolds, A. C. 2004 An Improved Implementation of the LBFGS Algorithm for Automatic History Matching. Paper SPE 90058 presented at SPE Annual Technical Conference and Exhibition, Houston, Texas, 26-29 September, DOI: https://doi.org/10.2118/90058-MS

Gao, G. H., Vink, J. C., Chen, C. H., Tarrahi, M. and Khamra, Y. E. 2016. Uncertainty Quantification for History Matching Problems with Multiple Best Matches Using a Distributed Gauss-Newton Method. Paper SPE 181611 presented at SPE Annual Technical Conference and Exhibition, Dubai, UAE, 26-28 September. DOI: https://doi.org/10.2118/181611-MS.

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

Hemmati-Sarapardeh, A., Shokrollahi, A., Tatar, A., Gharagheizi, F., Mohammadi, A. H. and Naseri. A. 2014. Reservoir Oil Viscosity Determination Using a Rigorous Approach. *Fuel*, 116, 39-48.

Holditch, S. A. 2010. Shale Gas Holds Global Opportunities. *The American Oil & Gas Reporter*, Editor's Choice.

Honorio, J., Chen, C. H., Gao, G. H., Du, K. F. and Jaakkola, T. 2015. Integration of PCA with a Novel Machine Learning Method for Reparameterization and Assisted History Matching Geologically Complex Reservoirs. Paper SPE 175038 presented at SPE Annual Technical Conference and Exhibition, Houston, Texas, USA, 28-30 September. DOI: https://doi.org/10.2118/175038-MS

Ilk, D., Rushing, J. A., Perego, A. D., and Blasingame, T. A. 2008a. Exponential vs. Hyperbolic Decline in Tight Gas Sands – Understanding the Origin and Implications for Reserve Estimates Using Arps' Decline Curves. Paper SPE 116731 presented at the SPE Annual Technical Conference and Exhibition, Dever, 21-24 September, DOI: 10.2118/116731-MS.

Joshi, K. and Lee, J. W. 2013. Comparison of Various Deterministic Forecasting Techniques in Shale Gas Reservoirs. Paper SPE 163870 presented at SPE Hydraulic Fracturing Technology Conference, The Woodlands, Texas, 4-6 February, 2013. DOI: http://dx.doi.org/10.2118/163870-MS.

Kaviani, D., Bui, T., Jensen, J. L. and Hanks, C. 2008. The Application of Artificial Neural Networks with Small Data Sets: An Example for Analysis of Fracture Spacing in the Lisburne Formation, Northeastern Alaska. *SPE Reservoir Evaluation & Engineering*, 11(3), 598-605. DOI: https://doi.org/10.2118/103188-PA

Khanal, A., Khoshghadam, M., Lee, W. J. and Nikolaou, M. 2017. New Forecasting Method for Liquid Rich Shale Gas Condensate Reservoirs with Data Driven Approach Using Principal Component Analysis. *Journal of Natural Gas Science and Engineering*, 38, 621-637. DOI: https://doi.org/10.1016/j.jngse.2017.01.014

LaFollette, R. F., Izadi, G. and Zhong M. 2014. Application of Multivariate Statistical Modeling and Geographic Information Systems Pattern-Recognition Analysis to Production Results in the Eagle Ford Formation of South Texas. Paper SPE 168628 presented at SPE Hydraulic Fracturing Technology Conference, the Woodlands, Texas, USA, 4-6 February, DOI: https://doi.org/10.2118/168628-MS.

Lafollette, R., Holcomb, W. D. and Aragon, J. 2012. Practical Data Mining: Analysis of Barnett Shale Production Results with Emphasis on Well Completion and Fracture Stimulation. Paper SPE 152531 presented at SPE Hydraulic Fracturing Technology Conference, the Woodlands, Texas, USA, 6-8 February, DOI: https://doi.org/10.2118/152531-MS.

Landa, J. L., Kalia, R. K., Nakano, A., Nomura, K. and Vashishta, P. 2005. History Match and Associated Forecast Uncertainty Analysis – Practical Approaches Using Cluster Computing. *International Petroleum Technology Conference, Doha, Qatar*, DOI: https://doi.org/10.2523/IPTC-10751-MS

Lolon, E., Hamidieh, K., Weijers, L., Mayerhofer, M., Melcher, H. and Oduba, O., 2016. Evaluating the Relationship Between Well Parameters and Production Using Multivariate Statistical Models: A Middle Bakken and Three Forks Case History. Paper SPE 179171 presented at SPE Hydraulic Fracturing Technology Conference, the Woodlands, Texas, USA, 9-11 February, DOI: https://doi.org/10.2118/179171-MS.

Makinde, I. and Lee, W. J. 2016. Statistical Approach to Forecasting Gas-Oil Ratios and Solution Gas Production from Shale Volatile Oil Reservoirs. Paper SPE 182933 presented at the Abu Dhabi International Petroleum Exhibition and Conference held in Abu Dhabi, UAE, 7-10 November 2016. DOI: https://doi.org/10.2118/182933-MS

Malayalam, A., Bhokare, A., Plemons, P., Sebastian, H. and Abacioglu, Y. 2014. Multi-Disciplinary Integration for Lateral Length, Staging and Well Spacing Optimization in Unconventional Reservoirs. *SPE/AAPG/SEG Unconventional Resources Technology Conference, Denver, Colorado, USA.* DOI: https://doi.org/10.15530/URTEC-2014-1922270

McLane, M. and Gouveia, J. 2015. Validating Analog Production Type Curves for Resource Plays. Paper SPE 175527 presented at SPE Liquids-Rich Basins Conference – North America, Midland, Texas, USA, 2-3 September, DOI: https://doi.org/10.2118/175527-MS

Oliver, D. S. and Chen, Y. 2011. Recent Progress on Reservoir History Matching: a Review. *Computational Geosciences*, 15(1), 185-221.

Oliver, D. S., Reynolds, A. C. and Liu, N. 2008. Inverse Theory for Petroleum Reservoir Characterization and History Matching. Cambridge University Press.

Pradhan, Y. and Xiong H. J. 2018. Additional Applications on Determining Optimal Lateral Lengths and Trajectories on University Lands Delaware Basin. *SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, Texas, USA*. DOI: https://doi.org/10.15530/URTEC-2018-2902309

Rafiee-Taghanaki, S., Arabloo, M., Chamkalani, A., Amani, M., Zargari, M. H. and Adelzadeh, M. R. 2013. Implementation of SVM Framework to Estimate PVT Properties of Reservoir Oil. *Fluid Phase Equilibria*, 346, 25-32. DOI: https://doi.org/10.1016/j.fluid.2013.02.012

Ramirez, A. M., Valle, G. A., Romero, F. and Jaimes, M. 2017. Prediction of PVT Properties in Crude Oil Using Machine Learning Techniques MLT. Paper SPE 185536 presented at SPE Latin America and Caribbean Petroleum Engineering Conference, Buenos Aires, Argentina. DOI: https://doi.org/10.2118/185536-MS

Ramsay, J. O. and Silverman, B. W. 2005. Functional Data Analysis, *Springer*

Ramsay, J. O., Hooker, G. and Graves, S. 2009. Functional Data Analysis with R and Matlab, *Springer*

Ruths, T., Zawila, J., Fluckiger, S. D., Miller, N. J. and Gibson, R. G. 2017. New methodology merging seismic, geologic, and engineering data to predict completion performance. *The Leading Edge*, 36(3), 220-226.

Rwechungura, R. W., Dadashpour, M. and Kleppe J. 2011. Advanced History Matching Techniques Reviewed. Paper SPE 142497 presented at SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 25-28 September, DOI: https://doi.org/10.2118/142497-MS

Schuetter, J., Mishra, S., Zhong, M. and LaFollette, R. 2015. Data Analytics for Production Optimization in Unconventional Reservoirs. *SPE/AAPG/SEG Unconventional Resources Technology Conference, San Antonio, Texas*. DOI: https://doi.org/10.15530/URTEC-2015-2167005

Shyeh, J. J., Hehmeyer, O. J., Gibbeson, J., Mullins, J. J. and Trujillo, D. 2008, Examples Right-Time Decisions from High Frequency Data. Paper SPE 112150 presented at Intelligent Energy Conference and Exhibition, Amsterdam, The Netherlands, 25-27 February. DOI: https://doi.org/10.2118/112150-MS

Spivey, J. P., Frantz, J. H., Williamson, J. R. and Sawyer, W. K., 2001. Applications of the Transient Hyperbolic Exponent. Paper SPE71038 presented at the SPE Rocky Mountain Petroleum Technology Conference held in Keystone, Colorado, 21-23 May 2001.

Subrahmanya, N., Xu. P., EI-Bakry, A. and Reynolds, C. 2014. Advanced Machine Learning Methods for Production Data Pattern Recognition. Paper SPE 167839 presented at SPE Intelligent Energy Conference & Exhibition, Utrecht, The Netherlands, 1-3 April. DOI: https://doi.org/10.2118/167839-MS

Tahmasebi, P., Javadpour, F. and Sahimi, M. 2017. Data Mining and Machine Learning for Identifying Sweet Spots in Shale Reservoirs, *Expert Systems with Applications*, 88, 435-447

Tarrahi, M. and Shadravan, A. 2016. Inverse Modeling for Fluid System Characterization Through Machine Learning Algorithms. Paper SPE 180034 presented at SPE Bergen One Day Seminar, Grieghallen, Bergen, Norway, 20 April, DOI: https://doi.org/10.2118/180034-MS

Tavakoli, R., Srinivasan, S. and Wheeler, M. F. 2014. Rapid Updating of Stochastic Models by Use of an Ensemble-Filter Approach. *SPE Journal*, 19(03), 500-513. DOI: https://doi.org/10.2118/163673-PA.

Tibshirani, R. J., Taylor, J. 2011. The Solution Path of the Generalized Lasso. *The Annals of Statistics*, 3, 1335-1371.

Valko P. P. and Lee W. J. 2010. A Better Way to Forecast Production from Unconventional Gas Wells. Paper SPE 134231 presented at SPE Annual Technical Conference and Exhibition, Florence, Italy, 19-22 September.

Vera, F., Lemons, C. R., Zhong, M., Holcomb, B. and Lafollette, R. F. 2015. Multidisciplinary Approach in the Permian Basin: A Geological, Statistical and Engineering Case Study to Production Results on the Wichita-Albany Formation, SPE 173352 presented at SPE Hydraulic Fracturing Technology Conference, the Woodlands, Texas, USA, 3-5 February, DOI: https://doi.org/10.2118/173352-MS

Vink, J. C., Gao, G. H. and Chen, C. H., 2015. Bayesian Style History Matching: Another Way to Under-Estimate Forecast Uncertainty? Paper SPE 175121 presented at SPE Annual Technical Conference and Exhibition, Houston, Texas, USA, 28-30 September, DOI: https://doi.org/10.2118/175121-MS

Wood, S. N. 2006. Low Rank Scale Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics*, 62(4), 1025-1036.

Yuan, G., Dwivedi, P., Kwok, C. K. and Malpani, R. 2017. The Impact of Increase in Lateral Length on Production Performance of Horizontal Shale Wells. Paper SPE 185768 presented at SPE Europec featured at 79th EAGE Conference and Exhibition, Paris, France. DOI: https://doi.org/10.2118/185768-MS

Zendehboudi, S., Ahmadi, M. A., James, L. and Chatzis, L. 2012. Prediction of Condensate-to-Gas Ratio for Retrograde Gas Condensate Reservoirs Using Artificial Neural Network with Particle Swarm Optimization. *Energy Fuels*, 26(6), 3432-3447. DOI: https://doi.org/10.1021/ef300443j

Zhang, H., Cocco, M., Rietz, D., Cagle, A., and Lee, W. J. 2015. An Empirical Exponential Decline Curve for Shale Reservoirs. SPE paper 175016 presented at the SPE Annual Conference & amp; Exhibition, Houston, TX, 28-30 September 2015, DOI: https://dx.doi.org/10.2118/175016-MS.

Zhang, H., Nelson, E., Olds, D., Rietz, D. and Lee, W. J. 2016. Effective Applications of Extended Exponential Decline Curve Analysis to both Conventional and Unconventional Reservoirs. SPE paper 181536 presented at the SPE Annual Technical Conference and Exhibition, Dubai, UAE, 26-28 September. DOI: https://doi.org/10.2118/181536-MS.

Zhong, M., Schuetter, J., Mishra, S. and Lafollette, R. F. 2015. Do Data Mining Methods Matter?: A Wolfcamp Shale Case Study. Paper SPE 173334 presented at SPE Hydraulic Fracturing Technology Conference, the Woodlands, Texas, USA, 3-5 February, DOI: https://doi.org/10.2118/173334-MS

Zhou, P., Sang, H. Y., Jin, L. Y. and Lee, W. J. 2017. Application of Statistical Methods to Predict Production From Liquid-Rich Shale Reservoirs. *SPE/AAPG/SEG Unconventional Resources Technology Conference, Austin, Texas, USA*. DOI: https://doi.org/10.15530/URTEC-2017-2694668

Zhou, P., Pan, Y. W., Sang, H. Y. and Lee, W. J. 2018. Criteria for Proper Production Decline Models and Algorithm for Decline Curve Parameter Inference. *SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, Texas, USA*. DOI: https://doi.org/10.15530/URTEC-2018-2903078

APPENDIX A

DERIVATION OF $\boldsymbol{k_{PT}}$ AND $\boldsymbol{k_{MBT}}$

In this appendix, we will show the derivation for Eq. (2.12). According to the definition of k_{MBT}, we have

$$k_{MBT} = -\frac{d\ln q}{d\ln t_{MBT}} = -\frac{d\ln q}{d\,t}\frac{dt}{d\ln t_{MBT}} = D(t)\ \frac{dt}{d\ln t_{MBT}} \tag{A.1}$$

where $D(t)$ denotes the production decline rate at time t.

Since

$$\frac{d\ln t_{MBT}}{dt} = \frac{1}{t_{MBT}}\left(1 - \frac{N_p}{q(t)^2}\frac{dq(t)}{dt}\right) = \frac{1}{t_{MBT}} + D(t) \tag{A.2}$$

where N_p denotes the cumulative production at time t.

Then

$$k_{MBT} = \frac{D(t)t_{MBT}}{1+D(t)t_{MBT}} \tag{A.3}$$

If the production decline model is two-segment hyperbolic model with $b_i > 1$ and $0 < b_f < 1$, then we have the equation for the decline rate as follows

$$D(t) = \begin{cases} \frac{D_i}{1+D_i b_i t}; & t \leq t_c \\ \frac{D_i}{1+D_i b_i t_c + D_i b_f (t-t_c)}; & t > t_c \end{cases} \tag{A.4}$$

Since the cumulative production and production rate at time $t \leq t_c$ is

$$N_p = \frac{q_i^b}{(b_i-1)D_i}\left[q^{1-b_i} - q_i^{1-b_i}\right] \tag{A.5}$$

$$q(t) = \frac{q_i}{(1+D_i b_i t)^{\frac{1}{b_i}}} \tag{A.6}$$

The equation of t_{MBT} at time $t \leq t_c$ is

$$t_{MBT}(t) = \frac{N_p}{q(t)} = \frac{1}{(b_i-1)D_i}\left[(1+D_i b_i t) - (1+D_i b_i t)^{\frac{1}{b_i}}\right] \tag{A.7}$$

Thus, at time $t \leq t_c$ we have

$$k_{MBT} = \frac{1-(1+D_i b_i t)^{\frac{1}{b_i}-1}}{b_i-(1+D_i b_i t)^{\frac{1}{b_i}-1}}; t \leq t_c \tag{A.8}$$

At $t > t_c$ the equation of cumulative production and production rate is

$$N_p = \frac{q_i^{b_i}}{(b_i-1)D_i}\left[q_{t_c}^{1-b_i} - q_i^{1-b_i}\right] + \frac{q_{t_c}^{b_f}}{(b_f-1)D_{t_c}}\left[q^{1-b_f} - q_{t_c}^{1-b_f}\right] \tag{A.9}$$

$$q(t) = \frac{q_{t_c}}{\left(1+D_{t_c}b_f(t-t_c)\right)^{\frac{1}{b_f}}} \qquad \text{(A.10)}$$

where q_{t_c} denotes the production rate at the switching time t_c given as follows

$$q_{t_c} = \frac{q_i}{(1+D_i b_i t_c)^{\frac{1}{b_i}}} \qquad \text{(A.11)}$$

Thus, the equation of t_{MBT} at $t > t_c$ is

$$t_{MBT} = \frac{\left(1+D_{t_c}b_f(t-t_c)\right)^{\frac{1}{b_f}-1}}{(b_i-1)D(t)}\left[1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}\right] + \frac{1}{(b_f-1)D(t)}\left[1-\left(1+D_{t_c}b_f(t-t_c)\right)^{\frac{1}{b_f}-1}\right] \qquad \text{(A.12)}$$

Then at $t > t_c$ the slope k_{MBT} is

$$k_{MBT} = \frac{\frac{b_f-1}{b_i-1}\left[1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}\right]-1+\quad\left(1+D_{t_c}b_f(t-t_c)\right)^{-\frac{1}{b_f}+1}}{\frac{b_f-1}{b_i-1}\left[1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}\right]-1+b_f\left(1+D_{t_c}b_f(t-t_c)\right)^{-\frac{1}{b_f}+1}} \qquad \text{(A.13)}$$

APPENDIX B

INEQUALITY RELATION BETWEEN $\boldsymbol{k_{PT}}$ AND $\boldsymbol{k_{MBT}}$

In this appendix, we will show that in the diagnostic plot, at any time t the slope of PT curve k_{PT} is always greater than or equal to the slope of MBT curve k_{MBT}. Here we assume $b_i > 1$ and $0 < b_f < 1$. The proof for the special case $b_i = 1$ or $b_f = 0$ follows the similar argument as below.

Proof:

Part 1:

When $t \leq t_c$, we have

$$k_{PT} = \frac{1}{\frac{1}{D_i t} + b_i} \tag{B.1}$$

$$k_{MBT} = \frac{1-(1+D_i b_i t)^{\frac{1}{b_i}-1}}{b_i-(1+D_i b_i t)^{\frac{1}{b_i}-1}} \tag{B.2}$$

When $b_i > 1$, we have

$$(1 + D_i b_i t)^{\frac{1}{b_i}-1} \leq 1 \text{ for all } t$$

Therefore, to prove that $k_{PT} \geq k_{MBT}$, it is equivalent to proof the following inequality

$$\left(\frac{1}{D_i t} + b_i\right)\left[1 - (1 + D_i b_i t)^{\frac{1}{b_i}-1}\right] \leq b_i - (1 + D_i b_i t)^{\frac{1}{b_i}-1} \tag{B.3}$$

Furthermore, the inequality (B.3) is equivalent to the following equality:

$$(1 + D_i b_i t - D_i t)(1 + D_i b_i t)^{\frac{1}{b_i}-1} \geq 1 \tag{B.4}$$

Let $g(t) = [1 + D_i(b_i - 1)t](1 + D_i b_i t)^{\frac{1}{b_i}-1}$, then the first derivative of $g(t)$ is

$$g'(t) = D_i^2(b_i - 1)(1 + D_i b_i t)^{\frac{1}{b_i}-2} t \geq 0 \text{ for all } t \tag{B.5}$$

Since the first derivative of $g(t)$ is non-negative, the minimum value of $g(t)$ is $g(0) = 1$. Therefore, the inequality (B.4) is true and we have $k_{PT} \geq k_{MBT}$ at time $t \leq t_c$.

Part 2:

When $t \geq t_c$, we have

$$k_{PT} = \frac{1}{\frac{1-D_{t_c} b_f t_c}{D_{t_c} t} + b_f} \tag{B.6}$$

$$k_{MBT} = \frac{\frac{b_f - 1}{b_i - 1}\left[1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}\right] - 1 + \quad [1 + D_{t_c} b_f (t - t_c)]^{-\frac{1}{b_f}+1}}{\frac{b_f - 1}{b_i - 1}\left[1-(1+D_i b_i t_c)^{\frac{1}{b_i}-1}\right] - 1 + b_f[1 + D_{t_c} b_f (t - t_c)]^{-\frac{1}{b_f}+1}} \tag{B.7}$$

Let

$$\frac{b_f-1}{b_i-1}\left[1-(1+D_ib_it_c)^{\frac{1}{b_i}-1}\right]-1=\alpha \tag{B.8}$$

$$\frac{1-D_{t_c}b_ft_c}{D_{t_c}}=\frac{1}{\beta} \tag{B.9}$$

Then when $0<b_f<1$, we have $\alpha<-1$, $\beta>0$ and

$$\alpha+\left[1+D_{t_c}b_f(t-t_c)\right]^{-\frac{1}{b_f}+1}<0$$

Therefore, to prove $k_{PT}\geq k_{MBT}$, we only need to prove the following inequality:

$$\left(\frac{1}{\beta t}+b_f\right)\left(\alpha+\left[1+D_{t_c}b_f(t-t_c)\right]^{-\frac{1}{b_f}+1}\right)\geq\alpha+b_f\left[1+D_{t_c}b_f(t-t_c)\right]^{-\frac{1}{b_f}+1} \tag{B.10}$$

Th inequality (B.10) is equivalent to the following inequality:

$$\frac{1}{\alpha}\leq\left(\beta t-b_f\beta t-1\right)\left[1+D_{t_c}b_f(t-t_c)\right]^{\frac{1}{b_f}-1} \tag{B.11}$$

Let

$$h(t) = \left(\beta t - b_f \beta t - 1\right)\left[1 + D_{t_c} b_f (t - t_c)\right]^{\frac{1}{b_f} - 1}$$

Then the first derivative of $h(t)$ is

$$h'(t) = \frac{D_{t_c}^2 t(1 - b_f)}{1 - D_{t_c} b_f t_c}\left[1 + D_{t_c} b_f (t - t_c)\right]^{\frac{1}{b_f} - 2} > 0 \text{ for all } t > t_c$$

The minimum value of $h(t)$ is $h(t_c) = \frac{D_{t_c} t_c - 1}{1 - D_{t_c} b_f t_c}$ and with some simple algebra we can easily show that it is greater than $\frac{1}{\alpha}$. Thus, the inequality (B.11) is correct and $k_{PT} > k_{MBT}$ when $t > t_c$. This complete the proof that $k_{PT} \geq k_{MBT}$ for all time t.

APPENDIX C

COMPARISON BETWEEN DIFFERENT APPROACHES

A comparison of different approaches (fPCA, modified Arps, extended exponential):

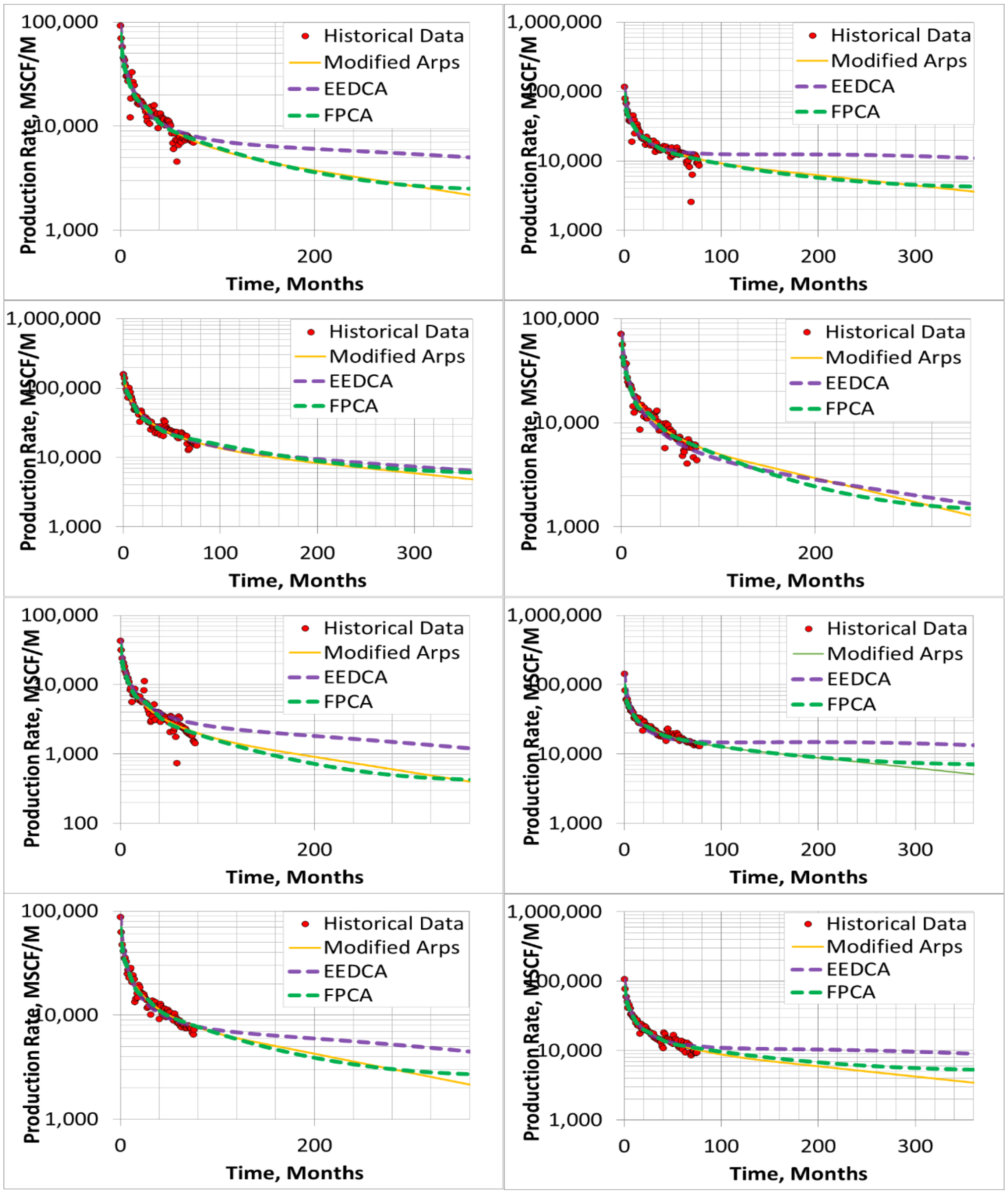


Figure 26: Production decline analysis of gas wells in Eagle Ford (Zhou *et al.* 2017)

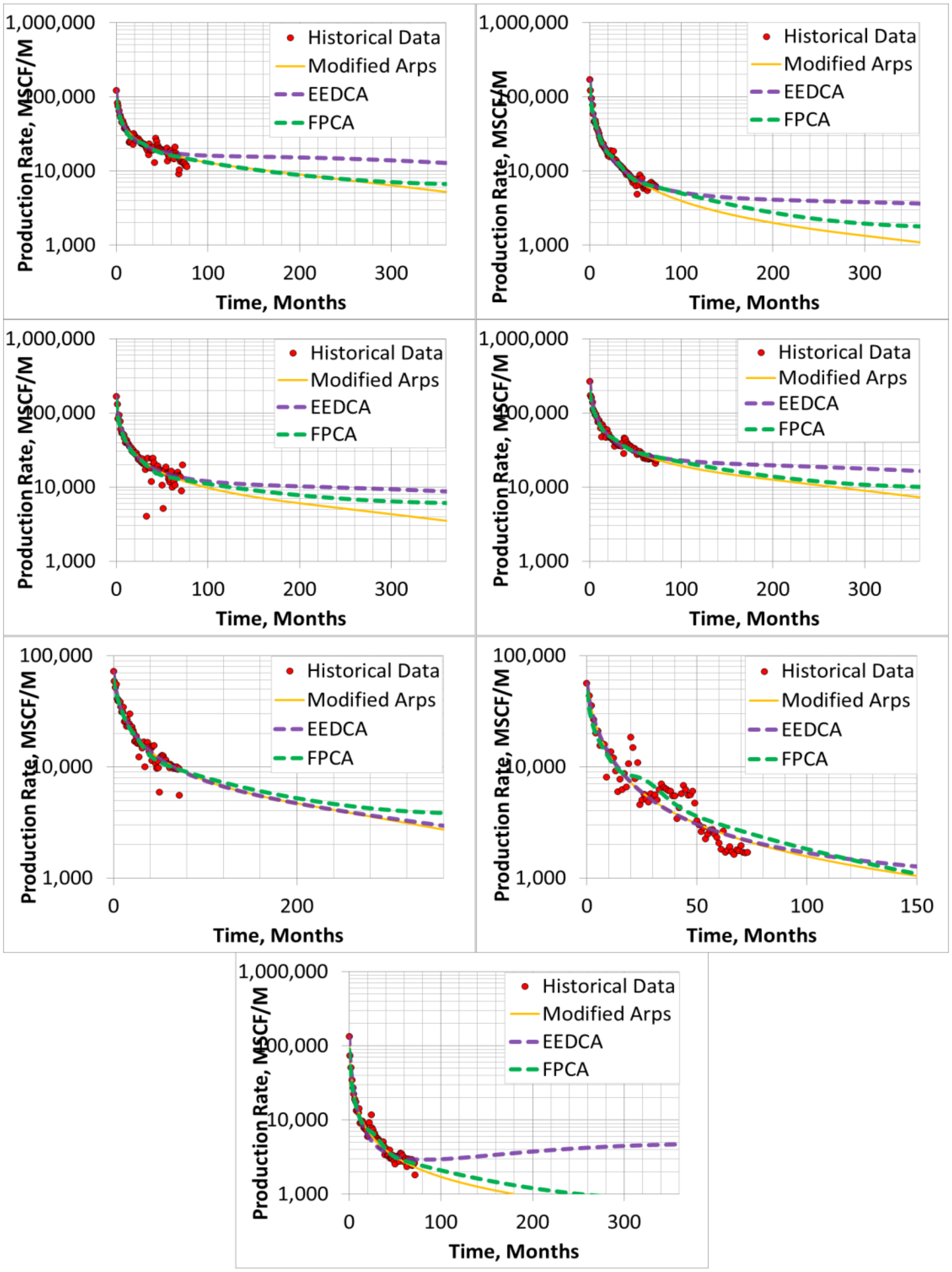


Figure 26: Continued

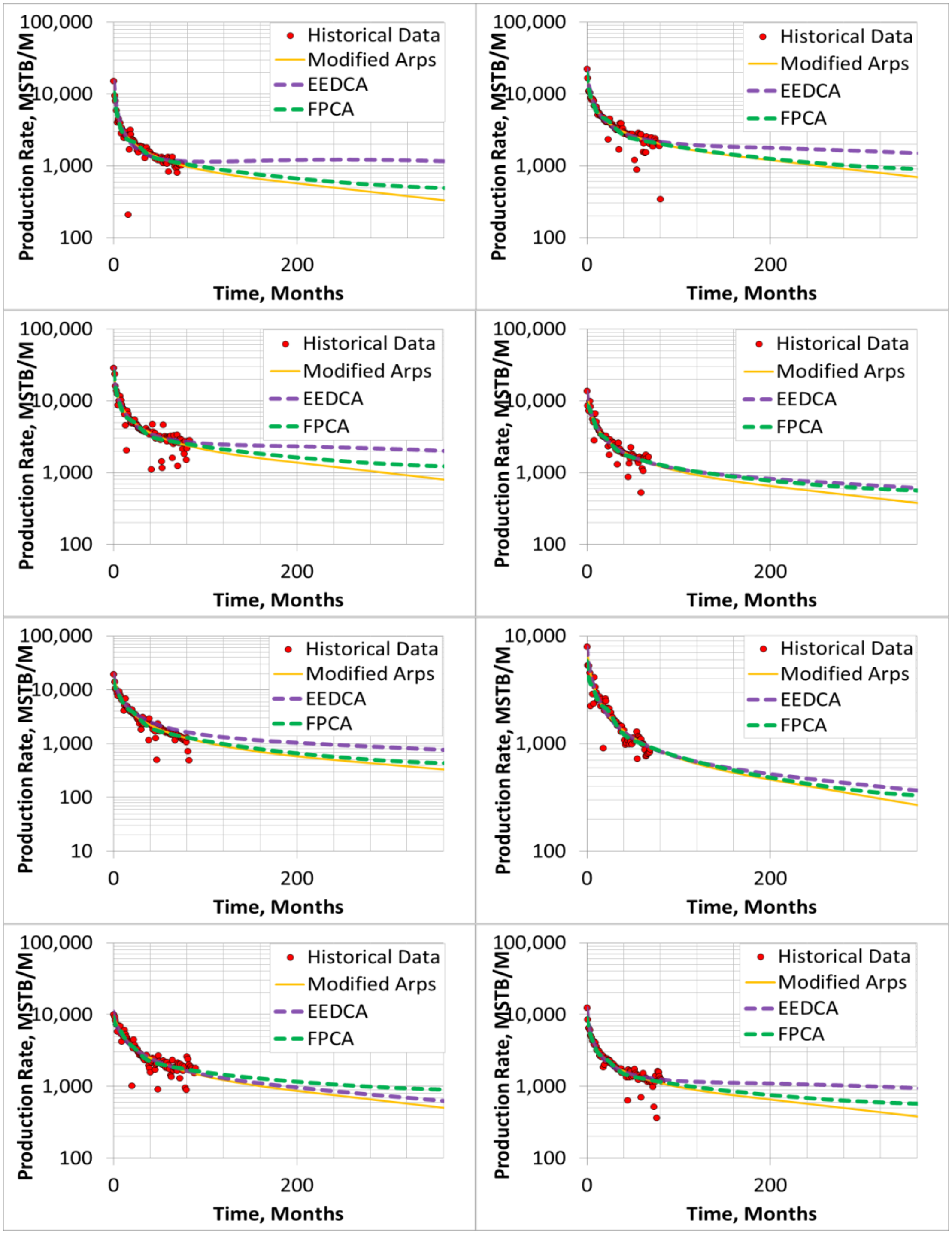


Figure 27: Production decline analysis of oil wells in Bakken (Zhou *et al.* 2017)

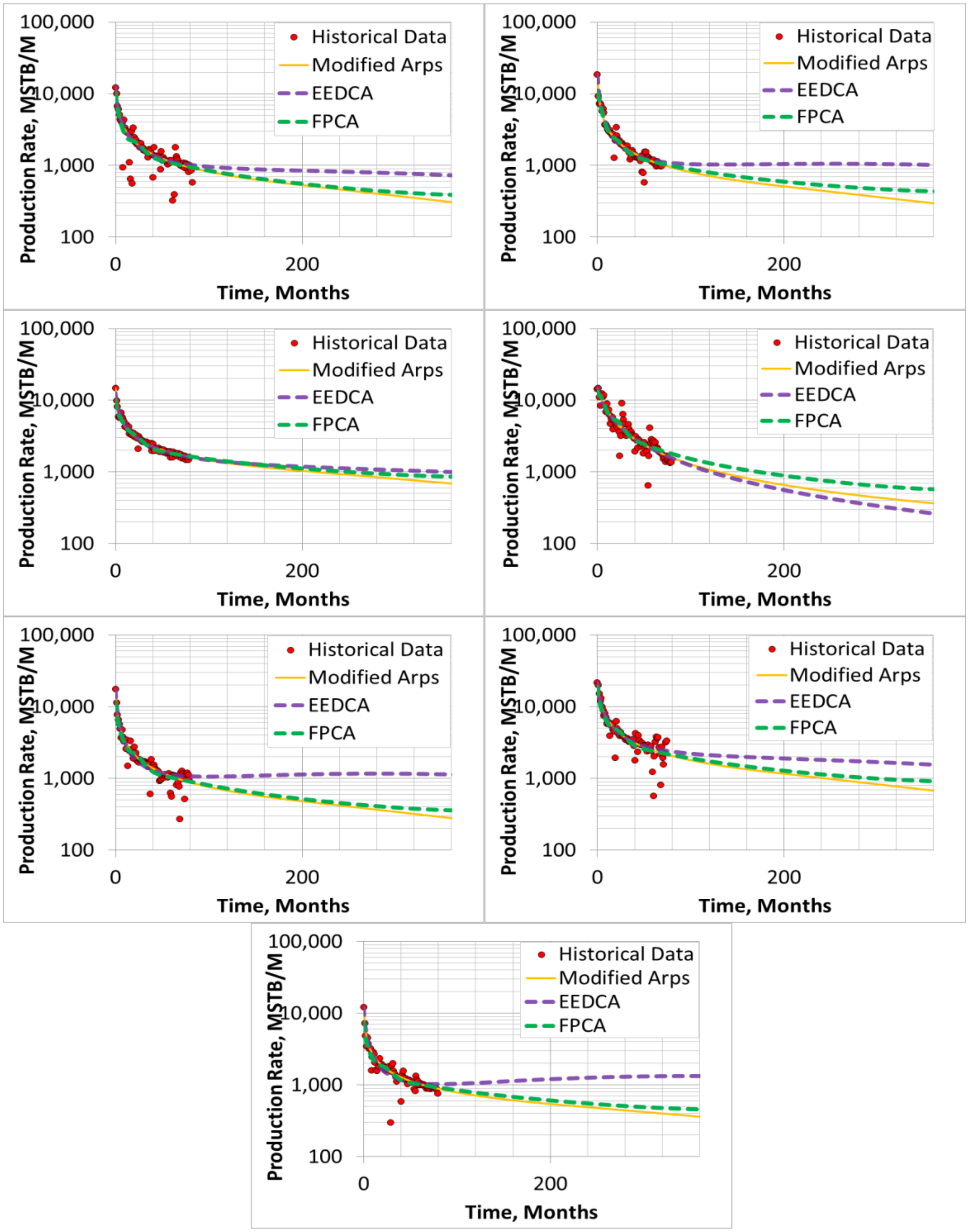


Figure 27: Continued