THE APPLICATION OF PRINCIPAL COMPONENT ANALYSIS IN PRODUCTION FORECASTING

A Thesis

by

YUYANG 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

MASTER OF SCIENCE

Chair of Committee, W. John Lee
Committee Members, Maria A. Barrufet
Duane A. McVay
Head of Department, A. Daniel Hill

December 2017

Major Subject: Petroleum Engineering

ABSTRACT

Current methods of production forecasting, such as Decline Curve Analysis and Rate Transient Analysis, require years of production data, and their accuracy is affected by the artificial choice of model parameters. Unconventional resources, which usually lack long-term production history and have hard-to-determine model parameters, challenge traditional methods.

This paper proposes a new method using principal components Analysis to estimate production with reasonable certainty. PCA is a statistical tool which unveils the hidden patterns of production by reducing high-dimension rate-time data into a linear combination of only a few principal components.

This paper establishes a PCA-based predictive model which makes predictions by using information from the first few months' production data from a well. Its efficacy has been examined with both simulation data and field data.

Also, this study shows that the K-means clustering technique can enhance predictive model performance and give a reasonably certain future production range estimate based on historical data.

DEDICATION

To my parents and my family

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. John Lee and my committee members, Dr. Duane McVay and Dr. Maria Barrufet, for their support of my research. Dr. Lee illustrated and led me to the gate of reserves estimation and machine learning. I truly benefited from his endless wisdom and kind personality. It is always my honor to be his student.

Special thanks to my parents for their encouragement and support. I can feel their love even though we are separated by the vast ocean.

I appreciate those who come to my life and then leave. I appreciate the support given by my lab mates, roommates, and friends here. I appreciate the kindness, warmness, and care that I feel in the United States.

It has been a unique and astonishing experience to spend the beginning of my 20s in the Harold Vance Department of Petroleum Engineering. Here, I managed to find out who I want to be and where I wish to live. Here, anyone like me has a chance to learn from world-class experts. Here, Aggie spirit and fearlessness on every front are no longer slogans, but guidelines that inspire me every day.

Always forward. Gig'em!

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professors John Lee and Duane McVay of the Harold Vance Department of Petroleum Engineering and Professor Maria A.Barrufet of the Artie McFerrin Department of Chemical Engineering.

The data analyzed for Chapters III and IV was provided by lab mate Yunan Li, Peng Zhou and DrillingInfo.

I independently completed all other work conducted for the thesis.

Funding Sources

Graduate study was supported by a fellowship from Harold Vance Department of Petroleum Engineering

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

The unconventional resources revolution is the biggest energy story in the 21st century (Wang et al. 2014). The application of hydraulic fracturing and horizontal drilling makes possible unconventional oil and gas extraction from extremely low-permeability reservoirs (Arthur, Langhus, and Alleman 2008). Rapidly developing commercial projects increase the need for proper production forecasting and reserves estimation techniques for unconventional resources (Walsh et al. 2009).

Current production forecasting approaches include decline curve analysis (DCA), type curve analysis, analytical/numerical reservoir simulation, and flow regime analysis (Clarkson 2013). DCA methods were first established to estimate conventional resources (Arps 1945). In attempts to modify Arps' model to accommodate unconventional resources (Long and Davis 1987), new models have been proposed such as Duong's model (Duong 2011).

Type curve methods introduced by Fetkovich (1987) compared pressure or decline curves with predefined type curves. Later engineers improved these curves to fit hydraulically fractured reservoirs (Agarwal et al. 1998, Araya and Ozkan 2002, Fraim 1987, Marhaendrajana and Blasingame 2013).

Numerical simulation generates production forecasting by simulating hydrocarbon flowing conditions. Oil companies widely use it in hydraulically fractured reservoirs,

coalbed methane and many other unconventional resources (Cipolla et al. 2010, Aanonsen et al. 2009, Fan, Thompson, and Robinson 2010, Floris et al. 2001, Soleng 1999).

Straight-line analysis or flow-regime analysis helps determine key formation parameters. By plotting the logarithm of production versus time, it can define the flow regime by analyzing a straight-line segment of the plot (Araya and Ozkan 2002, Clarkson 2013, Cox et al. 2015, Ilk et al. 2010, Lee, Rollins, and Spivey 2003) .

Recently, the trending concept of data mining and machine learning has gained attention in the oil and gas industry. Researchers have tried to use techniques such as fuzzy neural networks, a support vector machine, a K-nearest neighbor algorithm, and principal components Analysis to analyze production data, geological setting, and reservoir characteristics and do production forecasting (Bravo et al. 2014, Cao et al. 2016 , Denney 2015, Floris et al. 2001, Khazaeni and Mohaghegh 2013, Moridis et al. 2013, Soleng 1999, Bhattacharya and Nikolaou 2013, Duong 2011, Honorio et al. 2015)

This research applies the principal components analysis (PCA) algorithm into unconventional gas production data analysis. By applying this mature, widely applied algorithm into unconventional gas rate-time data, we can better understand data and make predictions. PCA can lower data dimensions, perform clustering, and do factor analysis. By combing PC linearly and doing a regression called principal components regression (PCR), we can capture the well's decline trend with much shorter production time data while not losing accuracy.

1.1 Production Data Analysis

Oilfield production can generate different types of data including rate-time data, pressure data, and well log data. Those data can be used to describe reservoir characteristic and further predict production, estimate reserves, and enhance well performance (Ilk et al. 2010).

Engineers can use analytic production data in either an analytical way or an empirical way (Clarkson 2013). Analytical ways including rate-transient analysis (RTA) and pressure-transient analysis (PTA) are performed on rate-time data/reservoir-flow pressure data. Empirical ways including decline curve analysis (DCA) and type curves are performed by fitting curves to past rate-time decline trends.

In this section, we focus on reviewing analysis methods on rate-time data, because rate-time data is the most available data. The first step of data processing begins at collecting consistently and reliable data. In some situations, such as short well life or initial production stages, certain types of data are not acquirable (Cheng et al. 2010). For example, in fields, pressure data is often incomplete, unreliable, and absent from daily records (Ilk et al. 2010). In comparison, rate-time data can have accurate, coherent and reliable records from daily or monthly sales sheets.

Decline curve analysis is a classical and common practice which analyzes the decline trending of a well's production history to predict future production and Estimated Ultimate Recovery (EUR). This method was first introduced by the U.S. Internal Revenue Service (Arnold and Darnell 1920) and then after several improvements (Lewis and Beal 1918, Cutler 1924, Johnson and Bollens 1927), was finally well established by Arps (1945).

$$q_t = \frac{q_i}{(1+bD_it)^{\frac{1}{b}}} \tag{1-1}$$

In Arp's equation, q_i is the initial production rate, q_t is time *t* production rate, and D_i and *b* are constant parameters. *D* is defined as the loss ratio and *b* is defined as a loss-ratio derivative.

$$\frac{1}{D} = -\frac{q}{dq/dt} \tag{1-2}$$

With *b* setting as 0, 1 or any value between them, the equation represents an exponential, harmonic, or hyperbolic curve.

$$b = \frac{d}{dt}\left[-\frac{q}{dq/dt}\right] \tag{1-3}$$

First set up for conventional oil and gas, the *b* value should not exceed the upper limit of 1. Different *b* value settings also reflect different reservoir characteristics and their drive mechanisms (Bhattacharya and Nikolaou 2013).

With the exploration of unconventional resources, the industry began to use larger-than-unit *b* values to fit the decline curve (Long and Davis 1987), in what is called a *superhyperbolic* equation. These adaptation attempts catch the unique decline patterns in shale reservoirs: the superhyperbolic equation fits the rapid, steep initial decline trend and

an exponential equation fits the gentle decline in later boundary-dominated-flow (BDF) stages. It somehow causes trouble and inconvenience for reservoir engineers to predict long-time well production.

A sample is shown below. The initial segment curve is fitted with a larger-than-unit *b* value to catch the steep decline trends. When shale gas is produced in transient flow, the rate declines rapidly. After it reaches the boundary, it goes to BDF. The production rate would turn shallower and can then be fitted with an exponential *b* value. This is shown in Figure 1-1.

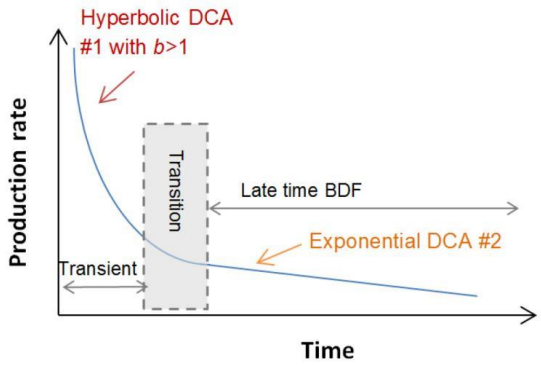


Figure 1-1: Two segment *b*-value in shale resource DCA (Zhang et al. 2015)

Ilk et al. (2008) proposed a power law exponential decline model (PLE). This model gives a new definition to the *D* and *b* values to fit the decline trends of shale resources. In this model, he used a power-law function to model the initial-stage decline-loss ratio. Then, the loss ratio would go constantly.

$$\boldsymbol{D = D_{\infty} + D_1 t^{-(1-n)}} \tag{1-3}$$

$$\boldsymbol{b = \frac{-D_1(n-1)t^n}{(D_{\infty}t+D_1t^n)^2}} \tag{1-4}$$

The production rate could be expressed as:

$$\boldsymbol{q = q_i \exp(-D_\infty t - \frac{D_1}{n} t^n)} \quad (1\text{-}5)$$

Valkó and Lee (2010) proposed stretched exponential decline model (SEPD). This model is designed to estimate technically recoverable hydrocarbons. It points out two new parameters: a dimensionless exponent *n* and the ratio of time, τ.

$$\boldsymbol{\tau = (\frac{n}{D_1})^{\frac{1}{n}}} \quad (1\text{-}6)$$

For rate-time, it has following equations:

$$\boldsymbol{q = q_i \exp[-\frac{t^n}{\tau}]} \quad (1\text{-}7)$$

For cumulative production, it has following equations:

$$\boldsymbol{Q = \frac{q_i \tau}{n}\{\Gamma\left[\frac{1}{n}\right] - \Gamma\left[\frac{1}{n}, \left(\frac{t}{\tau}\right)^n\right]\}} \quad (1\text{-}8)$$

Duong (2011) proposed a new DCA method that focuses on the transient flow period. He asserted that in traditional practice, Arps' model would not work for shale resources, but proposed instead a log-log plot method that generates a unit straight-line slope to determine initial rate q_i and infinite rate, q_∞. He defined new time parameters *t(a,m).* The equation of his methods is written as follows:

$$\boldsymbol{t(a,m) = t^{-m} \exp[\frac{a}{1-m}(t^{1-m} - 1)]} \quad (1\text{-}9)$$

The rate time equation could be written as:

$$\boldsymbol{q = q_1 t(a,m) + q_\infty} \quad (1\text{-}10)$$

Zhang et al. (2015) argued that Arps methods could generate more valid production forecasts in certain reservoir and flow regimes (e.g., tight reservoirs in BDF flow). Their practice is to apply a combination of methods. Before setting D_{min}, a superhyperbolic method or Duong method is applied to acquire an estimate in the transient flow regime. After the switching point, usually Arps' exponential model would be applied to estimate production in the BDF flow regime.

1.2 Computer-Assisted Production Data Analysis

Two revolutions happened at the same time: shale gas and Artificial Intelligence (AI). Shale gas evaluation presents challenges and opportunity to the traditional production data analysis. As was argued above, the classic Arps model has difficulties in proper evaluation if long-term well performance. At the same time, the characteristics of unconventional resources also present new challenges: heterogeneous formations, extremely low permeability, and unknown flow mechanism (Cao et al. 2016). New methods are needed to satisfy increasing need and new demands.

AI brings the world potentials for solving this. Big data, machine learning, deep learning, data mining—all those terms have become big hits in recent years. Businessmen, engineers, and operators wish to maximize the power of data into design, construction, drilling and production. AI can do those because it is strong at pattern recognition and automatic data processing where humans might require years-long experience and human bias.

The oil and gas industry is not unfamiliar with AI or computer-assisted production analysis approaches. In the last century, Dakshindas, Ertekin, and Grader (1999) proposed a way to combine AI with well testing. Bradley (1994) proposed a computer-assisted oil field economic forecasting method. Surguchev and Li (2000) and Alvarado et al. (2002) did extensive work in combing machine learning and neural networks with enhanced oil recovery (EOR). Brown (1991) used machine learning to study recovery efficiency.

AI methods have gained extensive attention from both academia and industry. Fuzzy logic and neural networks serve as powerful tools for analyzing unconventional resources (Grujic, Mohaghegh, and Bromhal 2010, Kulga 2010, Sondergeld et al. 2010, Clarkson et al. 2012, Keshavarzi and Jahanbakhshi 2013). Machine learning has also regained attention. Its power in recognizing patterns and its rapid processing have been acknowledged and accepted. As the Figure 1-2 indicates, there is a dramatically increasing trend in publishing machine learning-related papers in SPE associated conference and journals.

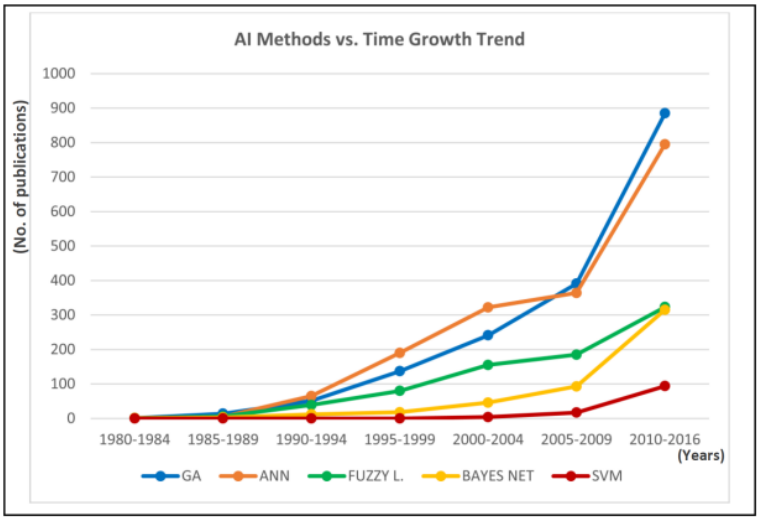


Figure 1-2: Each Year AI Related Literature in OnePetro (Ani et al. 2016)

A detailed literature review provides an understanding of future performance, reservoir characteristics, production forecasting and reserve estimation. This review includes the following factors: neural networks, machine learning/deep learning, time series analysis and multivariate statistical approaches.

1.2.1 Neural Networks

A neural network (NN) or artificial neural network (NN) is a computation model that enables the computer to mimic the way the human brain works. By connecting different processing neural units, the whole system can respond to input data and self-adapt its inner connection structure. This feature is beneficial for reservoir engineering to analyze the performance of hydrocarbon resources. NN can respond to the historical performance data

and catch hidden patterns or trends beneath the raw data. Further, it can be applied to do fitting, regression or predictions (Figure 1-3).

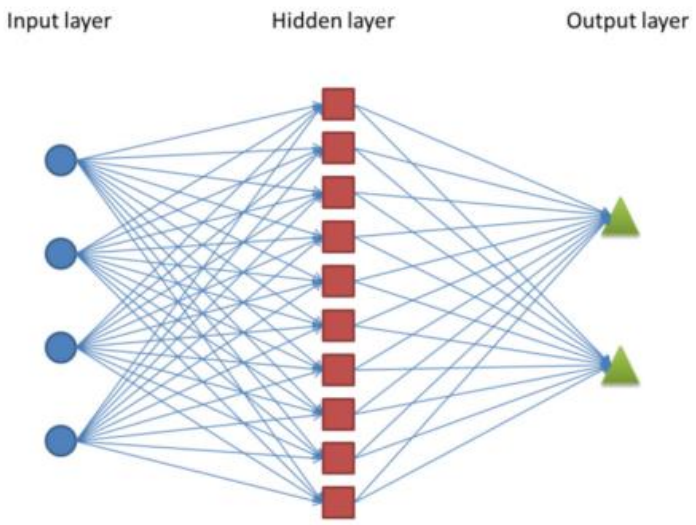


Figure 1-3: Layers Neural Network structure (Ma et al. 2015)

Mohaghegh (1995) asserted the potentials of neural networks in predicting well performance. Al-Fattah and Startzman (2001) first proposed three-layer neural network methods to predict U.S. natural gas production. At the same time, Texas A&M researchers, He et al. (2001) also introduced an NN method to forecast oil well performance based on historical performance. Queipo, Goicochea, and Pintos (2002) extended the application of ANN into steam-assisted gravity drainage (SAGD) production predictions. Lechner and Zangl (2005) combined Monte-Carlo simulation and ANN to assess the uncertainty of reservoir performance.

Bansal et al. (2013) applied ANN to predict well performance from discontinuous tight oil reservoirs. They found that it could help enhance tight oil development and avoid drilling less-productive wells. Rebeschini et al. (2013) improve ANN with nodal and time-series analysis to deal with real field data and acquire short-time production forecasts.

Combinations of methods also have been proposed to coordinate neural networks with other classic methods to gain deep insight into data. Jia and Zhang (2016) combined NN with traditional Arps decline curve analysis. Ma et al. (2015) combined ANN with principal components analysis, cluster analysis, and uncertainty analysis to predict SAGD well performance.

Like other data-mining techniques, NN does encounter some disadvantages at large data sets requirements. Mohaghegh et al. (2011), Oliver and Chen (2011), and Rwechungura, Dadashpour, and Kleppe (2011) gave some opinions on its limitations. The training of neural networks requires large amounts of effort and time to have the optimal parameters. The data need more than five years' production history and around 40 wells. Those disadvantages limit the real commercial application of neural networks to shale gas production forecasting.

1.2.2 Machine Learning/Deep Learning

Machine learning is a combination of algorithms that share a common characteristic: learning from data. It has supervised learning, unsupervised learning and reinforcement learning (Figure 1-4).

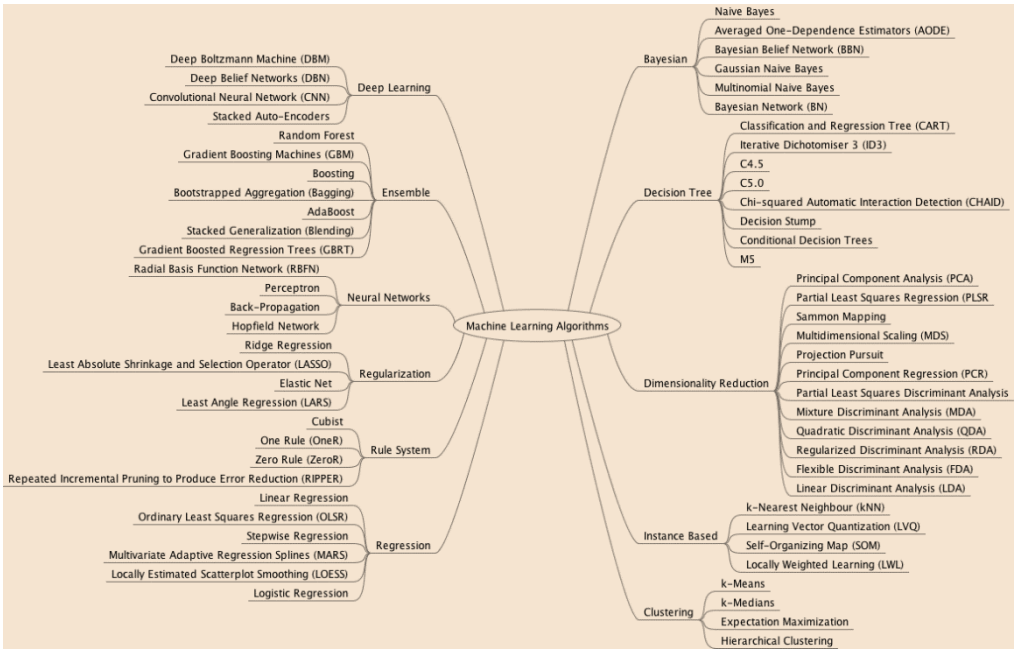


Figure 1-4: Machine learning algorithm overview (Brownlee 2013)

Fulford et al. (2016) proved machine learning is a reliable technology for evaluating rate-time performance in unconventional wells. Machine learning can serve as a reliable technology accepted by the U.S. Securities and Exchange Commission (SEC).

This work concludes that machine learning can largely enhance the production forecast process in a wide range of unconventional wells. This research soundly proves the great potential of machine learning for reserves engineers and oil companies.

Currently, academia is focused on some potential supervised learning algorithms. Researchers at Texas A&M University (Gong et al. 2014) combined Markov-chain Monte Carlo (MCMC) with probabilistic DCA to analyze forecast uncertainties. MCMC has served as a tool to examine the difference between prediction results and real data. The difference is measured by a statistical parameter called Bayesian inference.

Gonzalez, Gong, and McVay (2012) also applied MCMC techniques with probabilistic DCA to analyze shale gas reserves. This work used prior distribution to calibrate the posterior distribution. In this way, they could acquire desirable long-term production forecasting. They recommended production time of at least 18 months.

Crnkovic-Friis and Erlandson (2015) analyzed 800+ wells with more than 200,000+ geological data input using deep neural networks (DNN). The results, validated by data from the Eagle Ford Shale, were quite promising—significantly better than other methods in volumetric estimates and type curve predictions. These results demonstrate the potential of applying DNN into handling large amounts of data and give admirable predictions with a fast process.

Honorio et al. (2015) presented a way of integrating plur-principal components analysis (P-PCA) with piecewise reconstruction from a dictionary (PRaD) into assisted history matching. Their workflow would be practical to handle real problems.

1.2.3 Multivariate Statistical Analysis

The multivariate statistical approach applies statistical methods to analyze variable probability distribution, do classification, run clustering and analyze patterns. Those methods can be useful for recognizing data structures and unveil their hidden patterns. Traditional methods can combine with a multivariable statistical approach to have better process efficiency or accuracy Bhattacharya and Nikolaou (2013).

Multivariate statistical approaches have some advantages. First, their process does not require time-consuming and laborious human work to train and get optimal parameters. For most techniques, they can be done automatically, which is called *unsupervised learning*. For dealing with large sets of unconventional resources data, automatic processing is important to save time. Second, most multivariate statistical approaches can be made on an open-source platform and open-source community. Existing R packages can help users save time in reinventing wheels. Third, the multivariate approaches can cooperate with many existing reliable techniques such as probabilistic reserves estimation and probabilistic decline curve analysis (Sinha and Deka 2016).

Lawrence Berkeley National Lab (Moridis et al. 2013) has built a self-teaching system to dynamically process unconventional resources data. It applied principal components analysis (PCA) to lower data dimensions and do clustering to analyze which wells perform similarly and which wells perform distinctly from others. Their research was limited to a small number of well data; only 13 shale gas wells were analyzed.

Sinha and Deka (2016) presented a comprehensive analysis with the application of multivariate statistical analysis to Eagle Ford shale. Their data scope involved 1500 wells with light oil and different levels of mature gas. Their assessments included principal components analysis (PCA), clustering, regression and self-organizing maps (SOM). Those methods cooperating with traditional Arps model can have better predictions.

Lolon (2016) gave insights on how to judge the results of multivariate statistical approaches. He tested several regression methods to find the impact of completions and fracture stimulation on production. His data was based on field data in the North Dakota Three Forks formation in the Williston Basin. He argued that the best prediction model is often overfitted. Second, the best R-square score model had the worst prediction ability with some specific datasets. To overcome those disadvantages, he concluded that prediction methods should be tested on a “hold-out” dataset.

Singh (2017) introduced classification-and regression tree (CART) techniques to automatically diagnose gas well performance. This technique is based on a traditional decision tree but adds cross-validation to ensure to robustness and reliability of the prediction model. The testing was run on a gas well dataset and got good results (Figure 1-5).

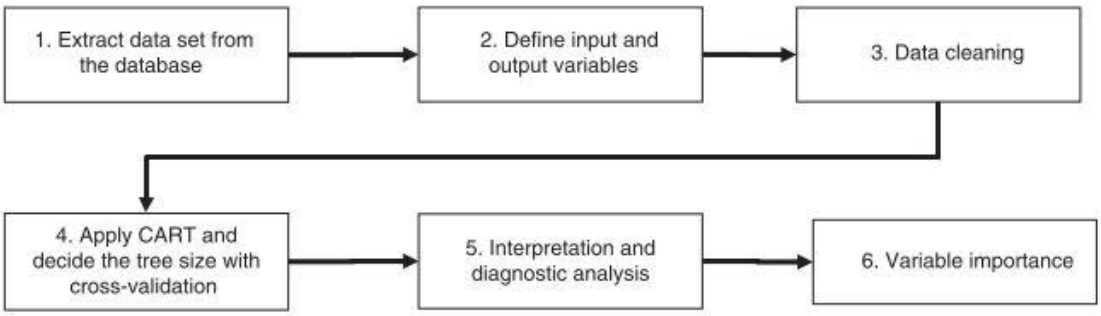


Figure 1-5: Workflow of CART (Singh 2017)

1.3 Motivation and Scope

This research investigated the potential of combing a multivariate statistical approach with a machine learning algorithm into production data. The production data type I focused on is production rate-time data. It provides an efficient and automatic approach to process data. I compare the results with an existing DCA model to illustrate the new method's advantages and disadvantages.

The main tool for analyzing production data was principal components analysis (PCA). PCA proposes a new automated regression model other than linear regression whose coefficient is artificial chosen. Its prediction result is compared with currently

applied methods, and commercial software can lower data dimensions and establish a regression model based on data patterns.

This research had the following objectives:

- Validate the potential application range of PCA methods in production forecasting
- Determine the critical amount production length and well amounts
- Propose a comprehensive workflow and procedure in applying PCA methods into production data analyzing
- Propose a practical automated regression model based on PCs
- Exam the potential of coupling pressure data, completion data with rate time data into enhancing prediction results.

CHAPTER II

PRINCIPAL COMPONENTS ANALYSIS

In this chapter, we first review the basic concept of principal components analysis (PCA). Second, we introduce the fundamental math knowledge of applying PCA into the analysis of multiple-well rate-time production history data analysis. Third, we explain the characteristics and applications of PCA.

2.1 Principal Components Analysis Concept

Principal components analysis (PCA) is a statistical method first published by Pearson (1901) and improved by Hotelling (1933). The concepts and applications have more recently been organized by Jolliffe (2002).

The concept of PCA is described like this:

Supposed we have a random sample $X_1, X_2, \cdots, X_n$, with standard deviation $S, S_2, \cdots, S_n$. We have:

$$PC_i = a_{i1}X_1 + a_{i2}X_2 + \cdots a_{in}X_n \; j = 1,2 \ldots n \tag{2-1}$$

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \tag{2-2}$$

1. If Var(PC_1) is the largest, then, we called it the first principal component (PC).
2. If $PC_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n$ and it is perpendicular to $(a_{11}, a_{12}, \cdots, a_{1n})$ and makes Var(PC_2) second large, it is called the second PC.
3. The following PC is limited up to *n*. For their characteristics, the most important one is:

$$Corr(PC_i, PC_j) = 0 \; 0 \leq i \leq n \; 0 \leq j \leq n \tag{2-3}$$

It can lower data dimensions by orthogonally transforming the correlated variables into linear uncorrelated variables. Those linear uncorrelated variables are called principal components (PC). Selecting large numbers of PCs accounts for the large variance of the original data matrix, which can be reduced to lower dimensions and reconstructed with the combinations of a few PCs. By plotting the original data with the first few PCs, the data can be graphically represented in a clear manner to show its inside patterns and hints for further analysis such as clustering and classification (Bhattacharya and Nikolaou 2013). By omitting those less important PCs (we usually apply Kaise's(1960) rule to drop any PC which variance less than 1.0), we can save important data characteristics without losing much accuracy. The multiple variables can be reconstructed to fewer new linear uncorrelated variables, helping to reduce data space dimensions.

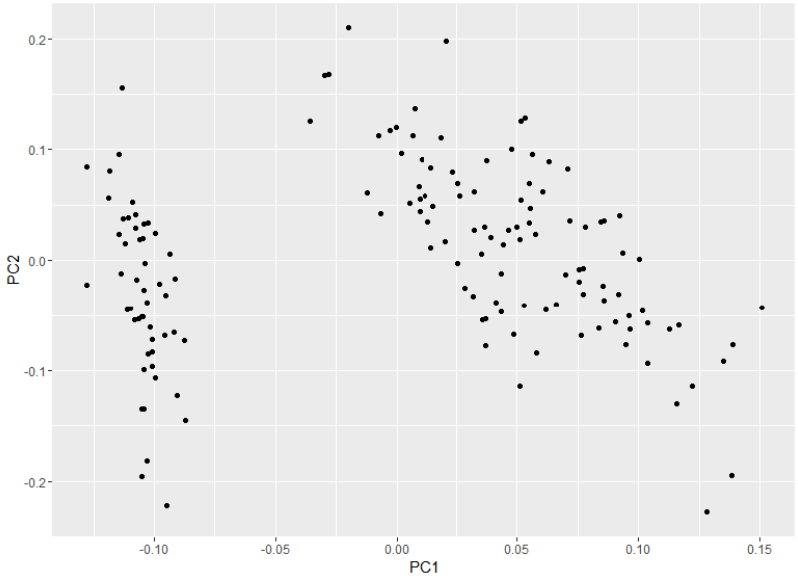


Figure 2-1: 2D plot of PCA-conducted iris data

Figure 2-1 was generated by R, an open-source statistical software. R contains a sample data set called iris flower data (Fisher 1936). This data contains 50 samples of 3 species of iris with 4 different variables (sepal length, sepal width, petal length, petal width). With these data, if we found a new iris flower but did not know what species it was, we might wish to study its sepal/petal length/width to predict. However, it's hard to generate a common regression rule with 50 samples, and each has 4 variables. This gives us a framework for plotting PCs, which can help us solve our iris flower problem.

A panel plot can be drawn by plotting the values of the greatest variance in accounting PC (PC1) on the X-axis and the second largest on the Y-axis, as in the figure above. If a user wishes to increase the dimensions to a 3D plot, they could plot third biggest variance in accounting PC (PC3) on the Z-axis, as in Figure 2-3.

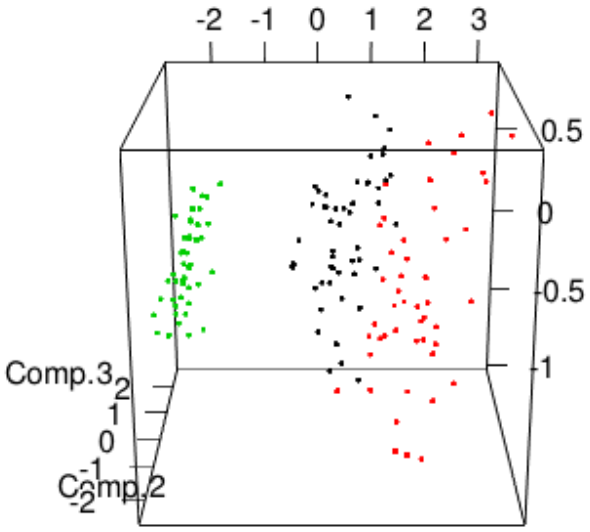


Figure 2-2: 3D plot of PCA-conducted with iris data

The different colors of dots show up as different species of iris. By rotating the 3D plot in R, we can view different panels showing the distribution of dots (samples). The clearest one without any overlap shows the linear uncorrelated combination of data, as in Figure 2-4.

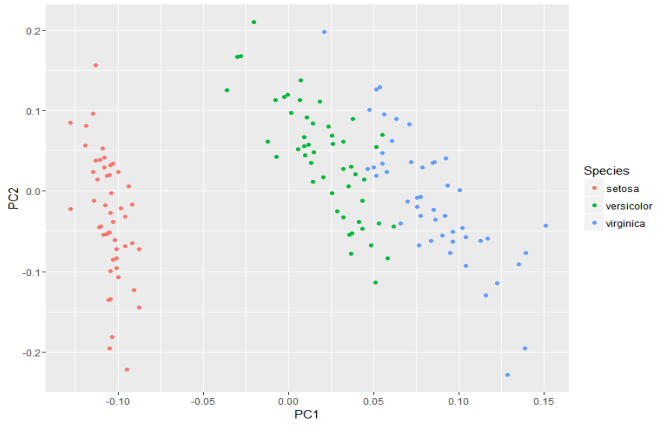


Figure 2-3: Labeled 2D plot of PCA-conducted iris data

By the example above, we can have a perceptual intuition toward to idea of PCA. PCA can lower the number of variables to one or two PCs. By reconstructing the data with 2 PCs, we can plot the PCA data, which will show the distinct difference of different iris flowers.

This exact the idea of PCA is that a single variable in an observation might be hard to use to define samples, but by reconstructing high-dimension variables with low-dimension PCs, we can have an aggregative indicator for defining the sample. Usually,

we hope this indicator has a distinct difference between different observations, as in Figure 2-5.

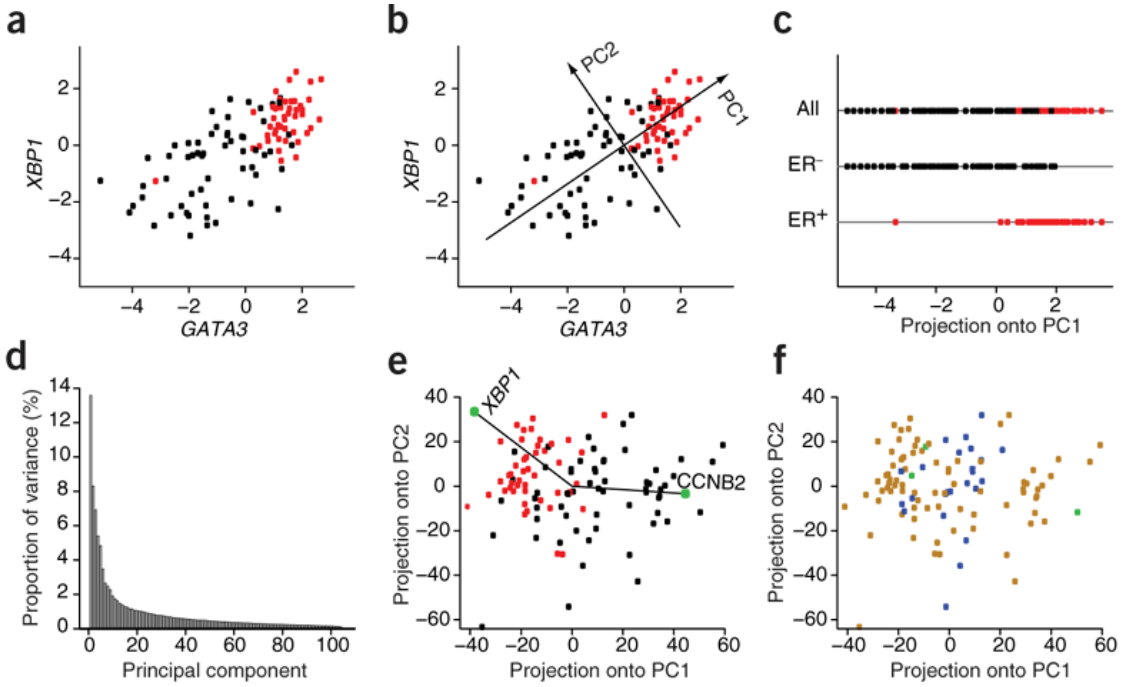


Figure 2-4: Explanation of data variable projection to PC (Ringnér 2008)

The concept of PCA can be described more clearly in math equations. Given the following data matrix *Z* (the following vectors are all row vectors),

$$\mathrm{Z} = \{\overrightarrow{Z_1}, \overrightarrow{Z_2}, \cdots, \overrightarrow{Z_n}\} \tag{2-5}$$

After centralization, it can be represented as

$$\mathrm{Z} = \{\overrightarrow{Z_1}, \overrightarrow{Z_2}, \cdots, \overrightarrow{Z_n}\} \tag{2-6}$$

$$= \{\overrightarrow{Z_1} - \vec{u}, \overrightarrow{Z_2} - \vec{u}, \cdots, \overrightarrow{Z_n} - \vec{u}\} \tag{2-7}$$

$\vec{u}$ is the average vector that is defined by the following:

$$\vec{u} = \frac{1}{n}\sum_{i=1}^{n}\overrightarrow{Z_i} \tag{2-8}$$

Here, considering the definition of PCA, with the transformation trying to find maximum variables to describe data, we are looking for variables as follow:

$$\frac{1}{n}\sum_{i=1}^{n}\left|\overrightarrow{X_i}\cdot\overrightarrow{u_i}\right| \tag{2-9}$$

which is the same as:

$$\frac{1}{n}\sum_{i=1}^{n}\left|\overrightarrow{X_i}\cdot\overrightarrow{u_1}\right|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\overrightarrow{X_i}\cdot\overrightarrow{u_1}\right)^2 \tag{2-10}$$

$$\overrightarrow{X_i}\cdot\overrightarrow{u_1} = {X_i}^T u_1 \tag{2-11}$$

So, the target function can be expressed as follows:

$$\frac{1}{n}u_1^T\left(\sum_{i=1}^{n}X_i\,{X_i}^T\right)u_1 \tag{2-12}$$

for $\sum_{i=1}^{n}X_i\,{X_i}^T$, because $X = [X_1 \quad X_2 \quad \cdots \quad X_n]$ and $X^T = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix}$

The function is finally written as:

$$\frac{1}{n}{u_1}^T X X^T u_1 \tag{2-13}$$

We have two approaches to find our maximum value and its directions.

2.1.1 Method 1 LaGrange Method

The target function and bound equation can be written as:

$$\max\{u_1{}^T XX^T u_1\} \tag{2-14}$$

$$u_1{}^T u_1 = 1 \tag{2-15}$$

if the matrix XX^T eigenvalue is λ and the corresponding eigenvector is ε.

We can construct a LaGrange function:

$$f(u_1) = u_1{}^T XX^T u_1 + \lambda(1 - u_1{}^T u_1) \tag{2-16}$$

Now we can take the derivative of u_1:

$$\frac{\partial f}{\partial u_1} = 2\, XX^T u_1 - 2\lambda u_1 = 0 \rightarrow XX^T u_1 = \lambda u_1 \tag{2-17}$$

Therefore, obviously, u_1 is the eigenvector corresponding to λ.

So, the function could be written into:

$$u_1{}^T XX^T u_1 = \lambda u_1{}^T u_1 = \lambda \tag{2-18}$$

And the proof is that, if we wish to find the biggest variance-explanation variable, it should also work with the biggest eigenvalue. The direction is the eigenvector direction of the biggest eigenvalue.

2.1.2 Single-Value Decomposition Methods

For single-value decomposition methods, the target equation is

$$u_1{}^T XX^T u_1 = (X^T u_1)^T (X^T u_1)$$
$$=< X^T u_1, X^T u_1 > \quad (2\text{-}19)$$
$$= \|X^T u_1\|_2^2$$

Recall that *A* is a random matrix:

$$\frac{\|Ax\|}{\|x\|} \leq \sigma_1(A) = \|A\|_2 \quad (2\text{-}20)$$

$\sigma_1(A)$ is the largest eigenvalue of Matrix A. So from this, we can find the maximum possible value of the Matrix A eigenvalue. Now we need to define the direction:

$$A^T A \in C^{n\times n} \quad (2\text{-}21)$$

Suppose we have the following as their eigenvalue:

$$\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_n \geq 0 \quad (2\text{-}22)$$

The corresponding eigenvector is

$$\xi_1, \xi_2, \cdots, \xi_n \quad (2\text{-}23)$$

Picking random vector *x*,

$$\mathrm{x} = \sum_{i=1}^{n} a_i \xi_i \quad (2\text{-}24)$$

So we have

$$\|x\|_2^2 =< x, x >= a_1^2 + \cdots + a_n^2 \quad (2\text{-}25)$$

$$\|Ax\|_2^2 =< Ax, Ax >= (Ax)^T Ax =< x, A^T Ax > \quad (2\text{-}26)$$

Substitute with $x = \sum_{i=1}^{n} a_i \xi_i$:

$$\begin{aligned} < x, A^T A x > &= < \sum_{i=1}^{n} a_i \xi_i \, , \sum_{i=1}^{n} \lambda_i a_i \xi_i > \\ &= \sum_{i=1}^{n} \lambda_i a_i^2 \end{aligned} \tag{2-27}$$

So we have:

$$\sum_{i=1}^{n} \lambda_i a_i^2 \leq \lambda_1 \left(\sum_{i=1}^{n} a_i^2 \right) = \lambda_1 \|x\|_2^2 \tag{2-28}$$

So:

$$\frac{\|Ax\|_2}{\|x\|_2} \leq \sqrt{\lambda_1} = \sigma_1 \tag{2-29}$$

It is obvious that when x=ξ_1, it picks it a maximum value σ_1.

So

$$u_1{}^T X X^T u_1 \tag{2-30}$$

u_1 is the biggest eigenvalue direction.

2.2 Applying PCA to Production Data Analysis

In a former section, I introduced the basic concept and math meaning of PCA and PC. With the ability to lower data dimensions, PCA can be illustrative for investigating production rate-time data. Bhattacharya and Nikolaou (2013) first introduced this technique into analyzing rate-time data. They tested PCA on Holly Branch unconventional gas wells with approximately 1100-ft effective production length. A regression model based on a linear combination of principal components was found to fit the decline curve. The fit results errors are less than 2%, which shows a high potential value of PCA application to production rate-time data. This is shown in Figure 2-6.

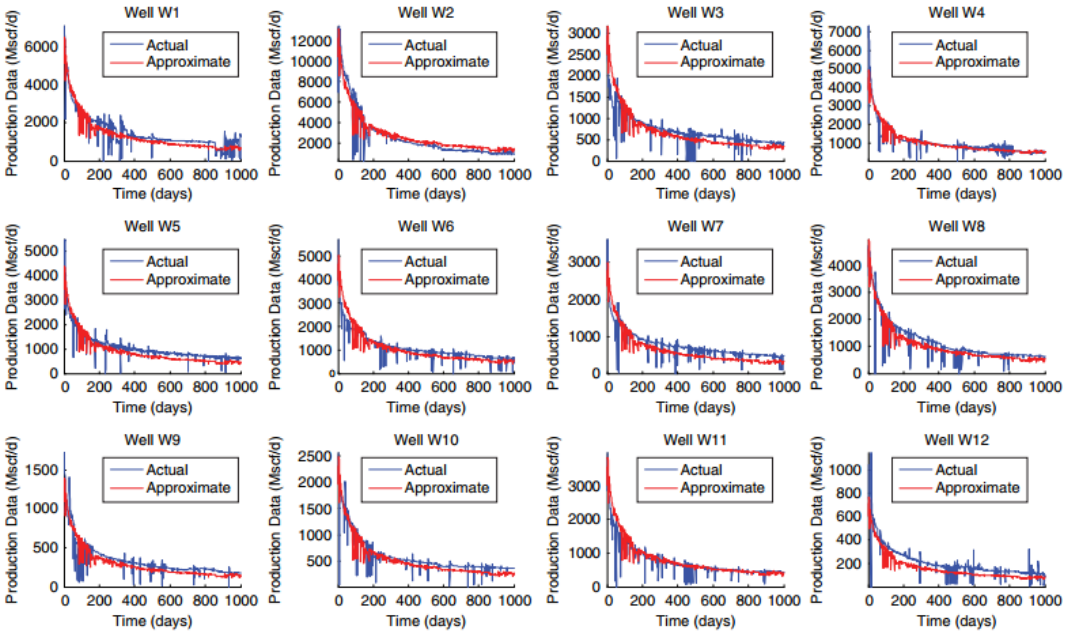


Figure 2-5: Actual data and PCA prediction data (Bhattacharya and Nikolaou 2013)

2.2.1 Rate-Time Data as Time-Series Data

Before applying PCA to rate-time data, first we discuss whether we can apply principal components analysis on rate-time data. It is improper to use PCA without discussing its proper application ranges. To discuss this, first, the data type of rate-time data need be clarified.

Rate-time data can be classified as time-series (TS) data. This type data has different statistical characteristics from traditional data. The traditional statistical data usually has following data structure (Table 2-1):

Table 2-1: Traditional data structure

Variables / Sample	X_1	⋯	X_m
1	x_{11}	⋯	x_{m1}
2	x_{12}	⋯	x_{m2}
⋮	⋮	⋯	⋮
n	x_{1n}	⋯	x_{mn}

In this table, for sample 1 to sample *n*, each sample has several variables, X_1 to X_m. What traditional PCA is trying to do is to lower the variables in *m* to *p* linearly uncorrelated PCs ($p < m$).

However, production rate-time data has a unique data structure for a stochastic sequence $\{\overrightarrow{Z_1}, \overrightarrow{Z_2}, \cdots, \overrightarrow{Z_n}\}$. Each of them can only acquire one observation at one arbitrary time step. Z_t is a time-related variable. This is shown in Table 2-2.

Table 2-2: Time-Series data structure

Samples \ Variables	...	X_1	...	X_t	...
1	...	x_1	...	x_t	...

Looking at a typical unconventional gas well decline curve, the production rate-time shows a typical time-series data characteristic. The rates at different days are only observed at the given date. This is shown in Figure 2-7.

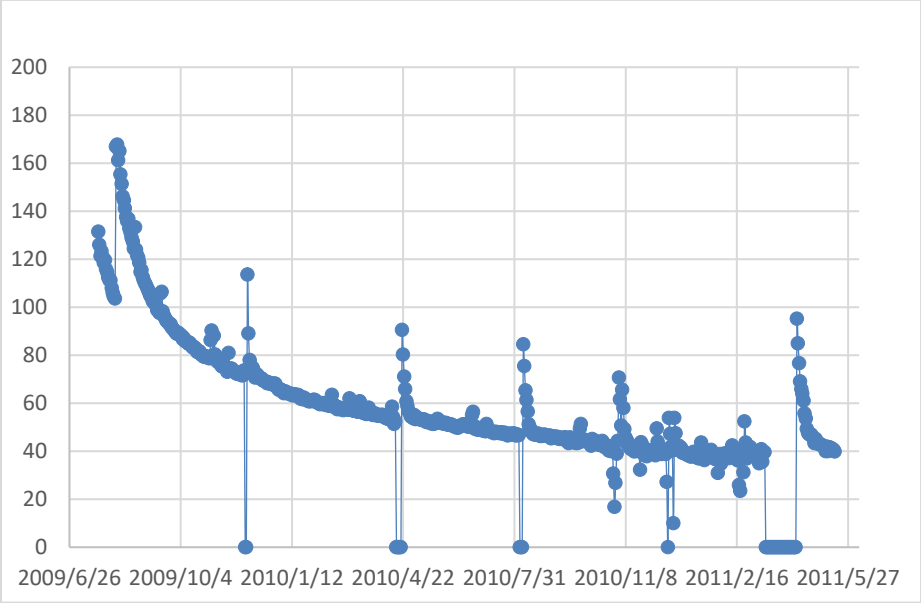


Figure 2-6: Rate-time data plot

Those rate time data could also be a list, like Table 2-3:

Table 2-3: Rate-Time data structure

Rate / Well	t_1	t_2	…	t_n
1	r(t_1)	r(t_2)	…	r(t_n)

The above plot and table prove that the O&G production rate-time data can be treated as time-series data when applying PCA.

2.2.2 Rate-Time Data Influence Factors

After defining the statistical category of rate-time data, we also need to investigate the data given other than rate-time data. Usually, when we assess a certain well in certain fields, we can locate the following data:

- Rate-time data. These data are usually continuous and reliable. They may contain some outliers and noise, which are caused by measurement error or instrumental error.
- Pressure data. As explained in the literature review, pressure data may not be continuous, and their value is doubtful. In some cases, due to maintenance or shut wells, those data may not be acquirable.
- Well-design data. Those data contain well depth, design diagrams, casing size, horizontal length, etc.

- Reservoir characteristics. Data usually contains initial pressure, reservoir temperature, net pay, porosity, gas saturation, permeability, fracture conductivity, and fracture half-length.

To understand these data and make physically reliable production forecasts, we need to review the drive mechanism that controls production rate and decline trends. Wang (2016) provided a comprehensive review and judgment on shale gas production patterns. Those factors/mechanisms include:

- Adsorption gas desorption
- Apparent permeability of the shale matrix
- Nonstimulated reservoir volume
- Fracture network conductivity

From those control factors, we find that, to understand the rate-time data, the reservoir data should include fracture half-length, fracture conductivity, permeability, porosity, and initial pressure. Those parameters control the decline trends and production rates.

2.2.3 Math Fundamentals of Applying PCA

This section introduces the math fundamentals of principal components analysis into production rate-time data analysis. Because well-production rate-time data can be treated as time series data, we can combine multiple wells into a data matrix.

Suppose we have *n* wells and *m* producing days. *m* days are simultaneous on each well among *n* wells. For each well we have:

$$[r(t_1) \quad r(t_2) \quad \dots \quad r(t_m)] \tag{2-31}$$

where $r(t_i)$ means the production rate on arbitrary *i* day.

Listing each well on each row, we have

$$\begin{bmatrix} r_1(t_1) & r_1(t_2) & \cdots & r_1(t_m) \\ r_2(t_1) & r_2(t_2) & \cdots & r_2(t_m) \\ \cdots & \dots & \vdots & \vdots \\ r_n(t_1) & r_n(t_2) & \dots & r_n(t_m) \end{bmatrix}_{n\times m} \tag{2-32}$$

We can write this matrix as follows:

$$[Z]_{n\times m} \tag{2-33}$$

From the concept of singular value decomposition, for arbitrary matrix A, we can write it as:

$$A = U\Sigma V^T \tag{2-34}$$

where $U_{n\times n}$ is an $n\times n$ unitary matrix, $\Sigma_{n\times m}$ is an $n\times m$ rectangular diagonal matrix (the element on diagonal is eigenvalue of matrix A), $V^T_{m\times m}$ is an $m\times m$ unitary matrix and also the conjugate transpose matrix.

In our case, data matrix *Z* can be express as:

$$[Z]_{n\times m} = U_{n\times n}\Sigma_{n\times m}V^T_{m\times \mathrm{m}} \tag{2-35}$$

$$\Sigma_{n\times m} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \tag{2-36}$$

λ_i is the i th eigenvalue of matrix Σ and:

$$\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_n \geq 0 \tag{2-37}$$

From the definition of PCA, we know that:

$$\begin{aligned}[Z]_{n\times m} &= \lambda_1[u_1]_{n\times 1}[V_1^T]_{1\times m} + \cdots \\ &\quad + \lambda_{\mathrm{n}}[u_{\mathrm{n}}]_{n\times 1}[V_n^T]_{1\times m} \\ &= PC_1[V_1^T]_{1\times m} + \cdots + PC_n[V_n^T]_{1\times m}\end{aligned} \tag{2-38}$$

$$n = rank\ (Z) \tag{2-39}$$

So, by selecting r eigenvalues that could explain enough variance, we have:

$$\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_r \geq 0 \tag{2-40}$$

$$\begin{aligned}[Z] \approx [K] &= \lambda_1[u_1]_{n\times 1}[V_1^T]_{1\times m} + \cdots + \lambda_{\mathrm{r}}[u_{\mathrm{r}}]_{n\times 1}[V_r^T]_{1\times m} \\ &= PC_1[V_1^T]_{1\times m} + \cdots + PC_r[V_r^T]_{1\times m}\end{aligned} \tag{2-41}$$

$$r = rank\ (K) \tag{2-42}$$

$$r < n \tag{2-43}$$

Therefore, the data matrix has been successfully lowered in dimensions from n to r.

For the i^{th} well in the data matrix until j time steps

$$\mathrm{well}_i = PC_1[V_1^T]_{1\times m} + \cdots + PC_p\left[V_j^T\right]_{1\times m} \tag{2-44}$$

2.3 PCA Functions to Rate-Time Data

For the function of PCA, Bhattacharya and Nikolaou (2013) proposed questions that could be answered with PCA.

- How to do factor analysis?
- How to do clustering?
- How can learning from existing wells be applied to predict new wells?

Besides those three questions, PCA also can detect outliers in time-series data. Outliers are observations performed differently than other observations, or they might be due to measurement error or human record error. However, those outliers need to be removed from the dataset. By combing *k*-nearest neighbor with PCA, this method can also remove outliers. So, PCA can also answer a new question:

- How to remove outliers/noise?

This section provides a fundamental and essential review of the algorithm for the next step. First, I review the factor analysis of PCA. Second, I explain the combination method of *k*-means clustering with PCA process data. Third, I introduce the linear regression model. In the end, I illustrate the approach of combing *k*-nearest neighbor and PCA.

2.3.1 Factor Analysis

Principal components analysis can include analysis of either the rate-time data alone or rate-time data combined with reservoir characteristics. Usually, we need a complex geological model or analytical model to investigate the influence of reservoir characteristics. PCA can give us a new way to investigate the dynamic influence of reservoir characteristics on production rate. This section provides the mathematic fundamentals of factor analysis with PCA.

Suppose PCA can generate *m* PCs from *p* original variables by *m* linear combinations. We have $\vec{x}$ as a $p\times1$ random vector; its mean equals μ, and its covariance matrix equals $\Sigma = (\sigma_{ij})$. $\vec{x}$ is affected by *k* factors. So $\vec{x}$ could be express by the following equation:

$$x = \mu + \Lambda\vec{f} + \vec{u} \tag{2-46}$$

where: $\Lambda = p\times k$ is a constant-number matrix.

$\vec{f} = p\times k, \vec{u} = p\times1$ is a random vector. $\vec{f}$ is called a public factor. $\vec{u}$ is called the factor-loading matrix. To do factor analysis, we need the relationship equation as follows:

$$\Sigma = \Lambda\Lambda' + \Psi \tag{2-47}$$

where Σ is the covariance matrix of the original data matrix and Ψ is the covariance matrix of the original factor loading matrix. So we have:

$$x = \mu + (\Lambda\Gamma)(\Gamma'\vec{f}) + \vec{u} \tag{2-48}$$

Here, $\Lambda\Gamma$ is the new loading factor matrix, and $\Gamma'\vec{f}$ is the new factor.

So we can transform our initial matrix *x* to generate a new, easy-to-explain factor and observe how those factors contribute to rate-time data performance.

2.3.2 K-means Clustering with PCA

In the application of Principal Components analysis, there is an important part called *clustering*. After PCA, rate-time data for multiple wells can be reduced to low-dimension space. To judge their distribution and find which wells perform similarly, we need to do clustering.

The definition of clustering varies by algorithm. In general, clustering means placing those that behave similarly in one cluster and those that are different in other clusters. A widely-applied clustering algorithm is called *k*-mean clustering. In *k*-mean clustering, we define *k* points in data space, and build clusters to categorize nearest observations. The steps can be listed as follows (MacQueen 1967):

1. Determine the number of clusters *k*
2. Generate arbitrary *k* cluster and determine the cluster's center
3. Calculate cluster's center for each observation
4. Recalculate the new cluster's center
5. Repeat these steps until cluster centers do not change

An important step in clustering is calculating the distance among points. In Euclidean space, we use Euclidean distance to judge their distance:

If $j = (j_1, j_2, j_3, \cdots, j_n)$ and $k = (k_1, k_2, k_3, \cdots, k_n)$ their distance is:

$$d(j,k) = d(k,j) = \sqrt{(j_1 - k_1)^2 + (j_2 - k_2)^2 + \cdots + (j_n - k_n)^2} = \sqrt{\sum_{i=1}^{n} (j_i - k_i)^2} \quad (2\text{-}49)$$

In our processing of rate-time data. Usually, we can represent data in a 2D panel (Bhattacharya and Nikolaou 2013), as in Figure 2-8.

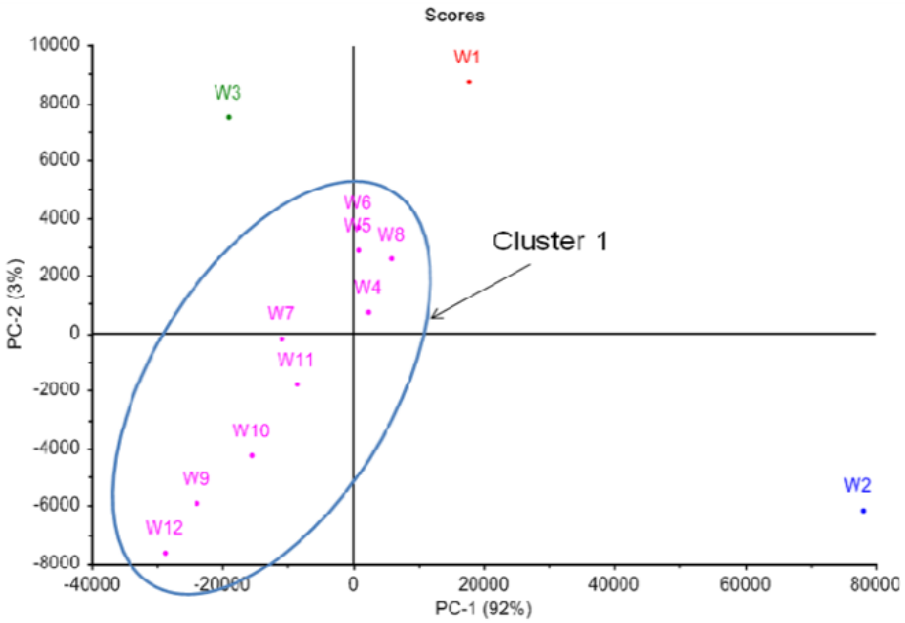


Figure 2-7: Scatter plot of 13 wells with 2 PC panel

Figure 2-8 shows a real case of 13 unconventional gas wells represented by first and second PCs in a scatter plot. Because of limited data, this plot shows a relatively concentrated distribution. Most wells could be clustered into 1 cluster. However, in most cases, we need figure out a more dispersed situation. So we need to investigate the math fundamentals of *k*-means clustering.

Suppose we have *N* data points divided by *k* clusters. Our target function is

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \tag{2-50}$$

It is same as finding the minimum of the center of clusters with

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \tag{2-51}$$

When we begin our calculation, the distance between points is:

$$D_i^{(t)} = \{x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \forall \mathrm{j}, 1 \leq \mathrm{j} \leq \mathrm{k} \tag{2-52}$$

On each updating step, the new center is calculated as:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2-53}$$

2.3.3 K-nearest Neighbor Outliers Detection

Outliers and the approach to remove them are important to data processing. Numerous technical papers have addressed the essentiality for removing outliers in data preparation (Chaudhary and Lee, 2016a, b, Seidle 2016). Researchers argue that outliers can decrease diagnostic value and prediction reliabilities. Therefore, in our process to evaluate unconventional gas well rate-time data, it is critical to remove them before we make predictions (Figure 2-9).

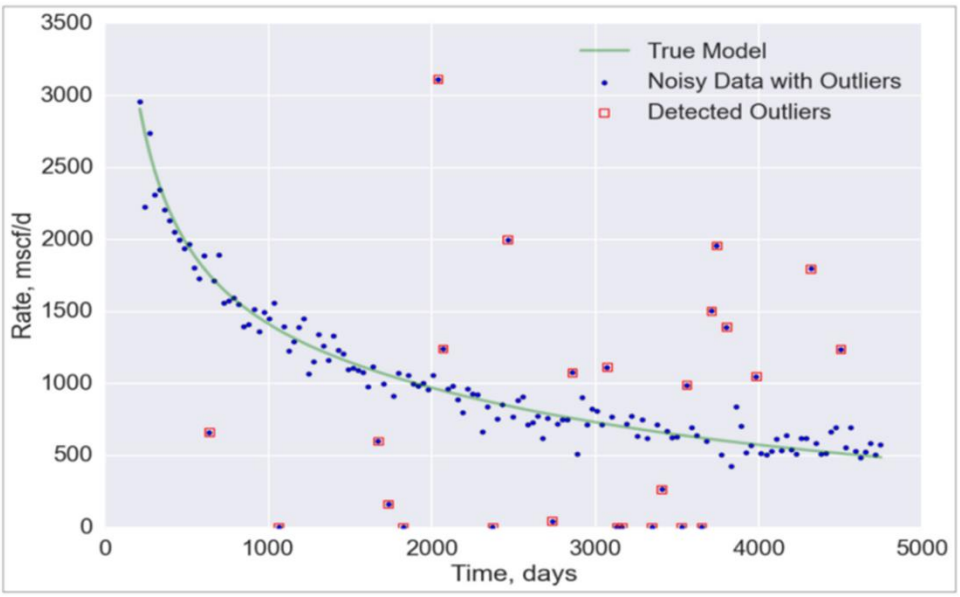


Figure 2-8: Field data containing outliers (Chaudhary and Lee 2016)

The definition of outlier given by Hawkins (1980) is an observation deviating so much from most other observations. It contains two points:

- It departs a great deal from mainstream data.
- It appears to have been generated by a mechanism other than random error.

We can remove and recognize those outliers by applying *k*-nearest neighbor (K-NN) together with PCA, as shown in Figure 2-10.

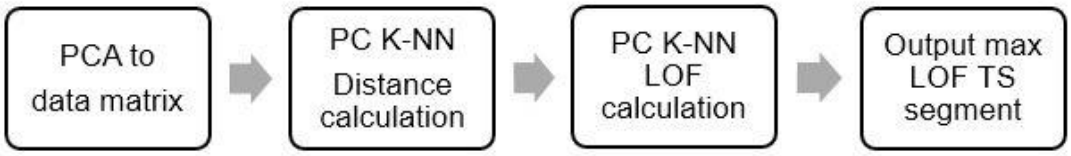


Figure 2-9: Workflow of PC K-NN outlier recognition

First, the distance of K-NN, *k*-dist*(q)*, can be defined as follows:

- At minimal *k* points $o \in D \setminus \{p\}, d(p,o) \leq k - \text{dist}(p)$
- At maximum *k*-1 points $o \in D \setminus \{p\}, d(p,o) \leq k - \text{dist}(p)$

This is illustrated in Figure 2-11:

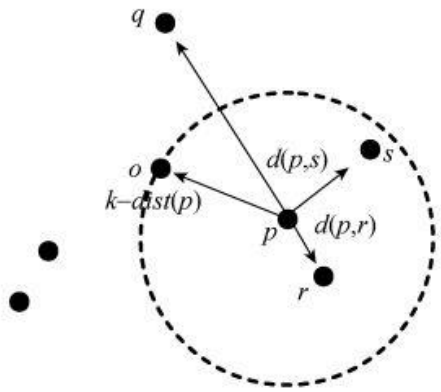


Figure 2-10: When *k*=3, k-dist(*p*)=*d*(*p*,o) (Guo, Li, and Song 2012)

Then, we can calculate the local reachability distance:

$$\text{lrd}(\text{q}) = \frac{k}{\sum_{p \in K(q)} r - \text{dist}_{\text{k}}(q,p)} \tag{2-54}$$

Finally, we need to calculate the local outlier factor:

$$\text{LOF}(q) = \frac{\sum_{p \in K(q)} \text{lrd}(q)}{k * \text{lrd}(q)} \tag{2-55}$$

LOF(q) value reflects the sparsity situation in q points of k domain. The higher the LOF(q) value, the higher sparsity will be in this local domain. After we calculate the LOF, we can apply the advantage of PCA to reduce the original data dimensions and generate a clear recognition of outliers (Figure 2-11).

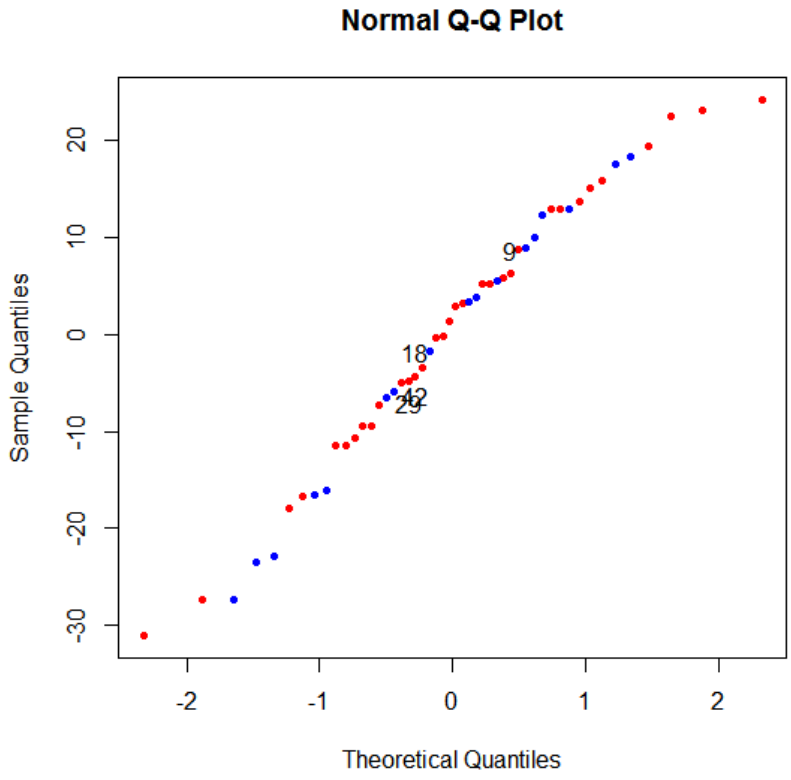


Figure 2-11: PC K-NN recognition on US arrest data

2.3.4 Linear Regression Model

The build-up of prediction models has following workflow (Figure 2-13):

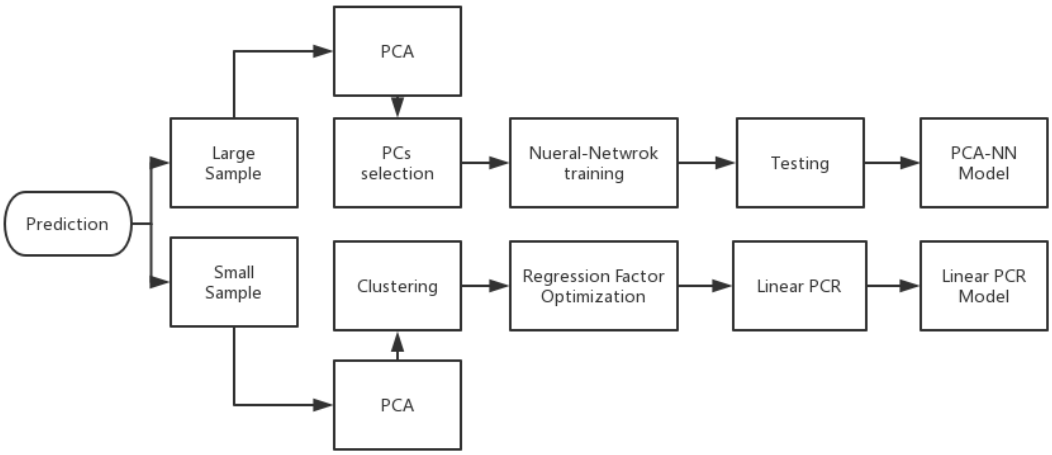


Figure 2-12: Workflow of PC prediction

When only limited amounts of rate-time data from producing wells are available, as less than 30 wells, we prefer using a linear regression model to catch to the pattern of production history. In this case, the model is efficient and precise.

As we have shown,

$$[Z]_{n\times m} \approx [K] = \lambda_1[u_1]_{n\times 1}[V_1^T]_{1\times m} + \cdots + \lambda_{\mathrm{r}}[u_{\mathrm{r}}]_{n\times 1}[V_r^T]_{1\times m} = PC_1[V_1^T]_{1\times m} + \cdots + PC_r[V_r^T]_{1\times m} \quad (2\text{-}56)$$

The production data matrix can also be expressed by several PC and loading matrixes. Therefore, it can be written as:

$$q_t = \beta_1 PC_1 V_1^T + \beta_2 PC_2 V_2^T + \beta_3 PC_3 V_3^T \quad (2\text{-}57)$$

The approximation is shown in Figure 2-14.

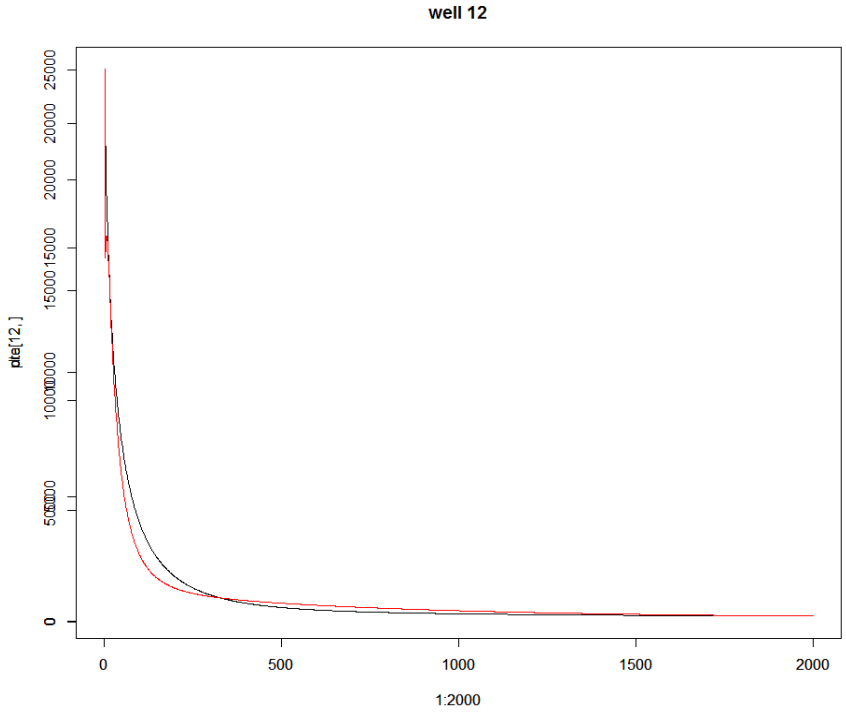


Figure 2-13: Prediction results of linear regression

In our processing of data, we have two ideas for testing the quality of prediction. In one case, the linear regression parameter is learned from the whole data matrix. In another case, the linear regression parameter is learned from the clustered data. By comparing the R square value, we can figure out the optimal regression parameter.

CHAPTER III

APPLYING PCA ON SIMULATION DATA

The first step is testing PCA on simulation data to establish a predictive model. The reason for using simulation data first is that it is usually smoother and has fewer outliers or noise than real field data.

The simulation data was generated by Kappa Ecrin. Kappa is a world-famous well testing interpretation software providing service from history matching to dynamic data analysis. Ecrin is a module built inside Kappa software that can be used for simulating wells performance. The operation window is shown in Figure 3-1.

Figure 3-1: Kappa Ecrin operation window

3.1 Simulation Data Generation

In our case, the simulation data was generated from the following parameters: permeability, porosity, half-length, fracture conductivity and formation pressure. The ranges of those parameters are listed in Table 3-1 and Figure 3-2.

Table 3-1: The range of setting parameter

	Permeability, md	Porosity	Half Length, ft	Frac Conduct, md-ft	Pressure, psia
Max	0.1125	0.095371	433.4508	406.1634	6000
Min	0.0017	0.069167	94.6663	84	3460

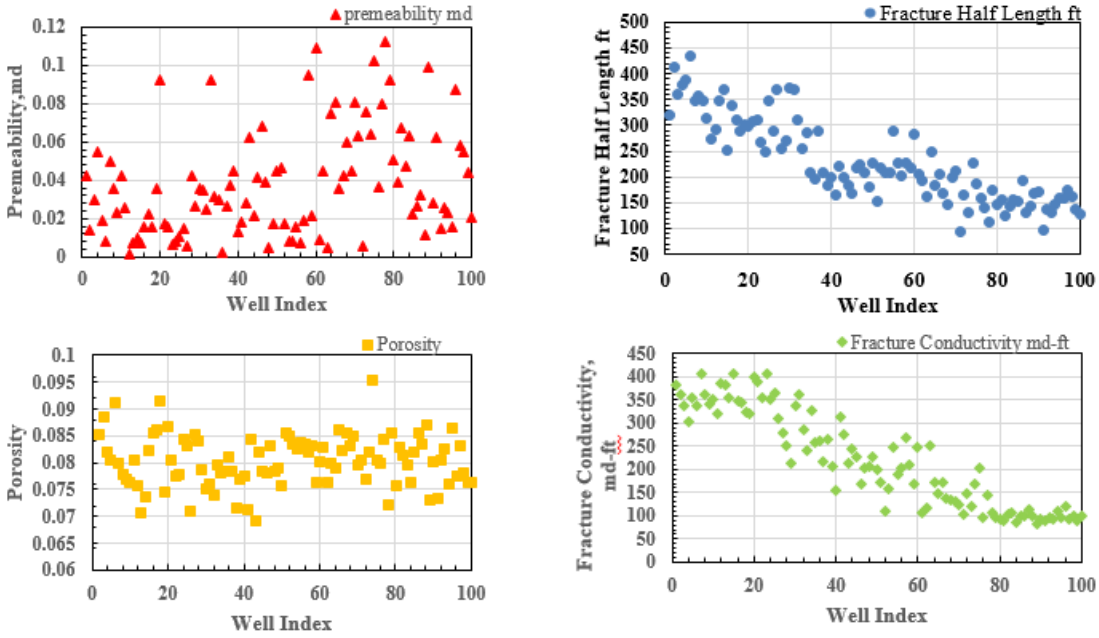


Figure 3-2: The setting parameters for each well

The simulation data contain 100 wells, and each has 2000 days of production history. That would be sufficient for building a testing interpretation and prediction center (Figure 3-3).

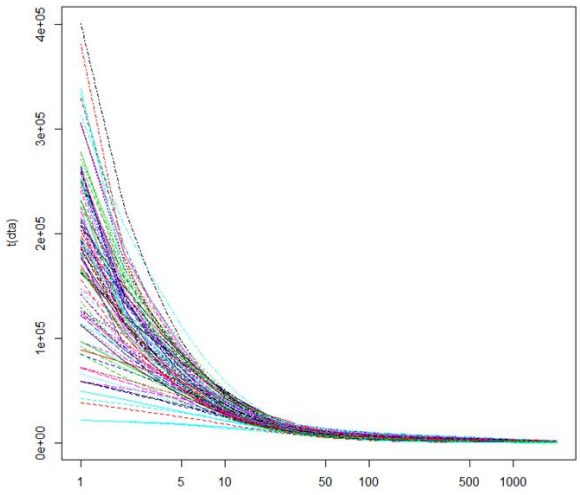


Figure 3-3: Production history for 100 wells plotted in semi-log plot

3.2 Workflow of Applying PCA

From these twp techniques, we can move to the establishment of our predictive model. The overall workflow can be expressed as in Figure 3-4.

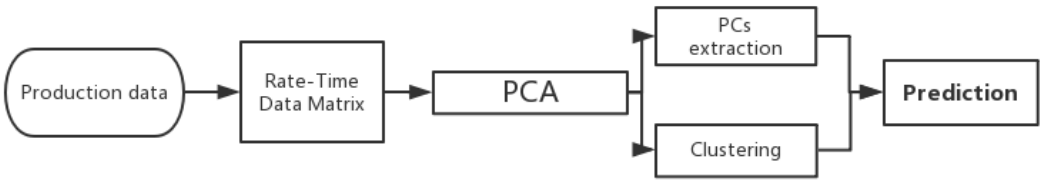


Figure 3-4: Workflow of simulation data

After we generated the data from Kappa Ecrin, the first step was to sort it into a data matrix. Because principal component analysis/singular value decomposition (PCA/SVD) can only be conducted with data as a matrix, in statistical software R, we can use the function **as.matrix()** to change the data type. The data matrix is established in the following manner: the *n* wells are listed by rows, and each daily production rate is listed by column. Currently, our model can only analyze the information from the rate-time matrix. However, other data such as pressure data, log data, wellbore design diagram and fracture procedure could also be used in data mining with PCA.

The next step is conducting PCA of the data matrix. In R, we can use function **svd()** to decompose the data matrix with three matrixes.

$$[Z]_{n\times m} = U_{n\times n}\Sigma_{n\times m}V^{T}_{m\times \mathrm{m}} \tag{3-1}$$

The workflow of PCA can be described as in Figure 3-5:

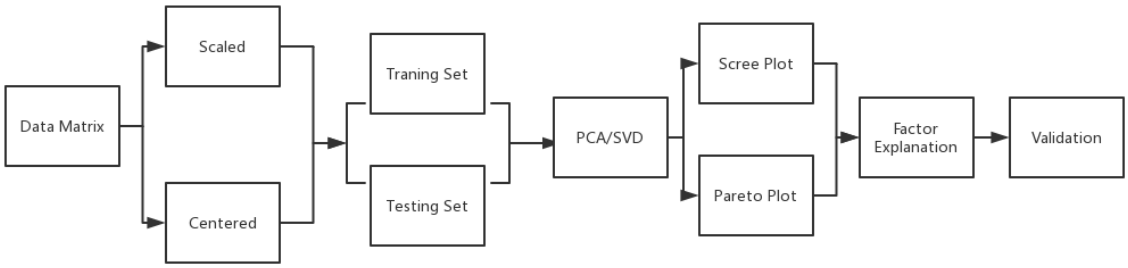


Figure 3-5: The work flow of PCA

The $V^{T}_{m\times m}$ is the eigenvector matrix consisting of principal components. We can extract PCs from this matrix by listing their variance and picking desirable amounts of PC. The amounts of PC are usually defined by a scree plot.

A scree plot can display the variance of PC on the y-axis and their indexes on the x-axis. By applying the elbow criterion, we can define the desired amounts of PC used in building up of the predictive model. This is shown in Figure 3-6.

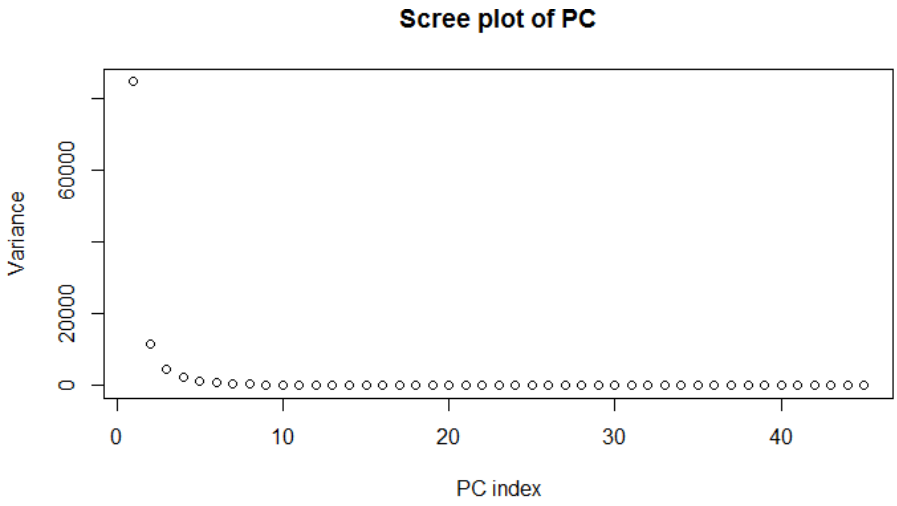


Figure 3-6: Scree plot of principal components

The elbow criterion (Ketchen and Shook 1996) is a way to pick up the proper number of clusters. It can also be applied in defining the number of principal components. The elbow means the point of the change in gradient—the point where not much variance would be added when adding more clusters or principal components. This is shown in Figure 3-7.

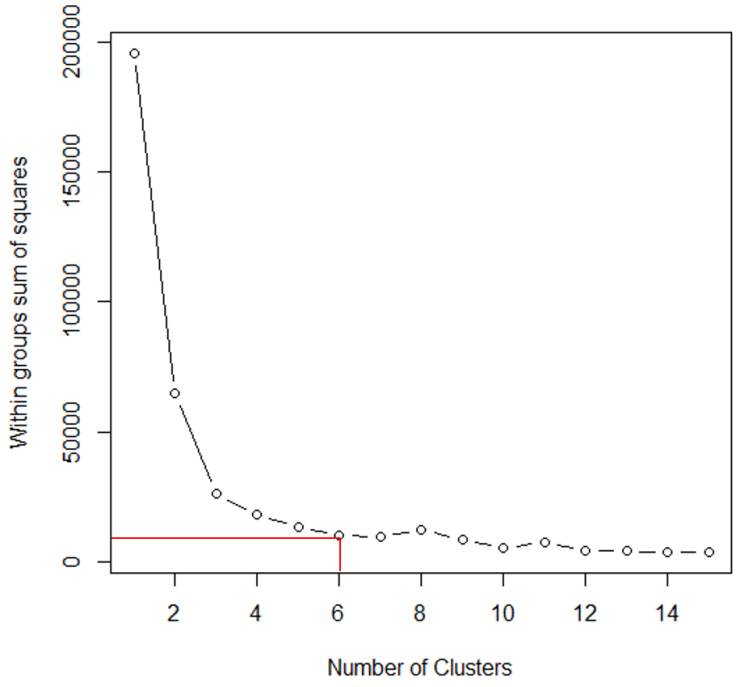


Figure 3-7: Illustration of elbow criterion

Another useful figure to define the number of principal components is a pareto plot. The y-axis is the accumulation of variance; the x-axis is the number of principal components or clusters. From Figure 3-8, we can see that when it goes to PC 4, not much variance would be added by increasing PC amounts.

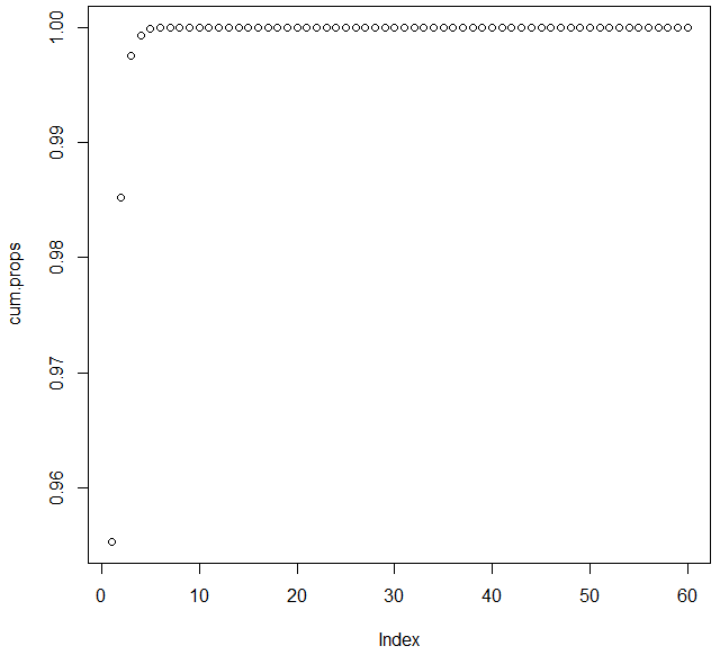


Figure 3-8: Example of pareto plot

Clustering is also an important technique that can be used with PCA. In our study, a clustering algorithm is chosen as *k*-means. The user can define the desired amounts of *k*, which is the reason for the name *k*-means. The number of *k* is usually defined by a WCSS (within-cluster sum of squares) plot and also with the elbow criterion.

The combination of PCA and *k*-means can be used for defining the performance distinction between different wells. In Figure 3-9, 100 wells are clustered into 4 clusters. Each cluster is represented by different color. Each different well in the plot is defined by its scores in PC1 and PC2.

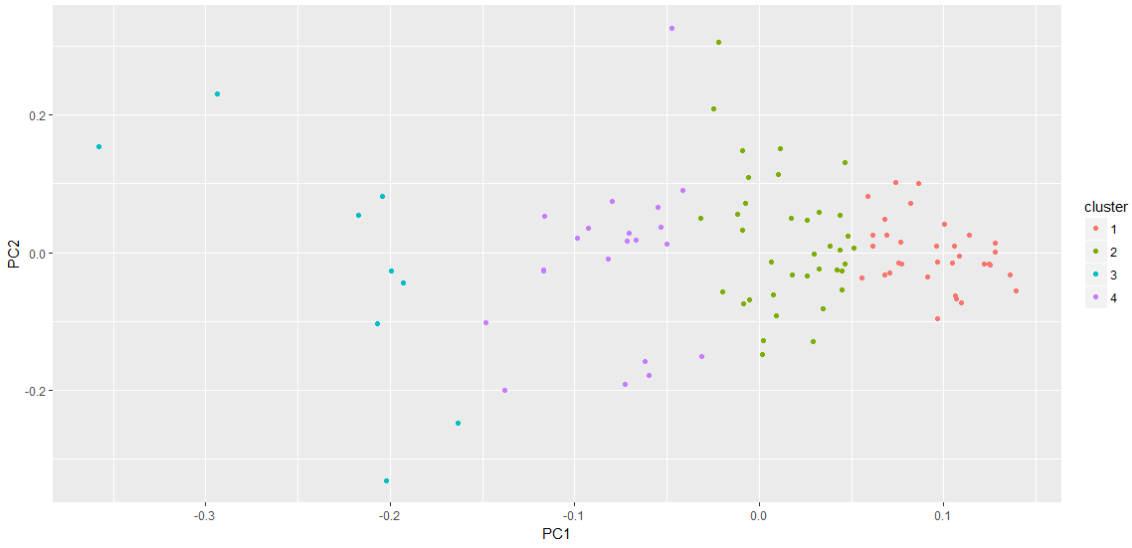


Figure 3-9: Illustration of *k*-means and PCA combination

3.3 Case Study of Simulation Data

The simulation data for 100 shale gas wells each with 2,000 days of production history can be used to establish a data matrix. The data matrix lists wells by row and production by column. This is shown in Table 3-2.

Table 3-2: Simulation data matrix

	Day1	Day2	...	Day 2000
Well 1	268369.7993	164624.7104	...	1332.686
⋮	⋮	⋮	⋱	⋮
Well 100	135382.0989	86176.6028	...	1303.603

After the establishment of the data matrix, we can conduct SVD on it. By plotting the eigenvalue of the data matrix, we create the scree plot (Figure 3-10).

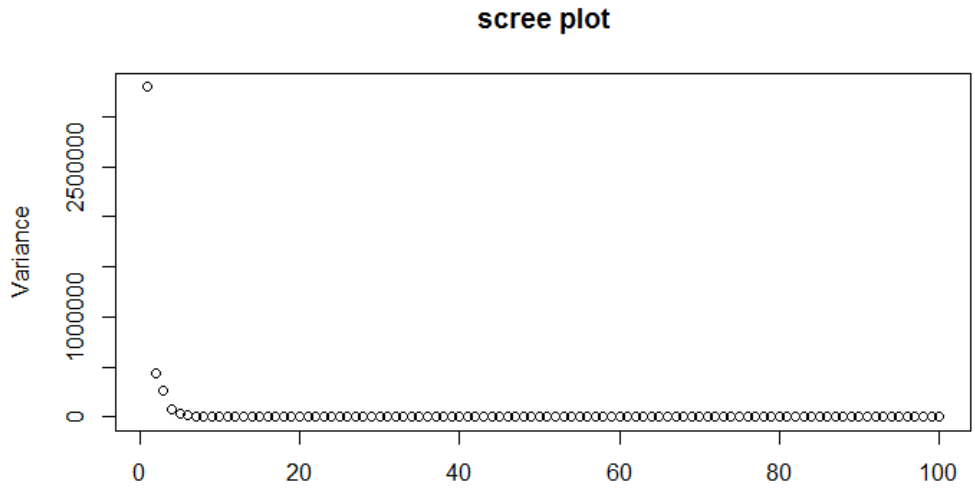


Figure 3-10: Scree plot of the simulation data matrix

We can also plot the pareto plot to look at the desirable number of PCs. This is shown in Figure 3-11.

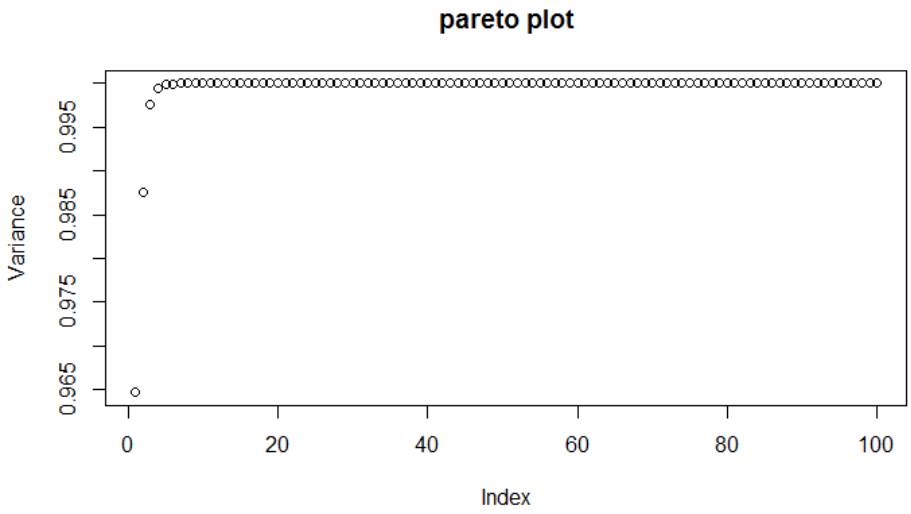


Figure 3-11: Pareto plot of the simulation data matrix

By applying the elbow criterion, choosing four principal components would be enough for explaining the variance of the whole matrix. The accumulation variance goes to 99.93% of the whole matrix variance. In this way, we can reduce the matrix from 100 dimensions to 4 dimensions while not losing its important features and information.

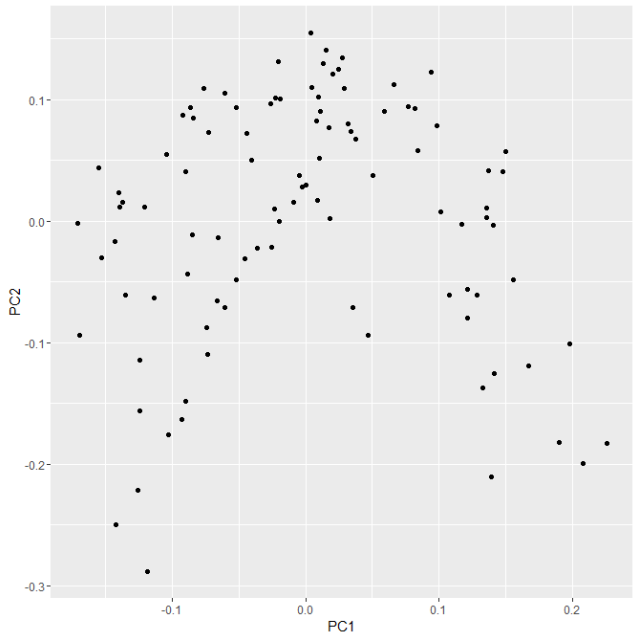


Figure 3-12: 2D visualization of PCA results

Figure 3-12 is a plot of 100 wells by their scores on PC1 and PC2. Two principal components would account for 98.75% of the variance of the matrix. This figure can give the audience a visualized understanding of well performance distribution.

We can also plot it in 3D; Figure 3-13 illustrates a similar distribution as Figure 3-12:

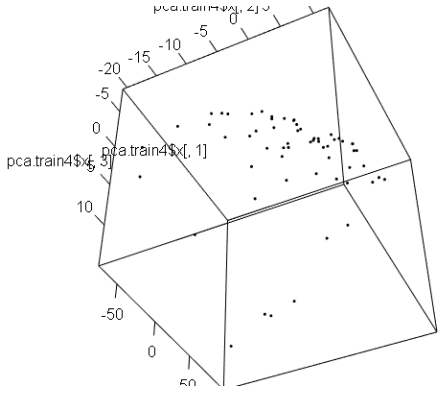


Figure 3-13: 3D visualization of PCA results

From these two figures, we find that parts of wells look closer than others. We can apply the *k*-means clustering algorithm on the dimension-reduced result to validate this assumption. Before applying *k*-means, we first need to determine the number of *k* from the WCSS plot in Figure 3-14:

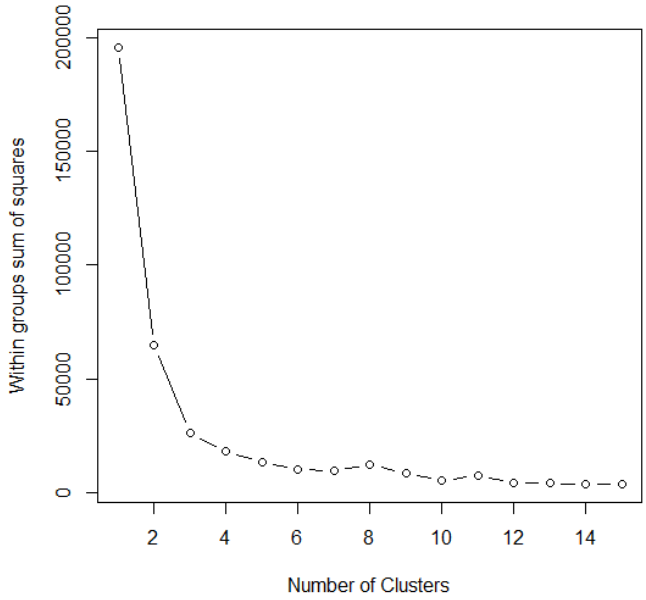


Figure 3-14: WCSS plot of simulation data

By applying the elbow criterion, we find that six would be a desirable number of components. The result of *k*-means is illustrated in Figure 3-15.

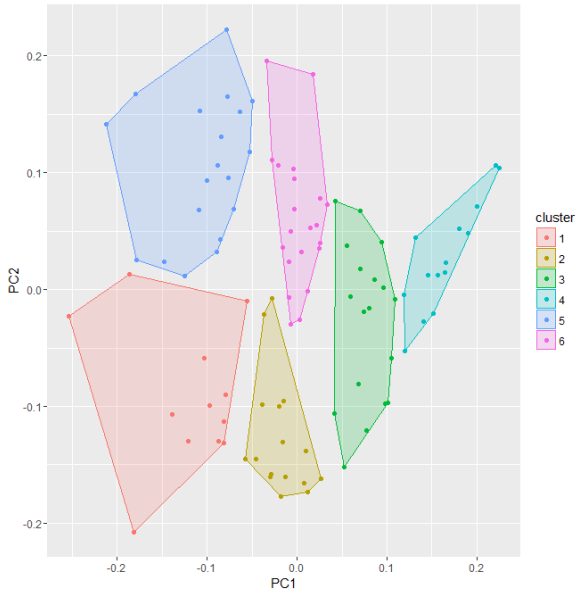


Figure 3-15: 2D *K*-means clustering of simulation data

The clustering result is displayed in Table 3-3.

Table 3-3: Clustering result of simulation data

1	19	32	42	45	57	59	63	64	69	70
	72	73	76	79	81	82	83	90	96	97
2	3	6	18	27	29	33	37	38	44	46
	49	50	61	65	66	67	68	75	80	85
	86	89	92	93						
3	1	4	8	13	15	17	21	58	87	56
	24	25	39	40	43	48	51	54		
4	2	7	9	10	16	20	84	91	94	99
	28	31	34	36	41					
5	5	11	12	14	22	23	26	60	62	71
	30	35	47	52	53	55				
6	74	77	78	88	95					

We could also plot this with 3 PCs and remain unchanged (Figure 3-16).

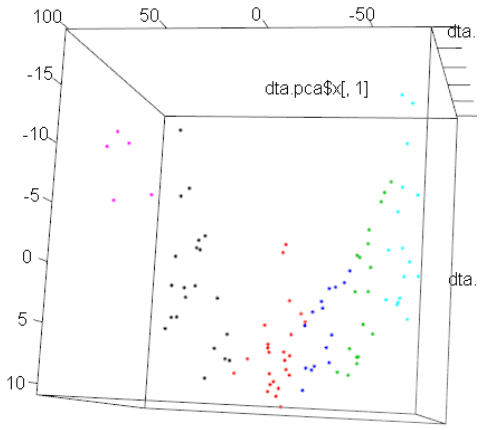


Figure 3-16: 3D *k*-means clustering of simulation data

With the PC dimension reduction, we can also investigate the effect of different variables on the performance of wells and acquire a visualization. This is called *factor analysis*. When we combine the input data (formation parameter) and output data (rate-time data) and explain it by PC, we have Figure 3-17.

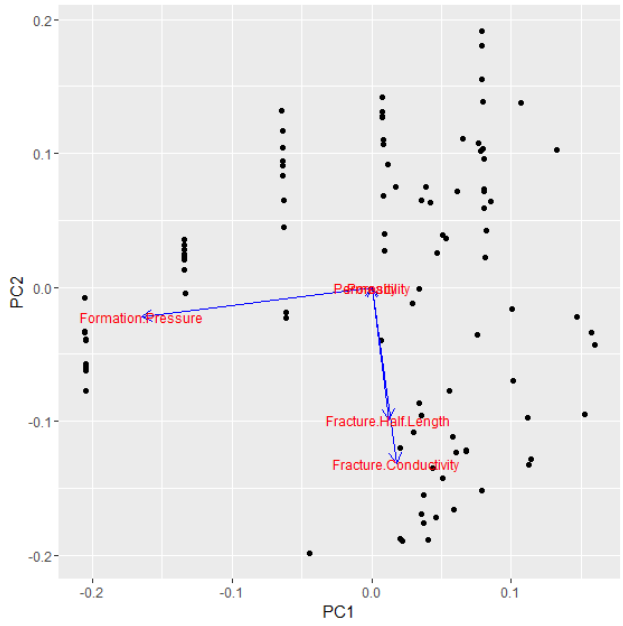


Figure 3-17: Factor analysis of simulation data

Now, we can build the predictive model with the clustering results and PCA data. From Chapter 2, we know that each well in the data matrix can be explained by several principal components that we have:

$$q_t = \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \cdots + \beta_n PC_n \tag{3-2}$$

In our case, the number of *n* is defined as 4, so each well from the data matrix can be explained by the following equation:

$$q_t \approx \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \beta_4 PC_4 \tag{3-3}$$

The prediction matrix can be explained as follows:

$$[Z_{\text{prediction}}] = [\beta_{\text{coefficent matrix}}][PC_{\text{eigenvector matrix}}] \tag{3-4}$$

The linear coefficient matrix is defined by the partial least squares (PLS) technique. It can be done in R using the **lm()** function. It allows the i^{th} row in data matrix Z to be written as follows:

$$Z_i \approx \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \beta_4 PC_4 + c \tag{3-5}$$

c is the intercept.

By calculating the intercept, we have the coefficient matrix as follows:

$$\beta_{\text{coefficent matrix}} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} & \beta_{1,4} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{100,1} & \beta_{100,2} & \beta_{100,3} & \beta_{100,4} \end{bmatrix}_{100\times 4} \tag{3-6}$$

The prediction result is validated by manual examination of the percentage of overlap with the original data and R^2. R^2 is an indicator reflecting how well the variance of dependent variables can be predicted from predictor variables. It is widely applied in the examination of linear regression.

First, we train this predictive model by fitting the data themselves. This validates the ability of principal components regression to reconstruct the data from only a few variables. The input data is the whole simulation data matrix, and we choose four principal components.

The prediction result is plotted in Figure 3-18:

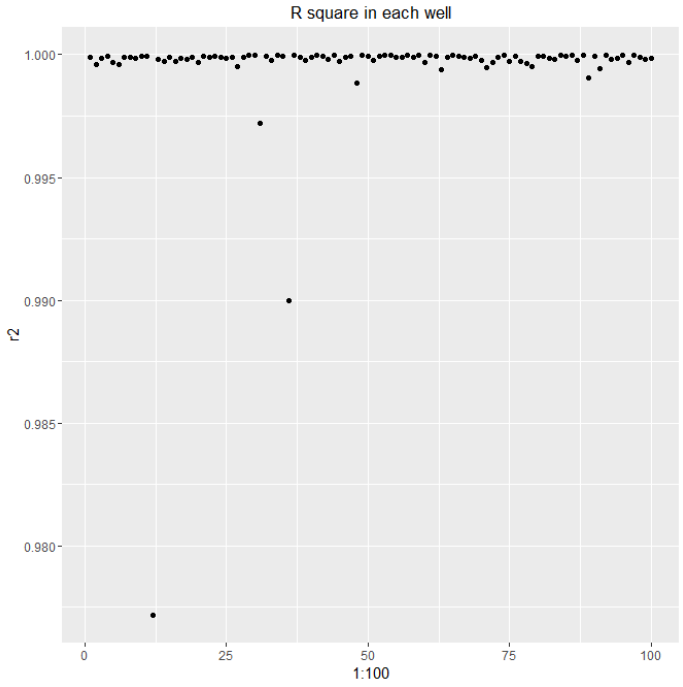


Figure 3-18: Self-fit principal components regression

We can see most wells get near 1.0 R^2, which means their variance can almost be explained with our four principal components. This is reasonable because most curves are smooth and without noise or outliers. We can plot the prediction results and original curve in the same plot to see their overlap ratio. Well 12 is the lowest scoring well in all 100 wells; it has 0.97 R^2.

The well 12 prediction results and the original curve are compared in Figure 3-19. From this figure, we can see that the prediction result fits with an original curve most of the time but has a difference in the transient period.

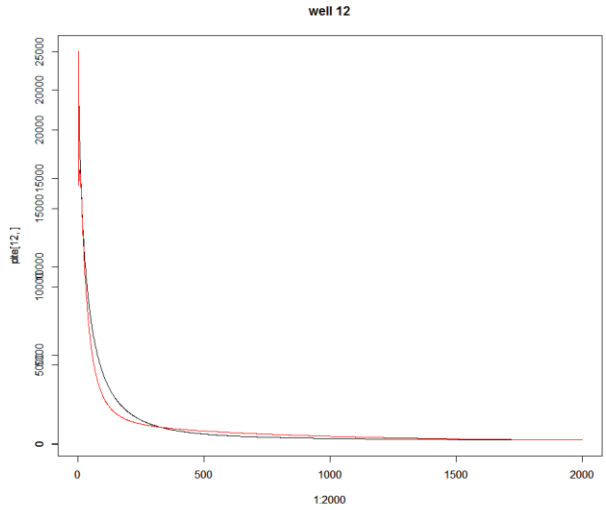


Figure 3-19: Prediction results and original curve of well 12

Then, as we formerly proposed, the addition of *k*-means clustering should enhance the prediction results and R squared score. This time, we trained the model with data from different clusters but all with 2000 days of production. The predictive model was then used to predict the original curve with four principal components.

Figure 3-20 is the predicted result.

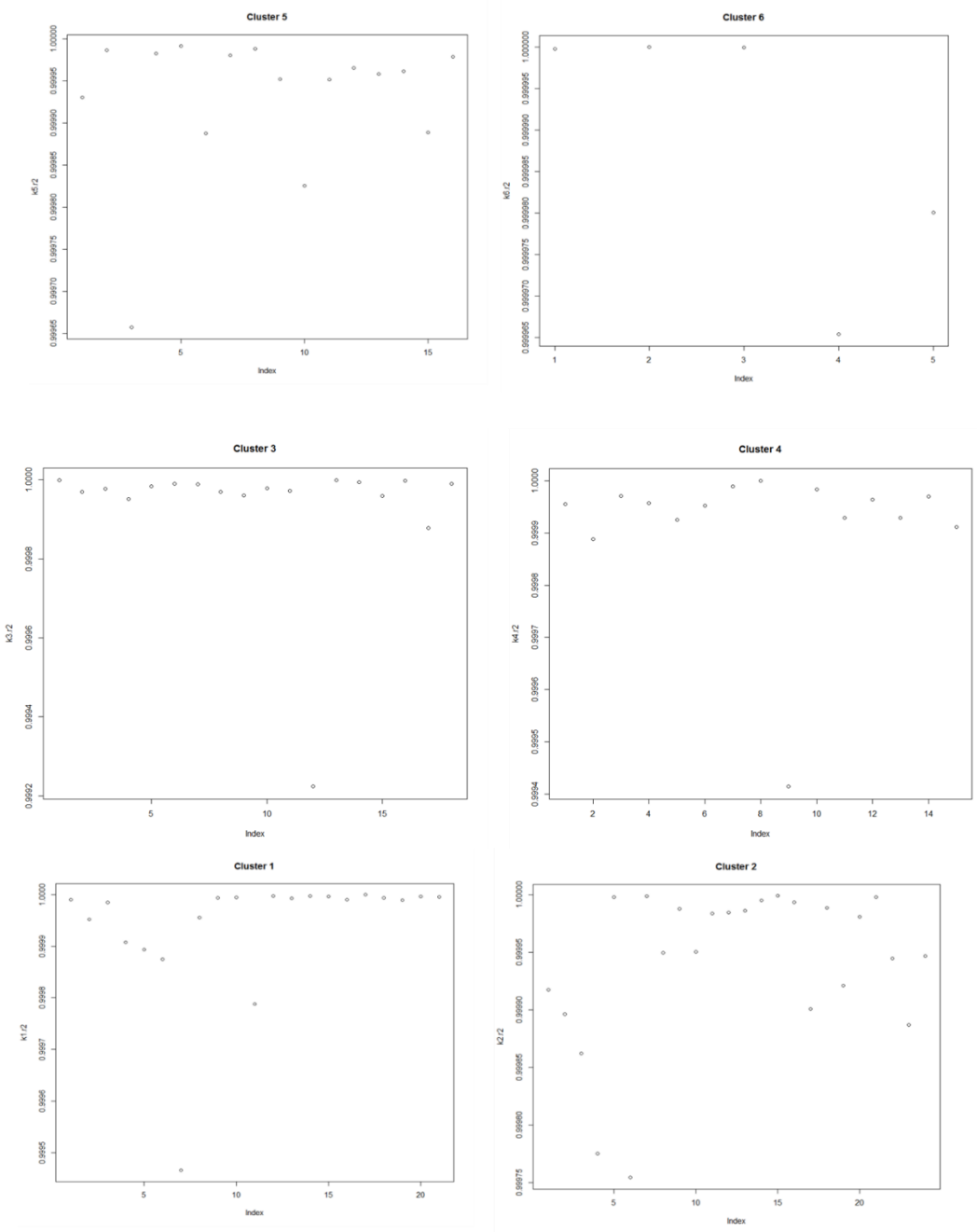
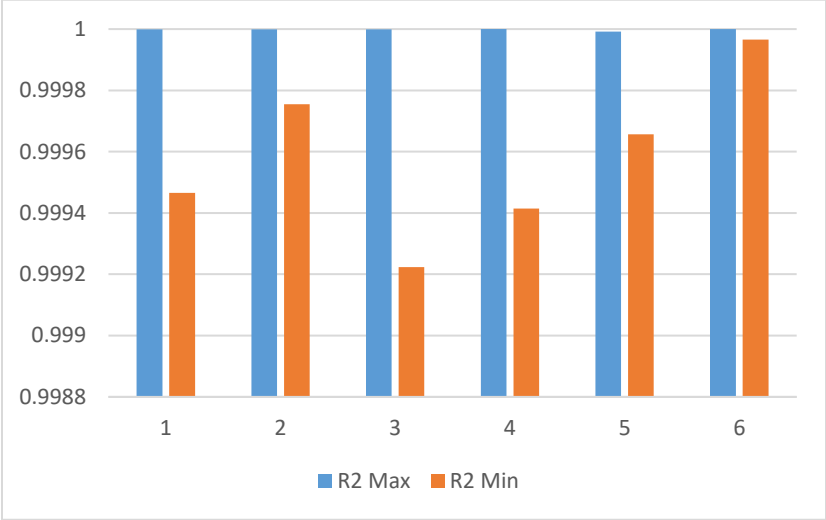


Figure 3-20: Prediction result at Clusters 1, 2, 3, 4, 5, and 6

We summarize the resultS of *k*-means plus principal components linear regression in Table 3-4.

Table 3-4: R^2 summary of *k*-means and PCR



From the summary figure, we find that:

- Using only a few principal components can precisely catch the hidden patterns of the unconventional gas well's decline curve.
- Linear regression with few principal components can reconstruct the decline curve, neatly fitting the original curve.
- With the addition of *k*-means clustering, the prediction result has an increasing R^2 score.

3.4 Performance Forecasting with PCA

Besides fitting history data with principal components, industries are interested in another ability of PCA: capturing the hidden patterns from wells with long production histories and forecasting new well performance.

As discussed above, the optimal number of principal components was chosen as 4. At first, by random sampling, we separated 80 wells as the training set. This was done with the **sample()** function of R. Then we extracted 4 PCs from the training data matrix (80 rows, 2000 columns).

Then, we calculated the coefficient matrix from the testing data matrix. This was done by applying least squares regression between the original data matrix and first 4 rows of the transposed eigenvector matrix. We used the **lm()** function in R. For the testing set, we used two conditions (300 days, 200 days) to do a sensitivity analysis.

3.4.1 Condition 1: Testing Set With 300 Days

The coefficient matrix of 300 days is listed in Table 3-5.

Table 3-5: Coefficient matrix of condition 1

	V1	V2	V3	V4
well 1	-418292	-5324.84	32257.76	5236.338
well 2	-382472	-40490	17620.8	1957.607
well 3	-445911	-24257.6	18944.29	3300.045
well 4	-340760	-75951.2	10693.21	-14780.7
well 5	-456313	-1026.13	50254.94	-3310.18
well 6	-270626	-91376.2	-21085.7	-19450.9
well 7	-207099	-52697.5	-52803.9	-1787.96
well 8	-358113	-34555.7	8854.801	6391.824
well 9	-489051	43370.92	19046.94	-7217.47
well 10	-222266	-20344.2	-8639.12	7589.838

Then we used the first four PCs to multiply it and get the prediction result. The fitting results are shown in Figures 3-21, 3-22, and 3-23.

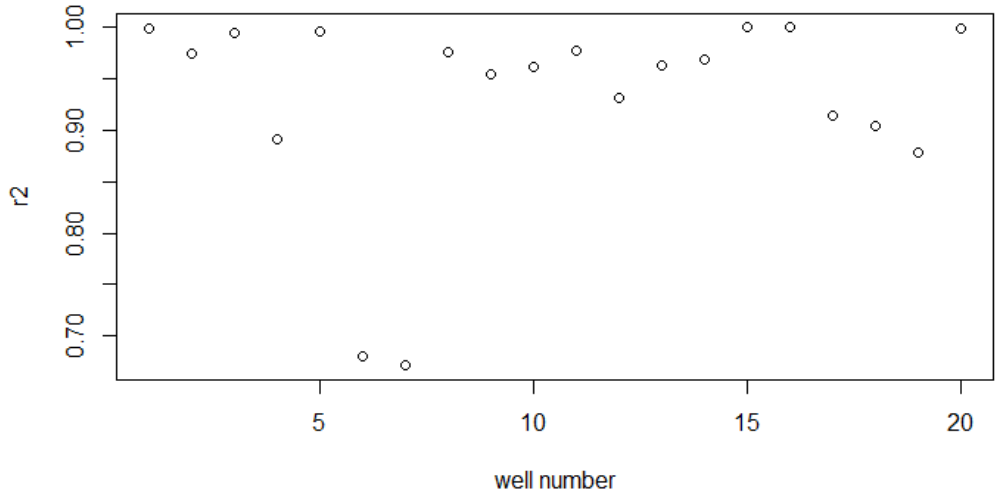


Figure 3-21: Prediction results from testing set condition 1

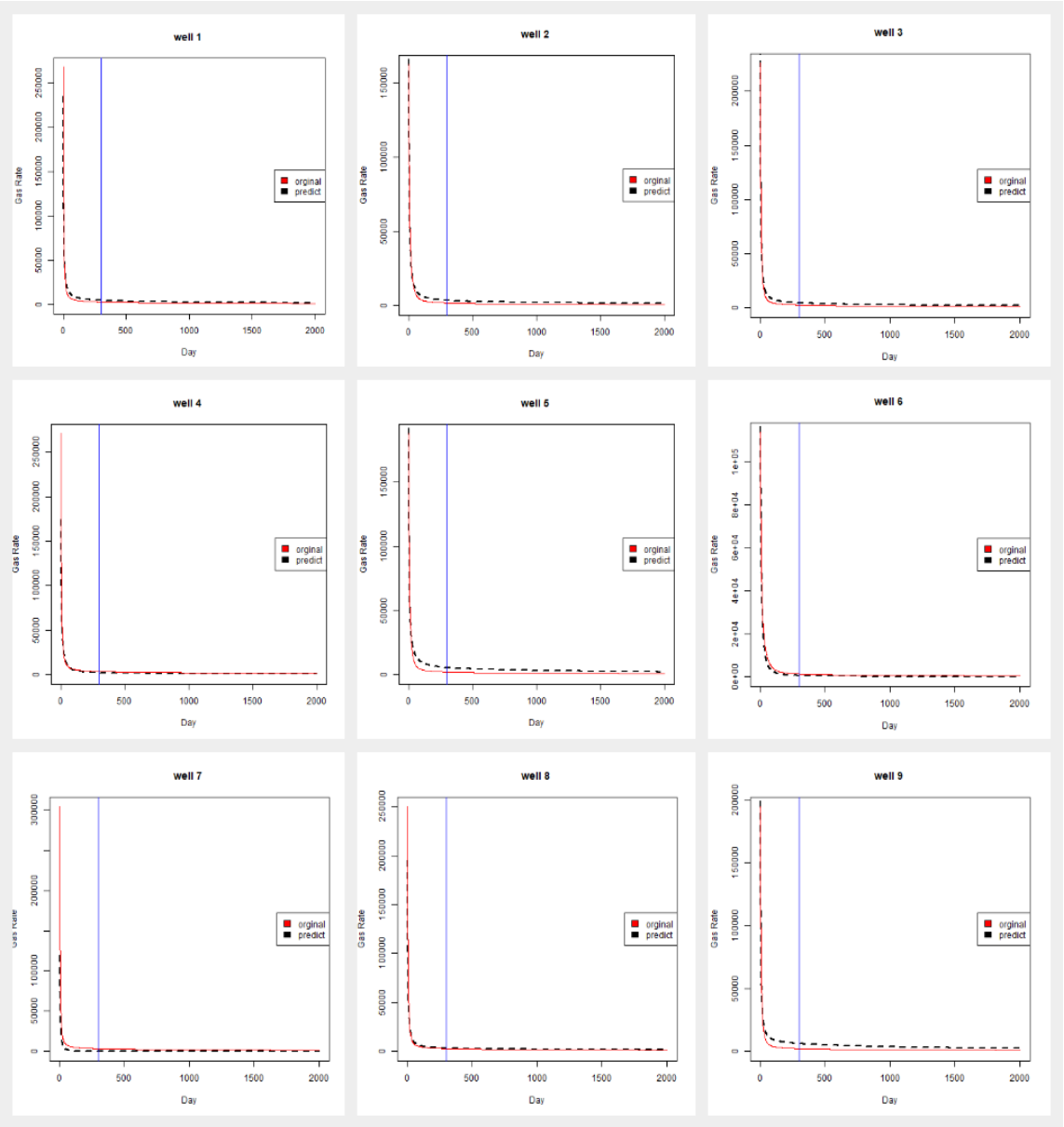


Figure 3-22: Wells 1-9 comparison

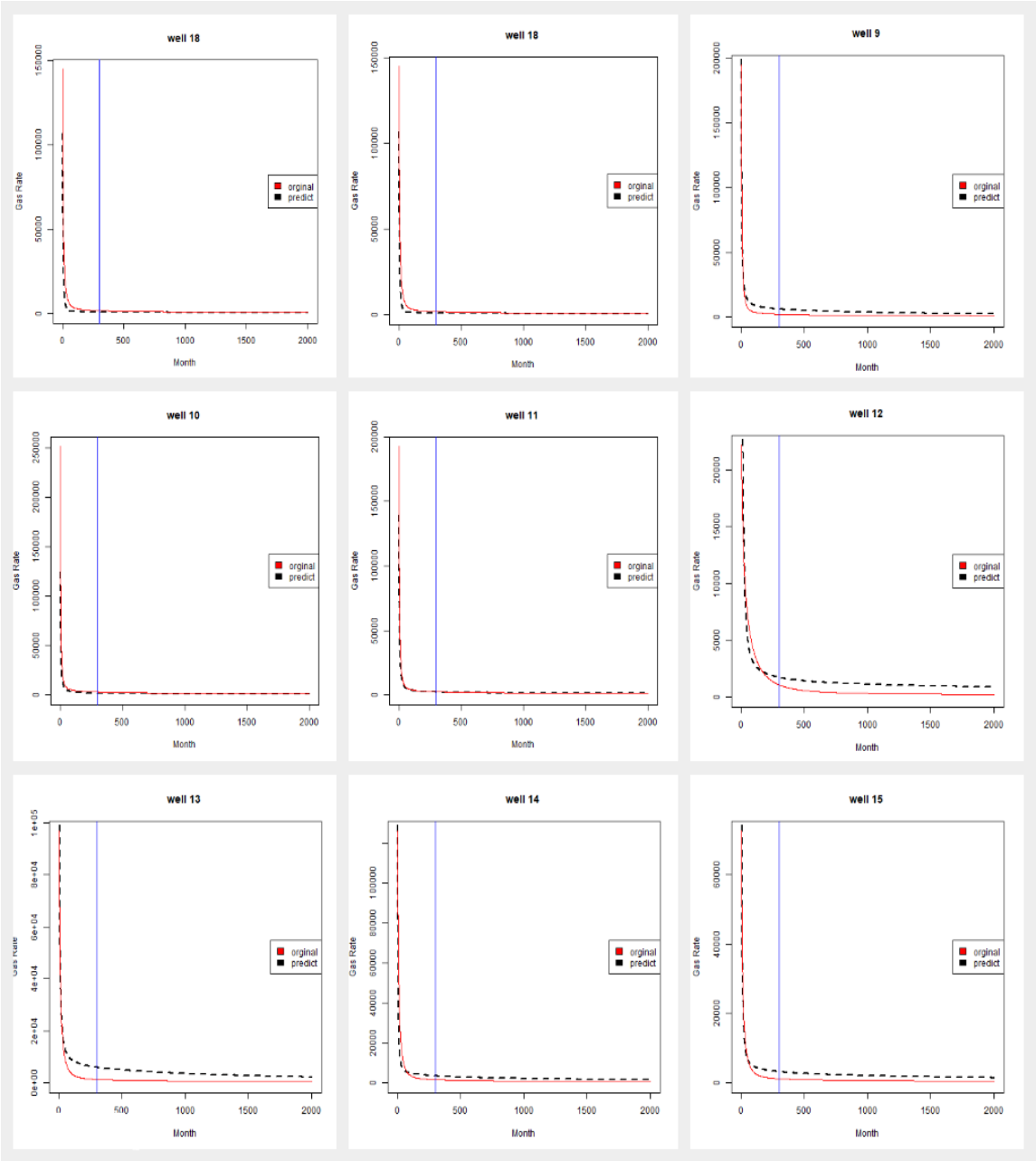


Figure 3-23: Wells 10-19 comparison

3.4.2 Condition 2: Testing Set With 200 Days

In condition 2, we reduced the testing set to 200 days. The training set is still the same 80 wells with 2000 days of production time. By reducing production history, we can test the robustness and capability of linear principal components regression. This is shown in Table 3-6.

Table 3-6: Coefficient matrix of condition 2

	V1	V2	V3	V4
1	-401467	-8814.78	-36794.7	-6839.53
2	-378519	27789.65	-14728.9	-832.165
3	-428589	12683.23	-29028.7	-3724.8
4	-359574	58508.95	19881.21	5425.119
5	-439935	-21909.2	-47467.1	-11648.6
6	-272817	81532.52	16321.76	-8095.79
7	-193059	64320.32	25034.27	-6021.51
8	-351469	26974.9	-11169.9	4903.497
9	-477806	-46541.8	8690.307	8664.429
10	-213364	22184.65	-1650.64	2609.787

The R-square scores were also calculated, as shown in Figure 3-24.

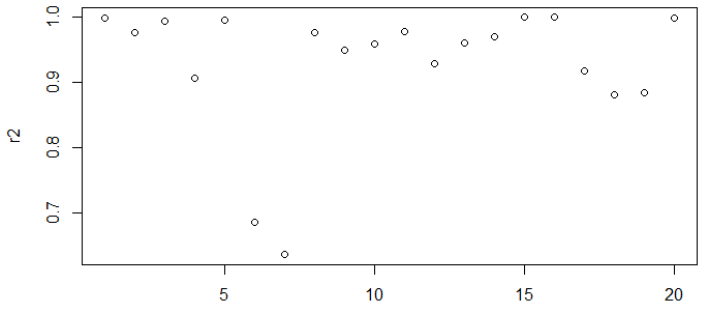


Figure 3-24: Prediction results of testing set condition 2

We can examine the results in Figures 3-25 and 3-26.

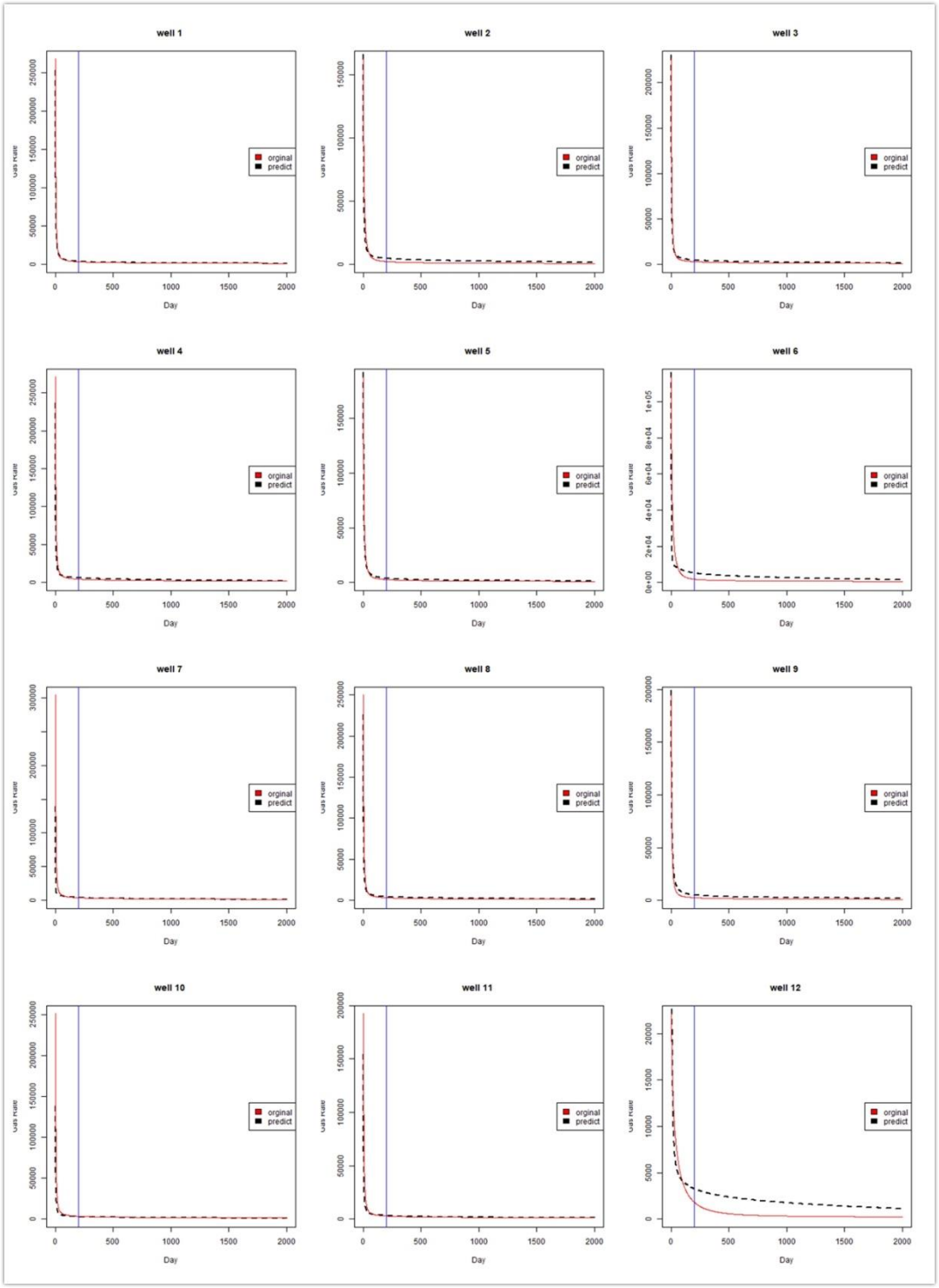


Figure 3-25: Wells 1-12 comparison condition 2

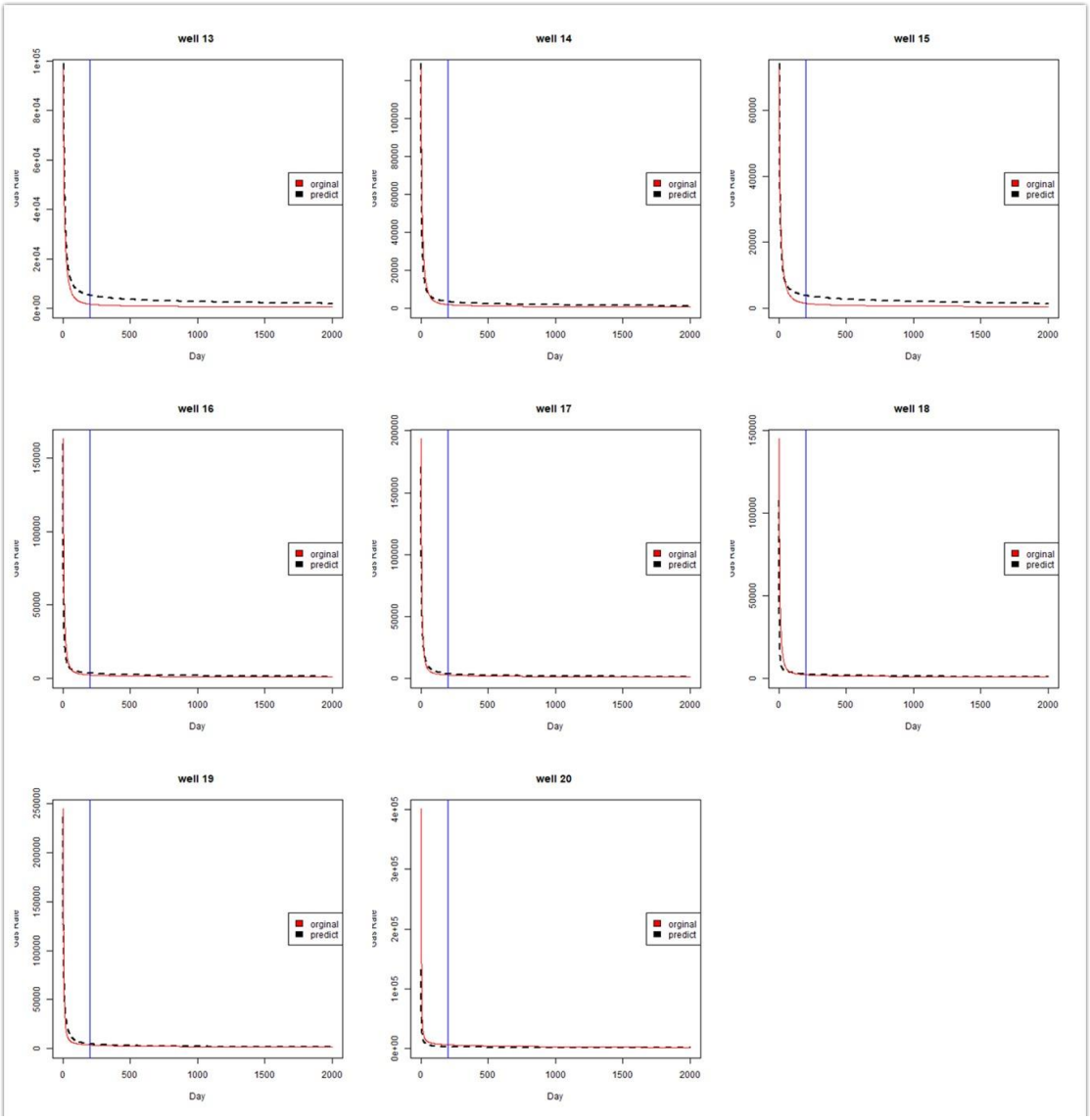


Figure 3-26: Wells 13-20 comparison condition 2

From the above figure, we can find that when learning history is decreasing, the prediction result is affected. The linear regression of Principal Components still shows the ability to catch the performance patterns and fit the original curve.

3.5 Conclusion

In this chapter, we tested the ability of PCA and linear regression with simulation data. The testing result proves that PCA can effectively reduce original data matrix dimensions and reconstruct the matrix from a few principal components.

We also tested the ability of linear regression to predict the performance of new producing wells with a short history. Linear regression of the principal components learned the curve of performance history from the pattern from producing wells with long histories. Even with short-term history, PCA extracted hidden patterns and found the coefficient matrix. This coefficient matrix can convey the information for predicting a well's longer future performance.

CHAPTER IV

APPLYING PCA ON FIELD DATA

After testing PCA and linear regression methods on simulation data, in this chapter, we verified the model on field data. Different from simulation data, field data usually are subject to real production situations and therefore harder to predict. Because we had successfully used simulation data to validate PCA, which learned from the very early production stage and acquired a good fit for prediction, we wanted to prove the ability of PCA to predict real production data.

The data decline curve is shown in Figure 4-1. Most of those wells are horizontal wells stimulated with a multistage hydraulic fracture.

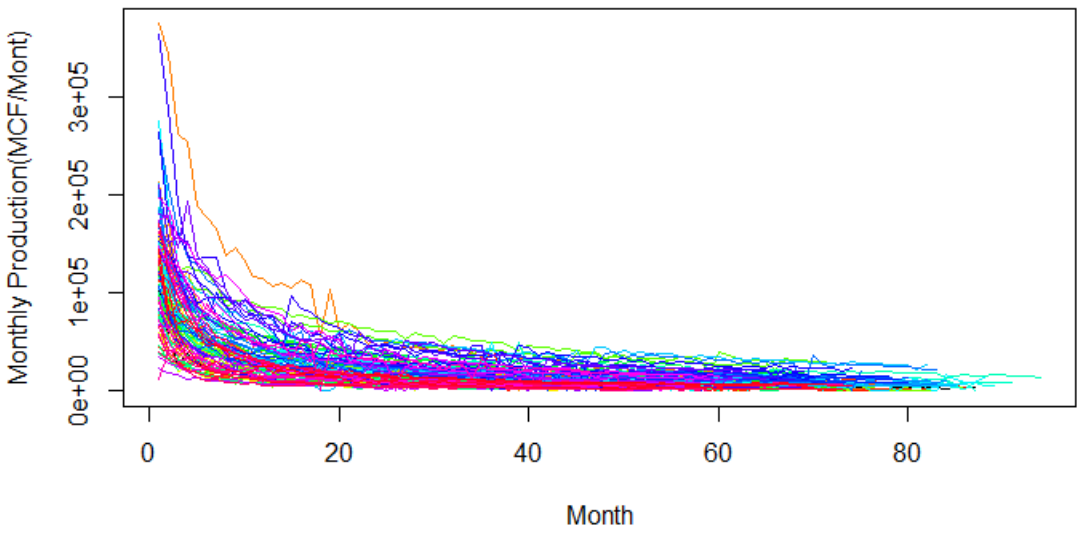


Figure 4-1: Field data decline curve

The producing length varied from a minimum of 45 months to a maximum of 93 months. The histogram is shown in Figure 4-2. Data were collected from online public access datasets.

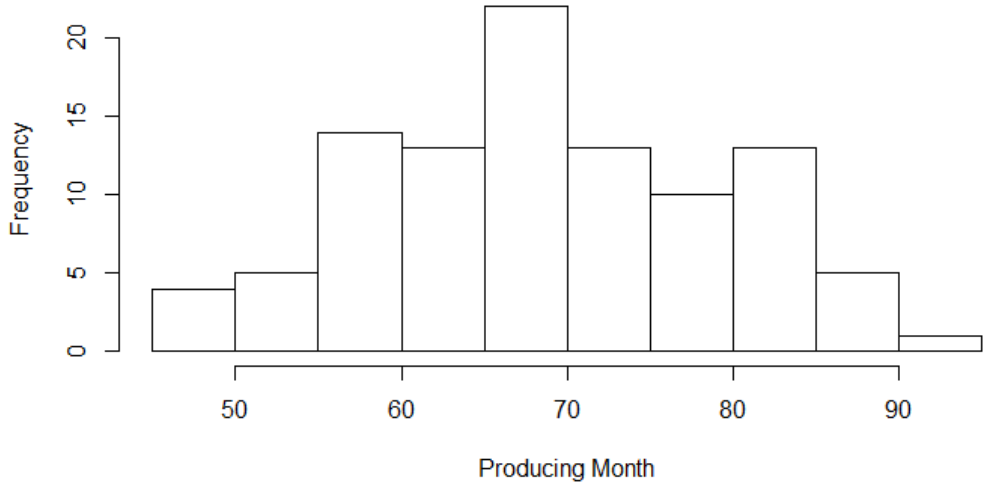


Figure 4-2: Histogram of producing length

The wells were selected from adjacent counties: McMullen, Webb, Dewitt, Bee, and Karnes. The data has the following information: monthly production, API gravity, location, operator, and owner. No further production design or geological information is available. The operating area is shown in Figure 4-3.

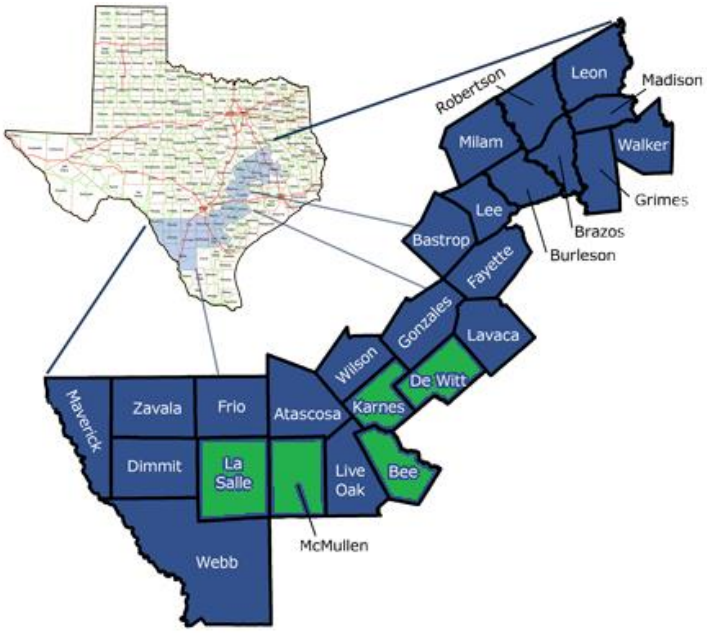


Figure 4-3: Wells located in adjacent counties

4.1 Date Pre-Processing

The original 100 wells each started producing at a different date. We eliminated the first few months' data before the wells reached their peak. The data matrix was established with only a partial decline history.

Before using PCA, we pre-processed the original data matrix to make it smoother. We chose a popular algorithm in well testing called the Bourdet derivative. It can approximate a day's production rate from monthly accumulated production. The comparison between the original wells' decline curves and the Bourdet derivative decline curves are illustrated in Figures 4-4 and 4-5.

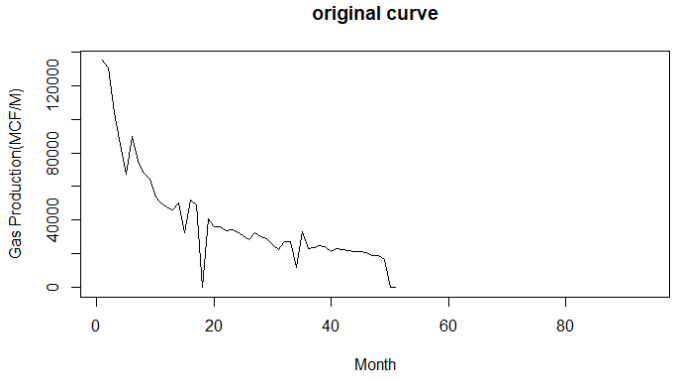


Figure 4-4: Original decline curve of well 88

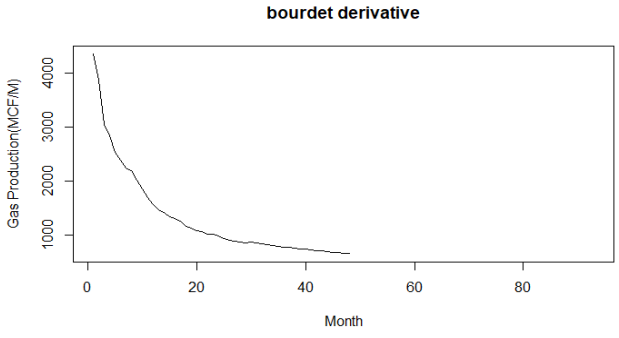


Figure 4-5: Bourdet derivative curve of well 88

After applying the Bourdet derivative algorithm to the original data matrix, we had a smoother data matrix, plotted as Figure 4-6.

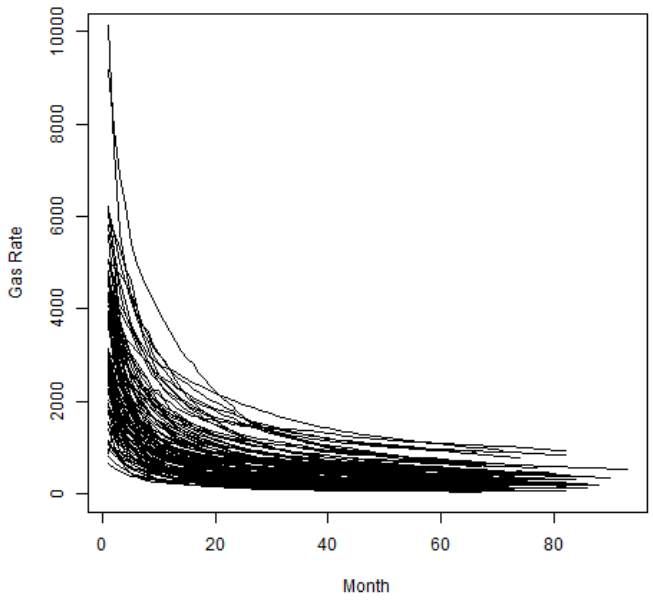


Figure 4-6: Bourdet derivative data matrix

4.2 Workflow of Applying PCA

The general workflow on field data is similar to that for simulation data. It has the following steps:

1. Split data between the training set and the testing set
2. Apply PCA to both the training set and the testing set
3. Decide the number of PCs
4. Establish a coefficient matrix from the testing set (linear regression)
5. Establish a V^t matrix from the eigenvector matrix in the training set
6. Predict results from $Z_{\text{estimate}} = \theta \times V^t$

However, there are still some differences between simulation data and field data. The first thing is the split of the training set and testing set. In simulation data, all wells have the same production history. The split is conducted by random sampling. In field data, only a few wells have the longest production history. This is shown in Figure 4-7.

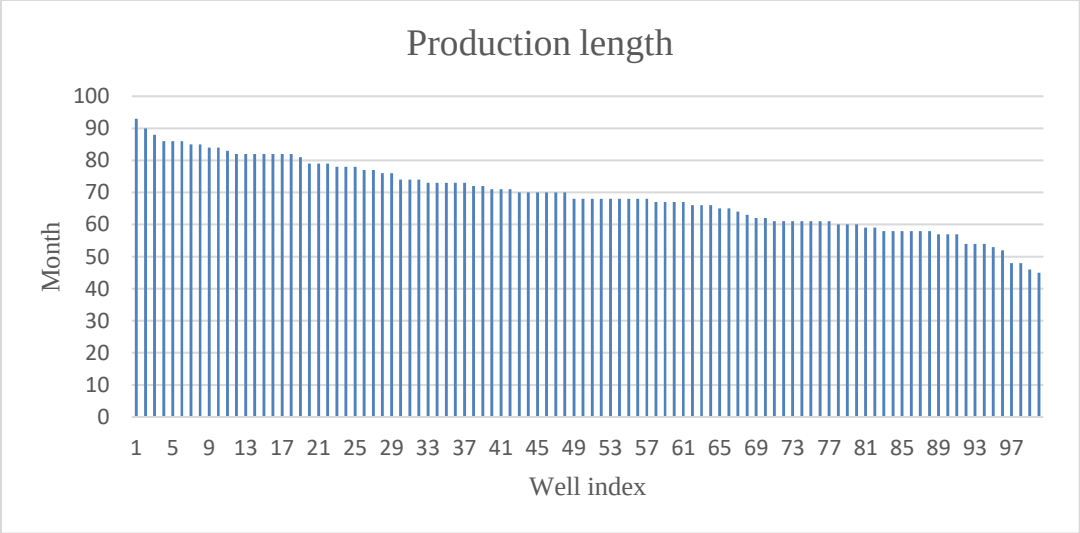


Figure 4-7: Field data matrix production length

In this case, the size of the training sets varied by the length of wells' production history. For example, the longest well has 93 months of production history. If we set this well as the only training set, many of testing wells would not have a good fit with the real situation. So, we need multiple training set samples to ensure we have a balance between the size of the training set and the length of production history.

- Training 1: 3 wells, 86 months
- Training 2: 20 wells, 79 months

We order all wells by their production length. Training set 1 contains the first 3 longest wells; all have 86 months of production history. Training set 2 contains the first 20 wells that all have 79 months of production history. Training set 1 has a relatively small sample for training (3% of the total wells) while training set 2 has a larger sample for training (20% of the total wells).

Also, we set up multiple testing sets with different production times. The industry wishes to have a production forecasting of new wells with limited production history. Therefore, we established five different testing sets to establish a sensitivity analysis.

- Testing 1: 100 wells (all), 45 months
- Testing 2: 100 wells (all), 24 months
- Testing 3: 100 wells (all), 18 months
- Testing 4: 100 wells (all), 12 months
- Testing 5: 100 wells (all), 6 months

4.3 PCA Prediction Model on Field Data

The decline curves of training set 1 are shown in Figure 4-8.

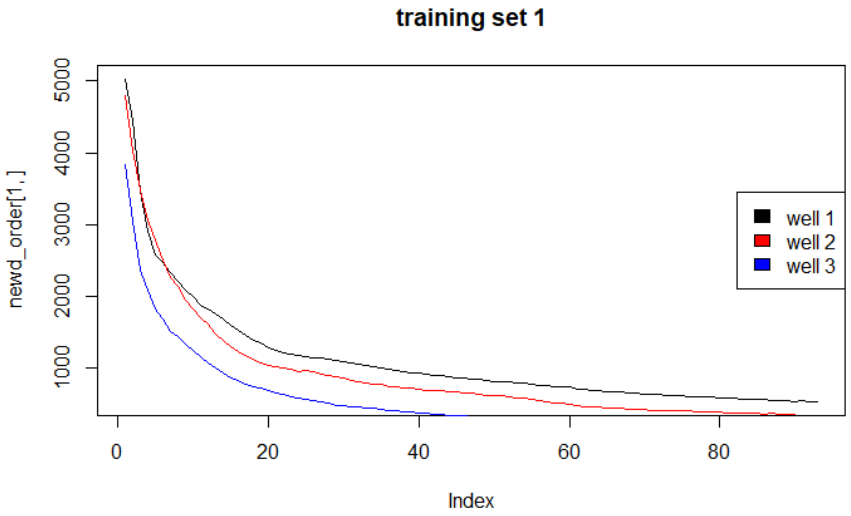


Figure 4-8: Decline curves of training set 1

The decline curves of training set 2 are shown in Figure 4-9:

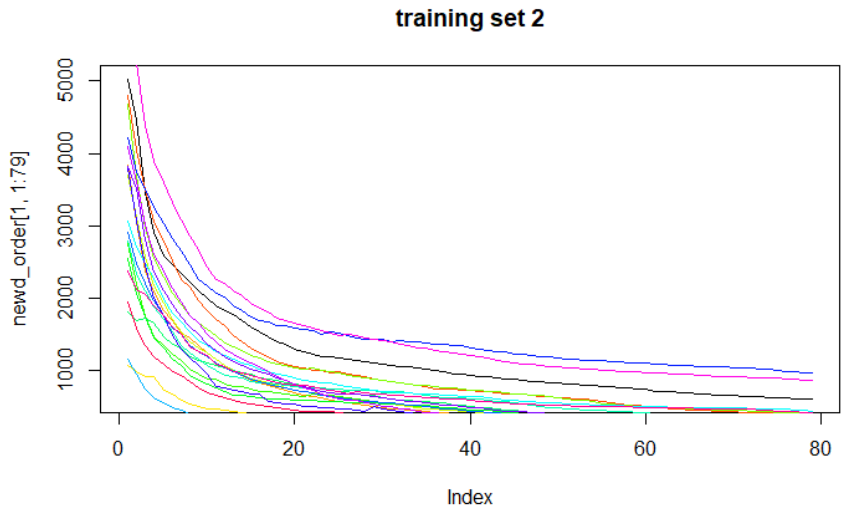


Figure 4-9: Decline curve of training set 2

Conducting PCA on both training 1 and training 2, we constructed the scree plot in Figures 4-10 and 4-11.

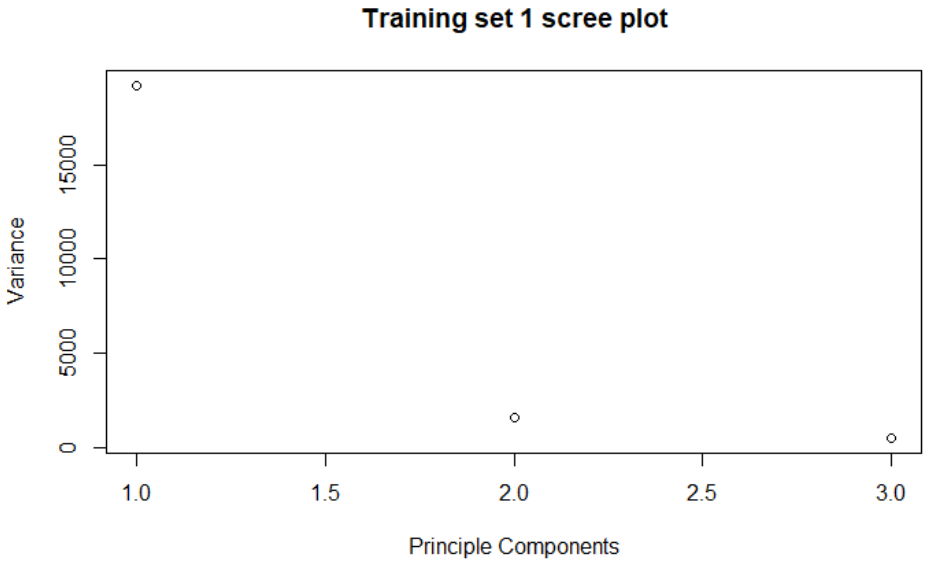


Figure 4-10: Scree plot of training set 1

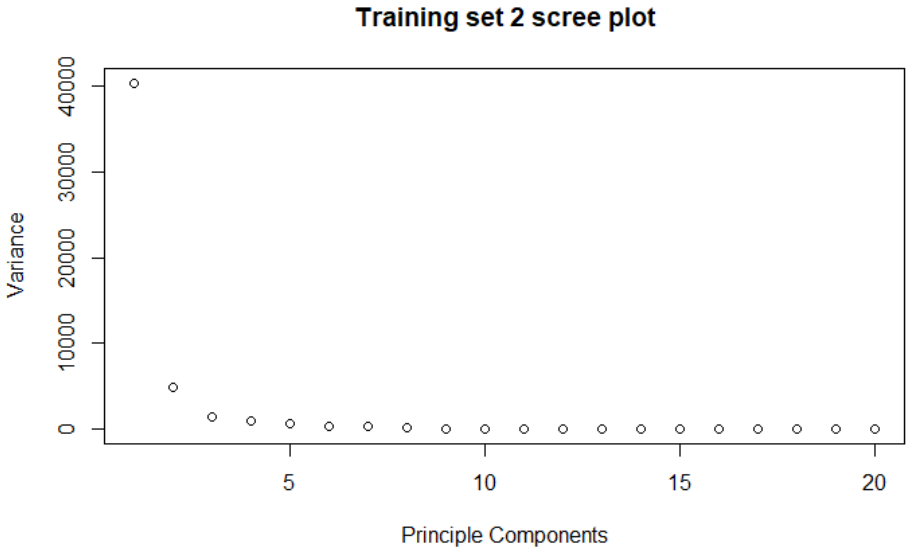


Figure 4-11: Scree plot of training set 2

So, for training set 1, which has only 3 principal components, we took all of them. For training set 2, we used the elbow criterion and took four principal components (95.32% variance).

We plotted five testing set decline curves (Figure 4-12).

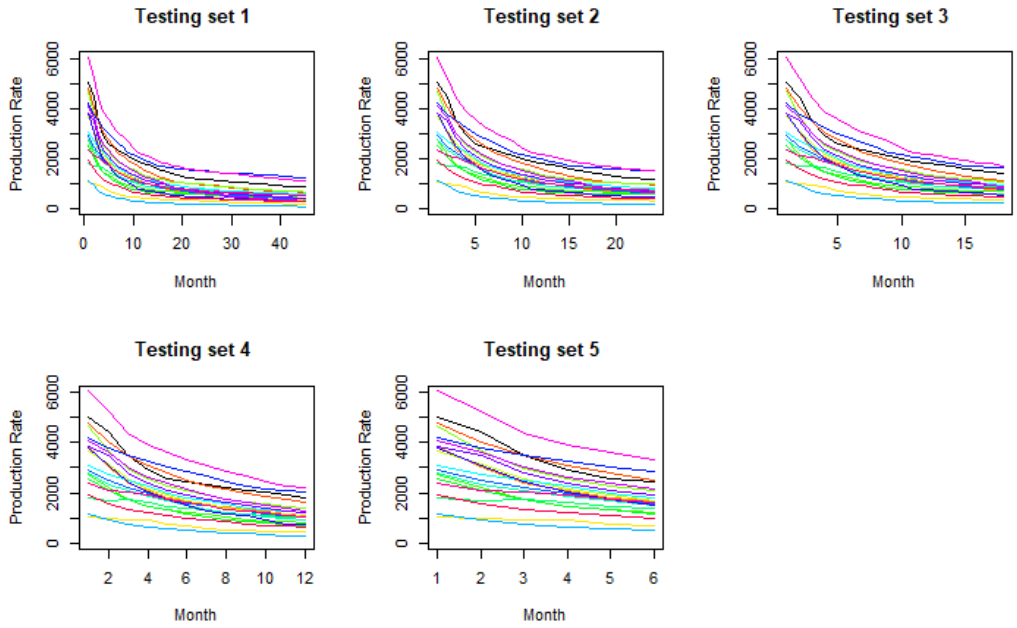


Figure 4-12: Decline curve of testing set 1 to 5

We conducted PCA on them and created the scree plots shown in Figure 4-13. We observe that from testing 1 to testing 5, four principal components would be enough to account for matrix variance.

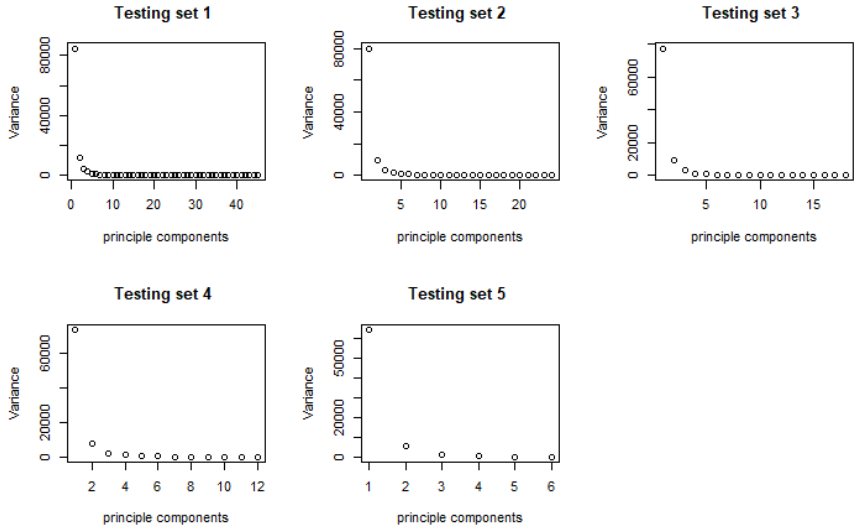


Figure 4-13: Scree plot of testing sets 1 to 5

4.3.1 Prediction from Training Set 1

The coefficient matrix of testing sets 1 to 5 was established by using linear least squares regression between the data matrix and the four principal components of the testing set matrix. The coefficient matrix was then multiplied with the principal components of the training set matrix, and the prediction results were generated.

$$[Z]_{\text{Test}} \approx [\beta]_{\text{coefficient}} \times \left[V^{T}_{\text{test},1} \quad V^{T}_{\text{test},2} \quad V^{T}_{\text{test},3}\right] \tag{4.1}$$

$$\begin{cases} [Z]_{\text{Train 1}} = U \times \Sigma \times \left[V^{T}_{\text{train 1,1}} \quad V^{T}_{\text{train 1 ,2}} \quad V^{T}_{\text{train 1,3}}\right] \\ [Z]_{\text{Train 2}} = U \times \Sigma \times \left[V^{T}_{\text{train 2,1}} \quad V^{T}_{\text{train 2 ,2}} \quad V^{T}_{\text{train 2 ,3}} \quad V^{T}_{\text{train 2 ,4}}\right] \end{cases} \tag{4.2}$$

$$\begin{cases} [Z]_{\text{predict 1}} = [\beta]_{\text{coefficient}} \times \left[V^{T}_{\text{train 1,1}} \quad V^{T}_{\text{train 1 ,2}} \quad V^{T}_{\text{train 1,3}}\right] \\ [Z]_{\text{predict 2}} = [\beta]_{\text{coef}} \times \left[V^{T}_{\text{train 2,1}} \quad V^{T}_{\text{train 2 ,2}} \quad V^{T}_{\text{train 2 ,3}} \quad V^{T}_{\text{train 2 ,4}}\right] \end{cases} \tag{4.3}$$

We conducted the same linear regression on 100 wells with the same training set (3 wells, 88-month production history) on 4 different testing sets (first 24/18/12/6 months). Because of the size of samples (100 wells, 4 different sets), the majority of comparisons are listed in Appendix A. Here, we only pick 8 wells for illustration. Those wells are well 5, well 20, well 30, well 50, well 65, well 75, well 80 and well 95. The red line is the original curve while the black line is the predicted curve. The fits are shown in Figure 4-14 to Figure 4-18.

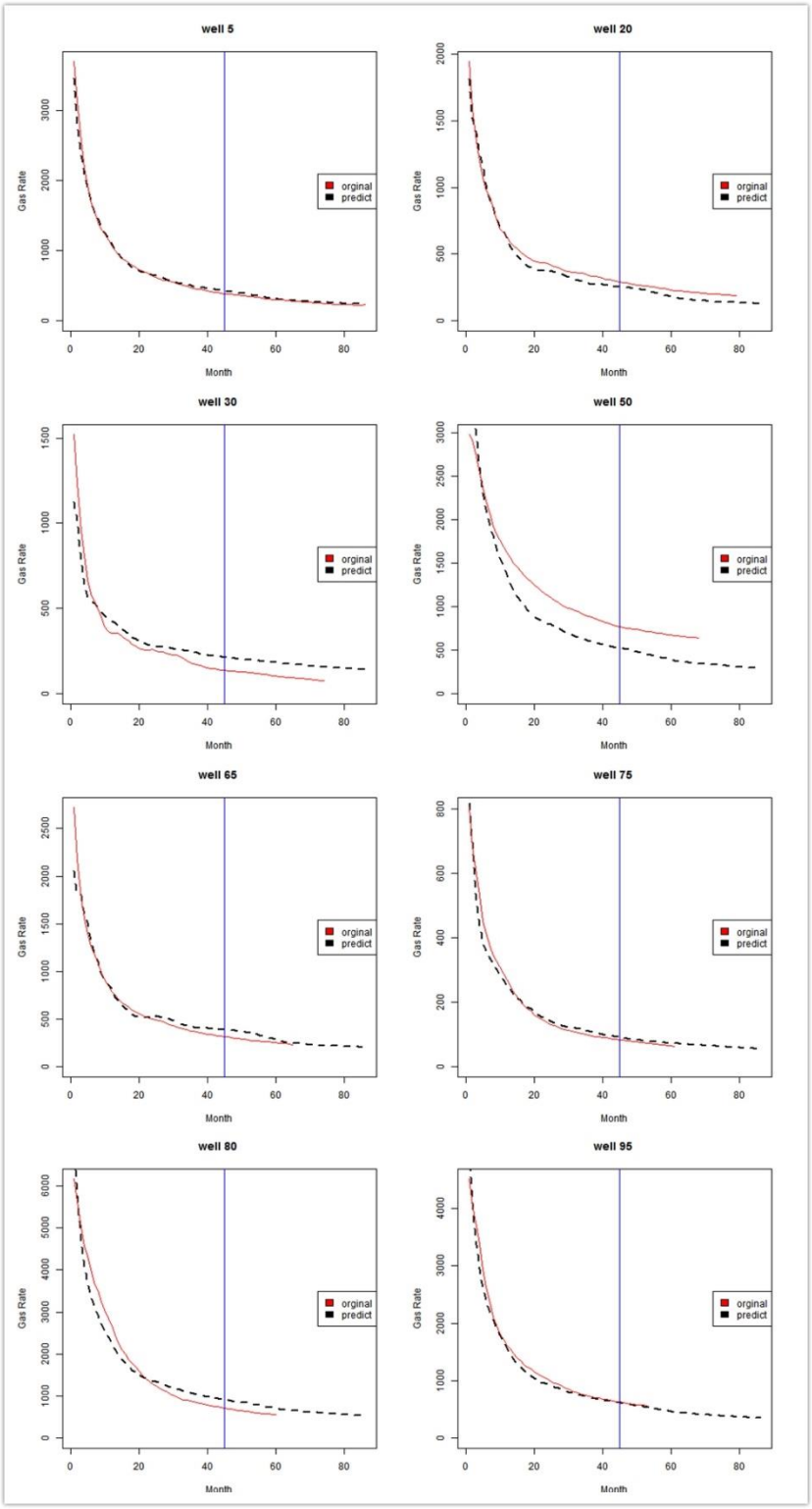


Figure 4-14: Testing set 1 (45 months)

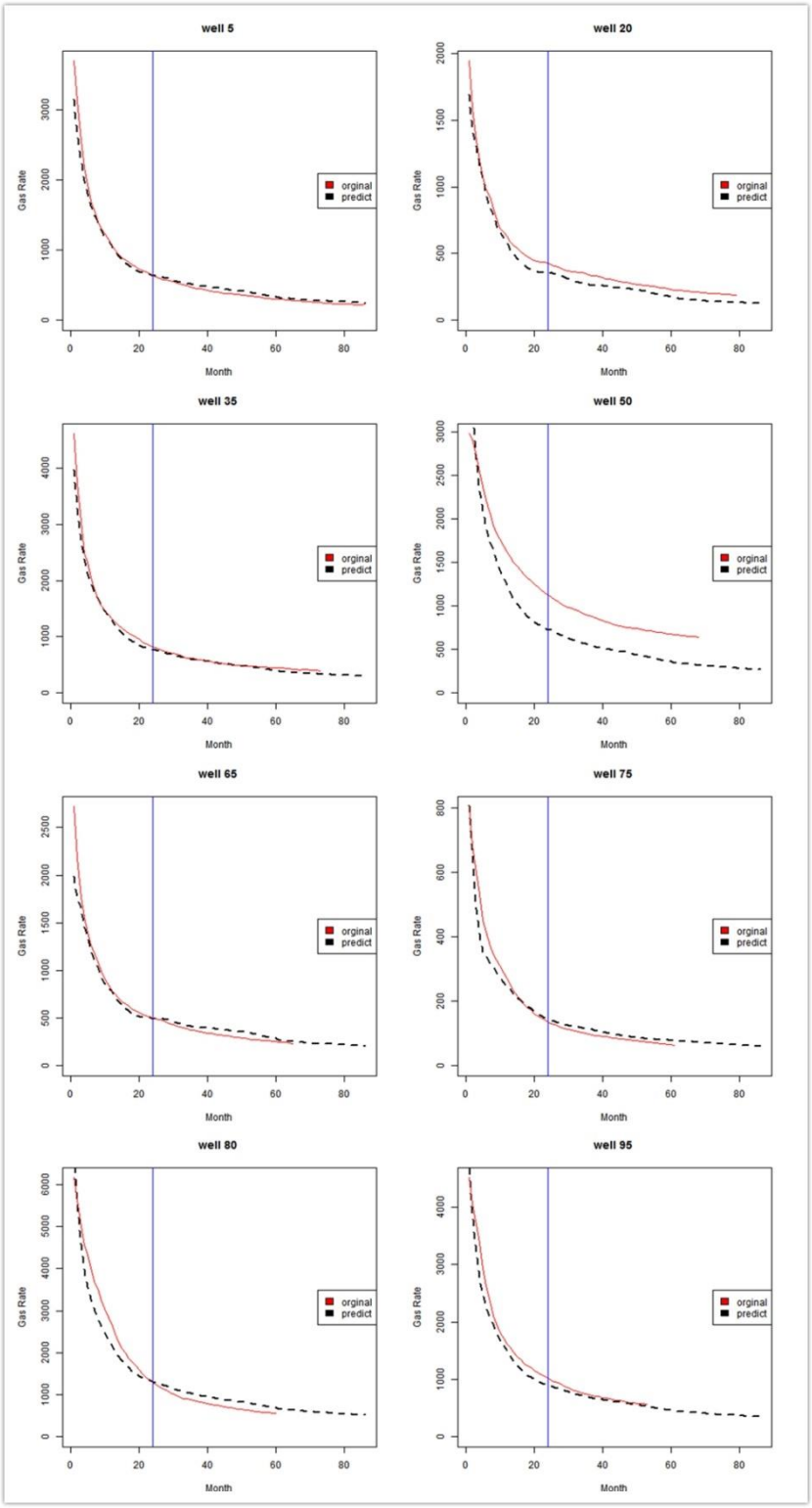


Figure 4-15: Testing set 2 (24 months)

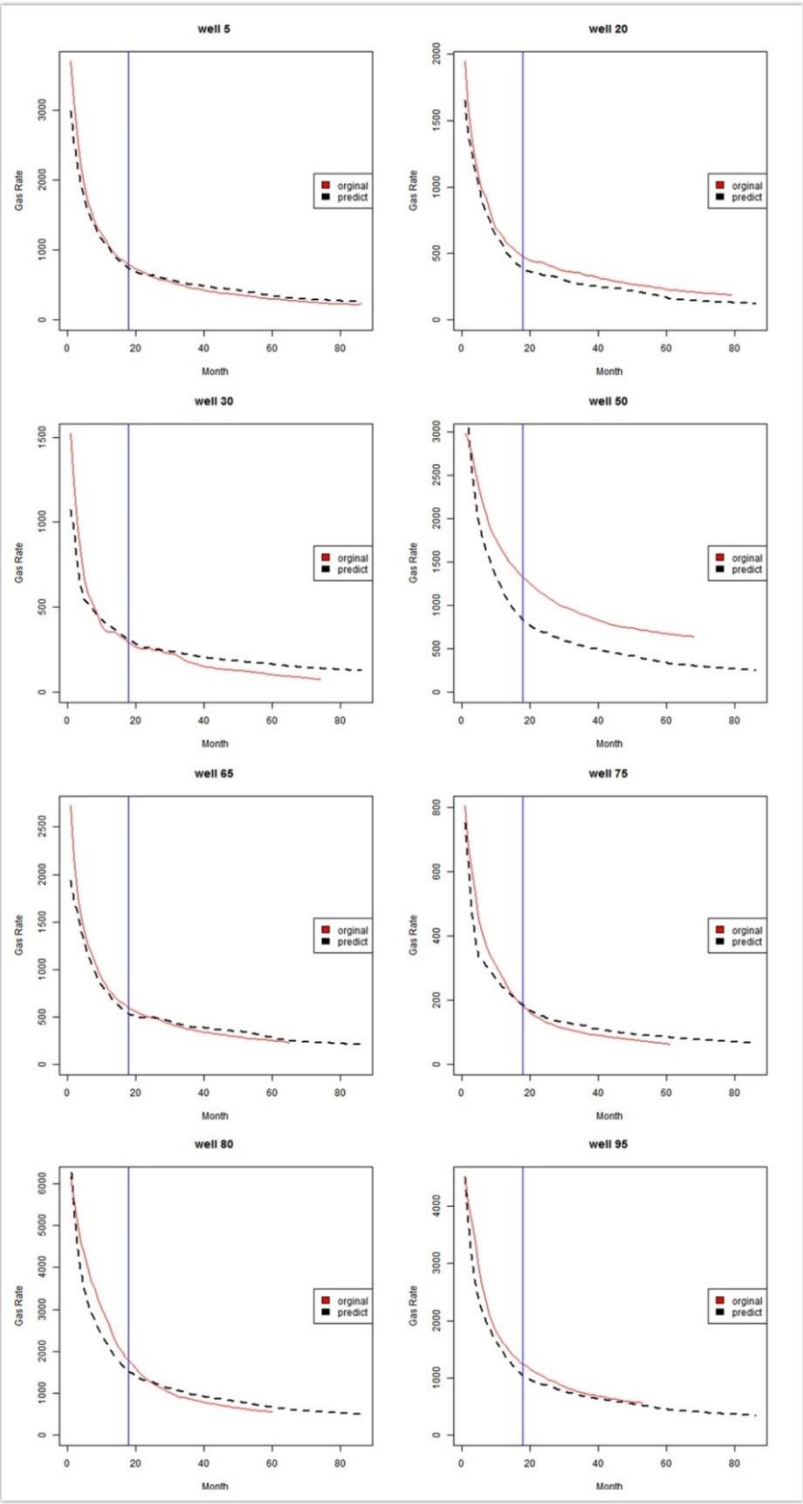


Figure 4-16: Testing set 3 (18 months)

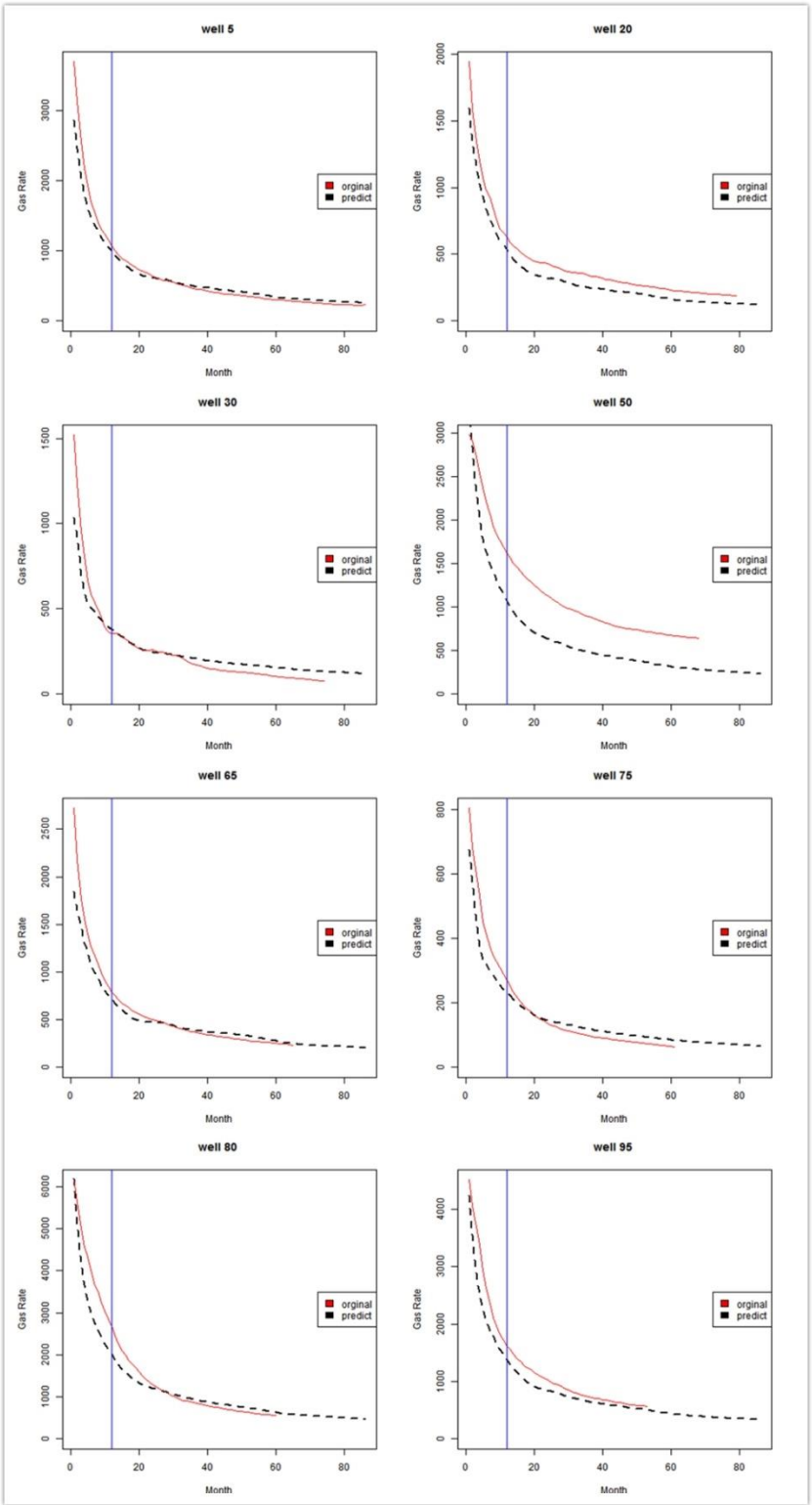


Figure 4-17: Testing set 4 (12 months)

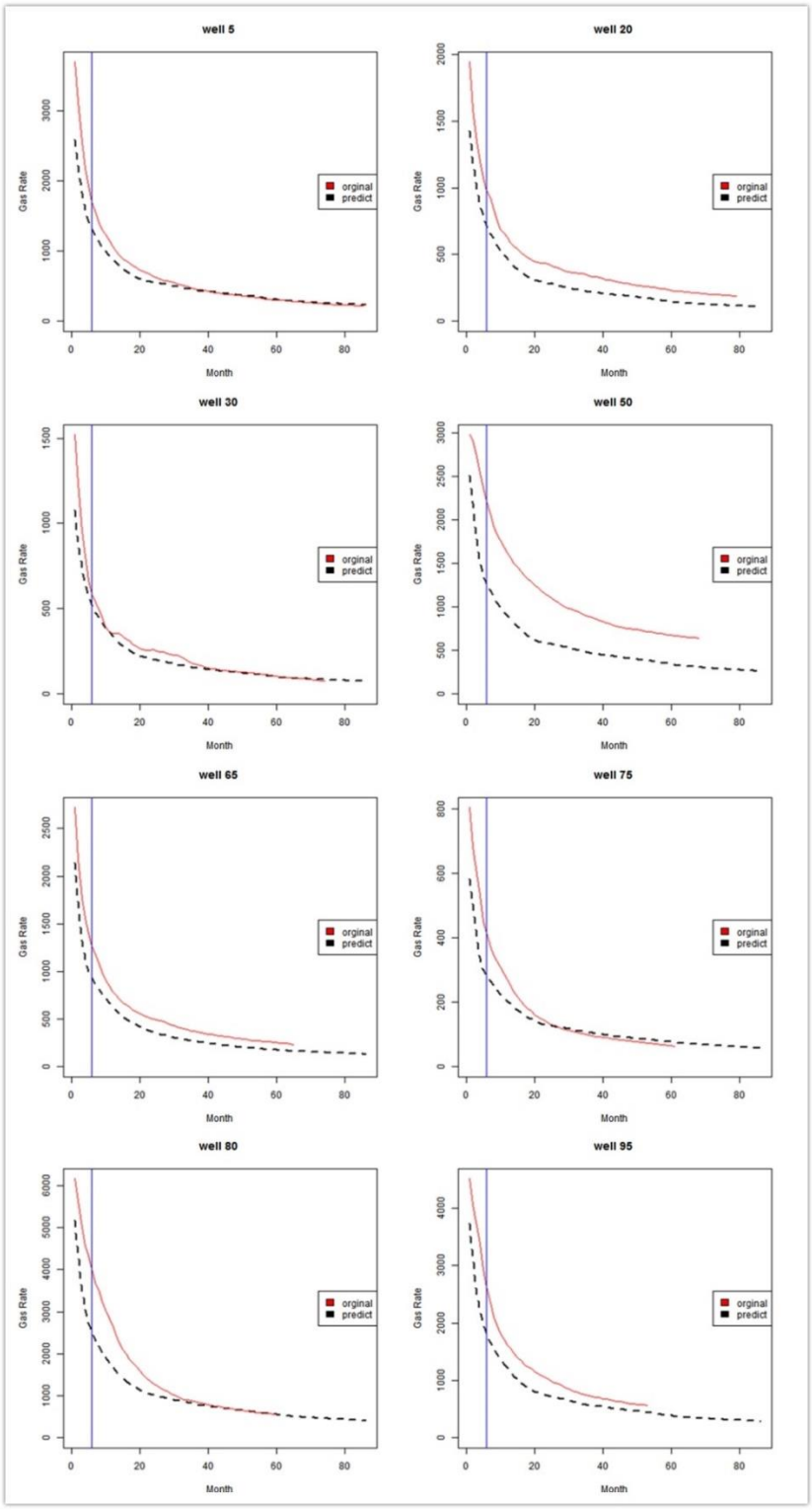


Figure 4-18: Testing set 5 (6 months)

Figure 4-19 is the log-log plot of the 8 wells.

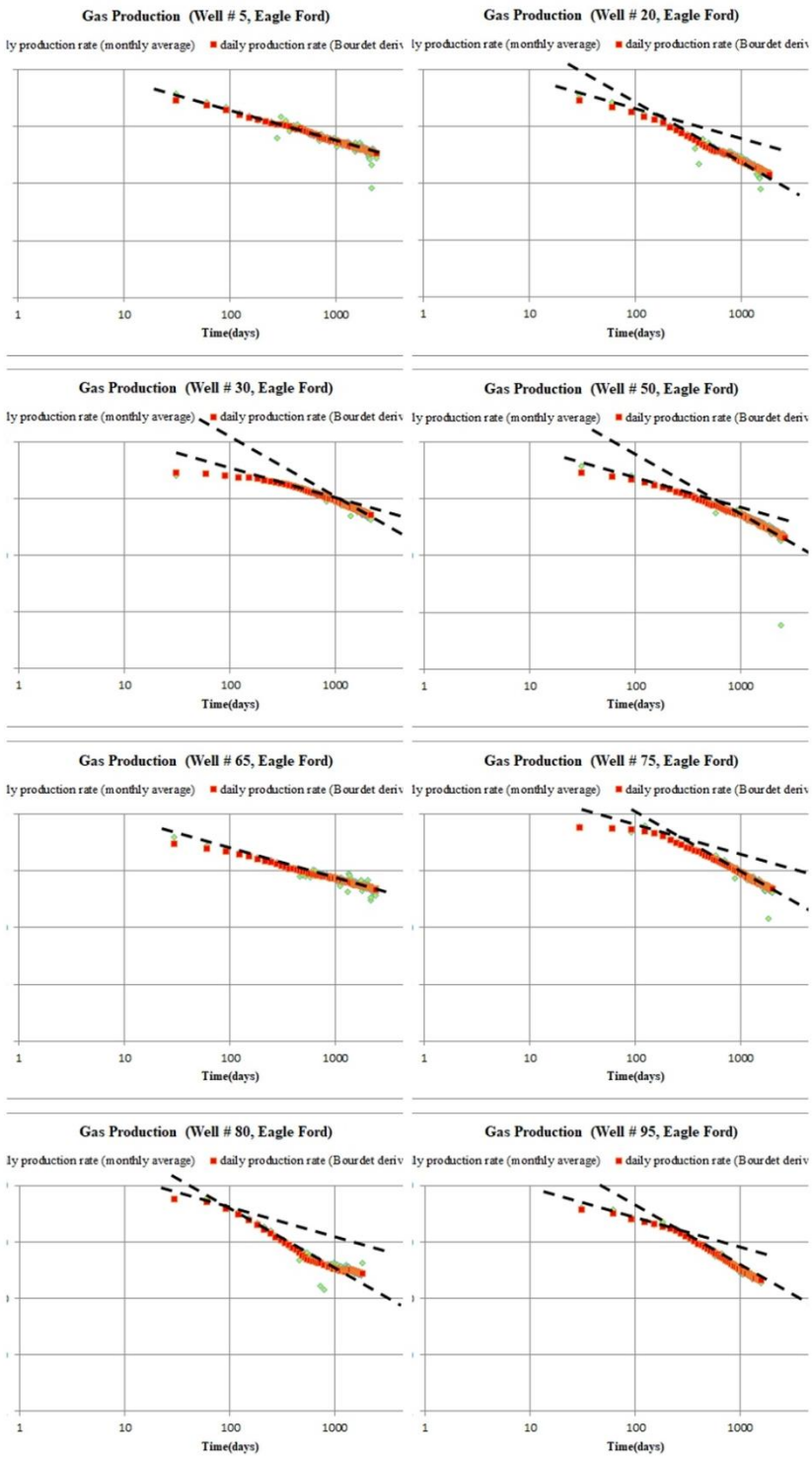


Figure 4-19: Log-log diagnostic plot of sample wells

By comparing the 400 cases (100 wells in 4 testing sets), we find that:

1. Prediction result accuracy has a positive correlation with the length of input time. With the increasing production time of the testing set, prediction results have visible improvement.
2. Testing only data in the linear flow period can also give PCA regression reasonable certain prediction accuracy.

4.3.2 Prediction from Training Set 2

Training set 1 included only the three longest-producing wells to learn their history. This might be too few samples. Therefore, in training set 2, we increased the sample to 20 wells (79 months of history).

Some of the sample wells and their prediction results appear in Figures 4-20 to 4-24.

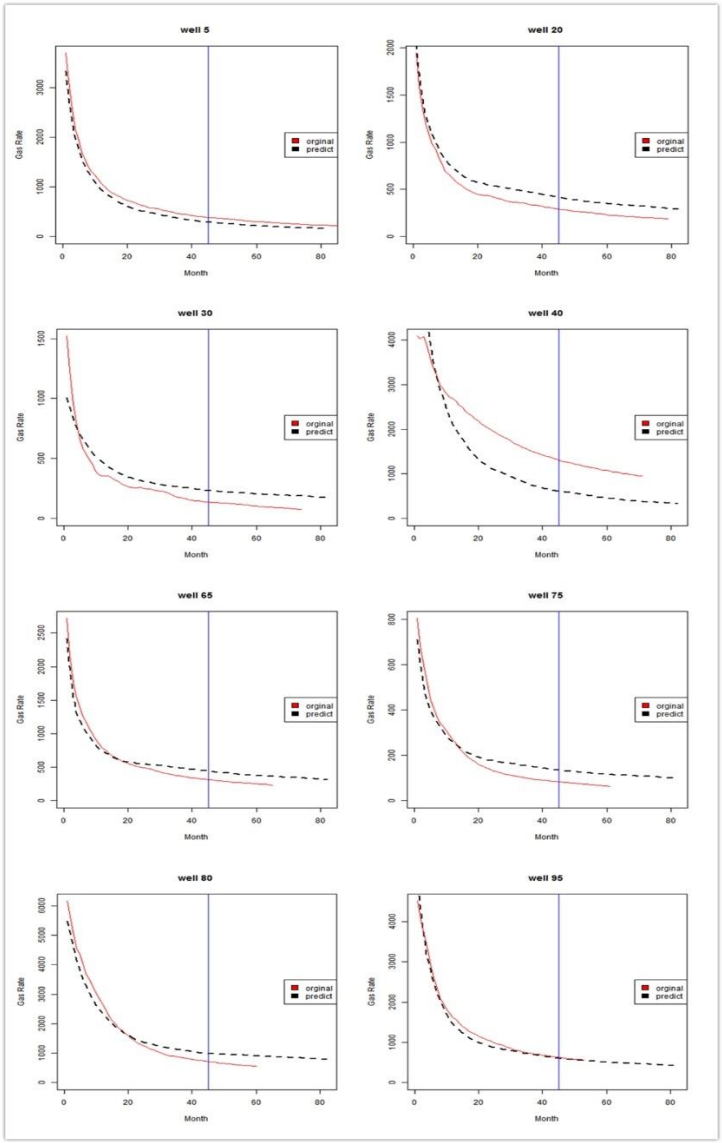


Figure 4-20: Testing set 1 (45 months)

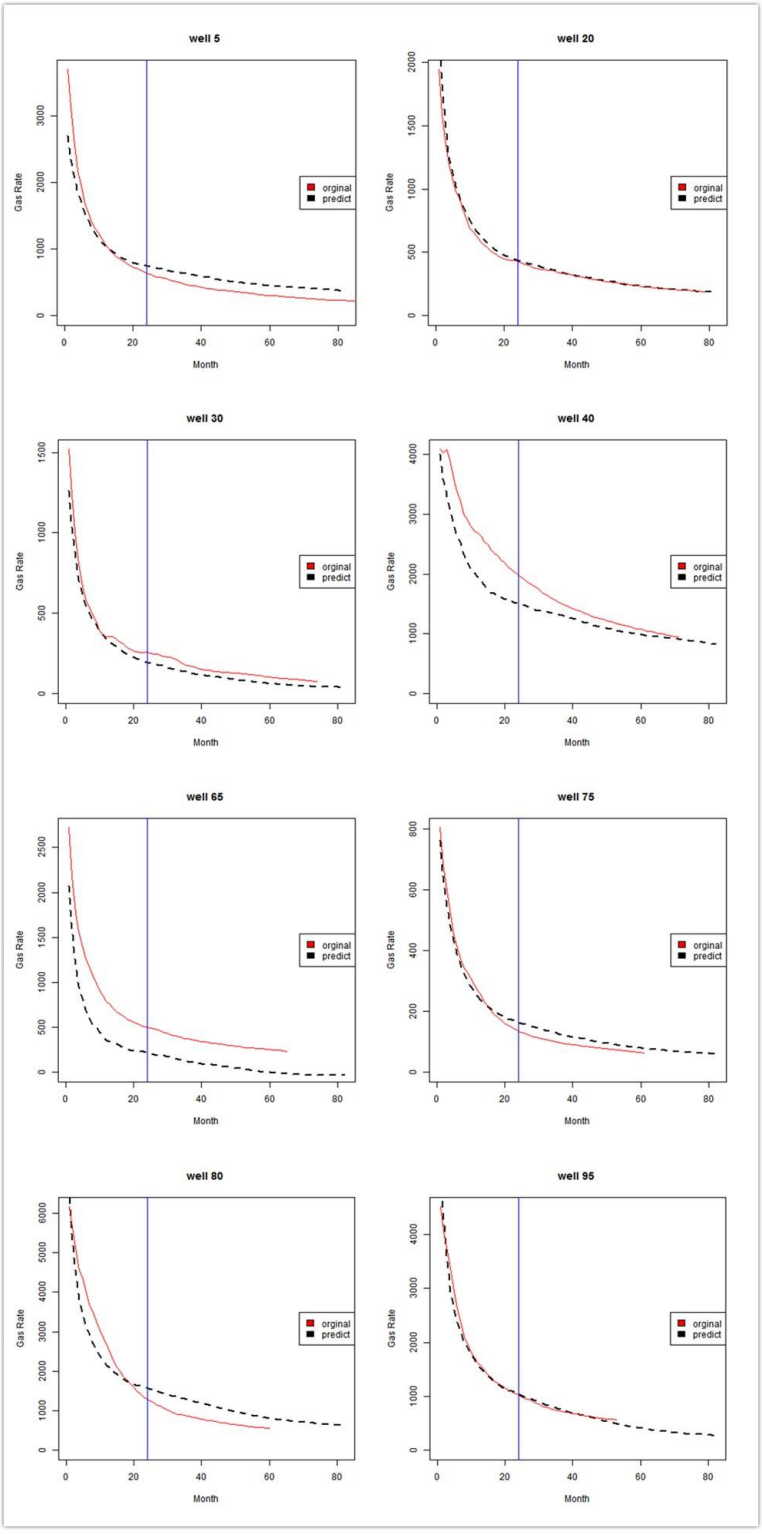


Figure 4-21: Testing set 2 (24 months)

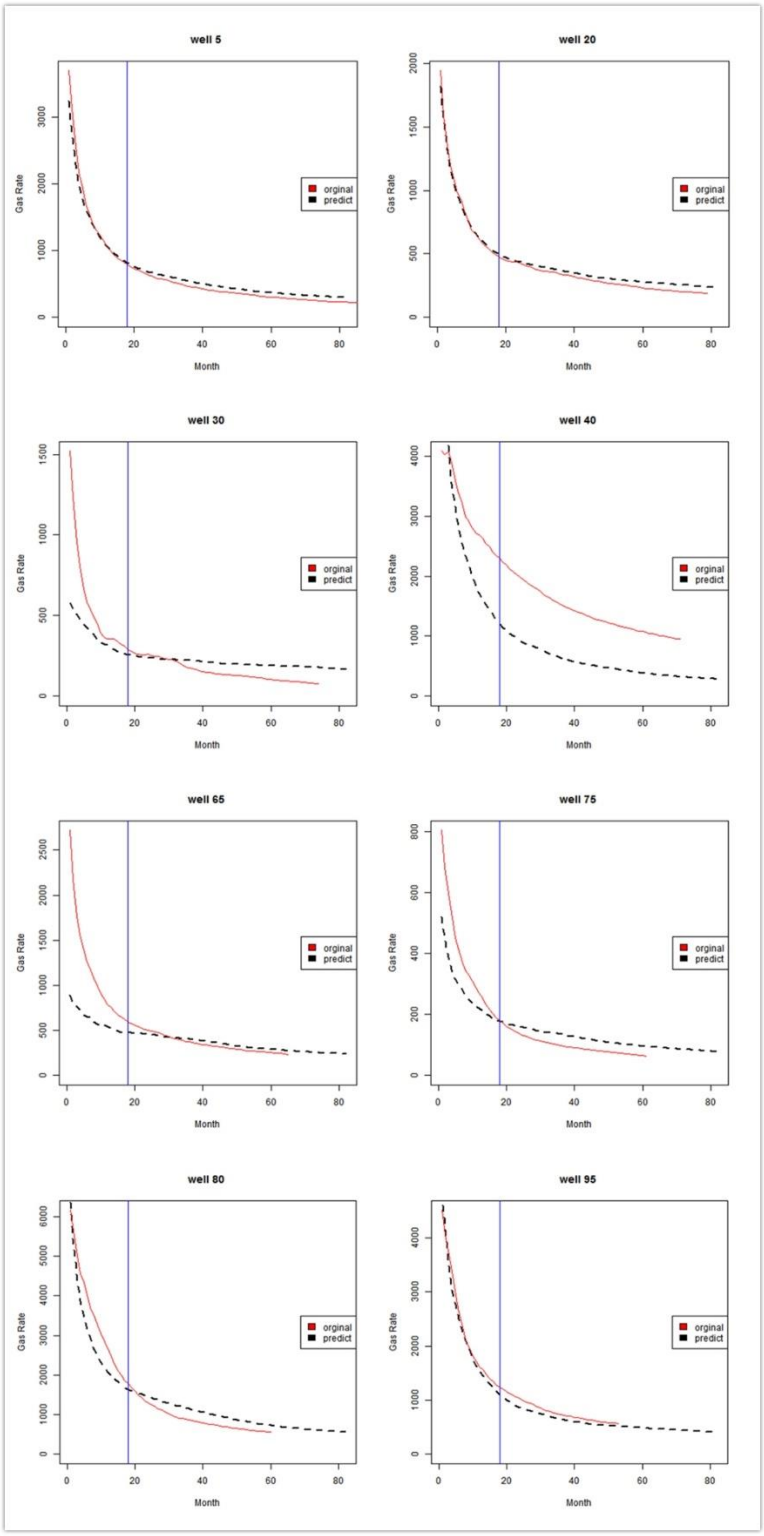


Figure 4-22: Testing set 3 (18 months)

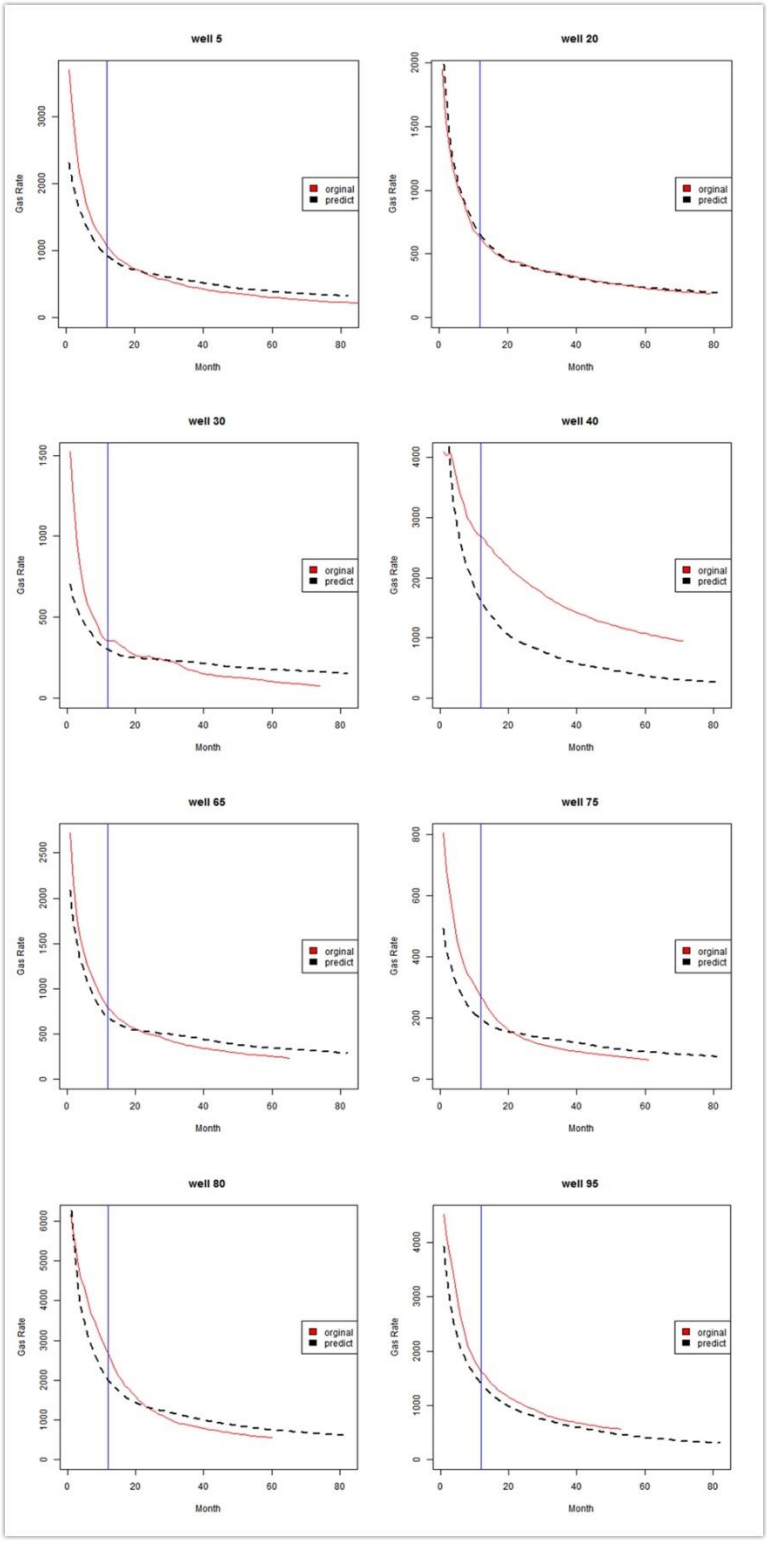


Figure 4-23: Testing set 4 (12 months)

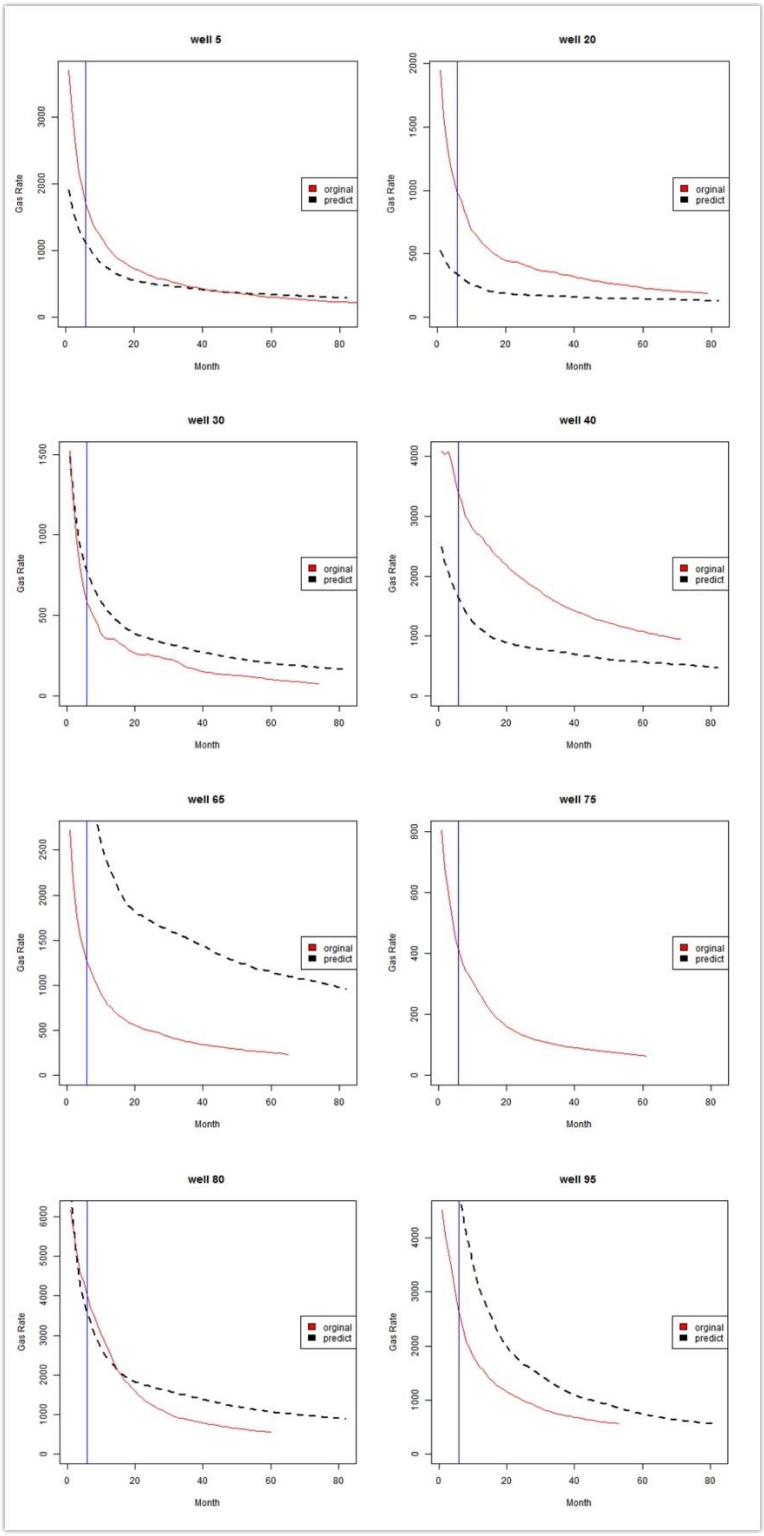


Figure 4-24: Testing set 5 (6 months)

Because of the increased number of wells, we supposed that the prediction results should have an enhanced performance. However, by only applying the linear regression of the PCA prediction method, some of the production forecasting for the test wells shows an even worse fit than results in training set 1. This is shown in Figure 4-25.

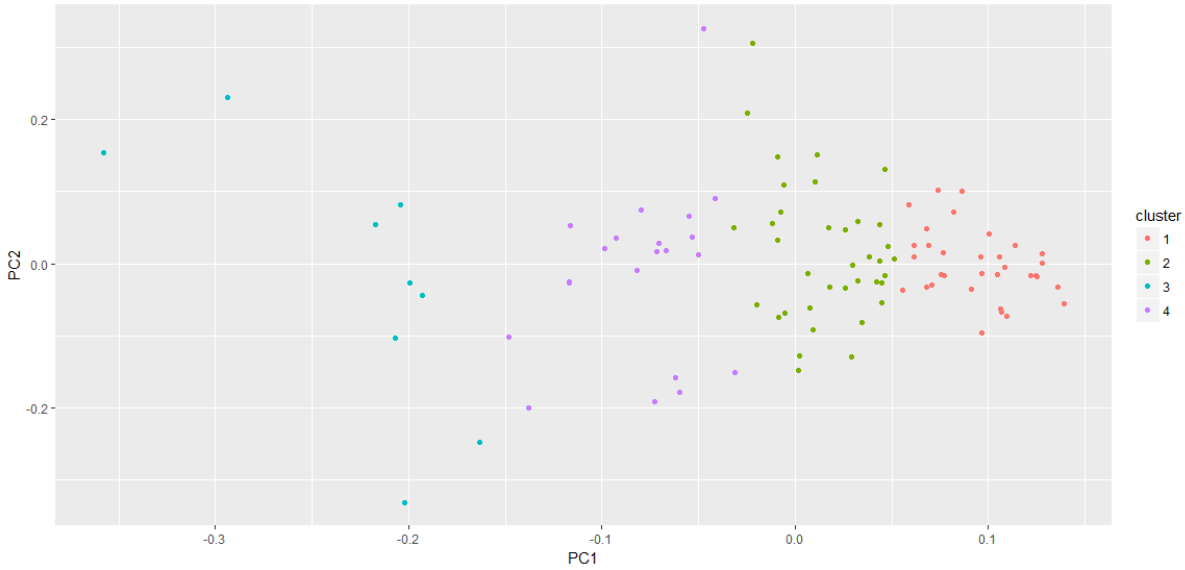


Figure 4-25: Field dataset *k*-means clustering

Their distribution in plots is calculated by *k*-means clustering. Different colors of points represent different clusters. By investigating the clustering result, we saw both training set 2 and the field dataset showing a varied distribution of each well's performance.

By only picking those wells with longest production history, the historical pattern would not fit the new wells' decline trends. If we wish apply linear regression of PCA to predict new wells, new wells should have history data similar to the same cluster (Figures 4-26 and 4-27).

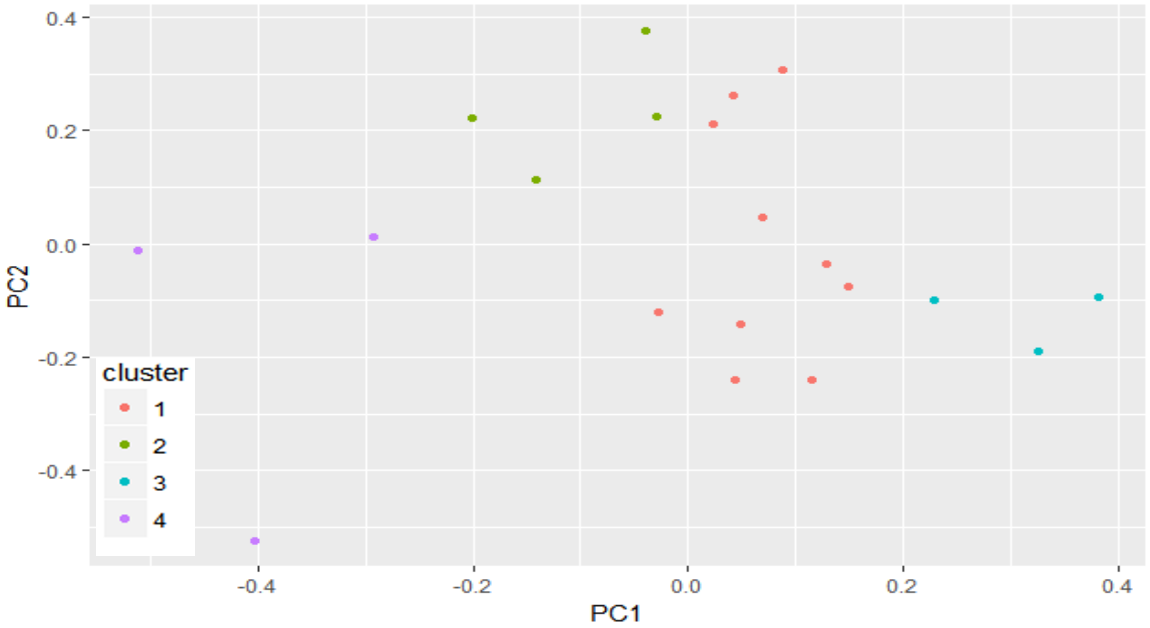


Figure 4-26: Training set 2 *k*-means clustering

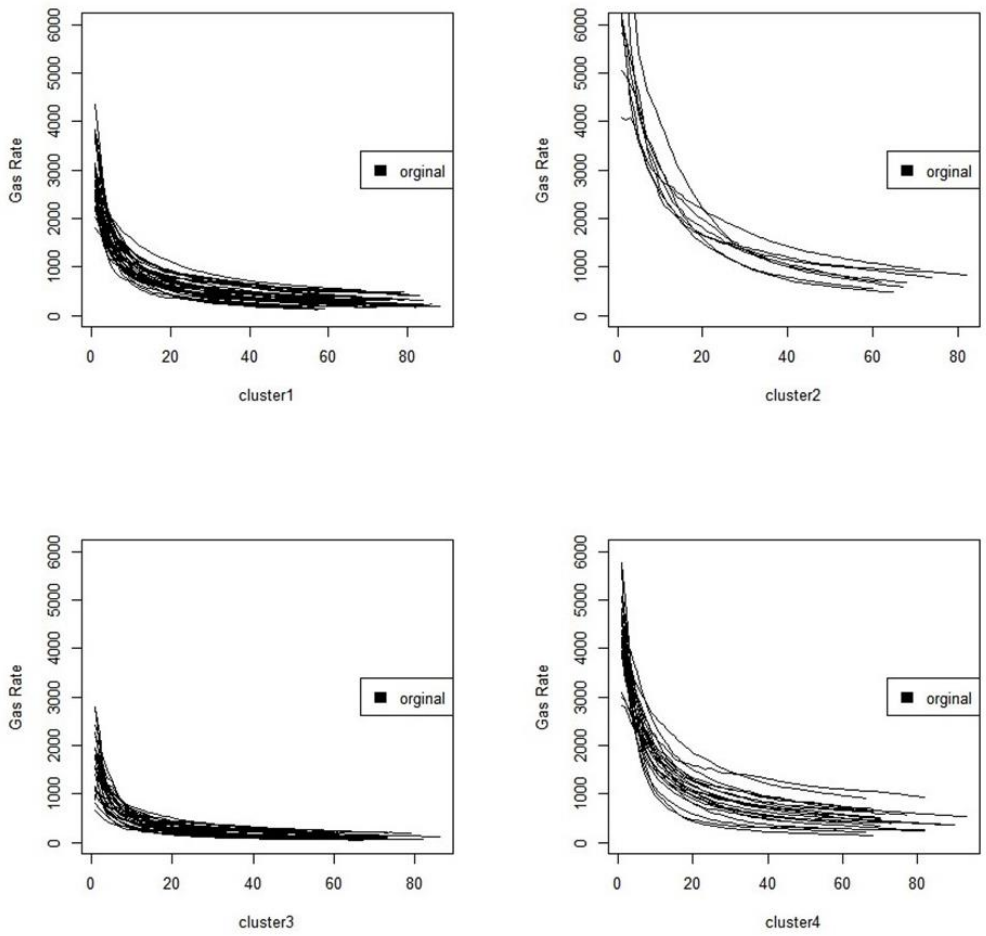


Figure 4-27: Each cluster wells decline curve

To verify this hypothesis, we conducted the prediction with clustered data. For each cluster, according to their size, we pick the first wells to have production history longer than 79 months as a training set, which included the longest 12, 3, 3, and 6 wells for clusters 1, 2, 3, and 4. The length of 79 months was picked to keep continuity with the former training set 2 predictions. For cluster 3, because of its relatively small size, we reduced the required training length from 79 months to 74 months so it could have 3 wells for training instead of only one well.

The same prediction method was applied to generate prediction results. Principal components were learned from the training set in each cluster. The linear regression coefficient of principal components was calculated from each different well. The only difference was that each cluster of the dataset had its own training set. In this way, new wells could be predicted with patterns learned from similar history datasets.

In the following comparison, we also did a sensitivity analysis with the numbers of principal components. In some cases, a scree plot might have had an ambiguous turn point, which makes it hard to use the elbow criterion. So we compared the prediction results after applying three principal components and five principal components. We found that increasing the number of principal components gave the prediction result an enhanced fit, especially to some wells that had fluctuations in early production history.

The prediction results are shown in Figures 4-28 and 4-29.

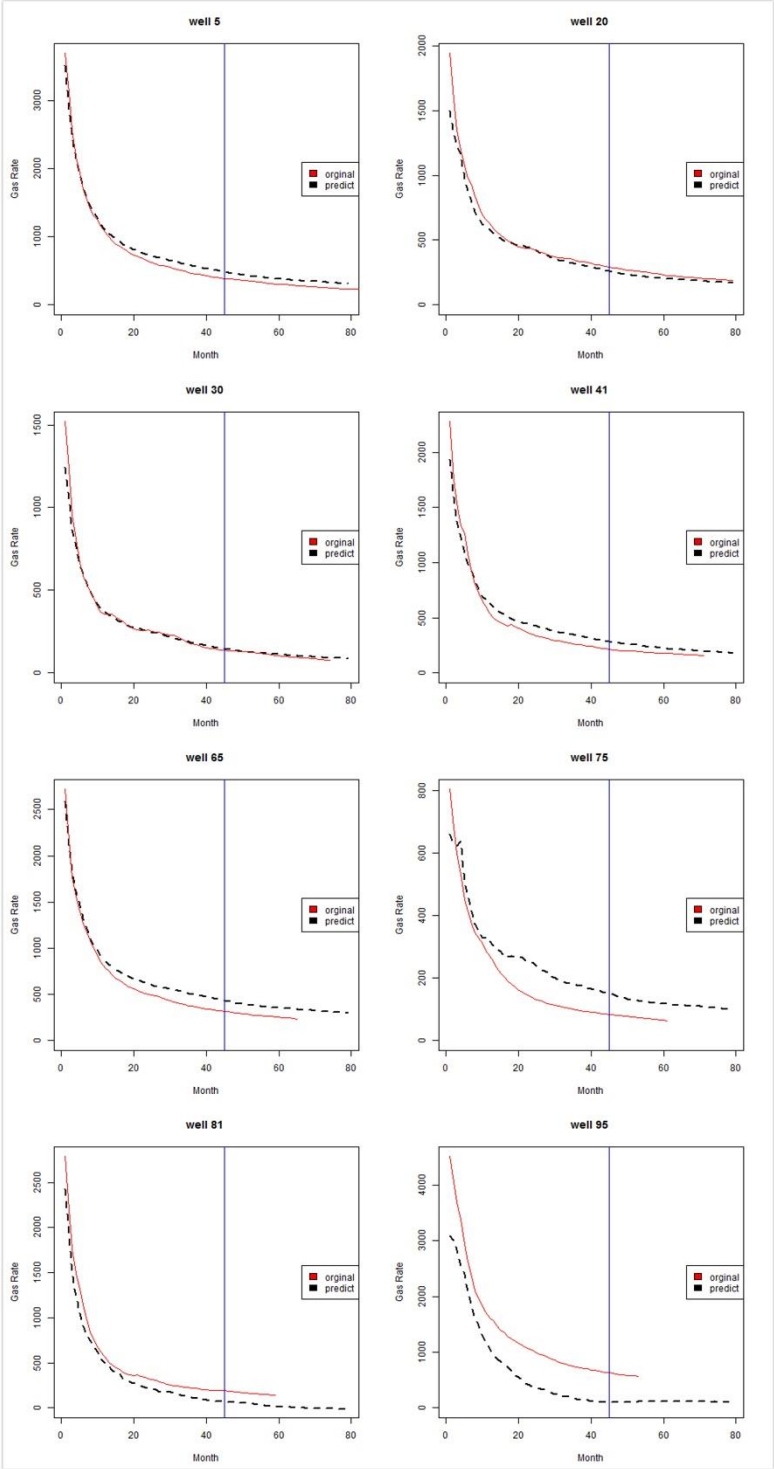


Figure 4-28: Learn 45 months history to predict 79 months (3PC)

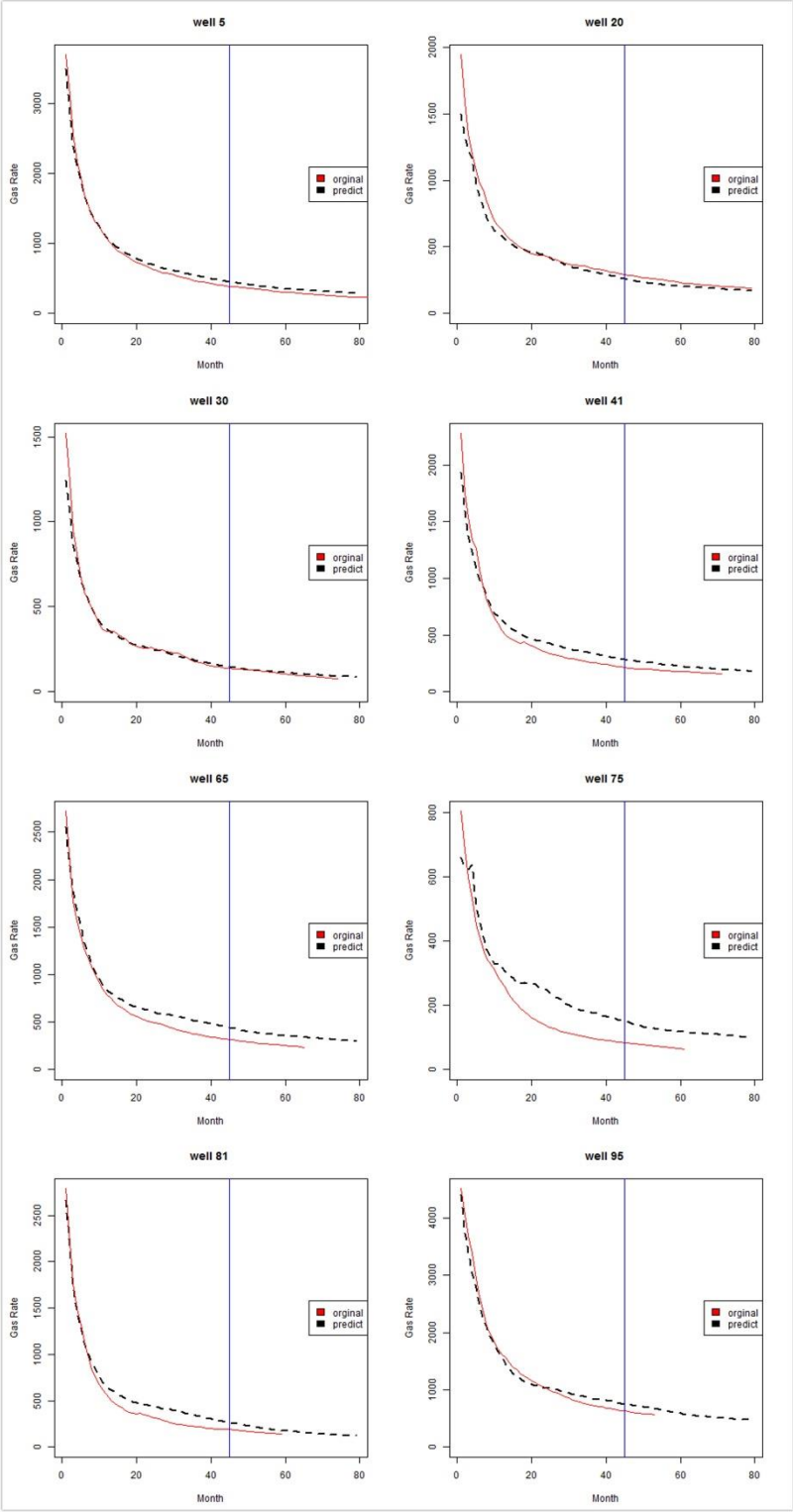


Figure 4-29: Learn 45 months history to predict 79 months (5PC)

4.4 Discussion of *k*-Means Physical Meaning

Both *k*-means clustering and PCA are well-recognized techniques in machine learning and statistics. In Chapter 2, we discussed their mathematicas meanings and workflow. In this section, I discuss their physical meaning and relationship with some petroleum engineering concepts.

Before the discussion, I would like to review some basic concepts of *k*-means and PCA. The dataset for analyzing is rate-time data. The rate at different time steps (day, month) is subject to the influence of certain physical parameters such as pressure, permeability, formation, half-length and so on. In reality, those physical parameters are difficult to measure accurately. Therefore, it brings importance to principal components analysis, which catches hidden patterns under production rate-time data when exact physical parameters are unknown.

Principal components are eigenvalues of the data matrix. They are not directly affected by certain physical parameters. Their value is defined by the overall variance of the data matrix, which is a linear combination of data matrix features. If proper numbers of PCs are picked, the original data matrix can be reconstructed and expressed by only a few PCs without losing much information. Based on this characteristic, in our analysis, we could usually reconstruct simulation or a field data set with only three to five PCs. For example, if we reconstruct a dataset with three PCs, each well (row) in the data matrix can be expressed with a coordinate system (PC1, PC2, and PC3).

k-means is a clustering technique that can be applied with after the data set has been reconstructed with PCA and the original data matrix dimensions have been reduced to only three to five dimensions.

Each well's performance distinction could be calculated by the Euclidean distance. This distance only matters with different wells score in each PC. Therefore, the distance does not have certain correlations with physical parameters. According to the clustering result, we can see each cluster's log-log plot as shown in Figures 4-30 to 4-37.

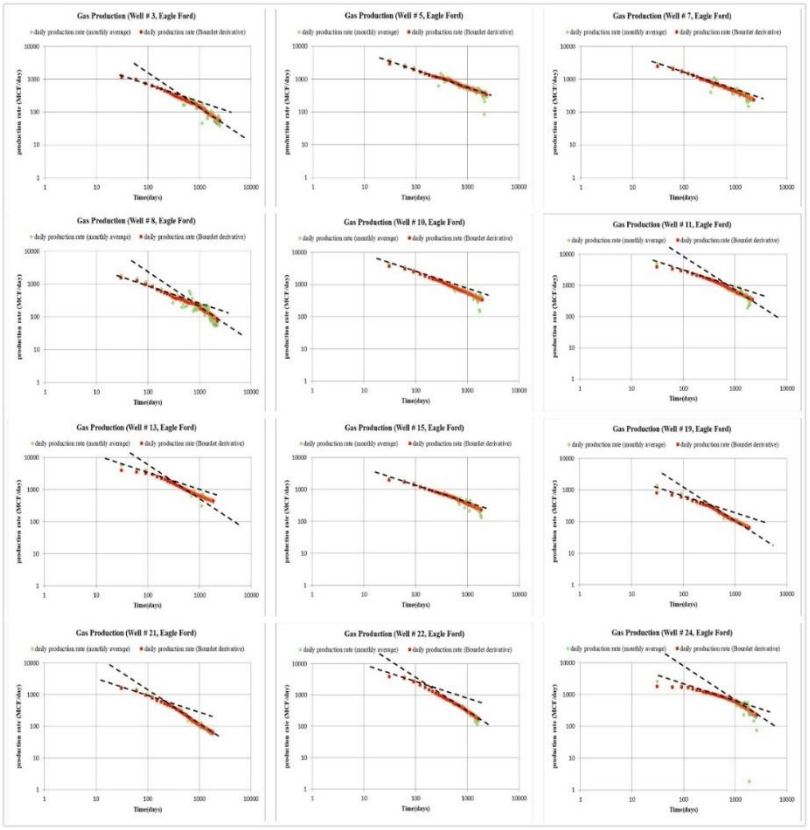


Figure 4-30: Cluster 1 log-log plot (part 1)

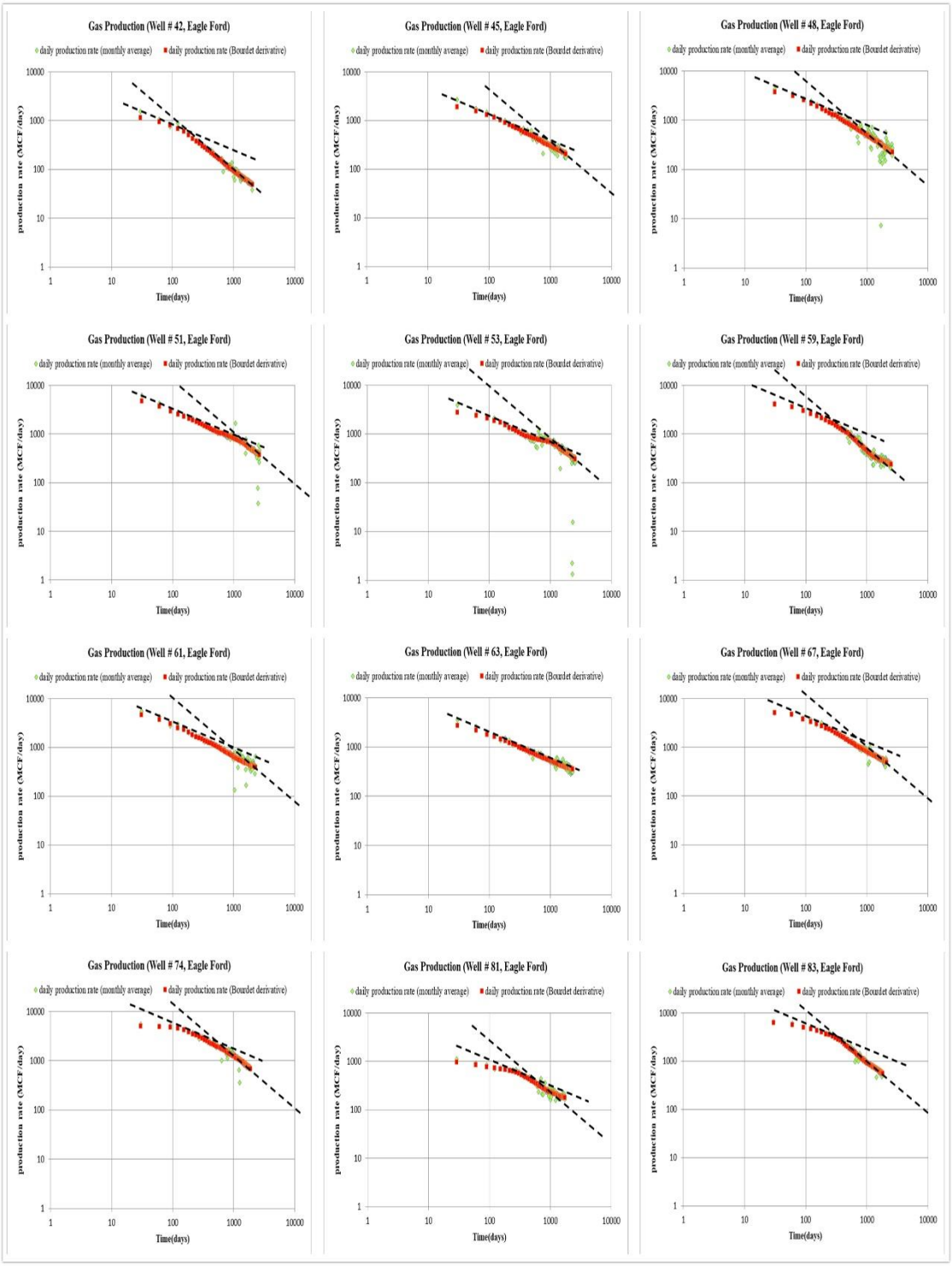


Figure 4-31: Cluster 1 log-log plot (part 2)

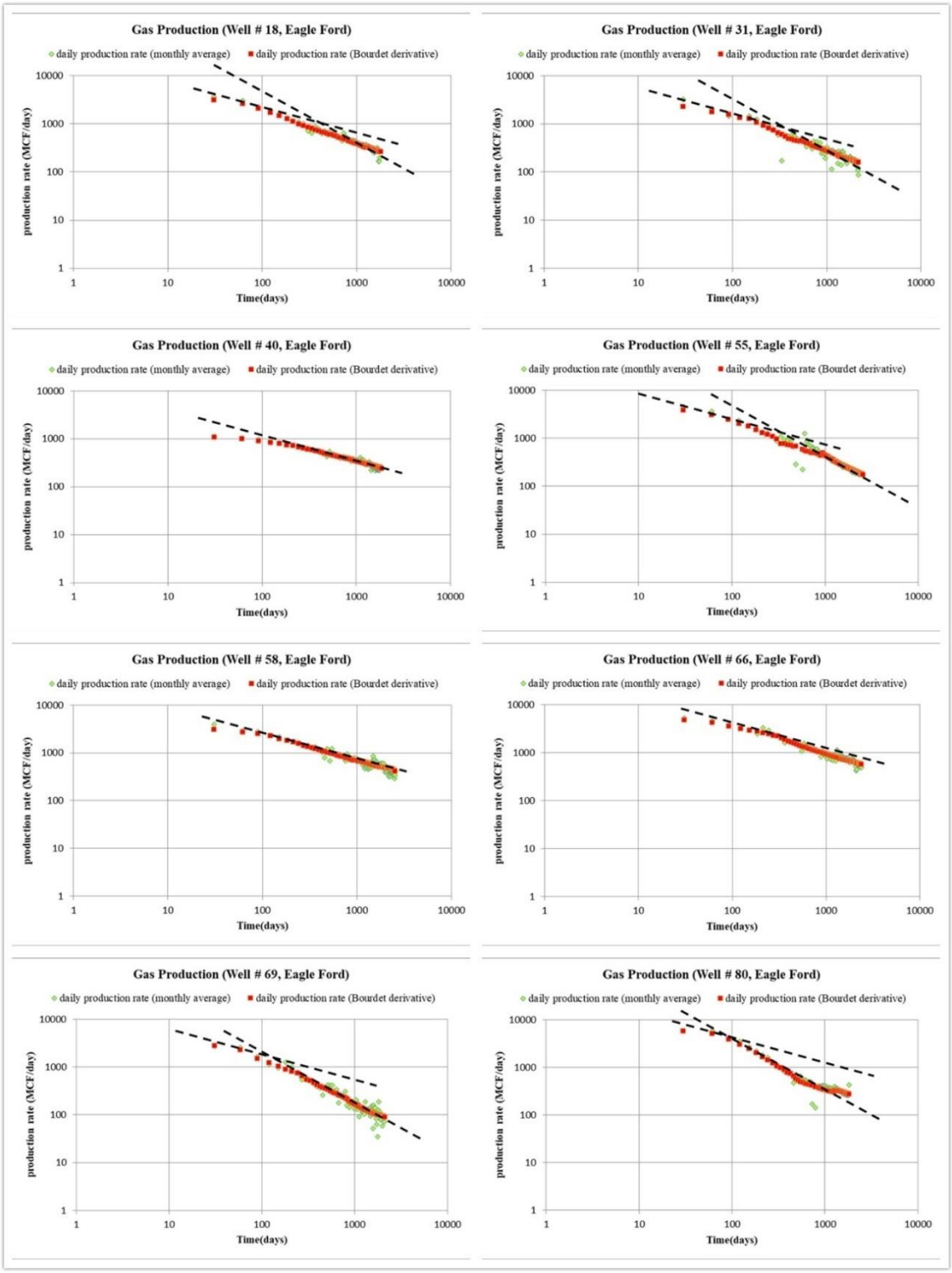


Figure 4-32: Cluster 1 log-log plot (part 3)

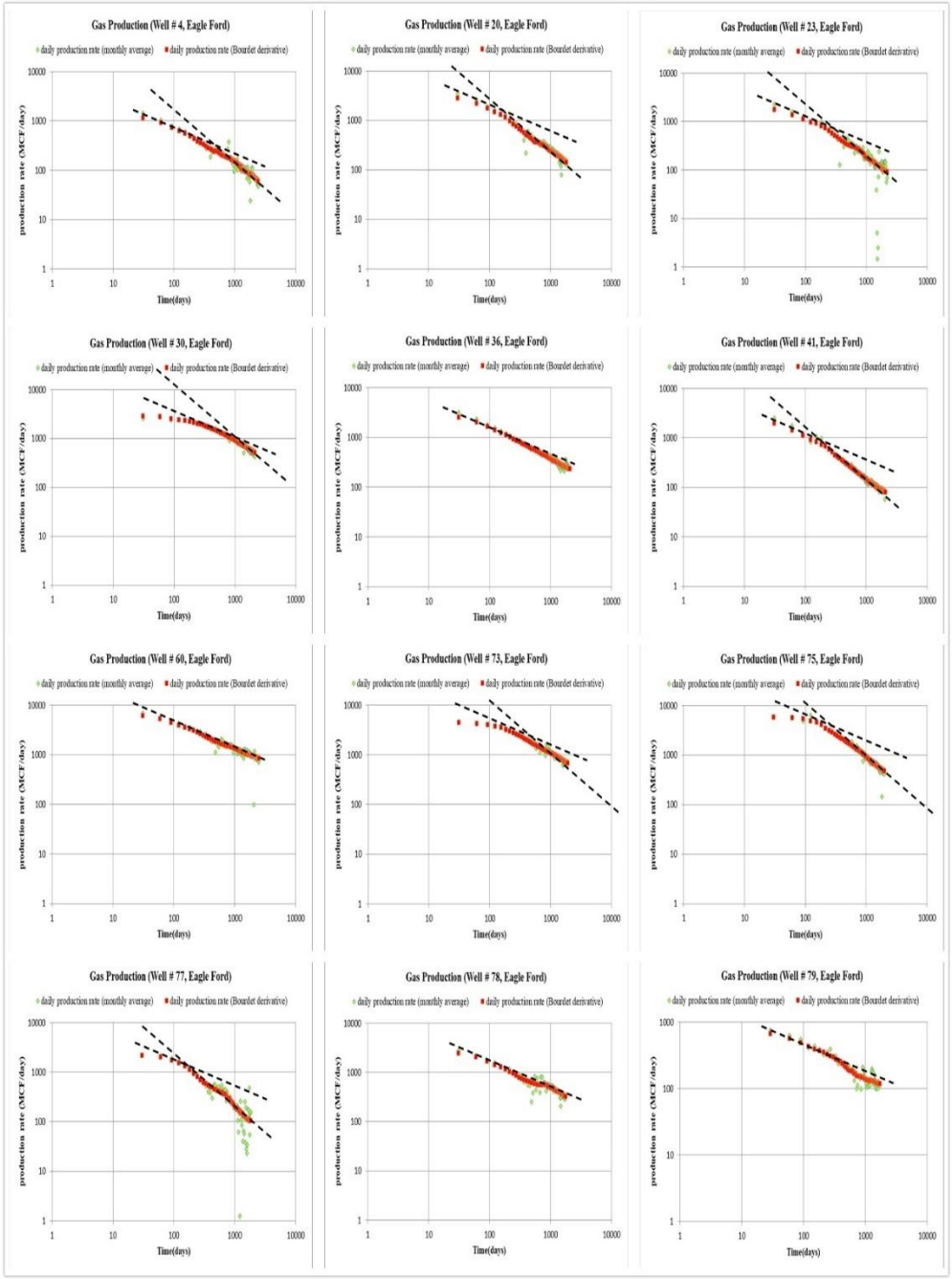


Figure 4-33: Cluster 2 log-log plot

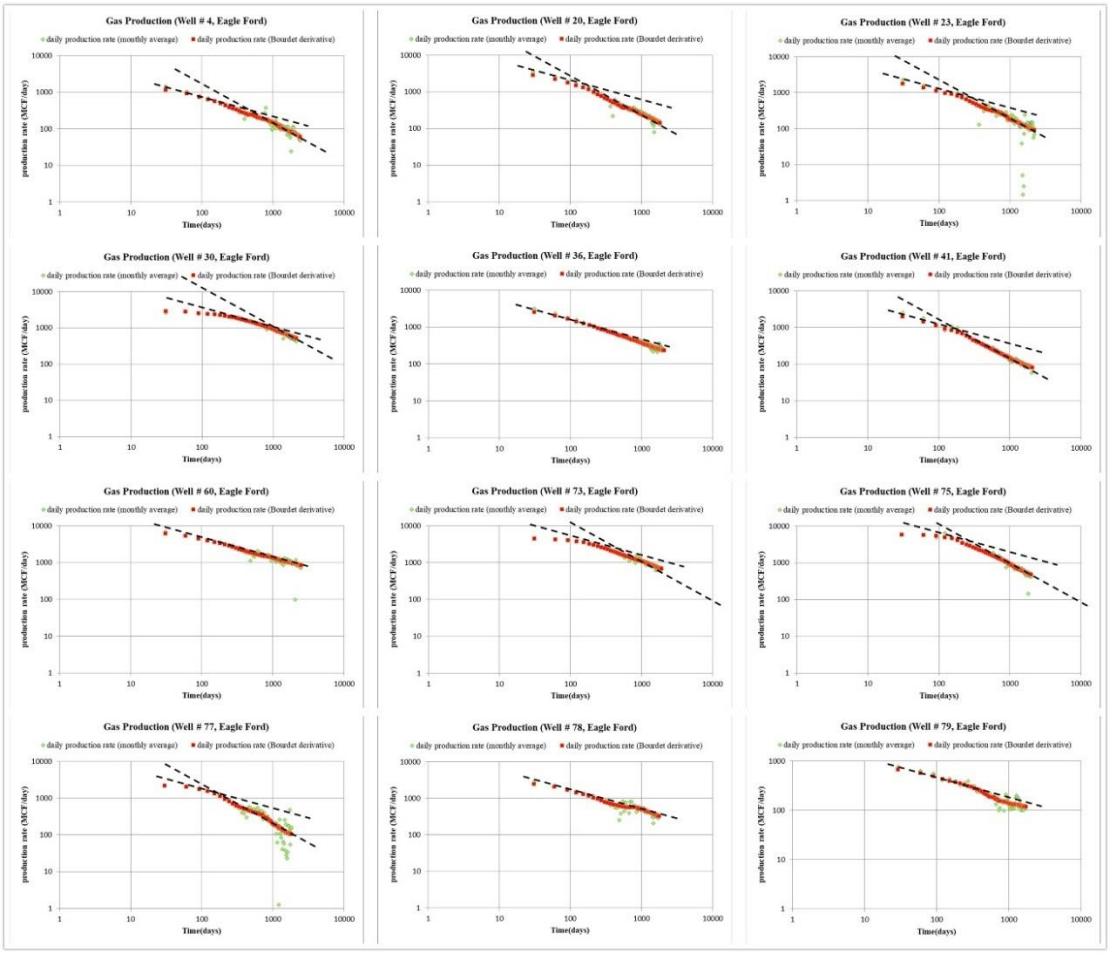


Figure 4-34: Cluster 3 log-log plot (part 1)

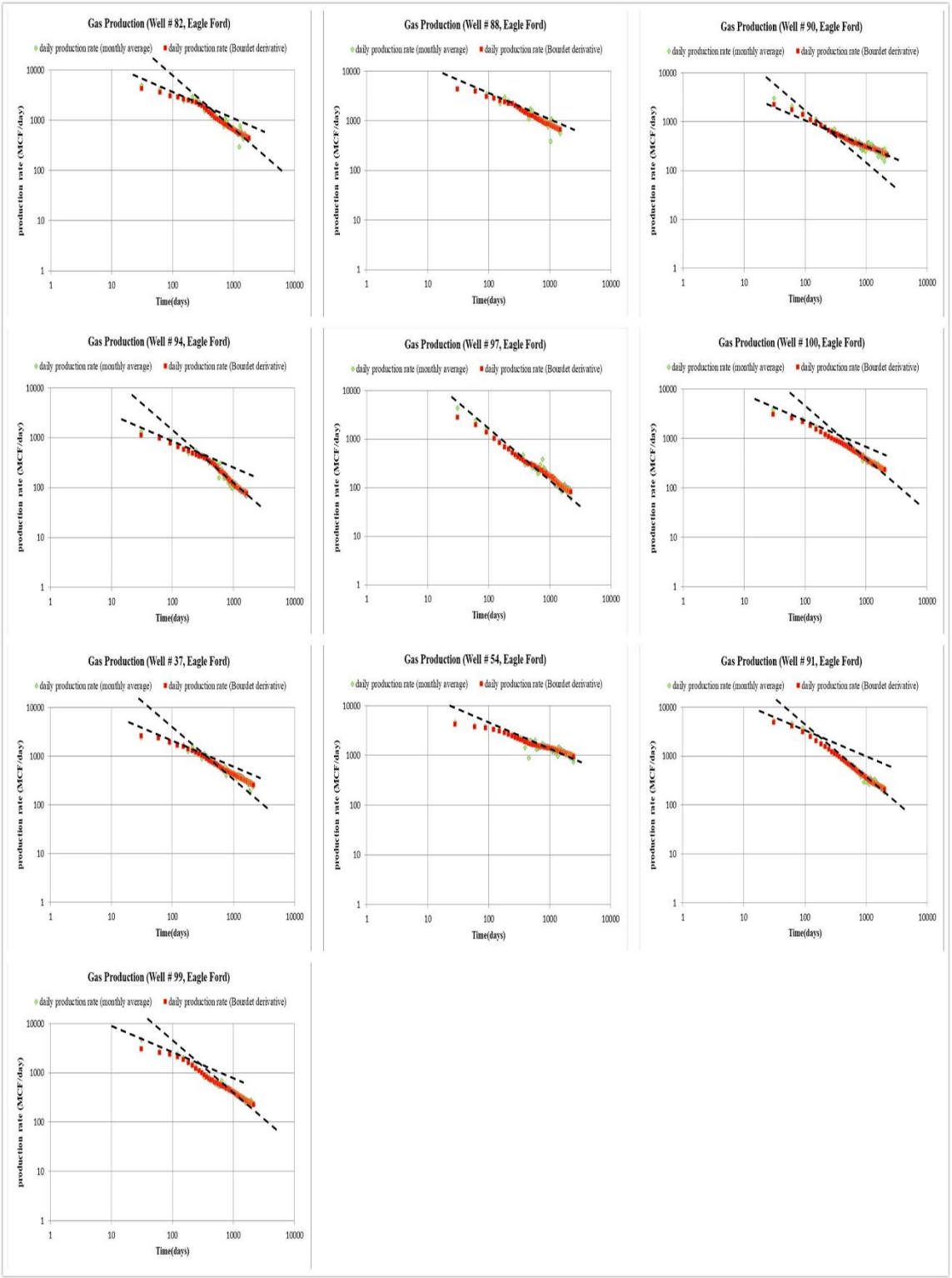


Figure 4-35: Cluster 3 log-log plot (part 2)

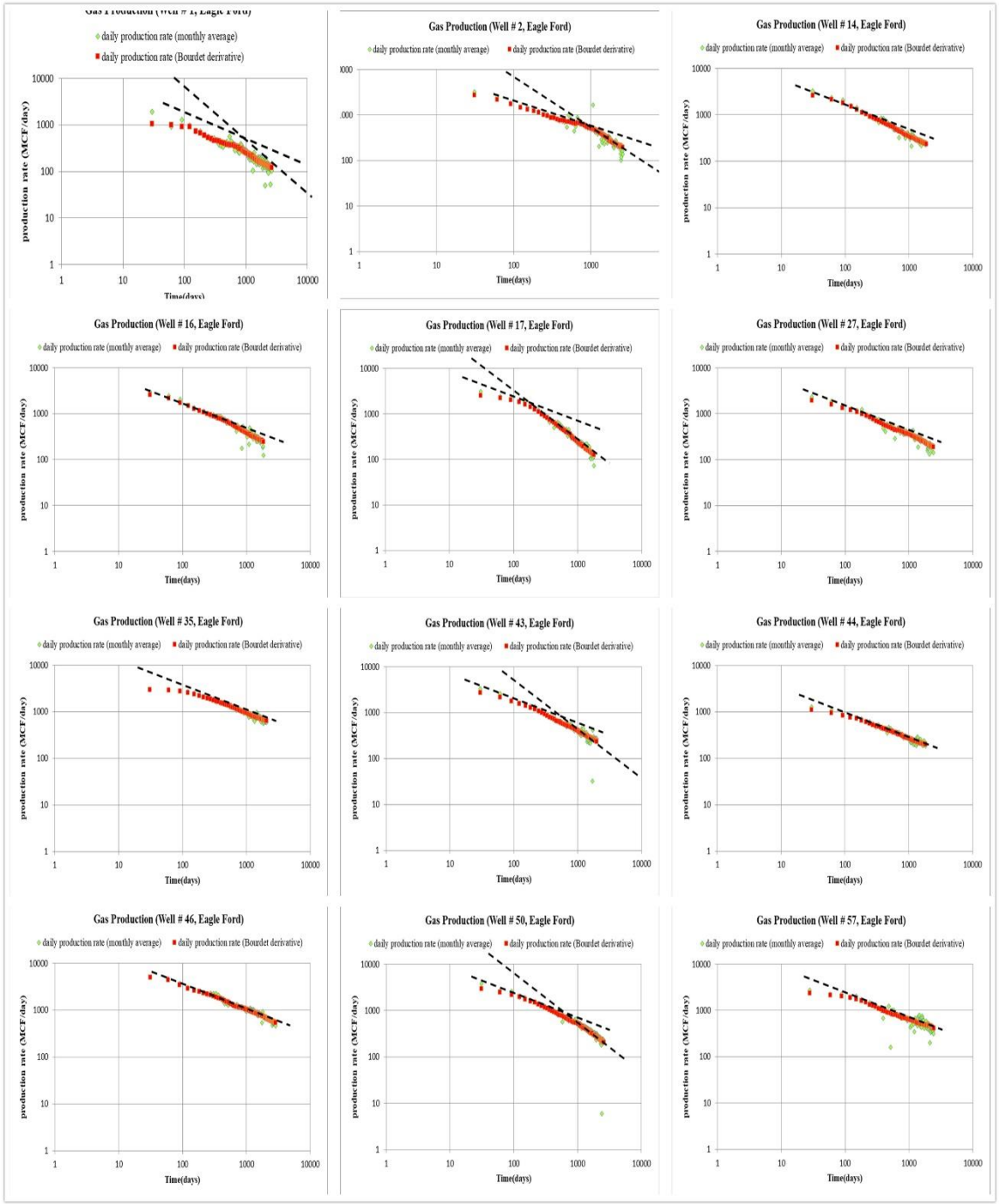


Figure 4-36: Cluster 4 log-log plot (part 1)

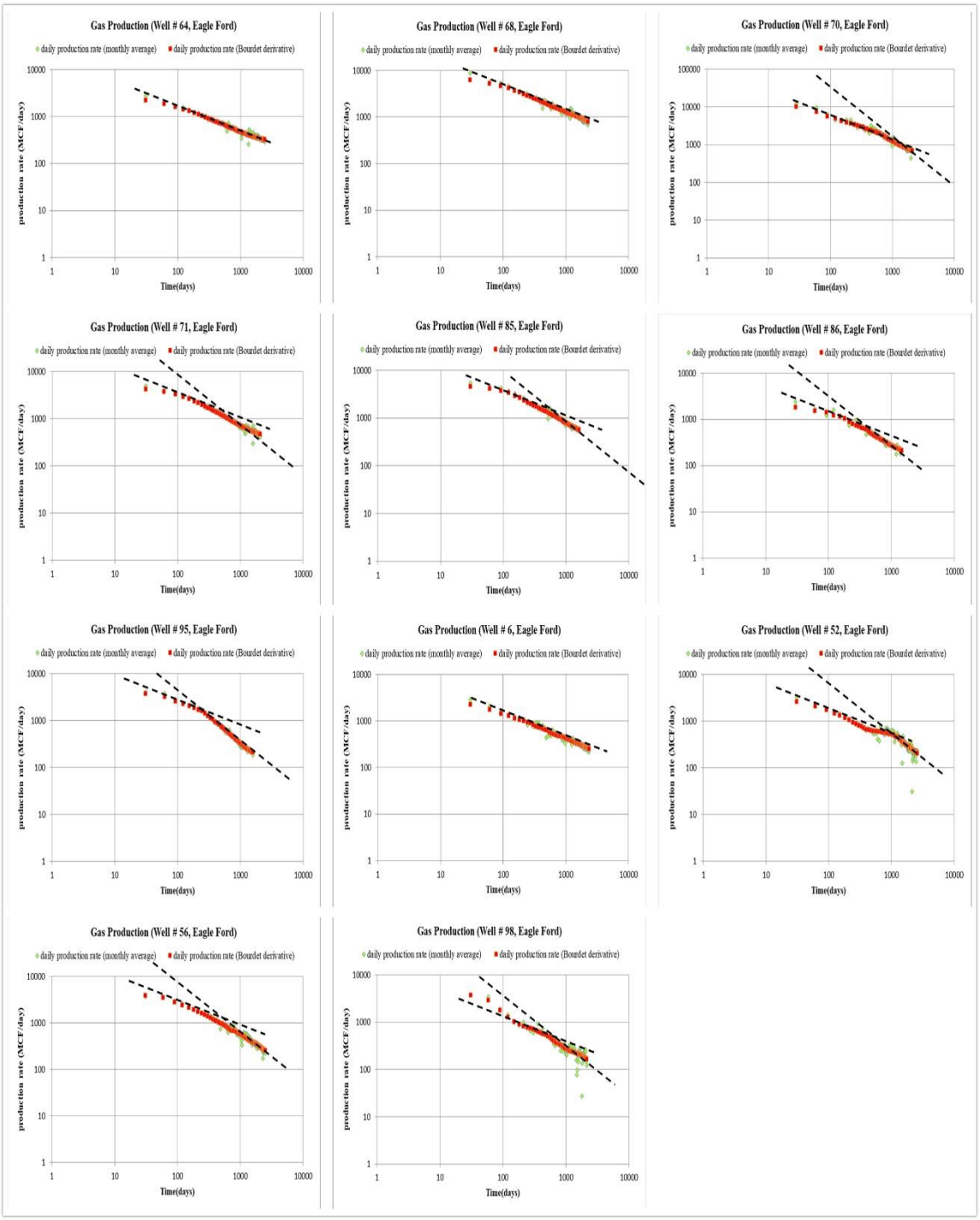


Figure 4-37: Cluster 4 log-log plot (part 2)

Another application of *k*-means is that it can forecast production based on Euclidean distance. For each well, it has a score on each principal component. Therefore, their similarity can be judged with Euclidean distance. One characteristic of *k*-means clustering is that it has stable clustering results with changing time.

Using simulation data, we plotted multiple 2D *k*-means plots with a PC1 and PC2 figure. We used 100 days, 250 days, 500 days, 1000 days, and 2000 days to verify this characteristic. We found that the distribution of each cluster changes a little bit with time, but it is basically kept unchanged. This is shown in Figures 4-38 to 4-42.

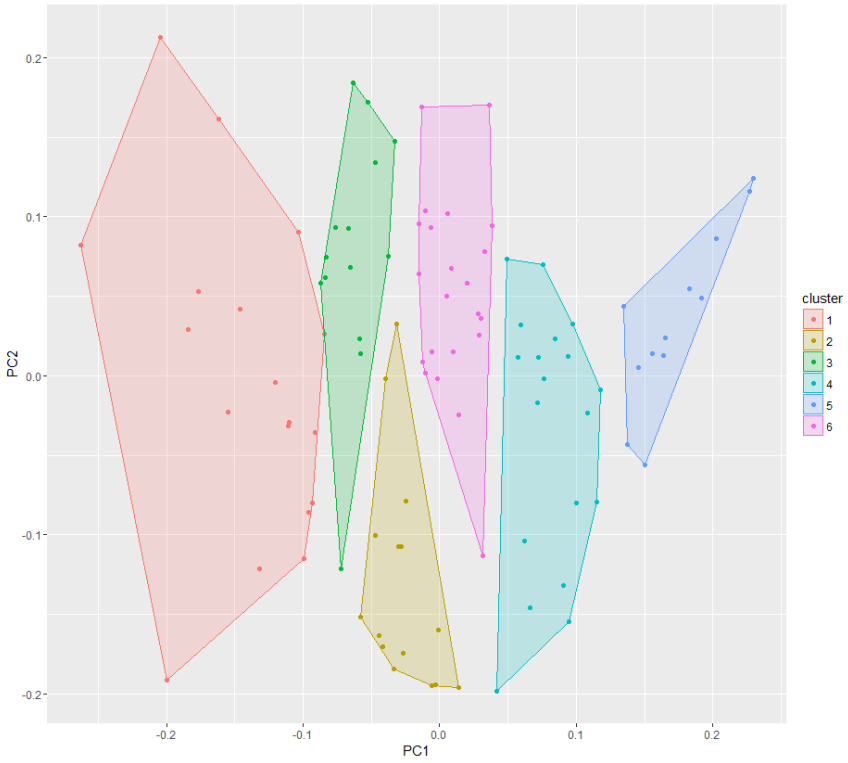


Figure 4-38: Simulation data K-means (100 Days)

Figure 4-39: Simulation data K-means (250 Days)

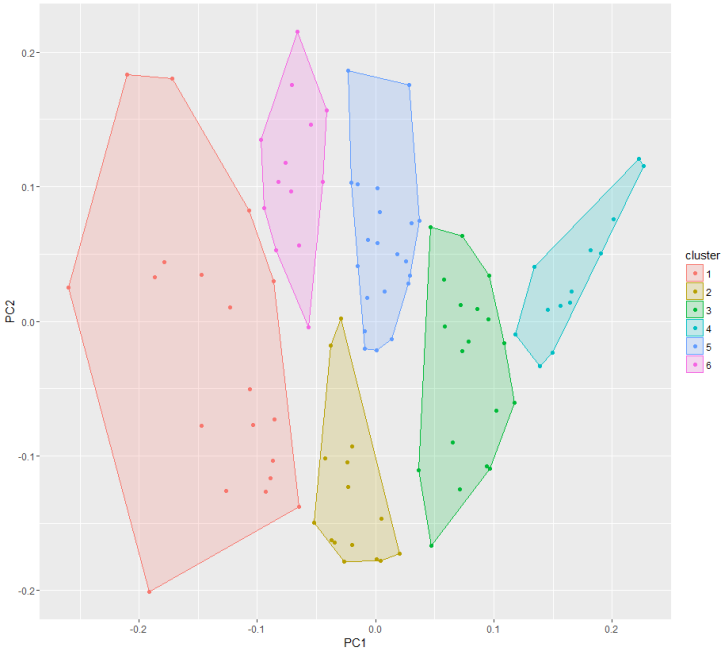


Figure 4-40: Simulation data K-means (500 Days)

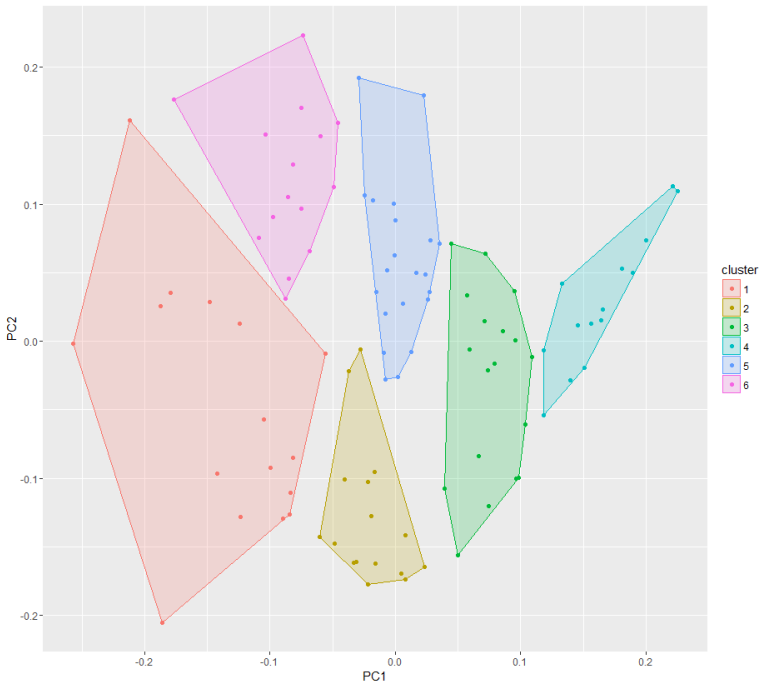


Figure 4-41: Simulation data K-means (1000 Days)

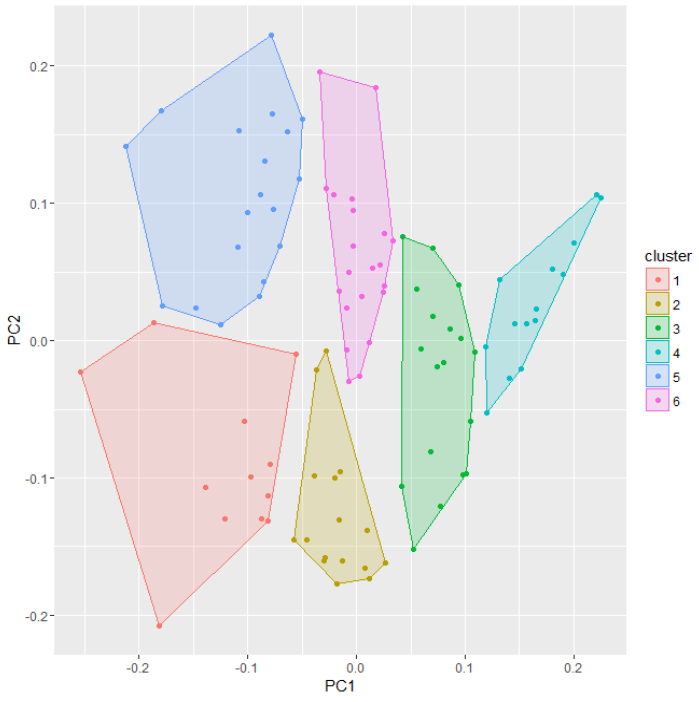


Figure 4-42: Simulation data K-means (2000 Days)

Therefore, in a field data set, we can apply *k*-means to give a range of future production from new wells (Figure 4-43). The upper limit and lower limit can be derived from wells similar to the new wells. This prediction can be made with as little as 6 months of data. It could give a reasonable certain estimation with new developing fields.

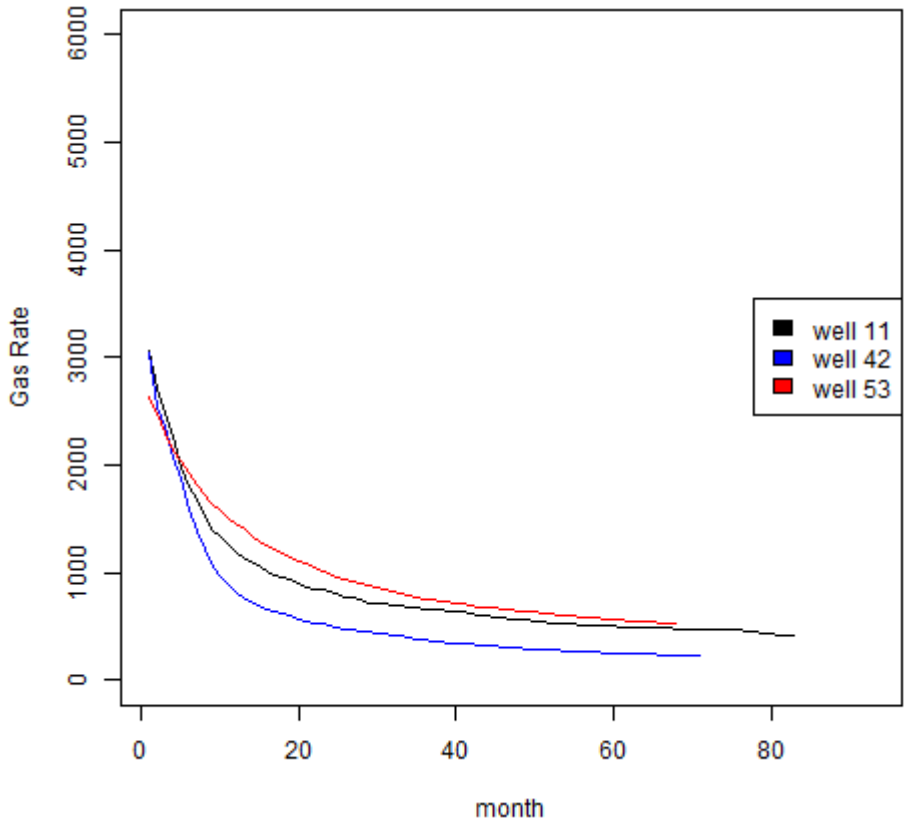


Figure 4-43: Estimation range of well 11 (6 months of history)

A more detailed comparison could be viewed in Figures 4-44 to 4-46.

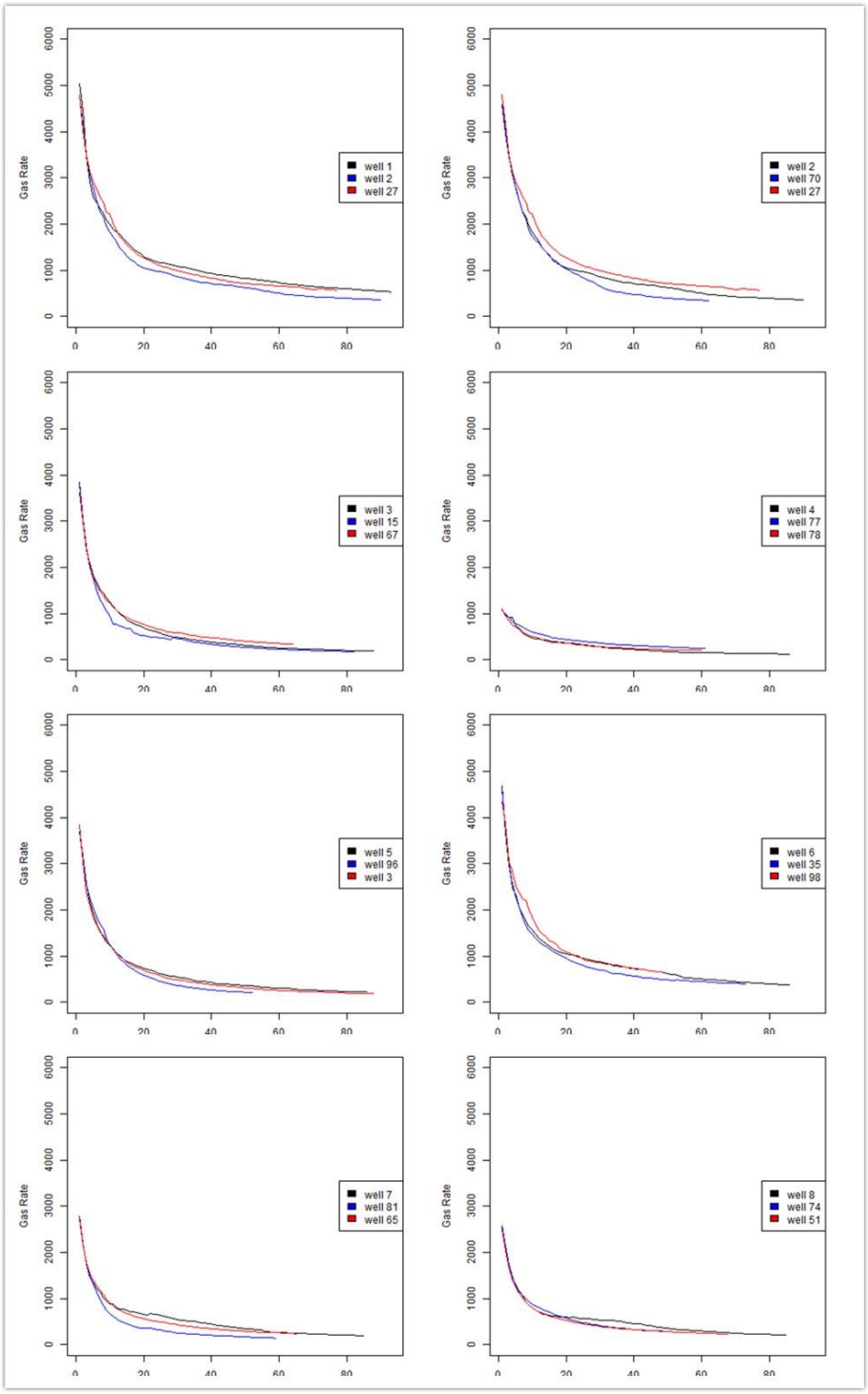


Figure 4-44: *k*-means clustering prediction range (6 months)

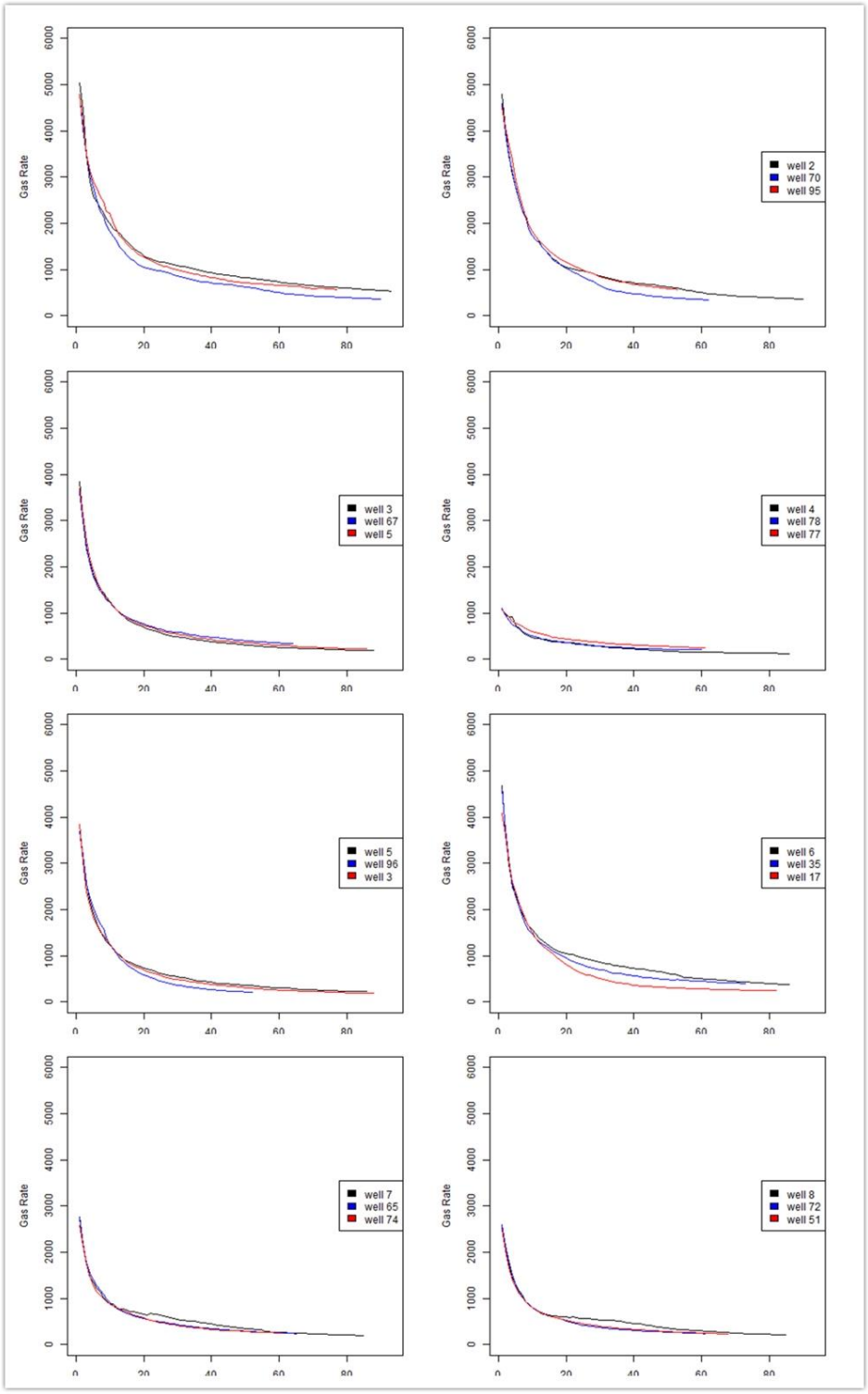


Figure 4-45: *k*-means clustering prediction range (12 months)

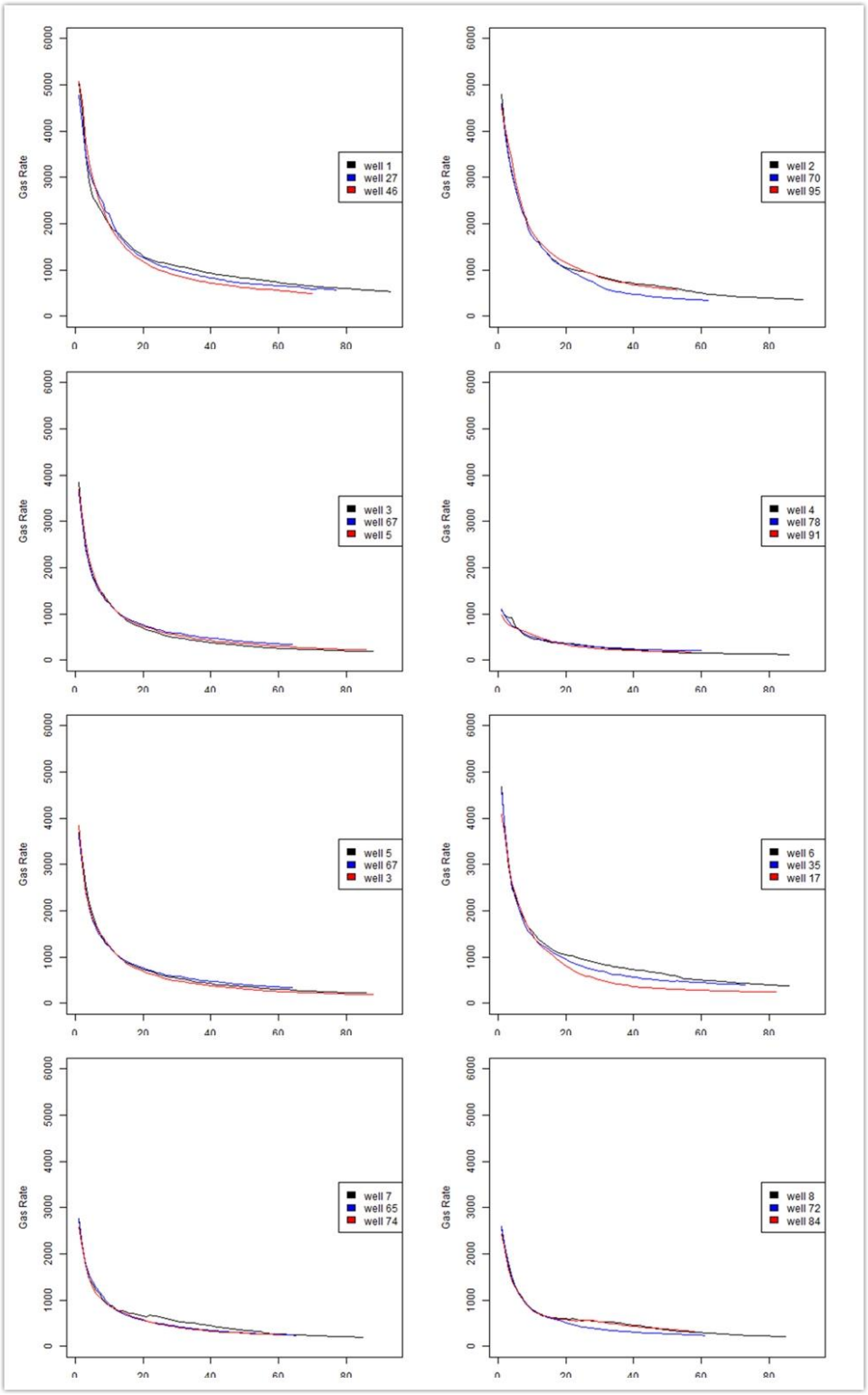


Figure 4-46: *k*-means clustering prediction range (18 months)

4.5 Discussion and Conclusion

In this chapter, we verified the applicability of PCA and linear regression in field data. 100 Eagle Ford gas wells served as a testing dataset. The overall prediction result is satisfying. The *k*-means clustering enhanced the prediction fit ratio.

We also discussed the physical meaning of principal components and *k*-means clustering techniques. A new method based on *k*-means to predict future performance of new wells has been established. It can learn from as little as 6 months of data to produce a reasonably certain forecast.

CHAPTER V

SUMMARY AND CONCLUSIONS

In this thesis, we first reviewed current practices on analyzing unconventional gas production data. Because many new, fast-growing unconventional fields are developing, they usually have a short history and limited production data. That causes difficulties for traditional practices in forecasting new wells.

Principal components analysis (PCA) and its predictive model were applied to solve this problem. In Chapter 2, we reviewed the mathematical proof of PCA, which can transform a data matrix to a linear combination of a few principal components. The prediction model can learn from old historic wells and apply their patterns to similar new wells. With only short history production data, it generated reasonably certain estimates.

In Chapter 3, we applied the predictive model on simulated gas well data. The simulation dataset was generated by Kappa Ecrin. It has 100 wells, each with 2000 days of production history. By learning performance history from the training set, a predictive model has an R-square 0.97 average prediction results.

In Chapter 4, we applied the prediction model as well as *k*-means on Eagle Ford field data. The dataset was picked from 6 adjacent counties in the Eagle Ford field. It has 100 gas wells with 45 to 83 months of production history. With as short as 6-month data, PCA and its predictive model generated satisfying prediction results. For a sample with a large set of wells, *k*-means could be applied to increase prediction performance. *k*-means could also make a future production range estimation based on a historical database.

In conclusion, PCA and its predictive model are promising for predicting production history in unconventional gas. Their prediction results match future performance of testing wells. *k*-means clustering could also give an estimation of a well's future performance with only limited data. PCA could also be illustrative for next step of research and investigation of a combination of reserves estimation with machine learning. Because of its ability to reduce the number of required dimensions, it could be set as a pre-processing step for further analysis through approaches such as neural networks, deep learning, and support vector machines.

GLOSSARY

AI	Artificial intelligence
ANN	Artificial neural network
DNN	Deep neural network
EOR	Enhanced oil recovery
EUR	Estimated ultimate recovery
MCMC	Markov-chain Monte Carlo
ML	Machine learning
NN	Neural network
PC	Principal components
PCA	Principal components analysis
PCR	Principal component regression
PRaD	Piecewise reconstruction from a dictionary
PTA	Pressure-transient analysis
RTA	Rate-time analysis
SAGD	Steam-assisted gravity drainage
SEC	Securities and Exchange Commission
SoM	Self-organizing maps
SVD	Singular value decomposition

NOMENCLATURE

A	Drainage area, acre/Random matrix
a	Duongs' model constant
b	Derivative of loss-ratio (Arps' decline exponent), dimensionless
β	Linear regression coefficients
D	Loss-ratio (Arps' decline constant), Days^{-1}
D_1	Loss-ratio at (t=1), Days^{-1}
D_∞	Loss-ratio at (t= ∞), Days^{-1}
D_i	Initial loss-ratio, Days^{-1}
$\Sigma_{n\times m}$	$n\times m$ rectangular diagonal matrix
S_i	Standard deviation
q	Flow rate, STB/Day or Mscf/Day
q_1	Flow rate at (t=1), STB/Day or Mscf/Day
q_∞	Flow rate at (t= ∞), STB/Day or Mscf/Day
q_c	Critical flow rate
q_i	Flow rate at (t=0), STB/Day or Mscf/Day
$r(t_i)$	production rate on arbitrary *i* day.
t	Time, days/months/years
$\text{t}(a,m)$	Duong's time function
t_1	First timestep, days/months/years
t_i	Aribitary timestep, days/months/years

$\sigma_1(A)$	Largest eigenvalue of arbitray matrix A
n	Time exponent (hyperbolic exponent)
Q	Cumulative production, Mscf or STB
Λ	constant-number matrix
λ	Eigenvalue of matrix
τ	the ratio of time
$\vec{u}$	Average vector
u_1	Largest eigenvalue directions
$U_{n\times n}$	$n\times n$ unitary matrix
$V^T_{m\times m}$	$m\times m$ unitary matrix
X_i	Arbitary random samples
Z	Data matrix

REFERENCES

Aanonsen, S.I. and Geir N.D., Oliver, D.S. et al. 2009. The Ensemble Kalman Filter in Reservoir Engineering—A Review. *SPE J.***14** (03): 393-412. SPE-117274-PA. https://doi.org/10.2118/117274-PA

Agarwal, R.G., Gardener, D.C., Kleinsteiber, S.W., et al. 1998. Analyzing Well Production Data Using Combined Type Curve and Decline Curve Analysis Concepts. *J Pet Technol* **50** (10): 76 – 77 SPE-1098-0076-JPT

Al-Fattah, S.M. and Startzman R.A. 2001. Predicting Natural Gas Production Using Artificial Neural Network. Presented at the SPE Hydrocarbon Economics and Evaluation Symposium, Dallas, Texas, 2-3 April, SPE-68593-MS. https://doi.org/10.2118/68593-MS

Alvarado, V. and. Ranson, A., Hernandez, K., et al. 2002. Selection of EOR/IOR Opportunities Based on Machine Learning. Presented at the European Petroleum Conference, Aberdeen, United Kingdom, 29-31 October, SPE-78332-MS. https://doi.org/10.2118/78332-MS

Ani, M., Oluyemi, G., Petrovski, A., et al. 2016. Reservoir Uncertainty Analysis: The Trends from Probability to Algorithms and Machine Learning. *Proc*. SPE Intelligent Energy International Conference and Exhibition, Aberdeen, Scotland, UK, 6-8 September, SPE-181049-MS, https://doi.org/10.2118/181049-MS

Araya, A. and. Ozkan, E. 2016. An Account of Decline-Type-Curve Analysis of Vertical, Fractured, and Horizontal Well Production Data. Presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 29 September-2 October. SPE-77690-MS. https://doi.org/10.2118/77690-MS

Arnold, R. and Darnell, J.L. 1920. *Manual for the Oil and Gas Industry Under the Revenue Act of 1918*. New York.

Arps, J.J. 1945. Analysis of Decline Curves. *Transactions of the AIME* **160** (01): 228-247.

Arthur, J.D., Langhus, B., and Alleman, D. 2008. An Overview of Modern Shale Gas Development in the United States. *All Consulting*. http://www.all-llc.com/publicdownloads/ALLShaleOverviewFINAL.pdf

Bansal, Y., Ertekin, T., Karpyn, Z., et al. 2013. Forecasting Well Performance in a Discontinuous Tight Oil Reservoir Using Artificial Neural Networks. Presented at the SPE Unconventional Resources Conference-USA, The Woodlands, Texas, USA, 10-12 April, SPE-164542-MS, https://doi.org/10.2118/164542-MS

Bhattacharya, S. and Nikalaou, M. 2013. Analysis of Production History for Unconventional Gas Reservoirs with Statistical Methods. *SPE J.***18** (05): 878 - 896. SPE-147658-PA. https://doi.org.ezproxy.library.tamu.edu/10.2118/147658-PA

Bradley, M.E. 1994. Forecasting Oilfield Economic Performance. *J Pet Technol* **46** (11): 965-971.

Bravo, C.E., Saputelli, L., Rivas, F., et al. 2014. State of the Art of Artificial Intelligence and Predictive Analytics in the E&P Industry: A Technology Survey. *SPE J.***19** (04): 547-563. SPE-150314-PA. https://doi.org/10.2118/150314-PA

Brown, J.P. A Machine Learning Approach to Studies of Recovery Efficiency. Society of Petroleum Engineers. Presented at the Petroleum Computer Conference, Dallas, Texas, 17-20 June. SPE-22304-MS. https://doi.org/10.2118/22304-MS

Brownlee, J. 2017. A Tour of Machine Learning Algorithms. *Machine Learning Mastery*. http://machinelearningmastery.com/a-tour-of-machine-learning-lgorithms/ (accessed 19 Sep 2017).

Cao, Q., Banerjee, R., Gupta, S., et al. 2016.Data Driven Production Forecasting Using Machine Learning. Presented at the SPE Argentina Exploration and Production of Unconventional Resources Symposium, Buenos Aires, Argentina, 1-3 June. 2016 SPE-180984-MS. https://doi.org/10.2118/180984-MS

Chaudhary, N.L. and Lee, W. J. 2016a. An Enhanced Method to Correct Rate Data for Variations in Bottom-Hole Rate-Pressure Deconvolution. Presented at the SPE/IAEE Hydrocarbon Economics and Evaluation Symposium, Houston, Texas, USA, 17-18 May. SPE-179959-MS. https://doi.org/10.2118/179959-MS

Chaudhary, N.L. and Lee, W.J. 2016b. Detecting and Removing Outliers in Production Data to Enhance Production Forecasting. Presented at the SPE/IAEE Hydrocarbon Economics and Evaluation Symposium, Houston, Texas, USA, 17-18 May. SPE-179958-MS. https://doi.org.ezproxy.library.tamu.edu/10.2118/179958-MS

Cheng, K., Wei, Y., Wu, W., et al. A Novel Optimization Model for Analyzing Production Data. Presented at the SPE Western Regional Meeting, Anaheim, California, USA, 27-29 May, SPE-132545-MS, https://doi.org/10.2118/132545-MS

Cipolla, C.L., Lolon, E.P., Erdle, J.C., et al. 2010. Reservoir Modeling in Shale-Gas Reservoirs. *SPE Res Eval & Eng* **13** (04): 638-653. SPE-125530-PA. https://doi.org/10.2118/125530-PA

Clarkson, C.R. 2013. Production Data Analysis of Unconventional Gas Wells: Review of Theory and Best Practices. *International Journal of Coal Geology* **109**: 101-146.

Clarkson, C.R., Nobakht, M., Kaviana, D., et al. 2012. Production Analysis of Tight-Gas and Shale-Gas Reservoirs Using the Dynamic-Slippage Concept. *SPE J.*17 (01): 230-242. SPE-144317-PA. https://doi.org/10.2118/144317-PA

Cox, S.A., Lee, J., Sutton, R.P., et al. 2015. A Comprehensive Approach to Rate-Time Production Analysis for Unconventional Resources. Presented at the SPE/CSUR Unconventional Resources Conference, Calgary, Alberta, Canada, 20-22 October, SPE-175993-MS. https://doi.org/10.2118/175993-MS

Crnkovic-Friis, L. and Erlandson, M. 2015. Geology Driven EUR Prediction Using Deep Learning. Presented at the SPE Annual Technical Conference and Exhibition September, Houston, Texas, USA, 28-30 September. SPE-174799-MS. https://doi.org/10.2118/174799-MS

Cutler, W.W. (1924). *Estimation of Underground Oil Reserves by Oil-Well Production Curves* (Vol. 225). Govt. print. off.

Dakshindas, S.S. 1999. Virtual Well Testing. Presented at the SPE Eastern Regional Conference and Exhibition, Charleston, West Virginia, 21-22 October, SPE-57452-MS. https://doi.org/10.2118/57452-MS.

Denney, D. 2011. Modeling, History Matching, Forecasting, and Analysis of Shale-Reservoir Performance with Artificial Intelligence. *J Pet Technol***63** (09): 60-63. SPE-0911-0060-JPT. https://doi.org/10.2118/0911-0060-JPT

Duong, A.N. 2011. An Unconventional Rate Decline Approach for Tight and Fracture-Dominated Gas Wells. Presented at the Canadian Unconventional Resources and International Petroleum Conference, Calgary, Alberta, Canada, 19-21 October, SPE-137748-MS, https://doi.org/10.2118/137748-MS

Fetkovich, M.J., Vienot, M.E., Bradley, M.D. et al. 1987. Decline Curve Analysis Using Type Curves: Case Histories. *SPE Form Eval* **2** (04): 637-656. SPE-13169-PA. https://doi.org/10.2118/13169-PA

Fisher, R.A. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics*, **7** (2), 179-188.

Floris, F.J.T., Bush, M.D., Cuypers, M., et al. 2001. Methods for Quantifying the Uncertainty of Production Forecasts: A Comparative Study. *Petroleum Geoscience* **7** (S): S87-S96.

Fraim, M.L. 1987. Gas Reservoir Decline-Curve Analysis Using Type Curves with Real Gas Pseudopressure and Normalized Time. *SPE Form Eval.* **2** (04): 671-682. SPE-14238-PA. https://doi.org/10.2118/14238-PA

Fulford, D.S., Bowie, B., Berry, M.E., et al. 2016. Machine Learning as a Reliable Technology for Evaluating Time/Rate Performance of Unconventional Wells. *SPE Econ & Mgmt.* SPE-174784-PA. https://doi.org/10.2118/174784-PA

Gong, X., Gonzalez, R., McVay, D.A., et al. 2014. Bayesian Probabilistic Decline-Curve Analysis Reliably Quantifies Uncertainty in Shale-Well-Production Forecasts. *SPE J.* **19** (06): 1,047 - 1,057 SPE-147588-PA https://doi.org/10.2118/147588-PA

Grujic, O.S., Mohaghegh, S.D., and Bromhal, G.S. 2010. Fast Track Reservoir Modeling of Shale Formations in the Appalachian Basin. Application to Lower Huron Shale in Eastern Kentucky. Presented at the SPE Eastern Regional Meeting, Morgantown, West Virginia, USA, and 13-15 October. SPE-139101-MS. https://doi.org/10.2118/139101-MS

Guo, X.F., Feng, L., and Song, XN. 2012. The Outlier Detection Approach for Multivariate Time Series Based on PCA Analysis. *Journal of Jiangxi Normal University (Natural Sciences Edition)* **36** (3): 280-283.

Hawkins, D.M. 1980. *Identification of Outliers* (Vol. 11). London: Chapman and Hall.

He, Z., Yang, L., Yen., J., et al. 2001. Neural-Network Approach To Predict Well Performance Using Available Field Data. Presented at the SPE Western Regional Meeting, Bakersfield, California, 26-30 March, SPE-68801-MS, https://doi.org/10.2118/68801-MS

Hotelling, H. 1933. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, **24**(6), 417.

Ilk, D., Rushing, J.A., Parego, A.D., et al. 2008. Exponential vs. Hyperbolic Decline in Tight Gas Sands: Understanding the Origin and Implications for Reserve Estimates Using Arps' Decline Curves. Presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 21-24 September, SPE-116731-MS. https://doi.org/10.2118/116731-MS

Ilk, D., Anderson, D.M., Stotts, G.W.J., et al. 2010. Production Data Analysis—Challenges, Pitfalls, Diagnostics. *SPE Res Eval & Eng* **13** (03): 538-552. SPE-102048-MS. https://doi.org/10.2118/102048-MS

Honorio, J., Chen, C., Gao, G., et al. 2015. Integration of PCA with a Novel Machine Learning Method for Reparameterization and Assisted History Matching Geologically Complex Reservoirs. Presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, USA, 28-30 September, SPE-175038-MS. https://doi.org/10.2118/175038-MS

Jia, X .and. Zhang, F. 2016. Applying Data-Driven Method to Production Decline Analysis and Forecasting. Presented at the SPE Annual Technical Conference and Exhibition, Dubai, UAE, 26-28 September, SPE-181616-MS, https://doi.org/10.2118/181616-MS

Johnson, R.H, and Bollens, A.L. 1927. The Loss Ratio Method of Extrapolating Oil Well Decline Curves. *Transactions of the AIME* **77** (01): 771-778.

Jolliffe, I. T. 2002. *Principal Component Analysis, Second Edition*. Springer.

Keshavarzi, R. and Jahanbakhshi, R. 2013. Real-Time Prediction of Complex Hydraulic Fracture Behaviour in Unconventional Naturally Fractured Reservoirs. Presented at the SPE Unconventional Gas Conference and Exhibition, Muscat, Oman, 28-30 January. SPE-163950-MS. https://doi.org/10.2118/163950-MS

Ketchen, D.J. 1996. The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic management journal*: 441-458.

Khazaeni, Y. and Mohaghegh, S.D. 2013. Intelligent Production Modeling Using Full Field Pattern Recognition. *SPE Res Eval & Eng.* **14** (06): 735-749. SPE-132643-PA. https://doi.org/10.2118/132643-PA

Kulga, I.B. 2010. *Development of an Artificial Neural Network for hydraulically fractured horizontal wells in tight gas sands*, MS thesis. Pennsylvania State University, State College, PA (Apr 2010).

Lechner, J.P. and Zangl, G. 2006. Treating Uncertainties in Reservoir Performance Prediction with Neural Networks. Society of Petroleum Engineers. *SPE J.* **58**(06):69-71 SPE-0606-0069-JPT. https://doi.org/10.2118/0606-0069-JPT

Lee, W.J., Rollins, J.B., and Spivey, J. P. 2003. *Pressure Transient Testing (Vol. 9)*. Henry L. Doherty Memorial Fund of AIME Society of Petroleum Engineers.

Lewis, J.O. and Beal, C.H. 1918. Some New Methods for Estimating the Future Production of Oil Wells. Transactions of the AIME **59** (01): 492-525.

Fan, L., Thompson, J.W., and Robinson, J.R. 2010. Understanding Gas Production Mechanism and Effectiveness of Well Stimulation in the Haynesville Shale Through Reservoir Simulation. Presented at the Canadian Unconventional Resources and International Petroleum Conference, Calgary, Alberta, Canada, 19-21 October, SPE-136696-MS, https://doi.org/10.2118/136696-MS

Lolon, E. and. Hamidieh K., Weijers, L., et.al. 2016. Evaluating the Relationship Between Well Parameters and Production Using Multivariate Statistical Models : A Middle Bakken and Three Forks Introduction to Statistical Modeling. Presented at the SPE Hydraulic Fracturing Technology Conference, 9-11 February, Woodlands, Texas, USA, 9-11 February, SPE-179171-MS. https://doi.org/10.2118/179171-MS

Long, D.R. and Davis, M.J.1987. A New Approach to the Hyperbolic Curve. Society of Petroleum Engineers. Presented at the SPE Production Operations Symposium, Oklahoma City, Oklahoma, 8-10 March, SPE-16237-MS, https://doi.org/10.2118/16237-MS

Ma, Z., Liu, Y., Leung, J.Y., et al. 2015. Practical Data Mining and Artificial Neural Network Modeling for SAGD. Presented at the SPE Canada Heavy Oil Technical Conference, Calgary, Alberta, Canada, 9-11 June, SPE-174460-MS, https://doi.org/10.2118/174460-MS

MacQueen, J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Proc*., the Fifth Berkeley Symposium on Mathematical Statistics and Probability **1** (14):281-297

Marhaendrajana, T. and Blasingame, T.A. 2013. Decline Curve Analysis Using Type Curves—Evaluation of Well Performance Behavior in a Multiwell Reservoir System. Presented at the SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, 30 September-3 October, SPE-71517-MS, https://doi.org/10.2118/71517-MS

Mohaghegh, S.D., Grujic, O., Zargari, S., et al. 2011. Modeling, History Matching, Forecasting and Analysis of Shale Reservoirs Performance Using Artificial Intelligence Top-Down, Intelligent Reservoir Modeling for Shale Formations. Presented at the SPE Digital Energy Conference and Exhibition, The Woodlands, Texas, USA. 19-21 April, SPE-143875-MS, https://doi.org/10.2118/143875-MS

Mohaghegh, S. 1995. Neural Network: What It Can Do for Petroleum Engineers. *J Pet Technol* **47**(01):42-42. SPE-29219-PA. https://doi.org/10.2118/29219-PA

Moridis, G.J., Kuzma-Anderson, H., Reagan, M.T., et al. 2013. A Self-Teaching Expert System for the Analysis, Design, and Prediction of Gas Production From Unconventional Gas Resources. *Proc*. Presented at the Canadian Unconventional Resources Conference, Calgary, Alberta, Canada, 15-17 November, SPE-149485-MS, https://doi.org/10.2118/149485-MS

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

Pearson, K. 1901. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2** (11): 559-572. http://dx.doi.org/10.1080/14786440109462720.

Queipo, N.V., Goicochea, J.V., and Pintos, S. 2002. Surrogate Modeling-Based Optimization of SAGD Processes. *Journal of Petroleum Science and Engineering* **35** (1): 83-93.

Rebeschini, J., Querales, M., G. A. Carvajal, G.A., et al. 2013. Building Neural-Network-Based Models Using Nodal and Time-Series Analysis for Short-Term Production Forecasting. Proc. Presented at the SPE Middle East Intelligent Energy Conference and Exhibition, Manama, Bahrain, 28-30 October, SPE-167393-MS. https://doi.org/10.2118/167393-MS

Ringnér, M. 2008. What Is Principal Component Analysis? *Nature Biotechnology* **26** (3): 303.

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

Seidle, J. 2016. SPEE Monograph 4—Estimating Developed Reserves in Unconventional Reservoirs: Knowledge Gained. Society of Petroleum Evaluation Engineers.

Singh, A. 2017. Application of Data Mining for Quick Root-Cause Identification and Automated Production Diagnostic of Gas Wells With Plunger Lift. *SPE Prod & Oper*.**32**(03):279-293.SPE-175564-PA. https://doi.org/10.2118/175564-PA

Sinha, S., Devegowda, D., and Deka, B. 2016. Multivariate Statistical Analysis for Resource Estimation in Unconventional Plays: Application to Eagle Ford Shales. Presented at the SPE Eastern Regional Meeting, Canton, Ohio, USA, 13-15 September, SPE-184050-MS, https://doi.org/10.2118/184050-MS

Soleng, H.H. 1999. Oil Reservoir Production Forecasting With Uncertainty Estimation Using Genetic Algorithms. *Evolutionary Computation,Vol. 2,* 1217-1223: IEEE.

Sondergeld, C.H., Newsham, K.E., Comisky, J.T., et al. 2010. Petrophysical Considerations in Evaluating and Producing Shale Gas Resources. Presented at the SPE Unconventional Gas Conference, SPE Unconventional Gas Conference, 23-25 February, 23-25 February, SPE-131768-MS. https://doi.org/10.2118/131768-MS

Surguchev, L. and Li, L. 2000. IOR Evaluation and Applicability Screening Using Artificial Neural Networks. Presented at the SPE/DOE Improved Oil Recovery Symposium, Tulsa, Oklahoma, 3-5 April, SPE-59308-MS, https://doi.org/10.2118/59308-MS

Valkó, P.P. and Lee, W.J. 2010. A better way to forecast production from unconventional gas wells. Society of Petroleum Engineers. Presented at the SPE Annual Technical Conference and Exhibition, Florence, Italy, 19-22 September, SPE-134231-MS, https://doi.org/10.2118/134231-MS

Walsh, M.R., Hancock, S.H., Wilson, S.J., et al. 2009. Preliminary Report on the Commercial Viability of Gas Production from Natural Gas Hydrates. *Energy Economics* **31** (5): 815-823.

Wang, H.Y. 2016. What Factors Control Shale-Gas Production and Production-Decline Trend in Fractured Systems: A Comprehensive Analysis and Investigation. *SPE J.* **22** (05): 562 – 581 SPE-179967-PA https://doi.org/10.2118/179967-pa

Wang, Q., Chen, X., Jha, A.N., et al. 2014. Natural Gas From Shale Formation–The Evolution, Evidences and Challenges of Shale Gas Revolution in United States. *Renewable and Sustainable Energy Reviews* 30: 1-28.