© Copyright Nathan Tran 2018

All Rights Reserved

**Application of Data Analytics to Prediction of Initial Production in Tight Oil Reservoir**

A Thesis

Presented to

The Faculty of the Department of Petroleum Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

In Petroleum Engineering

By

Nathan Tran

December 2018

**Application of Data Analytics to Prediction of Initial Production in Tight Oil Reservoir**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Nathan Tran

Approved: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Chair of the Committee

Dr. Ganesh Thakur, Distinguished Professor

Of Petroleum Engineering and director of UH

Energy Industrial Partnerships

Committee Members: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Guan Qin, Associate Professor

Department of Petroleum Engineering

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Adwait Chawathe, Unit Manager

Chevron Energy Technology Company

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Suresh K. Khator, Associate Dean Dr. Mohamed Soliman,

Cullen College of Engineering Professor and Chair,

Department of Petroleum Engineering

# **Acknowledgement**

First and foremost, I would like to thank my thesis advisor, Dr. Ganesh Thakur, for his guidance, support, and patience throughout this research process. Without his advice and direction, none of the work performed here would have been possible. I would also like to express my gratitude to Dr. Sriram Balasubramanian, Dr. Sushanta Bose, and Dr. Rahul Kumar. They were instrumental in assisting and critiquing various aspects of the project. A special mention to Dr. Guan Qin and Dr. Adwait Chawathe for accepting to be a part of my thesis committee. Their feedback is highly appreciated. I would also like to thank the client company who graciously provided the geo-model and the required field data.

Application of Data Analytics to Prediction of Initial Production in Tight Oil Reservoir

An Abstract

Of a

Thesis

Presented to

The Faculty of the Department of Petroleum Engineering

University of Houston

In partial fulfillment of the

Requirements for the Degree

Master of Science

In Petroleum Engineering

By

Nathan Tran

December 2018

# **Abstract**

The main objective of this study is to develop a quick predictive tool to forecast IP for wells in LT reservoir in the three fields J, K and N. Data includes the petrophysical properties, completion history, lab and field measurements (PVT, pressure transient test, production and pressure data). Missing pressure data is handled by pressure versus time correlation corresponding to the drive mechanisms in the reservoir. Thirty well samples with 39 possibly correlated parameters have been analyzed and transformed by principal component analysis (PCA) to construct a vector space of 29 orthogonal components. Then, the scores of the components are used for linear regression to develop a prediction model for IP.

After analyzing transforming data, two methods of picking components have been tested: Method (1) picking the first 10 components and Method (2) picking the 10 components most correlated to IP. Method (2) yields a better prediction model with an R2 equal to 0.76 (while method (1) has an R2 equal to 0.5). Robustness of the prediction model has been tested by reducing the data size from 30 to 25 samples. The reduced sample set generates a less accurate prediction during the blind test (only one out of seven testing wells falls within the tolerance range).

Some uncertainties for the model have also been addressed. Reservoir structure is not described in detail. In addition, drilling data and details of perforation are not considered in the data base. Operation aspects are not considered in this approach at this time.

# **Table of Contents**

[**Acknowledgement** v](#_Toc531857389)

[**Abstract** vii](#_Toc531857390)

[**Table of Contents** viii](#_Toc531857391)

[**List of Figures** ix](#_Toc531857392)

[**List of Tables** x](#_Toc531857393)

[**Chapter 1: Introduction** 1](#_Toc531857394)

[**Chapter 2: Background** 3](#_Toc531857395)

[**2.1: Overview of Machine Learning** 3](#_Toc531857396)

[**2.2: Previous Works of Machine Learning in Oil and Gas Industry** 4](#_Toc531857397)

[**2.3: Principal Component Analysis and Linear Regression** 6](#_Toc531857398)

[**Chapter 3: Workflow** 8](#_Toc531857399)

[**Chapter 4: Data Collection** 10](#_Toc531857400)

[**Chapter 5: Transforming Data** 17](#_Toc531857401)

[**Chapter 6: Performing Linear Regression Model** 20](#_Toc531857402)

[**6.1: Selecting Components for Regression** 20](#_Toc531857403)

[**6.2: Averaged versus Probabilistic Properties** 25](#_Toc531857404)

[**6.3: Adding New Properties-Averaged Skin** 26](#_Toc531857405)

[**6.4: Robustness Test** 28](#_Toc531857406)

[**6.5: Uncertainties** 30](#_Toc531857407)

[**Chapter 7: Conclusions** 32](#_Toc531857408)

[**Chapter 8: Recommendation for Future Studies** 34](#_Toc531857409)

[**References** 35](#_Toc531857410)

# **List of Figures**

[**Figure 1. Major Factors Influencing Initial Production** 9](#_Toc531857411)

[**Figure 2. Workflow to Develop Prediction Model for IP** 9](#_Toc531857412)

[**Figure 3. Drive Mechanisms in Field J** 12](#_Toc531857413)

[**Figure 4. SIBHP Trend in Block J1 of Field J** 13](#_Toc531857414)

[**Figure 5. SIBHP Trend in Block J2 of Field J** 13](#_Toc531857415)

[**Figure 6. Drive Mechanism in Field K** 14](#_Toc531857416)

[**Figure 7. SIBHP in Block K243, 332 and 341 in Field K** 15](#_Toc531857417)

[**Figure 8. SIBHP in Block K432 in Field K** 15](#_Toc531857418)

[**Figure 9. SIBHP Trend in Field N** 16](#_Toc531857419)

[**Figure 10. Actual IP versus Predicted IP by 10 Components Most Correlated to IP** 21](#_Toc531857420)

[**Figure 11. Actual IP versus Predicted IP by First 10 Components** 23](#_Toc531857421)

[**Figure 12. Correcting Data for Well 328 Moves the Prediction into Tolerance Range** 24](#_Toc531857422)

[**Figure 13. Actual IP versus Predicted IP Generated by Averaged Properties** 25](#_Toc531857423)

[**Figure 14. Actual IP versus Predicted IP Generated by Probabilistic Properties** 26](#_Toc531857424)

[**Figure 15. Actual IP versus Predicted IP Modeled with Skin Data** 28](#_Toc531857425)

[**Figure 16. Actual IP versus Predicted IP Generated by Modeling with 25 Well Samples** 30](file:///\\172.27.36.72\Graduate%20Research\JORAJAN\NATHAN\Presentations\Thesis%20Report%20V3.docx#_Toc531857426)

# **List of Tables**

[**Table 1. There Are Four Types of Machine Learning.** 4](#_Toc531864374)

[**Table 2. IP and Total Perforation Has Been Collected for the Selected Wells.** 10](#_Toc531864375)

[**Table 3. Thirty-Nine Parameters Are Used to Construct Prediction Model.** 17](#_Toc531864376)

[**Table 4. PCA Results in Eigen Values, Variability for 29 Components.** 18](#_Toc531864377)

[**Table 5. 10 Components with the Most Correlation to IP Are Used for Modeling.** 20](#_Toc531864378)

[**Table 6. 10 Components with the Most Variability Are Used for Modeling.** 22](#_Toc531864379)

[**Table 7. Actual Skin and Averaged Skin (in Orange Boxes) Are Used as New Parameters.** 27](#_Toc531864380)

[**Table 8. Training Set and Testing Set Are Used to Evaluate the Model's Robustness.** 29](#_Toc531864381)

# **Chapter 1: Introduction**

As operators explore and produce hydrocarbon from reservoirs, many types of data gradually increase. A good utilization of data can significantly provide helpful information for making business decisions. Machine learning is a growing technique which is applied to detect patterns, clusters and to predict unknown values of interest. In a study that involves big amount of data, machine learning is advantageous because it can integrate inputs and quickly perform calculations.

In the context of reservoir management, where the work requires geological, petrophysical, reservoir engineering, production engineering and other disciplinary input, incorporating many different types of data to make decision is critical. Therefore, an effective machine learning methodology has the potential to provide reservoir management teams necessary information to make decision in a dynamic and competitive oil and gas market environment.

The objective of this study is to apply machine learning algorithms to develop a predictive model for productivity of the wells penetrating LT reservoir. Therefore, this task is a data analytic problem. Most commonly used algorithm for dimensionality reduction is Principal Component Analysis (PCA). From there, a predictive model can be developed to forecast the performance of future wells operated to deplete LT sand. This is a regression problem. There are many regression algorithms to use for prediction. They could be anything of Linear Regression, Random Forest, Gradient Boosting, Support Vector Machine, etc. The size of the training data set will determine the best combination of algorithms working well in our particular case. The predictive model of this study can be used to forecast future performance of newly drilled wells. From there, reservoir management team can make future plans for development of the field. The author also hopes to find applicability of this workflow in other tight rock reservoirs in the U.S.A and around the world.

# **Chapter 2: Background**

## **2.1: Overview of Machine Learning**

ML is a programmed system designed to learn from data to perform a specific task. According to Tom M. Mitchell, ‘a computer program is said to learn from experience E with response to some class of task T and perform measure P, if its performance at tasks in T, as measured by P, improves with the experience E’ (1). In this research, E is that the computer learned the initial oil production influenced by various factors. Then it will perform a task T as a prediction of initial production for other wells. The number of correctly predicted wells can be seen as the performance measure P.

ML algorithms are mainly categorized into two group: Unsupervised learning and supervised learning. Unsupervised learning is used on data with no labels and the goal is to find relationship in the data. On the other hand, supervised learning algorithms are trained with labeled data or the data comprised of examples of answered wanted.

Most of the algorithms are supervised learning. They are applied for either classification or regression purposes. In classification problems, algorithms take input vector and decide which of the N classes they belong to, based on training from exemplars of each class.  classification provide discrete output. An example belongs to precisely one class, and the set of classes covers the whole possible output space. Classification is done by finding decision boundaries that can be used to separate out the different classes. On the other hand, a regression problem is essentially function approximation. The training data is assumed to come from some sort of function. The algorithms are to find out the function which passes as close to all the data as possible.

In unsupervised learning, the types of algorithm depend on the desired output. If the goal is to separate the data samples with their similarity into discrete groups, the learning is clustering. If the goal is to reduce the dimension of continuous data, the learning is dimension reduction.

**Table 1. There Are Four Types of Machine Learning.**

|  |  |  |
| --- | --- | --- |
|  | Supervised Learning | Unsupervised Learning |
| Discrete | **Classification** | **Clustering** |
| Continuous | **Regression** | **Dimensionality Reduction** |

## **2.2: Previous Works of Machine Learning in Oil and Gas Industry**

In an effort of evaluating production performance of a cold heavy oil reservoir, Cai et al. (2014) applied several machine learning algorithms in their analysis. This endeavor assessed petrophysical parameters with their impacts to production performance parameters such as peak value, effective life cycle and effective yield. Fuzzy logic was used to identify petrophysical parameters with greater impacts to the performance of 118 wells. After that, association rule was applied to classify the well performance as either low, medium or high. This research focused on discovering the unknown or useful patterns.

In another endeavor where, Alzahabi et al. (2016) applied machine learning algorithms to predict dew point pressures for wet gas and condensate gas reservoirs. The algorithms were trained with 667 sets of basic information such as reservoir temperature, molecular weight percent of C7+, reservoir gas gravity and fluid composition. After processing data, analytical steps include (1) screening out the significant predictor variables, (2) applying most commonly used algorithms to make the prediction and (3) using model selection criteria to decide the best prediction model. Pairwise interaction between variables have been also consider and resulted in 21 predictor variables. Then step (2) was carried out with independent application of 4 algorithms namely Linear Regression, Random Forest, Generalized Additive Model and Neural Network. For a straightforward model and the consideration of model selection criteria (AIC, AICc and BIC) utilized in step (3), Linear Regression model was decided as a predicting model for wet gas and condensate gas with a mean error of 2%.

A review of works utilizing machine learning techniques in oil and gas industry helps learn several good lessons. In situations where data is large enough and relevant to answer the question, machine learning algorithms can be used. In addition, depending on the desirable of the outcomes, some algorithms may be more favorable than others.

Previous study on field J, k and N indicated that infill drilling has been a strategy to provide incremental production in LT reservoir. Initial production, a stabilized rate during the first six month of oil production, is important because it is used for decline curve analysis to forecast cumulative production. The instantaneous rates and cumulative production are major parameters in economic analysis to assist decision making. It is desired to develop a quick tool to predict IP for future infill wells. In this particular thesis, the data includes 32 wells. Linear regression is a great candidate because it is very fast and the resulted model can be easy to understand. In addition, it is less prone to overfitting. However, linear regression cannot model complex relationships. Also, it cannot capture nonlinear relationships. Therefore, principal component analysis (PCA) will be used to transform the input to handle the nonlinear relationship and collinearity.

## **2.3: Principal Component Analysis and Linear Regression**

1. **Principal Component Analysis (PCA)**

Given a data set of ‘n’ observations and ‘p’ variables, it can be written in form of matrix X (nxp). Principal component analysis is a statistical technique that linearly transforms an original set of ‘p’ possibly correlated variables into a substantially smaller set of ‘k’ uncorrelated variables that represents most of the information in the original set of variables. Its goal is to reduce the dimensionality of the original data set.

PCA procedure:

* From original data matrix A(nxp), the covariance matrix can be calculated by
* Covariance matrix: , (1)

where . (2)

* Eigenvalues ( and eigenvectors can be obtained by solving:

(3)

and . (4)

The percentage variance of a variable Xi (i < = p) is :

. (5)

If the ‘p’ variables are correlated, they can be transformed into a relatively small set of k uncorrelated variables such that the ‘k’ derived variables. The ‘k’ derived variables which maximize the variance accounted for in the original variables are call principal components.

1. **Linear Regression**

* Score (Z) or the projection of the original data on the eigenvectors (X):

. (6)

* The targeted variable Y can be expressed in terms of the score:

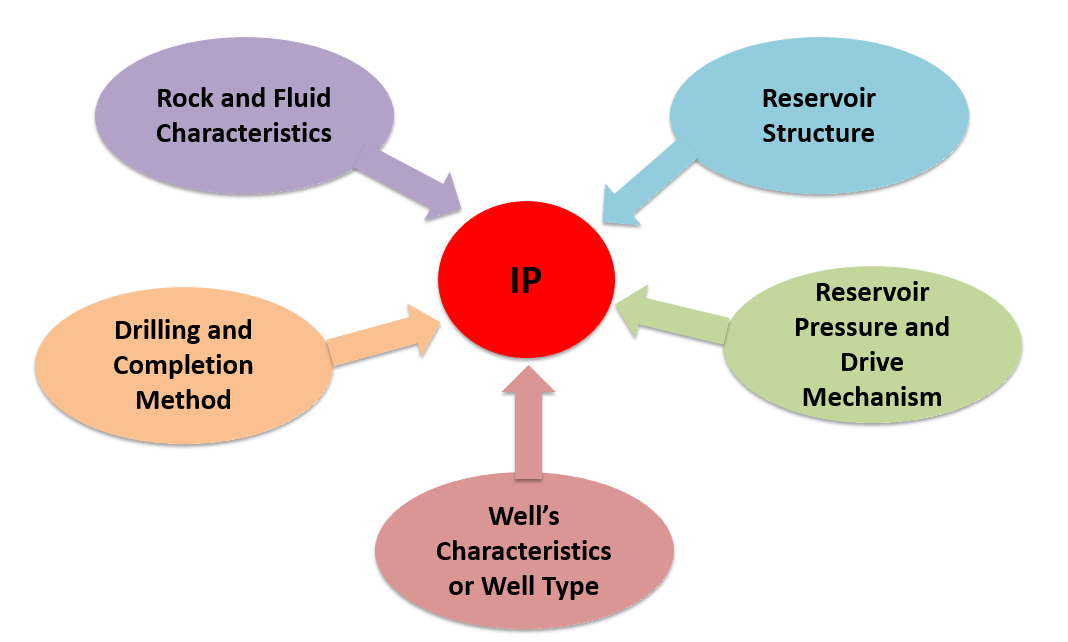
. (7)

* B can be calculated by :

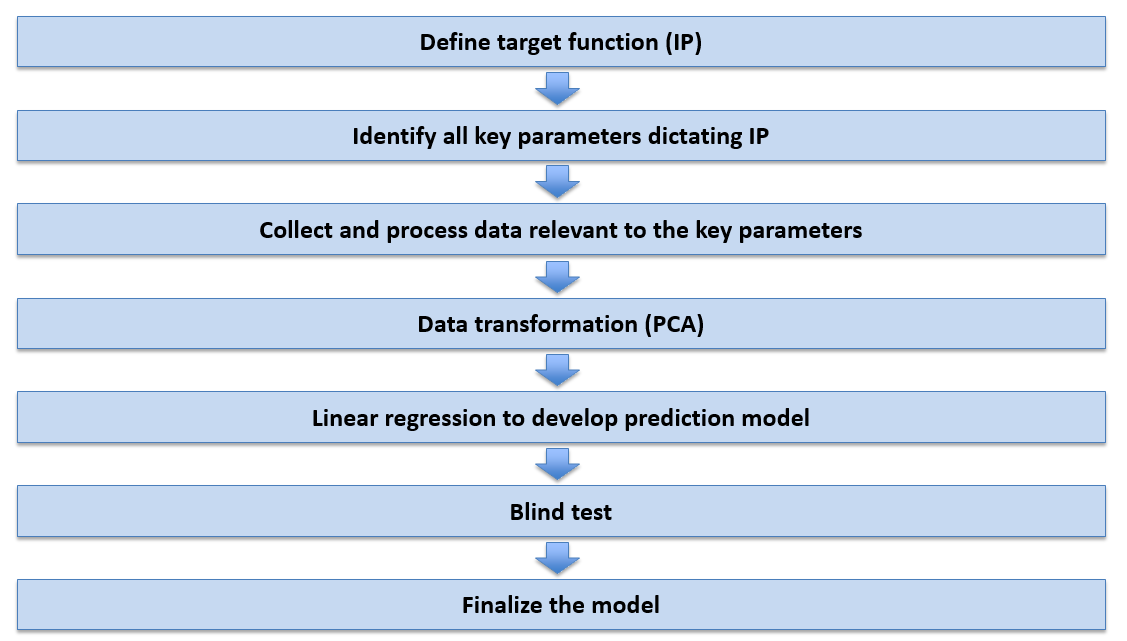
. (8)

# **Chapter 3: Workflow**

A machine learning study always starts with a clear identification for the targeted variable. The targeted variable in this prediction is the IP rate. Then its influencers are defined. When a well starts producing oil, its IP is the intersection of inflow performance relationship and well flow performance. Inflow performance relationship is dictated by the reservoir. In the other words, rock and fluid characteristics, reservoir structure, reservoir pressure and drive mechanism are the major influencers from the reservoir. The well flow performance is dictated by drilling and completion method as well as the well type. Figure XX summaries the IP’s dictators. After defining these influencers, data can be collected accordingly. Not all the data will be used to construct the prediction model. A major portion of data set will be used as a training set. The remaining set, named testing set, will be used for a blind test. Once the data matrix is complete, transformation technique namely PCA can be performed to reduce the data matrix dimension. The score generated by PCA will be utilized to regress for an IP prediction model. Blind test, robustness test and other parameter selection methods are also used to validate the model. Figure XX summaries the workflow to develop the prediction model.



**Figure 1. Major Factors Influencing Initial Production**



**Figure 2. Workflow to Develop Prediction Model for IP**

# **Chapter 4: Data Collection**

Field data and petrophysical properties from reservoir model are collected for each well to construct the data matrix. The targeted variable initial production is obtained as the stabilized rate within the first six-month of the oil producers. Other information are collected according to the key factors impacting IP.

Drilling and completion and well characteristics: All wells in three fields are vertical and mostly produce in primary recovery stage. Available data includes total length of the perforations. Second and third column of **Table 2** summarizes the initial production and completion in the wells of interest.

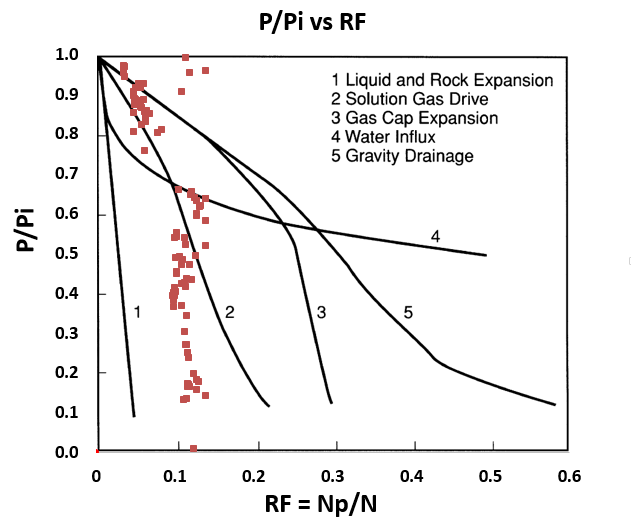
**Table 2. IP and Total Perforation Has Been Collected for the Selected Wells.**

|  |  |  |
| --- | --- | --- |
| **Completion ID** | **Initial  Production** | **Total Perf.** |
|  | **BOPD** | **ft** |
| **01E** | **439.090938** | **20.0** |
| **02D** | **681.81432** | **13.8** |
| **003** | **590.675118** | **18.0** |
| **008** | **1196.50865** | **19.7** |
| **013D** | **268.385766** | **9.8** |
| **014** | **278.763936** | **32.8** |
| **016D** | **30.82002** | **16.4** |
| **018** | **306.753546** | **29.5** |
| **019** | **396.634788** | **19.7** |
| **020** | **158.691654** | **39.4** |
| **023D** | **456.513684** | **59.1** |
| **025D** | **239.0124** | **16.4** |
| **026D** | **256.183554** | **16.4** |
| **028** | **549.665622** | **29.5** |
| **030** | **147.244218** | **59.1** |
| **032D** | **39.185454** | **16.4** |
| **243** | **591.429894** | **20.0** |
| **280D** | **51.136074** | **19.7** |
| **283D** | **220.709082** | **9.8** |

|  |  |  |
| --- | --- | --- |
| ***Table 2. (Continued)*** | | |
| **309** | **208.884258** | **29.9** |
| **310D** | **46.23003** | **19.7** |
| **328** | **985.863252** | **19.7** |
| **341** | **597.531** | **19.7** |
| **343A** | **95.91945** | **9.8** |
| **344** | **567.528654** | **19.7** |
| **346** | **442.17294** | **19.7** |
| **347D** | **392.420622** | **29.5** |
| **363D** | **72.64719** | **13.1** |
| **379** | **380.910288** | **29.5** |
| **387D** | **286.437492** | **13.1** |
| **392D** | **192.656574** | **16.4** |
| **396D** | **254.7369** | **13.1** |
| **415** | **17.863032** | **32.8** |
| **416** | **535.324878** | **19.7** |
| **418** | **242.031504** | **19.7** |
| **436D** | **117.30477** | **16.4** |
| **457D** | **165.170148** | **23.0** |
| **476D** | **6.2898** | **16.4** |
| **274** | **205.865154** | **20.0** |
| **298E** | **331.409562** | **9.8** |
| **367A** | **50.94738** | **29.5** |

Reservoir pressure and drive mechanism: Shut-in bottom hole pressure (SIBHP) is measure when the wells are shut-in. This pressure is close to reservoir pressure. SIBHP at initial production time were measure for a number of wells. The remaining missing pressure data can be estimated by the pressure trends corresponding to the drive mechanism in the reservoir.

* Field J: Two major drive mechanisms are Solution Gas Drive (early time) and a combination of Solution Gas Drive and Liquid and Rock Expansion (later time) (**Figure 3**). The two SIBHP trends could be the responses from the two drive mechanisms in the reservoirs (**Figure 4** and **Figure 5**). The trends are used to estimate the SIBHP for wells lacking this data.

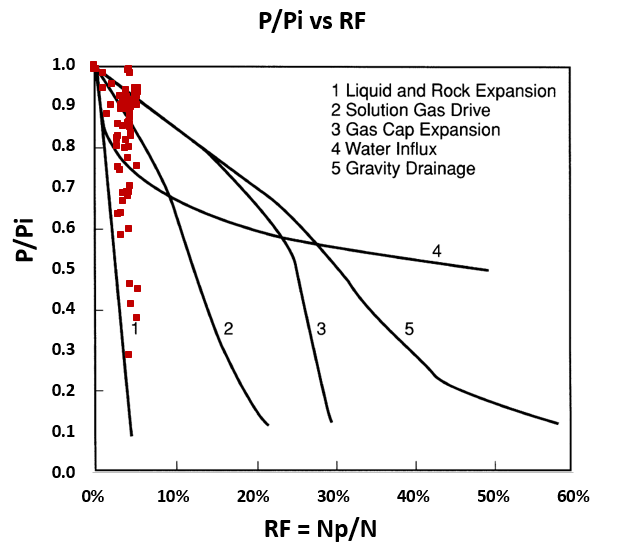


**Figure 3. Drive Mechanisms in Field J**

**Figure 4. SIBHP Trend in Block J1 of Field J**

**Figure 5. SIBHP Trend in Block J2 of Field J**

* Field K: LT reservoir in field K quickly approaches rock and fluid expansion mechanism (**Figure 6**). The single SIBHP trend corresponds to this single drive mechanism. The trends are used to estimate the SIBHP for wells lacking this data in field K.



**Figure 6. Drive Mechanism in Field K**

**Figure 7. SIBHP in Block K243, 332 and 341 in Field K**

**Figure 8. SIBHP in Block K432 in Field K**

* Field N: a well fit trend of SIBHP versus time in field N is used to estimate SIBHP for wells lacking this data in field N.

**Figure 9. SIBHP Trend in Field N**

# **Chapter 5: Transforming Data**

The data matrix of 30 well samples and 39 parameters (**Table 3**) is transformed by PCA. The resultant new vector space has 29 components. The first component reflects the highest variability of the original variables. As the order of the component increases, the variability on the components decrease. **Table 4** summarizes the eigen values, variability percentage and cumulative variability of the components.

**Table 3. Thirty-Nine Parameters Are Used to Construct Prediction Model.**

|  |  |
| --- | --- |
| **Variable** | |
| **Top of LT** | **SWI (P50)** |
| **Bottom of LT** | **SWI (P10 )** |
| **Reservoir Gross Thickness** | **SWI (P90 )** |
| **Total Perf.** | **SWE (Std\_Dev)** |
| **SIBHP** | **SWE (Mean)** |
| **FBHP** | **SWE (P50)** |
| **Viscosity** | **SWE (P10 )** |
| **Bo** | **SWE (P90 )** |
| **VSH (Std\_Dev)** | **PHIE (Std\_Dev)** |
| **VSH (Mean)** | **PHIE (Mean)** |
| **VSH (P50)** | **PHIE (P50)** |
| **VSH (P10 )** | **PHIE (P10 )** |
| **VSH (P90 )** | **PHIE (P90 )** |
| **SWT (Std\_Dev)** | **Vsh Cuttoff** |
| **SWT (Mean)** | **PERM (Std\_Dev)** |
| **SWT (P50)** | **PERM (Mean)** |
| **SWT (P10 )** | **PERM (P50)** |
| **SWT (P90 )** | **PERM (P10 )** |
| **SWI (Std\_Dev)** | **PERM (P90 )** |
| **SWI (Mean)** |  |

**Table 4. PCA Results in Eigen Values, Variability for 29 Components.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Component** | **Eigenvalue** | **Variability (%)** | **Cumulative %** |
| **F1** | **12.889** | **33.050** | **33.050** |
| **F2** | **8.827** | **22.633** | **55.683** |
| **F3** | **7.005** | **17.961** | **73.644** |
| **F4** | **3.302** | **8.466** | **82.111** |
| **F5** | **1.403** | **3.596** | **85.707** |
| **F6** | **1.206** | **3.091** | **88.798** |
| **F7** | **1.152** | **2.953** | **91.751** |
| **F8** | **0.865** | **2.219** | **93.970** |
| **F9** | **0.629** | **1.613** | **95.584** |
| **F10** | **0.520** | **1.334** | **96.917** |
| **F11** | **0.338** | **0.867** | **97.785** |
| **F12** | **0.267** | **0.686** | **98.470** |
| **F13** | **0.161** | **0.413** | **98.883** |
| **F14** | **0.117** | **0.299** | **99.183** |
| **F15** | **0.083** | **0.214** | **99.396** |
| **F16** | **0.067** | **0.171** | **99.567** |
| **F17** | **0.051** | **0.132** | **99.699** |
| **F18** | **0.035** | **0.090** | **99.789** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 4. (Continued)** | | | |
| **F19** | **0.026** | **0.067** | **99.857** |
| **F20** | **0.018** | **0.045** | **99.902** |
| **F21** | **0.015** | **0.039** | **99.941** |
| **F22** | **0.008** | **0.019** | **99.960** |
| **F23** | **0.007** | **0.018** | **99.978** |
| **F24** | **0.003** | **0.009** | **99.987** |
| **F25** | **0.002** | **0.006** | **99.993** |
| **F26** | **0.002** | **0.004** | **99.997** |
| **F27** | **0.001** | **0.002** | **99.999** |
| **F28** | **0.000** | **0.001** | **100.000** |
| **F29** | **0.000** | **0.000** | **100.000** |

# **Chapter 6: Performing Linear Regression Model**

## **6.1: Selecting Components for Regression**

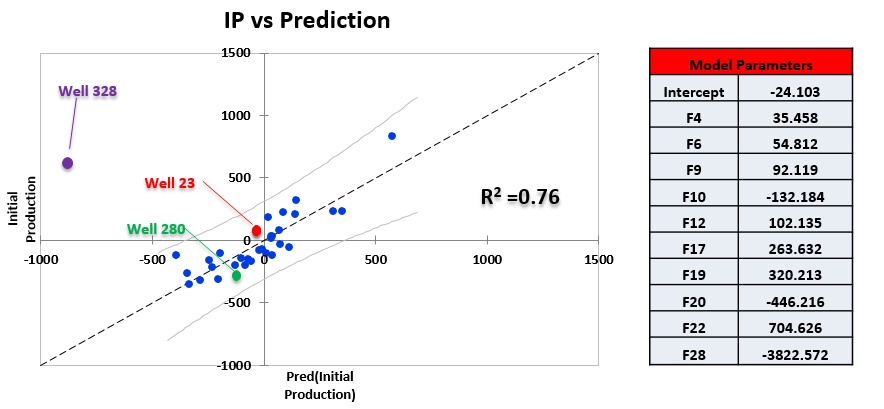
Two methods of selecting the components to build regression are performed. The results will be compared to determine the better method. A commonly used method is to include first components representing the most variability in the original data. A new proposed method is to components whose scores are correlated to IP the most.

**Table 5** includes 10 components most correlated to IP. A prediction model based on these components is created by linear regression technique. A plot of actual IP versus predicted IP and the model parameters are illustrated on **Figure 10**. The testing data set (3 wells) is used in the blind test.

**Table 5. 10 Components with the Most Correlation to IP Are Used for Modeling.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **IP** | **F4** | **F6** | **F9** | **F10** | **F12** | **F17** | **F19** | **F20** | **F22** | **F28** |
| **1E** | **79.44214** | **-0.144** | **2.489** | **1.360** | **0.540** | **-0.480** | **-0.190** | **-0.077** | **-0.005** | **0.044** | **0.001** |
| **2D** | **322.1655** | **1.032** | **0.801** | **0.821** | **-1.265** | **-0.306** | **-0.137** | **-0.269** | **-0.021** | **-0.007** | **0.002** |
| **3** | **231.0263** | **4.429** | **0.748** | **0.109** | **0.141** | **0.475** | **-0.234** | **-0.100** | **-0.048** | **-0.067** | **0.002** |
| **8** | **836.8599** | **1.303** | **0.715** | **0.692** | **-0.918** | **0.136** | **0.477** | **0.446** | **-0.026** | **0.009** | **-0.006** |
| **14** | **-80.8849** | **1.309** | **1.409** | **-0.011** | **0.989** | **0.669** | **0.203** | **0.114** | **0.136** | **-0.081** | **0.008** |
| **18** | **-52.8953** | **-3.025** | **0.497** | **1.160** | **-0.524** | **0.671** | **0.191** | **-0.105** | **0.164** | **0.069** | **0.006** |
| **19** | **36.98599** | **1.475** | **-1.929** | **-0.855** | **-0.642** | **-0.345** | **0.018** | **0.075** | **-0.023** | **-0.003** | **-0.028** |
| **20** | **-200.957** | **-1.532** | **-2.180** | **0.960** | **0.295** | **-0.260** | **0.078** | **-0.110** | **-0.044** | **0.059** | **-0.011** |
| **25D** | **-120.636** | **-4.814** | **0.861** | **-0.052** | **0.715** | **-0.013** | **-0.364** | **0.016** | **0.000** | **-0.116** | **-0.007** |
| **26D** | **-103.465** | **-2.107** | **0.117** | **-0.454** | **0.815** | **-1.259** | **0.098** | **0.154** | **-0.122** | **0.183** | **0.022** |
| **28** | **190.0168** | **0.567** | **-1.974** | **-0.367** | **-0.351** | **0.549** | **-0.113** | **-0.067** | **-0.085** | **0.135** | **0.005** |
| **30** | **-212.405** | **-2.597** | **-1.608** | **0.109** | **-0.485** | **0.381** | **-0.334** | **-0.019** | **0.038** | **-0.019** | **0.005** |
| **32D** | **-320.463** | **-0.543** | **-0.562** | **-1.256** | **-0.095** | **-0.450** | **0.273** | **0.049** | **0.054** | **-0.178** | **0.000** |
| **243** | **231.7811** | **1.252** | **-0.327** | **1.353** | **-0.239** | **0.481** | **-0.114** | **0.088** | **-0.038** | **0.032** | **-0.027** |
| **283D** | **-138.94** | **2.435** | **0.899** | **-1.499** | **0.903** | **-0.120** | **0.080** | **-0.236** | **-0.008** | **0.182** | **0.006** |
| **309** | **-150.765** | **0.491** | **-0.246** | **0.004** | **-0.456** | **-1.266** | **-0.054** | **0.046** | **0.138** | **-0.072** | **-0.034** |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Table 5. (Continued)*** | | | | | | | | | | | |
| **310D** | **-313.419** | **-1.044** | **0.017** | **0.131** | **-0.730** | **-0.384** | **-0.267** | **0.115** | **-0.069** | **-0.070** | **0.043** |
| **341** | **237.8822** | **-0.485** | **1.007** | **0.515** | **-0.443** | **0.009** | **0.030** | **-0.041** | **0.029** | **0.139** | **-0.028** |
| **343A** | **-263.729** | **-0.436** | **0.033** | **-0.832** | **-0.240** | **0.079** | **0.165** | **-0.397** | **0.327** | **-0.085** | **-0.004** |
| **344** | **207.8799** | **-0.664** | **-0.580** | **-0.226** | **-0.227** | **0.339** | **0.299** | **-0.034** | **-0.191** | **-0.032** | **-0.011** |
| **347D** | **32.77182** | **-0.010** | **0.421** | **-0.294** | **-0.437** | **-0.320** | **0.342** | **-0.224** | **-0.098** | **0.016** | **0.010** |
| **379** | **21.26149** | **-2.239** | **0.249** | **-0.405** | **0.441** | **0.954** | **0.320** | **0.105** | **0.013** | **0.009** | **0.000** |
| **387D** | **-73.2113** | **0.381** | **0.979** | **-1.747** | **-0.818** | **0.290** | **-0.519** | **0.288** | **0.199** | **0.106** | **-0.007** |
| **392D** | **-166.992** | **-0.455** | **-0.113** | **-0.884** | **-0.213** | **0.783** | **-0.075** | **0.012** | **-0.140** | **0.005** | **0.023** |
| **396D** | **-104.912** | **1.897** | **-0.205** | **0.518** | **0.413** | **-0.003** | **-0.154** | **0.095** | **-0.035** | **-0.056** | **-0.005** |
| **418** | **-117.617** | **-0.362** | **1.196** | **-0.599** | **0.520** | **0.019** | **-0.078** | **-0.062** | **-0.400** | **-0.135** | **-0.022** |
| **457D** | **-194.479** | **2.242** | **-0.808** | **0.741** | **-0.590** | **-0.313** | **0.043** | **0.017** | **0.039** | **-0.066** | **0.054** |
| **476D** | **-353.359** | **0.705** | **-0.856** | **0.331** | **1.759** | **-0.297** | **0.057** | **0.161** | **0.227** | **-0.015** | **0.004** |
| **274** | **-153.784** | **1.349** | **-1.760** | **0.675** | **1.572** | **0.237** | **-0.164** | **-0.056** | **-0.007** | **0.003** | **-0.001** |
| **298E** | **-28.2392** | **-0.411** | **0.712** | **0.001** | **-0.427** | **-0.259** | **0.122** | **0.018** | **-0.002** | **0.010** | **0.001** |



**Figure 10. Actual IP versus Predicted IP by 10 Components Most Correlated to IP**

Blind test:

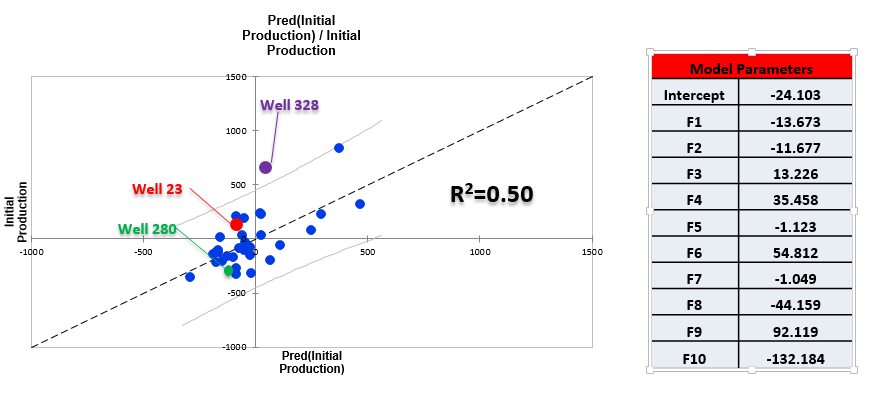
* Well 23, 280 is predicted within the tolerance range of the prediction model
* Well 328 is outside of the tolerance range of the prediction model. It is consider to be an outliner or there might be error in the data of this well.

**Table 6** includes the first 10 components with the most variability. A prediction model based on these components is created by linear regression technique. A plot of actual IP versus predicted IP and the model parameters are illustrated on figure XX. The testing data set (3 wells) is used in the blind test.

**Table 6. 10 Components with the Most Variability Are Used for Modeling.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **IP** | **F1** | **F2** | **F3** | **F4** | **F5** | **F6** | **F7** | **F8** | **F9** | **F10** |
| **1E** | **79.44214** | **-5.298** | **0.395** | **0.170** | **-0.144** | **-0.491** | **2.489** | **-0.538** | **-0.349** | **1.360** | **0.540** |
| **2D** | **322.1655** | **-2.443** | **-1.145** | **5.844** | **1.032** | **-1.771** | **0.801** | **1.203** | **-0.954** | **0.821** | **-1.265** |
| **3** | **231.0263** | **-8.625** | **-0.776** | **-2.464** | **4.429** | **1.559** | **0.748** | **-1.238** | **-0.724** | **0.109** | **0.141** |
| **8** | **836.8599** | **-1.955** | **-2.655** | **3.514** | **1.303** | **-0.758** | **0.715** | **0.210** | **-0.472** | **0.692** | **-0.918** |
| **14** | **-80.8849** | **-2.294** | **-3.034** | **-3.087** | **1.309** | **-0.624** | **1.409** | **-0.587** | **1.608** | **-0.011** | **0.989** |
| **18** | **-52.8953** | **-0.162** | **-2.645** | **-0.275** | **-3.025** | **0.311** | **0.497** | **0.525** | **-0.180** | **1.160** | **-0.524** |
| **19** | **36.98599** | **-3.204** | **-1.452** | **-1.599** | **1.475** | **1.276** | **-1.929** | **-0.434** | **-1.295** | **-0.855** | **-0.642** |
| **20** | **-200.957** | **-3.380** | **1.548** | **-2.102** | **-1.532** | **-0.436** | **-2.180** | **-0.034** | **0.103** | **0.960** | **0.295** |
| **25D** | **-120.636** | **-1.309** | **-2.012** | **-0.957** | **-4.814** | **0.899** | **0.861** | **0.773** | **-1.152** | **-0.052** | **0.715** |
| **26D** | **-103.465** | **-1.224** | **-4.210** | **-1.640** | **-2.107** | **1.105** | **0.117** | **-0.127** | **-0.723** | **-0.454** | **0.815** |
| **28** | **190.0168** | **-3.872** | **-2.128** | **-0.912** | **0.567** | **-0.532** | **-1.974** | **1.176** | **0.395** | **-0.367** | **-0.351** |
| **30** | **-212.405** | **-2.613** | **-2.348** | **-0.733** | **-2.597** | **-1.400** | **-1.608** | **-0.928** | **2.372** | **0.109** | **-0.485** |
| **32D** | **-320.463** | **-2.597** | **-4.937** | **1.107** | **-0.543** | **-0.594** | **-0.562** | **1.753** | **0.352** | **-1.256** | **-0.095** |
| **243** | **231.7811** | **6.017** | **-0.199** | **-4.263** | **1.252** | **0.857** | **-0.327** | **1.107** | **-0.104** | **1.353** | **-0.239** |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Table 6. (Continued)*** | | | | | | | | | | | |
| **283D** | **-138.94** | **3.595** | **-1.793** | **-0.326** | **2.435** | **-0.618** | **0.899** | **0.381** | **0.255** | **-1.499** | **0.903** |
| **309** | **-150.765** | **1.516** | **1.655** | **0.030** | **0.491** | **-0.521** | **-0.246** | **-2.304** | **0.647** | **0.004** | **-0.456** |
| **310D** | **-313.419** | **4.985** | **2.214** | **0.219** | **-1.044** | **-0.138** | **0.017** | **-0.867** | **-0.497** | **0.131** | **-0.730** |
| **341** | **237.8822** | **3.604** | **1.233** | **-1.931** | **-0.485** | **-0.306** | **1.007** | **-0.519** | **0.260** | **0.515** | **-0.443** |
| **343A** | **-263.729** | **3.353** | **-0.017** | **-0.823** | **-0.436** | **0.984** | **0.033** | **0.165** | **-1.140** | **-0.832** | **-0.240** |
| **344** | **207.8799** | **-0.350** | **4.910** | **-0.522** | **-0.664** | **-0.244** | **-0.580** | **-0.497** | **-0.967** | **-0.226** | **-0.227** |
| **347D** | **32.77182** | **1.433** | **1.602** | **0.190** | **-0.010** | **-1.200** | **0.421** | **-0.789** | **1.246** | **-0.294** | **-0.437** |
| **379** | **21.26149** | **2.438** | **0.157** | **1.682** | **-2.239** | **0.529** | **0.249** | **-2.222** | **-0.817** | **-0.405** | **0.441** |
| **387D** | **-73.2113** | **0.827** | **1.495** | **1.483** | **0.381** | **-0.370** | **0.979** | **-0.063** | **0.180** | **-1.747** | **-0.818** |
| **392D** | **-166.992** | **-1.022** | **5.325** | **0.808** | **-0.455** | **-0.594** | **-0.113** | **0.313** | **-0.845** | **-0.884** | **-0.213** |
| **396D** | **-104.912** | **8.209** | **-2.745** | **-0.454** | **1.897** | **-0.342** | **-0.205** | **1.849** | **-0.191** | **0.518** | **0.413** |
| **418** | **-117.617** | **2.495** | **0.227** | **-0.802** | **-0.362** | **-0.668** | **1.196** | **0.255** | **0.808** | **-0.599** | **0.520** |
| **457D** | **-194.479** | **3.342** | **-0.181** | **-3.634** | **2.242** | **0.736** | **-0.808** | **-0.551** | **0.066** | **0.741** | **-0.590** |
| **476D** | **-353.359** | **-2.470** | **8.028** | **-0.505** | **0.705** | **-2.029** | **-0.856** | **1.479** | **-0.423** | **0.331** | **1.759** |
| **274** | **-153.784** | **3.142** | **-2.189** | **9.149** | **1.349** | **1.221** | **-1.760** | **-1.287** | **0.271** | **0.675** | **1.572** |
| **298E** | **-28.2392** | **-2.139** | **5.677** | **2.833** | **-0.411** | **4.160** | **0.712** | **1.797** | **2.272** | **0.001** | **-0.427** |



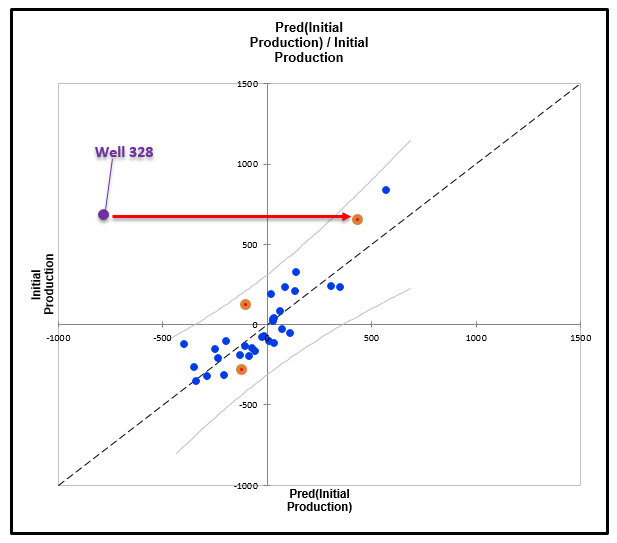
**Figure 11. Actual IP versus Predicted IP by First 10 Components**

Blind test:

* Well 23 is predicted within the tolerance range of the prediction model
* Well 328 is outside of the tolerance range of the prediction model. It is consider to be an outliner or there might be error in the data of this well.

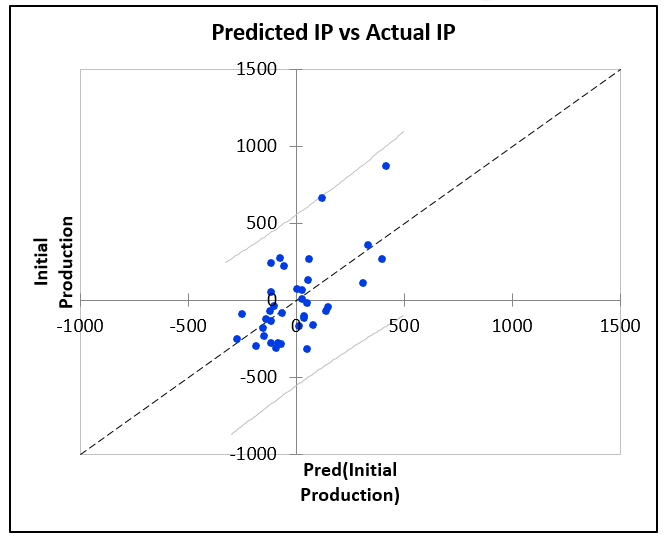
Correcting Data for Well N328

Changing the data values doesn’t help model, built by first 10 components, to bring prediction into the tolerance range. The model, built by 10 correlated components, is most sensitive to the parameter SWT(P50). Changing it from 0.69 to 0.63 will make the prediction in the tolerance range. Further confirmation from operator is needed to confirm the change.

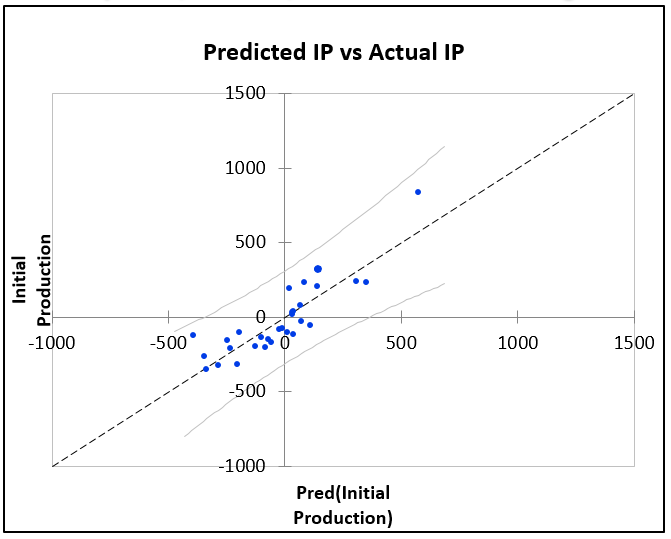


**Figure 12. Correcting Data for Well 328 Moves the Prediction into Tolerance Range**

## **6.2: Averaged versus Probabilistic Properties**

**Figure 13** is a plot of actual IP versus prediction from a model built from the averaged petrophysical parameters (effective porosity, water saturations, shale volume and permeability). The R2 value is 0.45 with a wide tolerance range. **Figure 14** shows actual IP versus prediction from a model built from the probabilistic of petrophysical parameters (their standard deviation, mean, P10, P50, and P90). The R2 value is 0.76 with a narrower tolerance range. This indicates that probabilistic of the petrophysical parameters yield a better prediction model. 

**Figure 13. Actual IP versus Predicted IP Generated by Averaged Properties**



**Figure 14. Actual IP versus Predicted IP Generated by Probabilistic Properties**

## **6.3: Adding New Properties-Averaged Skin**

Skin values are from well test reports of 8 wells. The other values are estimated by averaging the skin of other wells in the reservoir. R2 value of the prediction model is reduced from 76% to 66%.

**Table 7. Actual Skin and Averaged Skin (in Orange Boxes) Are Used as New Parameters.**

|  |  |
| --- | --- |
| **Completion** | **Skin** |
| **01E** | **5.1500** |
| **02D** | **2.6000** |
| **003** | **25.1000** |
| **008** | **8.5000** |

|  |  |
| --- | --- |
| ***Table 7. (Continued)*** | |
| **014** | **15.7000** |
| **018** | **11.0833** |
| **019** | **11.0833** |
| **020** | **11.0833** |
| **25D** | **11.0833** |
| **26D** | **11.0833** |
| **028** | **11.0833** |
| **030** | **11.0833** |
| **32D** | **9.4500** |
| **243** | **6.3400** |
| **283D** | **2.1500** |
| **309** | **2.1500** |
| **310D** | **2.1500** |
| **341** | **2.1500** |
| **343A** | **2.1500** |
| **344** | **2.1500** |
| **347D** | **2.1500** |
| **379** | **-2.0400** |
| **387D** | **2.1500** |
| **392D** | **2.1500** |
| **396D** | **2.1500** |
| **418** | **2.1500** |
| **457D** | **2.1500** |
| **476D** | **2.1500** |
| **274** | **8.8500** |
| **298E** | **8.8500** |

**Figure 15. Actual IP versus Predicted IP Modeled with Skin Data**

## **6.4: Robustness Test**

For the robustness test, training set is reduced from 30 to 25 wells. The testing set will have 7 wells. The same procedure is used to construct prediction model for the reduced training set. This reduced model will be compare with the original model to assess the effect of data size to modeling prediction.

Even though the R2 value is very high but the model couldn’t predict IP for the training set. Therefore, the training set of 25 samples may not be enough as a training data set.

**Table 8. Training Set and Testing Set Are Used to Evaluate the Model's Robustness.**

|  |  |
| --- | --- |
| **Training Set** | **Testing Set** |
| **JRN001E** | **JRN023D** |
| **JRN003** | **JRN002D** |
| ***Table 8. (Continued)*** | |
| **JRN008** | **JRN028** |
| **JRN014** | **NHK310D** |
| **JRN018** | **NHK328** |
| **JRN019** | **NHK274** |
| **JRN020** | **NHK418** |
| **JRN025D** |  |
| **JRN026D** |  |
| **JRN030** |  |
| **JRN032D** |  |
| **NHK243** |  |
| **NHK283D** |  |
| **NHK309** |  |
| **NHK341** |  |
| **NHK343A** |  |
| **NHK344** |  |
| **NHK347D** |  |
| **NHK379** |  |
| **NHK387D** |  |
| **NHK392D** |  |
| **NHK396D** |  |
| **NHK457D** |  |
| **NHK476D** |  |
| **NHK298E** |  |

**Figure 16. Actual IP versus Predicted IP Generated by Modeling with 25 Well Samples**

## **6.5: Uncertainties**

If a piece of information which is related to IP is missing in this analysis, it is the cause of uncertainty for this predictive model. Missing information includes the following,

* Reservoir structure is not described in detail (fault and natural fracture system): Previous studies about oil fields J, K and N indicate that the fault system in the area is complex as the result of strong tectonic activities. Faults can be compartmentalizing or leakage path established in the reservoir. Understanding the fault system in the reservoir and incorporating them into the data matrix could be a great uncertainty reducer.
* Drilling data: Drilling method can make great impact to the production of wells. Especially when rock lithology is the concern, choosing mud type, mud weight and the casing design are very important factor because it could determine the integrity of the borehole hence IP. In LT reservoir, smectite, a swelling clay, is identified. If fresh water mud is used as the drilling fluid, smectite will swell and prevent the flow of reservoir fluid into the wellbore. Therefore, data related to drilling process can also have an important impact on IP.
* Details of perforation: wellbore integrity can be impacted by perforation. Damage to the well can be caused by charge debris in perforation tunnel and crushing of rock and/or completion fluid invasion into the perforations. In addition, low production can be caused by not enough perforations or the perforations are shallow and not penetrating beyond damage zone. Therefore, detail of perforation data can be helpful to improve the accuracy of the predictive model.
* Wellbore damage (skin): skin can be caused by mechanical or chemical damage in near wellbore region. As positive skin will reduce the effective wellbore radius, the production will be reduced. In LT reservoir, about a quarter of the wells have been tested and skin values fluctuate from 8 to 98. In the developed predictive model, skin data is absent due to limited number of well test available. However, if skin can be reasonably estimated for the un-tested wells, its value can be included in the data matrix. It is highly expected that skin can be a significant parameter affecting IP.
* Surface and downhole operation: downhole and surface pressure directly influence the inflow performance relationship and vertical flow performance. Therefore, any operation affecting surface and bottom hole pressure could help reduce the uncertainties in this prediction model.

# **Chapter 7: Conclusions**

* PCA and multivariate linear regression can be used to develop a prediction model for IP. In this study, PCA reduces a data matrix with dimension (30 x 39) to smaller dimension (30x29). The 29 components are orthogonal (or independent) to each other. Ten out of 29 components are selected to construct a prediction model using multivariate linear regression technique.
* In this case, components with the most correlation to IP contribute to a better prediction model than do the first components with greatest variance. With components mostly correlated to IP, the model yields an R2 value of 0.76 and a narrow tolerance range. With components mostly represent the variability, the model yields R2 value of 0.50 and a wider tolerance range.
* Thorough analysis and quality check on data are important before incorporating them into the predictive model. SIBHP data is collected by using the correlations corresponding to the reservoir drive mechanism. FBHP data is collected by using an identified relationship between SIBHP and FBHP.
* The model can be used to predict future wells’ initial production with reasonable accuracy (0.76). In the earth model for LT reservoir, after identifying potential infill drilling locations, petrophysical parameters can be obtained. Other operational parameters like pressures and completion can be estimated. Then the predicted IP can be done with the developed model.
* Missing data should be handled by understanding the behavior of various reservoir parameters. Reasonable estimation for SIBHP can be estimated by understanding the reservoir drive mechanisms. As a result, SIBHP greatly contributes to the model’s accuracy (0.76). On the other hand, averaged skin values (instead of actual or physic-based estimation) reduces the model’s accuracy (0.66). By considering the above, the proposed model predicts IPs more accurately.
* The more data samples in the training set, the more robust the prediction model.

# **Chapter 8: Recommendation for Future Studies**

The model accuracy can be improve by considering the following in order of priority,

1. Normalize IP with respect to perforation interval. This will require additional information in completion history to analyze the perforated intervals in LT reservoir.
2. Once IP has been normalized, the new target parameter will be IP per foot (IP/ft). Then, redundancy in the data matrix can be reduced by using engineering judgement. For example, the parameter Total Perforation Length can be taken out of the data matrix after the IP normalization. Furthermore, Bottom of Reservoir can also be removed from the data matrix since Top of Reservoir and Gross Thickness are available.
3. Robustness of the model could be improved by using training, testing and validation data sets – whereby validation data is never seen by the model. In addition, the training set can be created by selecting wells whose performance has been well understood.
4. Incorporate learnings from the reservoir engineering analysis done at the beginning to improve the quality of the model, e.g., incorporate the information that wells close to faults produce better. This task will require additional data from reservoir model to be accomplished.

# **References**

1. Bell, Jason. *Machine Learning Hands-on for Developers and technical Professionals*. Indianapolis: John Wiley & Sons, 2015.
2. Yongxiang Cai, Xin Wang, Kezhen Hu, Mingzhe Dong. "A data mining approach tofinding relationships between reservoirproperties and oil production for CHOPS." 2014.
3. Alzahabi, Ahmed, Soliman, Mohamed. "A regression model for estimation of dew point pressurefrom down-hole fluid analyzer data." *Journal of Petrolum Exploration and Production Technology* (2016).