Politecnico di Torino

Department of Environmental, Land and Infrastructure Engineering

Master of Science in Petroleum and Mining Engineering

Investigation of the Relationship between Operational Parameters and Decline Curve Characteristics in Shale Gas Wells using Data Analytics and Machine Learning

Supervisor:

Prof. Francesca Verga

Co-Supervisor:

Assoc. Prof. Emre Artun (Istanbul Technical University)

Candidate:

Ahad Jafarov

October 2022

Thesis is submitted in compliance with the requirements for the Master of Science degree in Petroleum and Mining Engineering

Abstract

In the current technologically advancing world, the role of data science and data analytics has been increasing throughout the last decade. Machine learning has become increasingly important in different disciplines and its application spheres include also the petroleum industry. There have been studies that revealed how the operational parameters affect the production and well performance which leads to more and more studies to be dedicated to them. One of the methods to predict the production and well performance for the future and its potential lifespan, especially in the decline phase is the decline curve analysis. This study involves an investigation of the relationship between operational parameters and decline curve characteristics based on the dataset consisting of 53 shale gas wells data provided by SPE. Via use of well data, decline curves were fit onto the production history for all 53 shale gas wells, and decline curve characteristics (which are q_i, D_i, b) were obtained accordingly. As a main subsequent step, the development and application of different machine learning algorithms such as Multiple Linear Regression and a tree-based method of Random Forest, has been performed for the determination of prediction models using operational parameters as an input and decline curve characteristics as an output. As the additional second part of the project, new predictive models of aforementioned types were developed for the prediction of the cumulative production after 0.5 and 1 year. The conclusion reached regarding the relationship between operational parameters and decline curve characteristics is that there is some correlation although the lack of data has complicated the decision-making procedure a lot. With a much higher amount of data, it would have been more precise to define to what extent the correlation is. When it comes to the comparison between the distinct types of models, it has been concluded that Random Forest model performed better wholistically despite in the second part of the study, the Linear Regression model outperformed the former one. Furthermore, feature importance analysis was conducted to disclose the influence level of input parameters on the output ones after the predictive models have been developed. Parameters making the most significant contribution to the results were different based on the case being analysed.

Keywords: Data science, Data analytics, Machine learning, Decline curve characteristics, Shale gas wells, Operational parameters, Variable importance analysis

Acknowledgements

I would like to express my gratitude to people who had a crucial contribution to this project. Initially, I am thankful to the co-supervisor of this project, Assoc. Prof. Emre Artun from Istanbul Technical University for accepting to be my supervisor. He was continuously supervising and providing worthwhile comments and feedback throughout this study despite being occupied with other responsibilities. Moreover, I am grateful to SPE professionals for the supply of the dataset.

Many thanks to my internal supervisor Prof. Francesca Verga for her supervision and support providing me with an opportunity to benefit from the outgoing mobility program which became an irreplaceable and unique experience in my life. Also, I would like to thank Polytechnic University of Turin officials for supplying and supporting students with such an awesome chance to have an additional international experience. Furthermore, I am grateful to professors of the Department of Environment, Land and Infrastructure Engineering for teaching and giving indispensable instructions throughout the whole didactic period.

Ultimately, I am thankful to my precious family and associates for their support, love and motivation.

Contents

List of Tables

List of Figures

List of Abbreviations

DCA	Decline Curve Analysis
PLE	Power Law Exponential
SEDM / SEPD	Stretched-Exponential Production Decline Model
NN	Neural Network
ML	Machine Learning
EDA	Exploratory Data Analysis
OLS	Ordinary Least Squares
MLR	Multiple Linear Regression
SLR	Simple Linear Regression
RF	Random Forest
SVR	Support Vector Regression
GBM	Gradient Boosting Machine
AAE	Average Absolute Error
MSE	Mean Squared Error
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
VIA	Variable Importance Analysis
EUR	Estimated Ultimate Recovery
BDF	Boundary-Dominated Flow
BHP	Bottom-Hole Pressure
ASE	Absolute Squared Error
SSE / RSS	Residual Sum of Squares or Sum of Squared Errors
TSS	Total Sum of Squares
KDE	Kernel Density Estimation
TVD	True Vertical Depth
MDI	Mean Decrease in Impurity

List of Symbols

q	Flowrate
D_i	Nominal decline rate
b	Hyperbolic exponent
G_p	Cumulative gas production
n	Exponent
q_0	Initial production rate
τ	Time-characteristic parameter
Γ	Incomplete gamma function
m	Slope of log-log plot
a	Intercept coefficient
$\hat{n}$	Exponent of time
D_1	Decline constant at initial time
D_∞	Decline constant at infinite time
$\widehat{D}_i$	Initial decline constant
$\hat{q}_i$	Rate 'intercept'
x_i	Individual sample point of independent variable
y_i	Individual sample point of dependent variable
s	Standard deviation
$n-1$	Degree of freedom
β_0	Intercept
β_1	Slope
$\hat{y}$	Predicted dependent variable
t	Time

Chapter 1: Introduction

Since the time when oil and gas has become a significant source of energy supply, a huge amount of works has been dedicated to estimate, predict and investigate the relationship between major parameters contributing to the performance of petroleum reservoirs. After several years of petroleum extraction when the production started to demonstrate declining behavior, back to 1944, a review of the development of decline-curve analysis has been suggested by J. J. Arps. The mathematical correlations between cumulative production, time, production rate and decline parameters have been studied and common sorts of decline curves have been discussed. (J. J. Arps, 1944)

From the end of 1980s, the substantial contribution to energy supply by shale gas has been considerable with the aid of hydraulic fracturing and horizontal wells. Challenges in predicting the shale gas production arise because of complicated fracture networks and complex mechanisms (gas slippage and gas desorption) in shale. Decline Curve Analysis is known for its cons being efficient and simple in the forecast of hydrocarbon production despite the flexibility of several simulation techniques as well as analytical models. Currently, the energy supply by natural gas constitutes a quarter of the total energy consumption around the globe, which is illustrated in the Figure 1 below. (Lei Tan, Lithua Zuo and Binbin Wang, 2018)

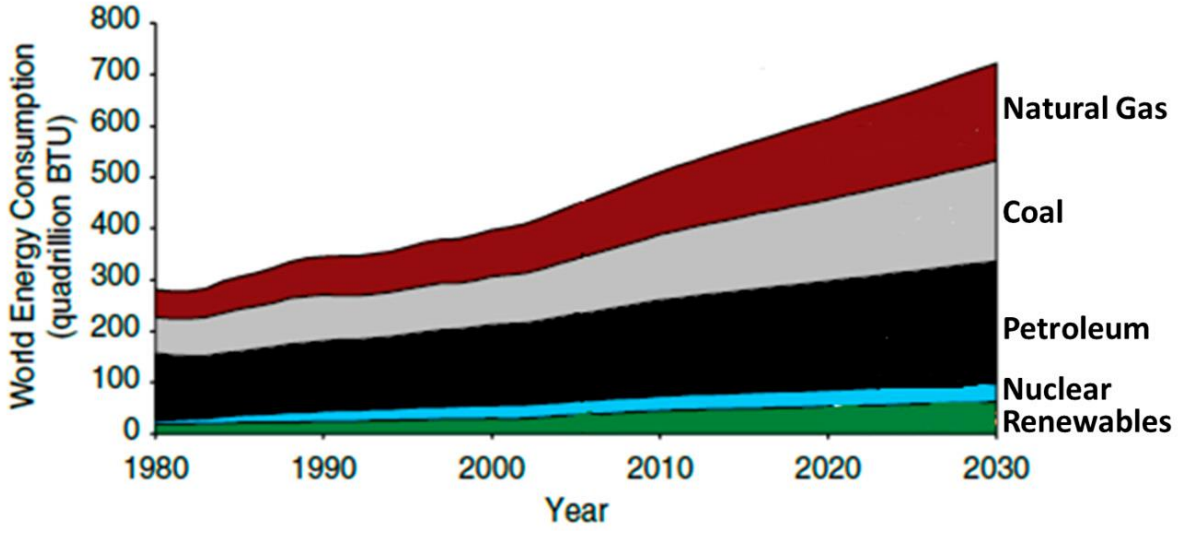


Figure 1. Map of world energy consumption in terms of shares by energy sources. (Lei Tan, Lithua Zuo and Binbin Wang, 2018)

Recently, data-driven models and big data analytics have become significant, specifically, in regard to the analysis of production behaviour in petroleum reservoirs. The acquisition and

management of various large volume data, and utilizing statistical learning methods to explore the data and reveal unseen relationships and associations in complex and huge multivariate datasets are demonstrating the mounting application of big data and data science concepts as new technologies develop from day to day. In order to better understand and optimize performances of unconventional reservoirs, the ultimate purpose is the development of data-driven perceptions in the sphere of oil and gas industry. Nevertheless, due to usage of 'black box' algorithms and statistically heavy vocabulary, the topic of data analytics is remaining indefinite to engineers in the oil and gas industry regardless of its achievements in the well-known spheres such as cyber security, marketing and medicine. (Srikanta Mishra & Luan Lin, 2017)

Petroleum industry is facing different problems and challenges when it comes to dealing with data and its processing. There is a need for appropriate technical analysis of the database in order to improve the performance in oil and gas industry. At this point, machine learning (ML) and artificial intelligence (AI) techniques emerge to tackle such problems showing promising accomplishments and benefits in efficiently providing numerical computations and capability to store high-volume data. To inspire the use of data mining and analytics, supervised and unsupervised learning, AI and other methods, a framework is organised. Stages used in ML are illustrated in the following figure, Figure 2. (A. Sircar, K. Yadav, K. Rayavarapu et al., 2021)

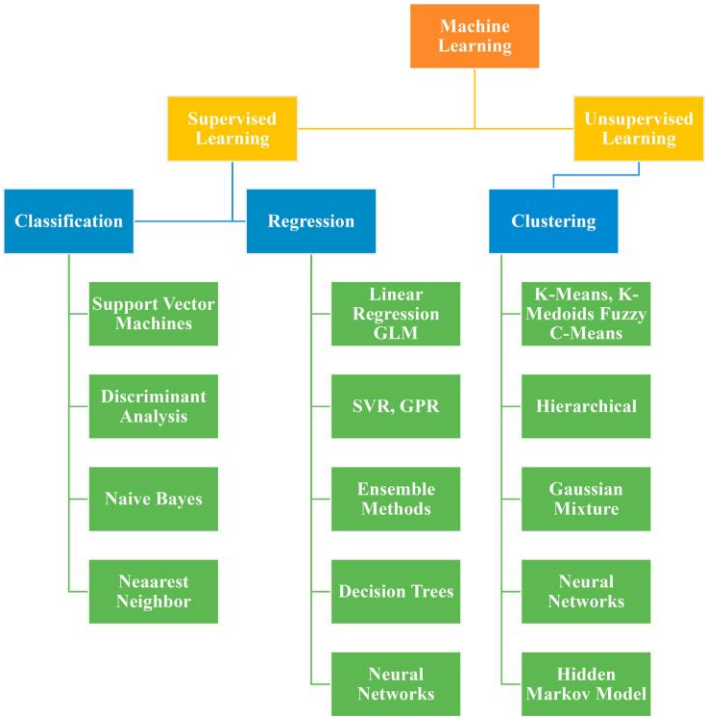


Figure 2. Stages involved in ML. (A. Sircar, K. Yadav, K. Rayavarapu et al., 2021)

Chapter 2: Literature Review

As unconventional reservoirs have a significant role in production of hydrocarbon (HC) in US, in the B. Nelson et al. (2014) work, several decline-curve analysis (DCA) models were discussed and proposed for such reservoirs, specifically analysing Marcellus shale formations. The reliability and applicability of different DCA models, such as Arps, PLE and Duong models, were taken into consideration. This study has brought about a method developed to evaluate DCA model parameters for the prediction of the production relying on the production history. The work has revealed that the Arps model is also applicable and reliable to fit the decline curves. R-squared (R^2) value of 0.98 for Arps model has been accomplished. It has been concluded in the study that Arps hyperbolic type decline curve, which is one of the simplest techniques, has provided consistent prediction results, too. (B. Nelson et al., 2014)

Later, in 2021, a study by Sulaiman A. Alarifi involved an all-inclusive productivity outline and analysis comprising 1216 abandoned wells from different shale formations in the US. The study reports the use of two DCA techniques for the history match of production data utilizing least-squared fitting approach to identify best fit parameters for further predicting the production reliably. By the conduct of the history matching process, the study exposed that matches of high accuracy between two DCA techniques and actual production data have been achieved with a correlation coefficient being equal to 0.99. Two DCA methods used were Arps hyperbolic type decline curve and SEPD model. These DCA models were applied for different early production durations such as 0.25, 0.5, 1 and 2 years having a major target of evaluation of optimal parameters for further predictions and estimation of estimated ultimate recovery (EUR) for those formations like Haynesville, Eagle Ford and etc. The outcome obtained, specifically for 6 months and 1 year was high enough having correlation coefficients varying from 0.85 up to 0.94 for Arps decline method, which demonstrated that hyperbolic decline results are promising in the matching procedure. (Sulaiman A. Alarifi, 2021)

Throughout the literature, for example, in Wilson (2015), several decline-curve analysis models – such as Arps, PLE, Duong, and SEPD – have been studied and contrasted for unconventional reservoir. The conclusions made were towards the idea that most DCA techniques generate great results with regard to history matching up to 4-8 years, however, they differ quite enough when it comes to forecasting. (Sulaiman A. Alarifi, 2021)

As new technologies and methodologies emerge in the recent years, more and more studies and papers are being dedicated to the application of machine learning, especially in the oil and gas

industry. In another study conducted in 2021, Gang Hui and his colleagues have investigated hundreds of wells in an unconventional resource of shale gas in Canada, to analyze shale gas production through machine-learning techniques making the conduct comprehensive adding operational and geological considerations to the analysis. Thirteen operational and geological parameters have been taken into account as input while targeting to estimate 1-year production utilizing four ML methods consisting of linear regression, neural network and two tree-based approaches. The results of the study revealed that parameters mostly influencing the production were mass of total proppant used, permeability, gas saturation, porosity, stage quantity, pressure of formation, horizontal length and etc. The outcome of the machine learning methods has shown quite high results with the highest outcome in case of Extra Trees technique showing R-squared (R^2) value equal to 0.81. (G. Hui et al., 2021)

In addition, in the study Y. Li et al. (2017), different types of machine learning methods have been used to forecast production and to obtain decline curve fit parameters based on production data matching. Neural network (NN) approach has been utilized to investigate the relationships and certain patterns of reservoir and hydraulic fracture characteristics with decline curve ones. The fit of production data obtained through NN method gave high-accuracy results of mean squared estimation and R value equal to 0.013 Mscf/D and 0.92, respectively. (Y. Li & Y. Han, 2017)

In conclusion, there are sufficient number of studies dedicated on the analysis and comparison of different decline-curve analysis models resulting in applicability of Arps DCA model with a hyperbolic exponent (b) being higher than 1 when required. There are a few papers published in the recent years involved in the investigation and prediction of the production data based on data analytics and machine learning techniques providing quite high outcome in accuracy. However, there is still a need for the investigation of the correlation between operational/reservoir parameters and DCA characteristics. As the world technology develops, the application of ML and NN becomes of indisputable significance and studies using such technological developments could make improvements and contribution to the entire science.

2.1 Exploratory Data Analysis (EDA)

In the current world of growing data, it is not quite trivial to manually process the data. To reach even a deeper understanding, data analytics and visualization programs come in to support. The programming language Python having an easy-to-follow syntax becomes a powerful open-source tool as alternative to conventional applications and methods.

Exploratory Data Analysis (EDA) comprises summarizing the data taking into consideration the major characteristics and visualizing them through appropriate representation. A fast description of the dataset in terms of row/column numbers, data type, data missing and preview; as well as visualizing distributions through histograms, bar charts and box plots, correlations and their calculation are performed via EDA. (K. Sahoo et al., 2019)

EDA has been used in several research studies for production prediction and optimization. In the research study of J. Schuetter & S. Mishra et al. (2018), EDA has been mentioned as a standard initial step in exploring the dataset, and EDA has been performed by examining predictors and response variables with a representation of pairwise scatterplots to identify if there is any effect of input variable on the output ones. The scatterplot matrix, shown in the Figure 3, illustrates possible correlations between predictors and response parameters, as well as among input variables themselves, providing empirical histograms for parameters on the diagonal.

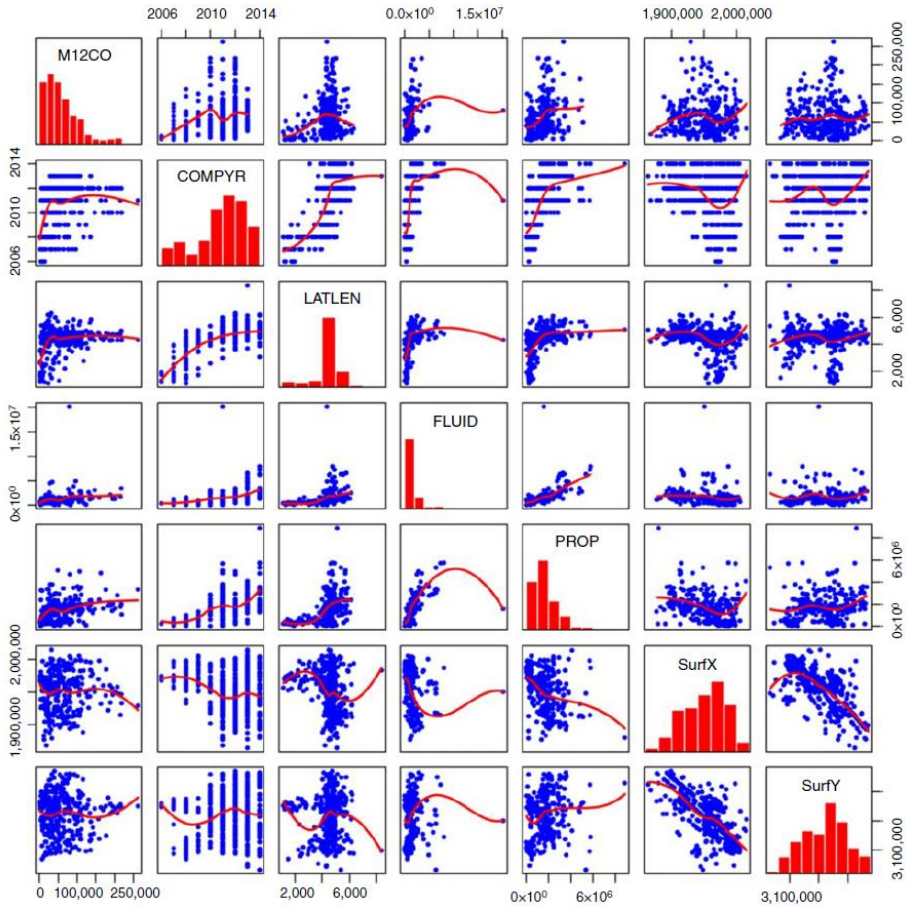


Figure 3. The matrix of scatterplots. (Schuetter & Mishra et al., 2018)

In a similar way, a couple of studies, which are Jebb, A. T., et al. (2016) and K. Sahoo & A. K. Samal et al. (2019), includes performing EDA through the application of univariate and bivariate analysis to visualize and understand the variables separately as well as their relationship by means of box, group, scatter and distribution plots, which is represented in the following figures (Figures 4, 5, and 6).

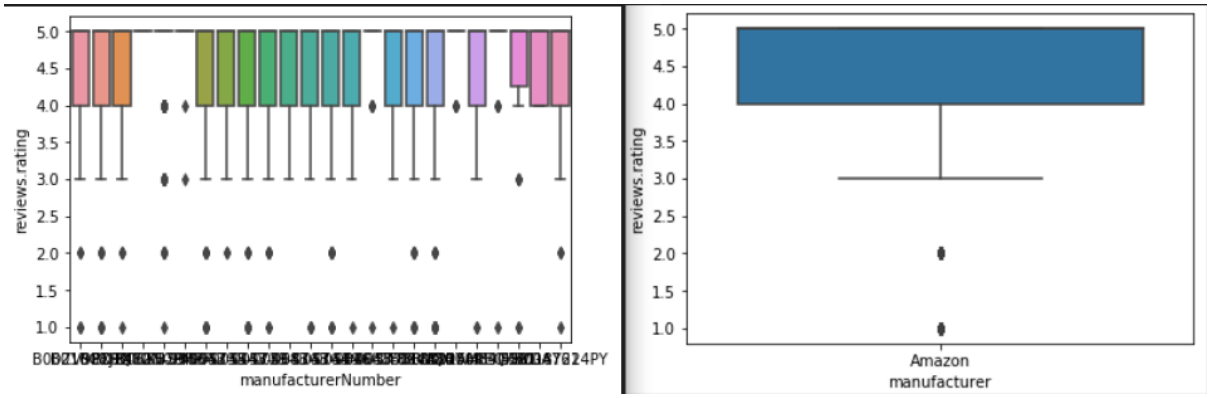


Figure 4. Box plots from EDA. (K. Sahoo & A. K. Samal et al., 2019)

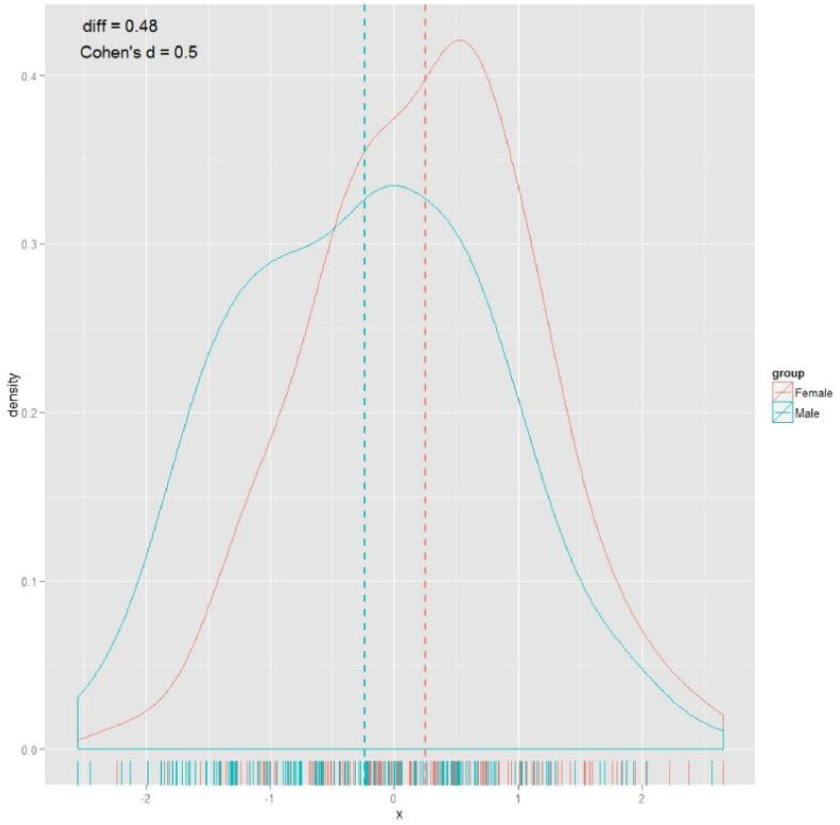


Figure 5. Group-means plot from EDA. (Jebb, A. T., et al., 2016)

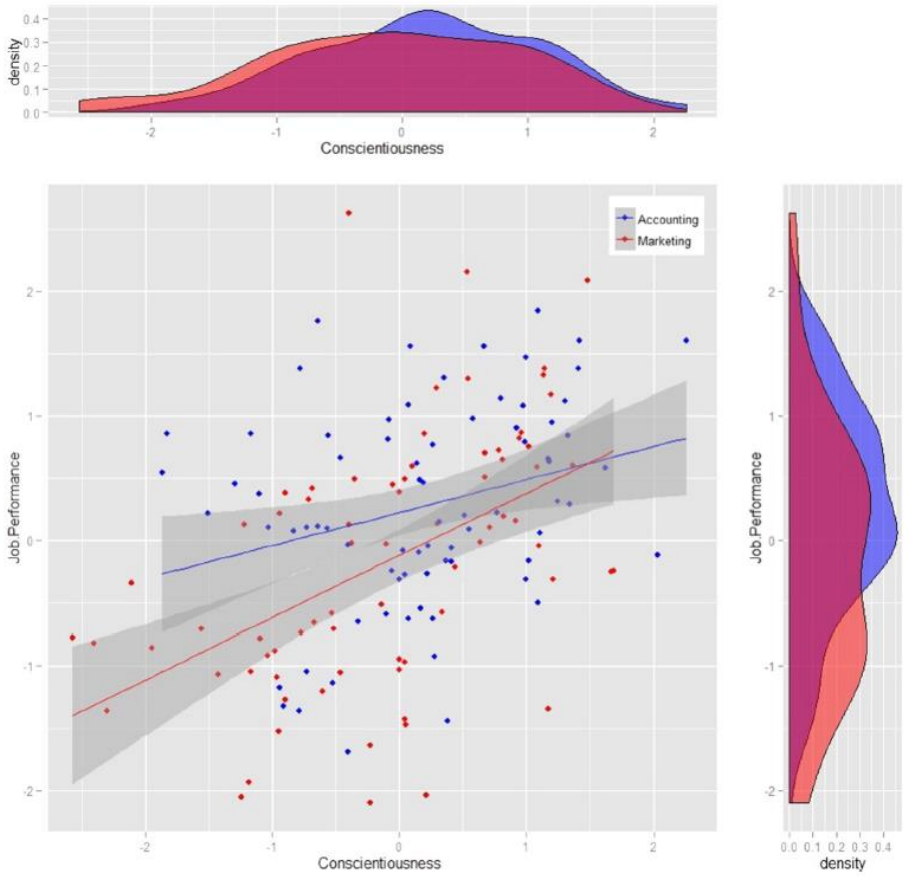


Figure 6. Combined scatter and distribution plots from EDA. (Jebb, A. T., et al., 2016)

Nevertheless, techniques applied in the aforementioned studies are instances of the possible methods, providing graphical representations. Assessing distributions of variables and investigating huge correlation matrices including coefficients are examples of several ways used in EDA procedures. The majority of such methods are discussed and analyzed in the following chapters.

2.2 Predictive Input-Output Modelling

In several studies such as M. Kuhn et al. (2013) and J. Schuetter et al. (2018), predictive input-output modelling is defined to be a development of a math-based tool or a model accomplishing precise predictions.

The major purpose of the applications in oil and gas field moving forward after EDA is the predictive model building. The forecast of cumulative production in unconventional reservoirs using operational parameters has been performed in this work creating particular types of regression models, fitting model parameters in order to estimate how the goodness of the model is assess the ability to accurately predict the future data. Utilizing the available data, it is

possible to forecast well performance in case the model fits well. (S. Mishra & M. Zhong et al., 2015)

In the work of J. Schuetter et al., (2018), a thorough investigation has been conducted to analyze relatively the advantages and cons of predictive modelling techniques. The decision-making process to select the type of model to be used is not obvious and satisfaction by training dataset is required in some models. Some modelling methods with consideration of regression and classification problems such as Ordinary-Least-Squares (OLS) regression, decision trees, Random Forest (RF), Support Vector Regression and Gradient Boosting Machine (SVR and GBM) have been analyzed in this paper and a detailed explanation is given in the following chapters. (J. Schuetter et al., 2018)

Moreover, in the study of S. Mishra & M. Zhong et al. (2015), the distinct forecasting abilities of predictive models have been discussed and the results reveal that the outcome of predictions can be different and sometimes even contradictory based on datasets. The results have shown that tree-based techniques like RF and GBM were less time-requiring in terms of initial processing and less susceptible to the quality of data providing also better results in predictions.

2.3 Variable Importance Analysis and Model Evaluation

The estimation of the goodness of model fit is one of the uppermost aspects to be considered in the model selection procedure and its significance is sometimes disregarded. Plotting the output predicted by the model versus the actual data using scatterplots is a conventional method to assess the goodness-of-fit. A highly acceptable fit to the training data is achieved when all points are situated close to the 45-degree line meaning one-to-one correspondence of the actual values to the predicted ones, which is demonstrated in the Figure 7 (left plot). Nevertheless, this is not necessarily meaning if the model will work well also for the future data because there is a concept of overfitting when the model tries extremely to fit the training data achieving a reduced flexibility for new datasets. The example of overfitting is demonstrated in the Figure 7 (right plot). (S. Mishra & M. Zhong et al., 2015)

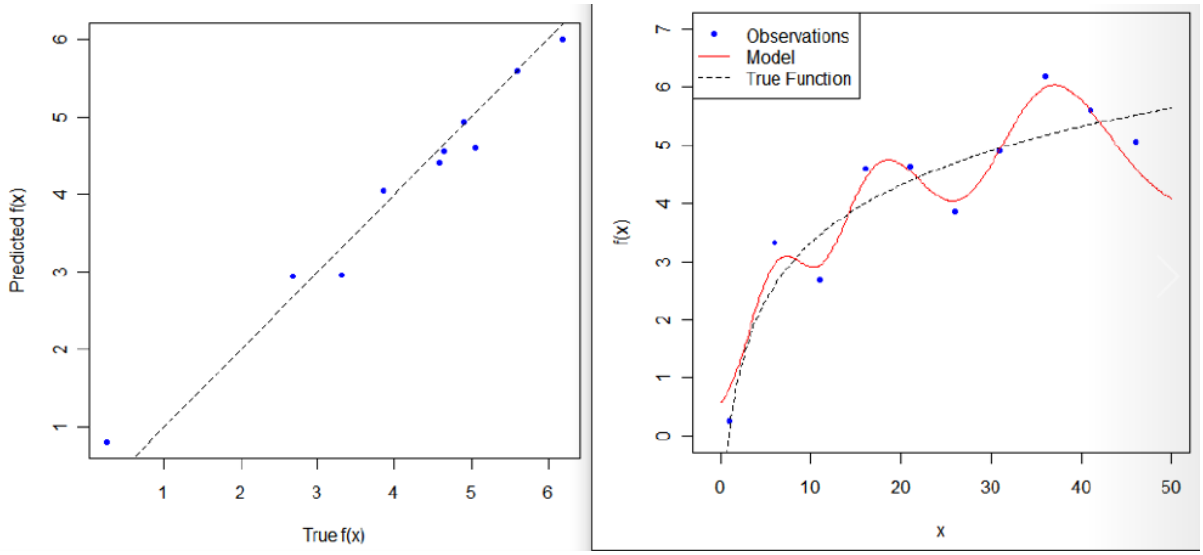


Figure 7. Model evaluation on training data (left plot) and model fit (right plot). (S. Mishra & M. Zhong et al., 2015)

Furthermore, there are several different techniques to evaluate and measure the model fit. Some commonly utilized metrics to quantify the goodness-of-fit have been provided in the study of S. Mishra & L. Lin, (2017) which are Average Absolute Error (AAE), Mean Squared Error (MSE), and R-squared. Such metrics are conceptually similar in assessment of the prediction quality of the model. Evaluation metrics are thoroughly analyzed and the underlying basis is discussed in the methodology chapter. After models have been assessed and selection process, the variables affecting the model and the results are required to be identified through Variable Importance Analysis (VIA).

In the majority of cases, VIA is specific for a model and the expression of corresponding metrics can be relative or absolute. Variable influence is measured relatively in Random Forest models which comprises the method calculating the change in RMSE (Root Mean Square Error) when a variable is introduced provided that the other parameters are kept unchanged. The meaning of this procedure is that it reveals to what extent the model and the results are affected by a predictor when it is removed or added thereby determining the strength of a parameter. (S. Mishra & L. Lin, 2017) (J. Schuetter et al., 2018) (S. Mishra & M. Zhong et al., 2015)

Chapter 3: Problem Statement

With the development of the technology, unconventional resources of petroleum have become increasingly important in the recent years. More and more studies are conducted to analyse significant parameters, to match the production history efficiently and to predict accurately future production using different models. There is considerable amount of research on different decline-curve analysis models and decline curve fitting. However, there is some insufficiency in the number of studies investigating correlation of operational parameters and DCA characteristics. To improve the situation there is a need for additional studies and for identification of main driving parameters.

Several research works have focused on the analyzing and comparison of distinct DCA models such as L. Tan et al. (2018), B. Nelson et al. (2014), S. A. Alarifi (2021) further predicting the production. Nevertheless, these studies do not include use of different machine learning techniques when it comes to the forecast. There are some studies involving the utilization of ML methods to predict the production, for instance G. Hui et al. (2021), Y. Li et al. (2017), S. Mishra et al. (2017), A. Sircar et al. (2021), but they are mainly centered on K-Means Clustering, K-Nearest Neighbor, and Neural Network methods. It is still required to apply regression and tree-based methods to obtain a comprehensive understanding of the situation regarding the application of up-to-date technology to oil and gas industry.

The problems of insufficient studies and issues related with the situation of the research in these fields have become an incredible inspiration in making the decision to conduct new research trying to achieve more understanding of the correlation between operational parameters (such as amount of total proppant used, gas saturation, reservoir temperature, sandface temperature, stages, clusters and etc.) and DCA characteristics (q_i, D_i, b) through use of ML models. In the following sections, the full list of those parameters has been indicated. Moreover, the second part of this study incorporates the use of same machine learning methods to forecast the estimated ultimate recovery (EUR) after six months and one year. According to our knowledge, such analysis of aforementioned type using shale gas wells would be one of first studies.

3.1 Research Question and Objectives

Investigation of the relationship between operational parameters are decline-curve characteristics, and to forecast cumulative production based on shale gas wells from different formations are the major goals of this study. In order to perform these objectives, the first duty is to fit decline curves to production history of all 53 wells both manually and through Excel SOLVER software, and obtain Arps hyperbolic decline-curve parameters which are Hyperbolic Exponent (b), Nominal Decline Rate (D_i) and flow rate (q_i). Afterwards, two different ML methods (Multiple Linear Regression, Random Forest) are applied to analyze the correlation taking operational parameters – indicated in Table 2 in the following section – as input, and predicting DCA constants (q_i, D_i, b) as output.

The aims of study have been listed below in the following way:

- Fitting decline curves for all wells
- Obtaining DCA constants manually and by Excel SOLVER
- Creating a dataframe with operational and DCA parameters
- Performing univariate and multivariate data analysis to reveal any patterns
- Quantification of correlation and creation of correlation matrix between independent and dependent variables
- Forecasting DCA constants using ML models
- Performing Variable Importance Analysis (VIA)
- Creating a new dataframe based on operational parameters and actual cumulative production values
- Forecasting EUR for 0.5 and 1 year duration using ML models
- Performing VIA

Regarding research questions have been emphasized below in the following way:

- Is there a correlation between input and output parameters?
- To what extent is the relationship?
- What is the prediction accuracy of ML models?
- Is the data available sufficient for the analysis conducted?
- Which are the major driving parameters influencing the results?
- Does the addition of different formations into the investigation change the results?
- What is the precision of the EUR forecast?
- What variables affect considerably the prediction outcome?

3.2 Description of the Dataset

The dataset used in this study has been provided by SPE Dataset Repository which includes spreadsheets with data from 53 shale gas wells collected from 5 different formations collected from unconventional reservoirs in the United States. Five different formations are Eagle Ford, Haynesville Shale, Bossier Shale, Marcellus, and Marcellus-Upper. The number of wells in the dataset corresponding to each formation is given in the Table 1 below. Well data, deviation survey, production data, calculated data and essential graphs have been provided in each spreadsheet of wells in the dataset.

Formations	**Number of wells in the dataset**
Eagle Ford	11
Haynesville Shale	14
Bossier Shale	1
Marcellus	11
Marcellus-Upper	16

Table 1. Number of wells provided in the dataset for each formation

There are several parameters listed in the well data of each well from which 25 relevant characteristics have been selected to create the main dataframe. These variables are given in the following Table 2.

	Selected relevant parameters	**Units of measurement**
1	Lease	-
2	Well Number	-
3	State	-
4	Formation / Reservoir	-
5	Initial Pressure Estimate	psi
6	Reservoir Temperature	degree F
7	Net Pay	ft
8	Porosity	-
9	Water Saturation	-
10	Oil Saturation	-
11	Gas Saturation	-

12	Gas Specific Gravity	-
13	CO_2	N/A
14	N_2	N/A
15	TVD	ft
16	Spacing	N/A
17	Number of Stages	-
18	Number of Clusters	-
19	Number of Clusters per Stage	-
20	Amount of Total Proppant	lbs
21	Lateral Length	ft
22	Top Perforation	ft
23	Bottom Perforation	ft
24	Sandface Temperature	degree F
25	Static Wellhead Temperature	degree F

Table 2. List of selected relevant parameters from well data

Moreover, in the calculated data sheets of spreadsheets, corresponding gas volume, gas production and time have been provided in the dataset.

3.3 Overall Workflow Plan

A general summary of the procedure utilized to accomplish the research goals and to find answers to questions raised is illustrated in the Figure 8 provided below. The procedure is highlighted considering the framework of data analytics based on statistical and machine learning techniques.

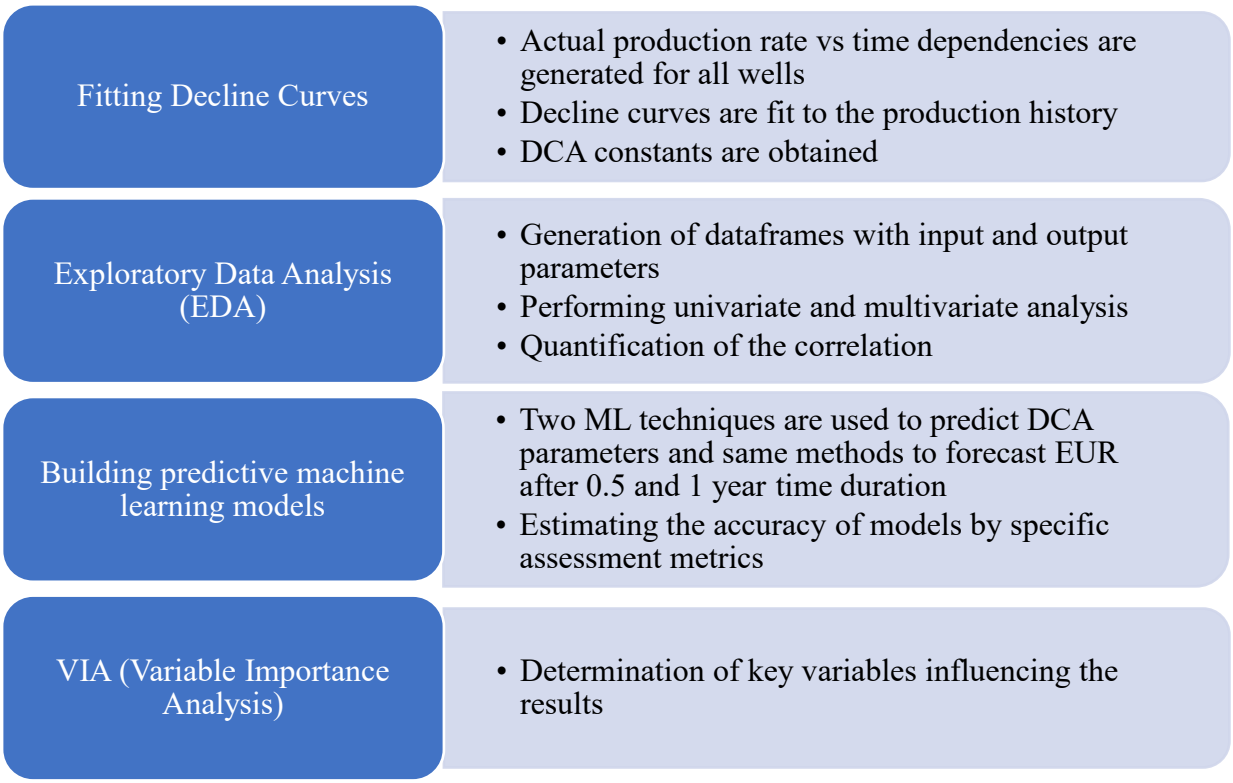


Figure 8. Brief summary of the procedure followed throughout the study.

Chapter 4: Methodology

4.1 Approach of the Methodology

The goals in the current research study are to investigate the relationship between operational and DCA parameters, to predict cumulative gas production after some time duration, and to identify key variables for each model built for unconventional reservoirs. The software programs utilized in this analysis are Excel, Excel SOLVER machine and Python. The methodology considered to find answers to the research problems is based on data analytics and statistical learning. The data analytics cycle starting from data and ending up with evaluation and visualization has been provided in the Figure 9.

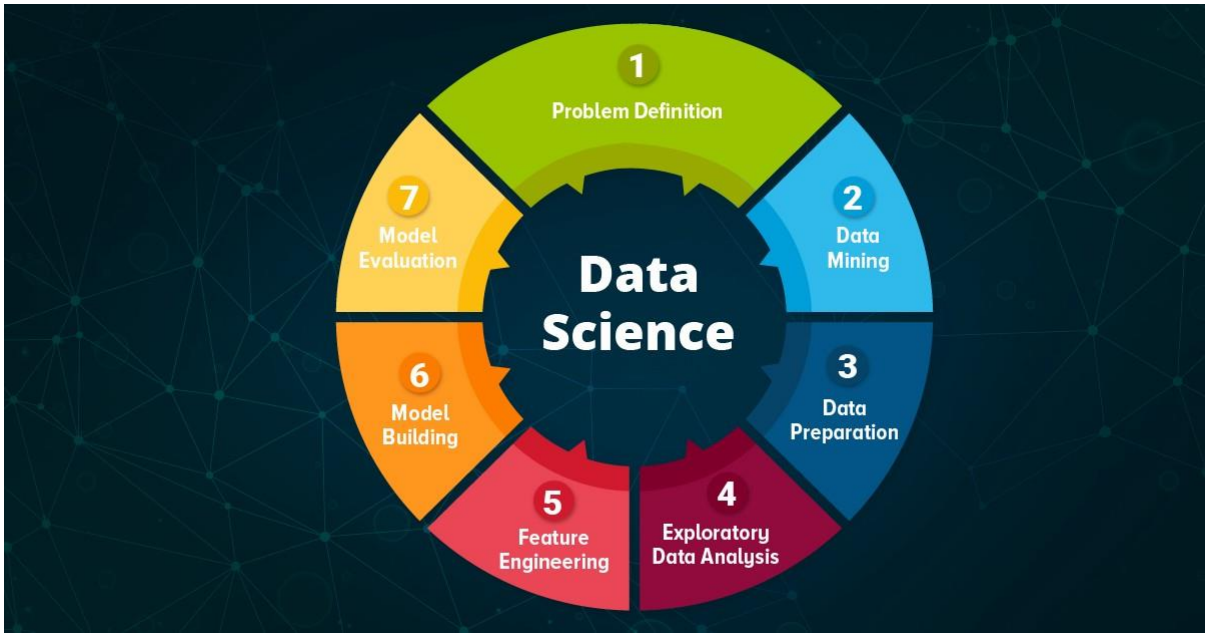


Figure 9. Data Analytics cycle. (I. Tanriverdi, 2021)

4.2 Decline Curve Analysis (DCA)

A graphical method utilized for the analysis of production rate declines and prediction of future performance of oil and gas wells is called Decline Curve Analysis (DCA). Reservoir pressure decrease, change in volumes of produced hydrocarbons are some of causes for the decline of oil and gas production rates. To fit a line to historical production performance constitutes the basis of the DCA. Since its introduction in the 1940s, decline curve analysis has been one of the most used approaches for estimating the production potential of wells for the future. The decline features of wells in a field may be identified and extrapolated to estimate oil and gas

reserves. Based on current trends, the approach is simple and is presently utilized to assess the future potential of oil and gas wells in both conventional and unconventional reservoirs. Oil and gas wells demonstrate a distinct diminishing pattern in rates when a reservoir is drained throughout production, which may be extrapolated for the future and examined to get important information. (A. Satter & G. M. Iqbal, 2015)

4.2.1 Advantages of DCA

The pros of DCA are listed in the following way:

- Based on the simple but yet powerful approach of empirical modelling. History matching and extrapolation are performed using graphical methods.
- Being an intuitive and rapid technique in predicting future production and ultimate recovery. Less time-consuming as in some instances, it is possible to conduct analysis of hundreds of wells in a short time duration.
- To accurately forecast future performance, DCA can include application of multiple modelling in distinct phases of production.
- Recognizing different flow regimes in complicated geological settings, for example, tight shale with natural and induced fractures has become one of recent advances in DCA.
- Cash-flow analysis for a well or field is eased as predictions for annual and monthly production are available.
- Possibility to apply not only for an individual well but also for the field as an aggregate trend. With an inclusion of all producing wells, the ultimate recovery can be computed for the entire field.
- It is also possible to predict water cut for a well based on the trend.
- Flexibility of the analysis makes it feasible to conduct further analysis in case of unanticipated trends.
- Not resource intensive as reservoir simulation. In a relatively short period of time, the analysis can be performed through use of available software. (A. Satter & G. M. Iqbal, 2015)

4.2.2 Assumptions in Traditional DCA

Traditional DCA application to conventional reservoirs has several assumptions as listed below:

- The well is produced by depletion drive mechanism. Production rate may decline in an unidentifiable manner due to water/gas injection, water flux from aquifer, or presence of gas cap.
- The flow regime concerned is BDF (boundary-dominated flow).
- The well is producing from its own drainage area without interfering with others.
- The well is producing at constant BHP (bottom-hole pressure). This condition may not be observable in reality. (A. Satter & G. M. Iqbal, 2015)

4.2.3 Limitations

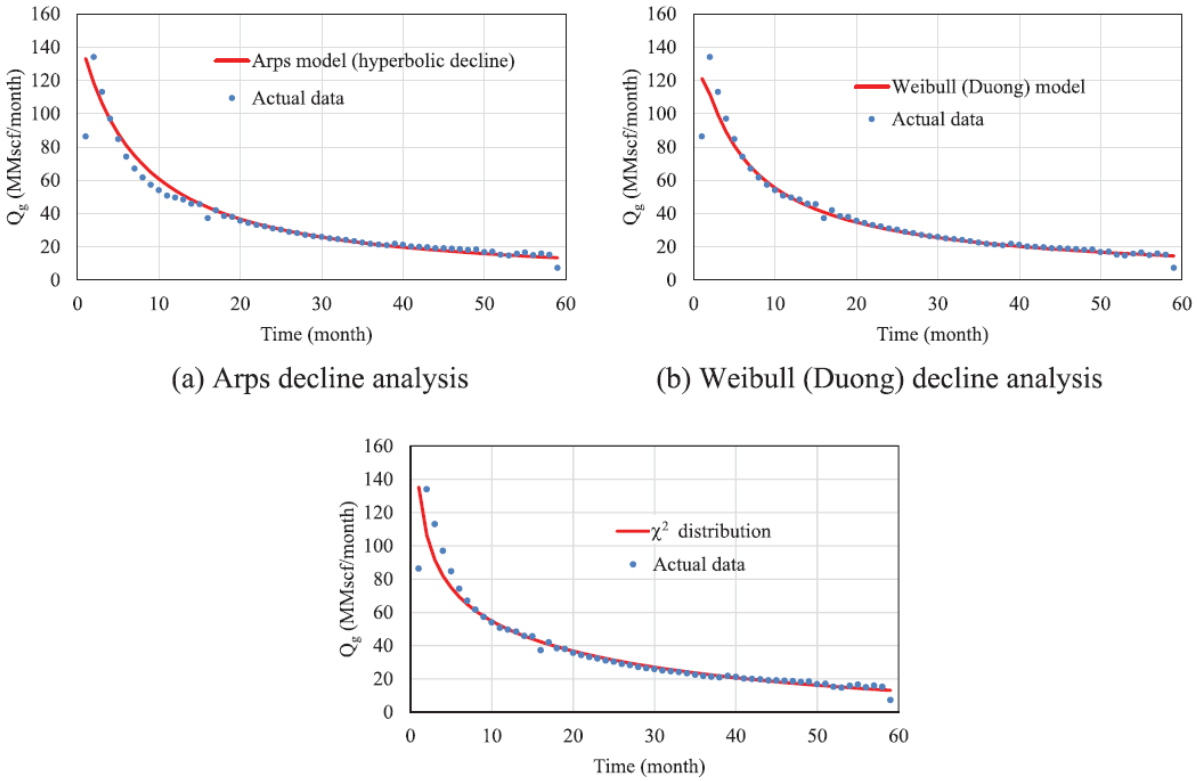
The technique is applicable in case a recognizable trend of declining production rate can be observed despite being a simple and straightforward method. Enough well rate data (from several months to a year) is required for the analysis to confidently forecast future performance. However, for some instances, the trend may not be identifiable because of fluid injection to reservoir for pressure maintenance processes, two-phase flow, hydraulic fracturing, stimulation, water breakthrough and etc. (A. Satter & G. M. Iqbal, 2015)

The emergence of unconventional resources, such as shale gas reservoirs, has revealed that classic decline curve analysis is insufficient for estimating ultimate recovery or reserves. The fluid flow properties of shale gas can differ significantly from those of conventional gas production. Shale has extremely low permeability, and production occurs through a vast and complicated network of induced and natural fractures. It is critical to detect the presence of distinct flow regimes (linear, transient, boundary dominated) over the productive life of the well. The decline pattern of wells produced from shale formations, as usually noted, alters substantially after the first phase of production. Extrapolating the features of the first decline to the economic limit of the well in the future may result in overestimating or underestimating the final recovery. (A. Satter & G. M. Iqbal, 2015)

4.2.4 Review of Different DCA Models

According to past performance, DCA models forecast future rates and are empirical models. Through determination of one or more unknown constants in the formula, a best fit is required by the model using mathematical and graphical methods. Well-known DCA model types are exponential, harmonic and hyperbolic decline curves. These models are jointly called Arps

model where exponential and harmonic decline curves are the specific cases of the hyperbolic model (b=0 is exponential, b=1 is harmonic). Unconventional reservoir analysis has become a reason for the modification of traditional DCA models in recent years. Some new models such as SEDM/SEPD (Stretched Exponential Decline Model), Duong, Modified Hyperbolic Decline, PLE (Power Law Exponential) have obtained a popularity in the application of decline curves for shale gas reservoirs. The examples of different decline curve model applications have been provided in the following Figure 10. (M. Meyet et al., 2013) (A. Satter & G. M. Iqbal, 2015)



(a) Arps decline analysis

(b) Weibull (Duong) decline analysis

(c) χ^2 distribution decline analysis

Figure 10. DCA using different types of decline curve models. (Y. Yuan et al., 2020)

Arps Decline Model

Based on Arps (1945), traditional DCA models are often mentioned as Arps model. The general formulae relating production rate and time with decline rate is expressed in the following way:

$$D = kq^b = -\frac{1}{q}\frac{\mathrm{d}q}{\mathrm{d}t}$$

where t is production time period, q is production rate of well, D is instantaneous decline rate, k and b are constants based on decline characteristics. Traditional DCA is divided into 3 types

depending on the b value in a way such that b=0 shows exponential decline, b=1 is defined as harmonic decline and others values of b (0<b<1) correspond to hyperbolic decline which is the generalized form. It is worth to mention that in the modern DCA when dealing with unconventional reservoirs, the hyperbolic DCA also include values of b greater than 1. Overestimation and underestimation of reserves may be observed when using classic hyperbolic decline with b greater than 1, however, according to several research studies mentioned in the previous chapters, the quality of fitting decline curves is affected very little by that. (L. Tan et al., 2018) (A. Satter & G. M. Iqbal, 2015)

SEDM/SEPD Model

Valko (2009) has suggested SEDM model to evade the arbitrariness related with estimations for long time periods in modified hyperbolic model. The relationship of production rate versus time is given in the following equation:

$$q = q_0 \exp\left[-\left(\frac{t}{\tau}\right)^n\right]$$

where q_0 = parameter of initial production rate, q is production rate varying with time, n and τ are exponent and time-characteristic parameters. (S. Mishra, 2012)

The cumulative gas production (G_p) is obtained through integration and expressed as follows:

$$G_p = \frac{q_0\tau}{n}\left\{\Gamma\left[\frac{1}{n}\right] - \Gamma\left[\frac{1}{n}, \left(\frac{t}{\tau}\right)^n\right]\right\}$$

where Γ is incomplete gamma function (M. Abramowitz & I. A. Stegun, 1972) (S. Mishra, 2012)

An example dependency plot of SEDM has been provided in the Figure 11 below.

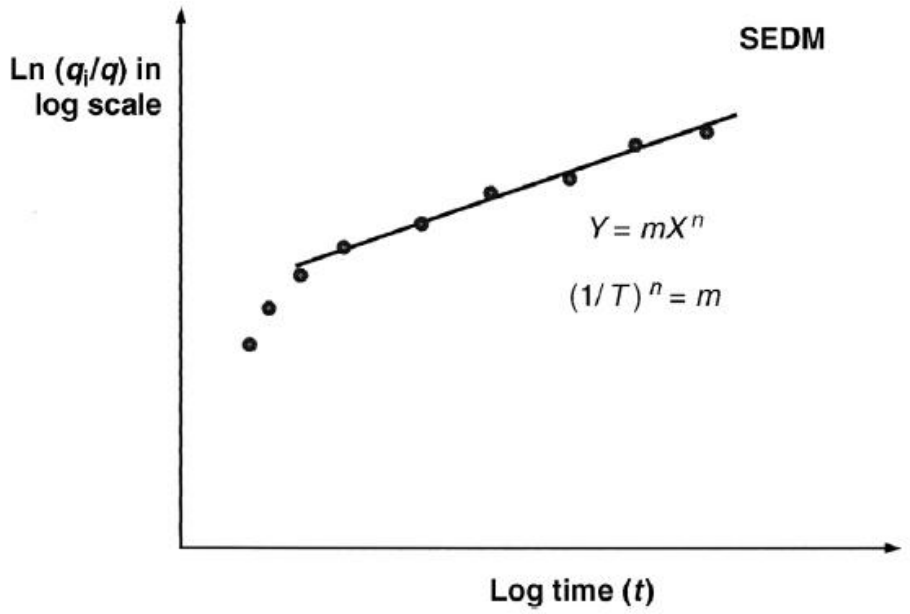


Figure 11. Plot of SEDM/SEPD model. (A. Satter & G. M. Iqbal, 2015)

Duong Model

Based on an empirical equation according to which the log-log plot of q/G_p versus t is a straight line, the Duong model is expressed by the formula below provided by the author Duong (2011):

$$\frac{q}{G_p} = at^{-m}$$

where m is slope of log-log plot and a is intercept coefficient. The derived equations of cumulative production and production rate have been given below, respectively.

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

$$q = q_i t^{-m} e^{\frac{a}{1-m}(t^{1-m}-1)}$$

According to Lee (2012), a good fit to the data from field is also achieved using the Duong model and can be a worthy substitute to the Arps hyperbolic decline model. (L. Tan et al., 2018) (S. Mishra, 2012)

The plot of dependence in Duong model is given in the following Figure 12.

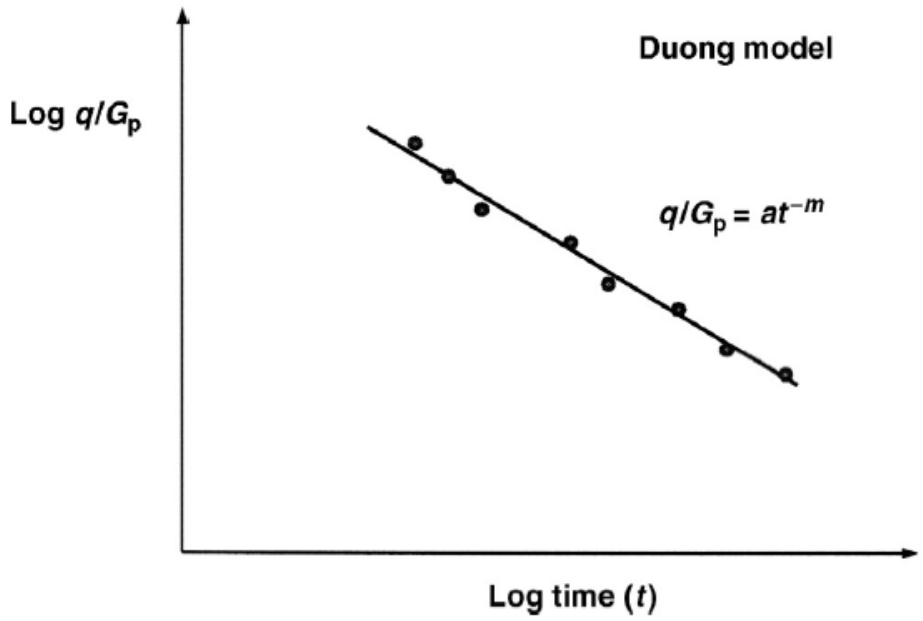


Figure 12. Plot of Duong model. (A. Satter & G. M. Iqbal, 2015)

PLE Model

A different approach has been proposed by Ilk et al. (2008) to express b parameter equation of which is provided as follows:

$$b = D_\infty + D_1 t^{-(1-\hat{n})}$$

where $\hat{n}$ is exponent of time, D_1 and D_∞ are decline constants at initial and infinite time. The derived equation of the production rate is represented below:

$$q(t) = \hat{q}_i e^{[-D_\infty t - \hat{D}_i t^{\hat{n}}]}$$

where $\hat{D}_i$ is initial decline constant, and $\hat{q}_i$ is rate 'intercept'. The relation between decline constants and the exponent of time is as follows:

$$\hat{D}_i = \frac{D_1}{\hat{n}}$$

This model shown here is referred as PLE model, uses power law approximation and bases on Arps decline model. This model has also wide applications in shale gas wells as Duong model. (L. Tan et al., 2018)

4.2.5 Fitting Decline Curves

Decline-curve analysis is the preferred method to use in the investigation of declining behavior of the production and its forecast mainly due to simplicity, effectiveness and requirement of much less data compared to other methods. First of all, to fit decline curves there is a need to calculate actual gas production rate to generate its dependency versus time. In order to do so, gas volume produced and time values are used which were provided in each spreadsheet. After actual gas production rates for all wells have been calculated, predicted production rates based on Arps hyperbolic decline curve model are also calculated using the equation 1 provided below.

$$q = q_i(1 + bD_i t)^{-\frac{1}{b}} \quad (1)$$

In the equation 1, q is the predicted gas production rate, q_i is one of DCA characteristics selected for the better matching, b is the hyperbolic exponent (second DCA parameter), and D_i is the nominal decline rate (third DCA parameter).

Actual and predicted gas production rates versus time graphs have been plotted considering that the predicted rate involves the interval of time when an apparent decline is observed in the actual data based on the DCA rules. Afterwards, absolute squared errors (ASE) for each time value are calculated based on actual and predicted production rates using the equation 2 provided below.

$$ASE = (q_{predicted} - q_{actual})^2 \quad (2)$$

In the equation 2, ASE is the absolute squared error calculated for time value, $q_{predicted}$ is the predicted gas production rate, and q_{actual} is the actual gas production rate. Then, decline curve is fitted to the production history, firstly, using Excel SOLVER software. This software allows to obtain a relatively desired decline curve based on the constraints set. The major limitation for this software is selected to be sum of squared errors (SSE) or similarly called residual sum of squares (RSS). The sum of squared errors is calculated based on the equation 3 given below, which is also the sum of absolute squared errors obtained from equation 2.

$$SSE\ or\ RSS = \sum_i^n (q_{predicted,i} - q_{actual,i})^2 \quad (3)$$

In the equation 3, RSS is the residual sum of squares, $q_{predicted,i}$ is the predicted production rate for each time value, and $q_{actual,i}$ is the actual production rate for each time value. The objective of the Excel SOLVER is set to minimize the RSS using decline curve parameters. However,

the match obtained from the software is not the desired one, so to achieve a reasonably better match, slight manual adjustments are made. After obtaining the desired decline curve fit, the corresponding DCA constants are saved for the upcoming step. In the following Figure 13, the qualitative Excel SOLVER environment with objective setting window is provided.

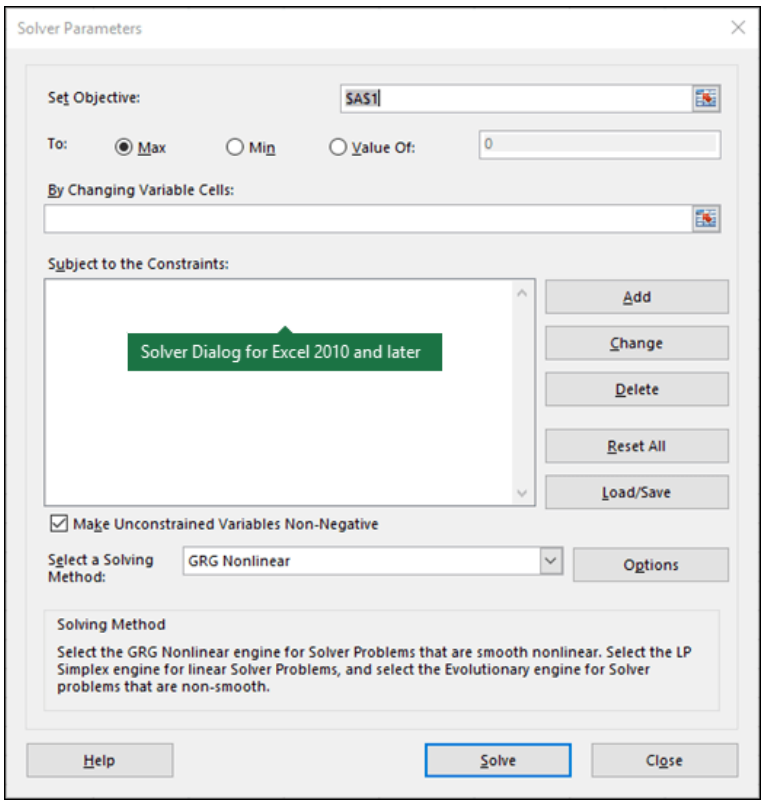


Figure 13. Excel SOLVER environment.

4.3 EDA (Exploratory Data Analysis)

EDA is, principally, a technique to observe the communication of the data apart from testing hypothesis and formal modelling. Meaningly, EDA is the analysis to investigate what exactly the data can tell us. The statistical properties of datasets – these are analyzed by EDA – are summarised through a focus on 4 core aspects such as central tendency measurements (the mode, mean and median), spread measurements (variance and standard deviation), distribution shapes of the dataset, and outlier presence. The key concepts of the EDA have been described in the following sections. Data visualisation and data analysis are being broadly utilized at each phase of the ML process. In the stage of data exploration, the properties, size and contents of the dataset cab be known. Any missing part of the data and correlations among the data can be

found. The use of data in tabular form and understanding the properties constitute the data visualisation. (W. L. Martinez et al., 2017)

The graphical equivalent to the traditional numerical EDA is the graphical EDA which are basically same as both are used in analysis of datasets to overall obtain the statistical properties in terms of 4 main characteristics: the distribution shape, measurements of central tendency and spread, and the presence of outliers. The categorization of graphical EDA includes univariate, bivariate and multivariate analysis. (K. Sahoo & A. K. Samal et al., 2019)

After decline curve parameters have been obtained for all wells using decline curve fitting, a dataset including input parameters mentioned while describing the dataset in the previous chapters, and 3 variables (q_i, D_i, and b) has been built for EDA.

4.3.1 Univariate Analysis

Univariate analysis in EDA represents some sort of a summary of the raw dataset and considers only 1 variable at a time. Cumulative and probability density distributions, box plots, histograms and violin plots are instances of univariate type of analysis in EDA. Some of such plots are discussed in this section. (W. L. Martinez et al., 2017) (K. Sahoo & A. K. Samal et al., 2019)

Histograms

Histogram is one of the means of describing and graphically summarising the dataset through visual transmission of its distribution utilizing vertical bars. Histograms can be used in analysis of massive datasets due to the ease of creating them and being computationally practicable. In the case of a histogram, the whole range of values of a parameter is divided into several intervals and they are used for the continuous data. The representation of a histogram as a frequency distribution takes place via rectangles areas of which represent frequencies accordingly, widths of which correspond to the class interval, and heights of which signify the density of appearance in the data. (W. L. Martinez et al., 2017)

Generally, histograms are graphed in a way that:

- It is possible to have empty bins depending on the distribution of the data.
- The quantity of bins is user-dependent.
- Bins should have equal widths.
- In case of an absent empty bin, there should not be empty spaces between bars. (P. Bruce et al., 2020)

An example of a histogram has been provided in the following Figure 14.

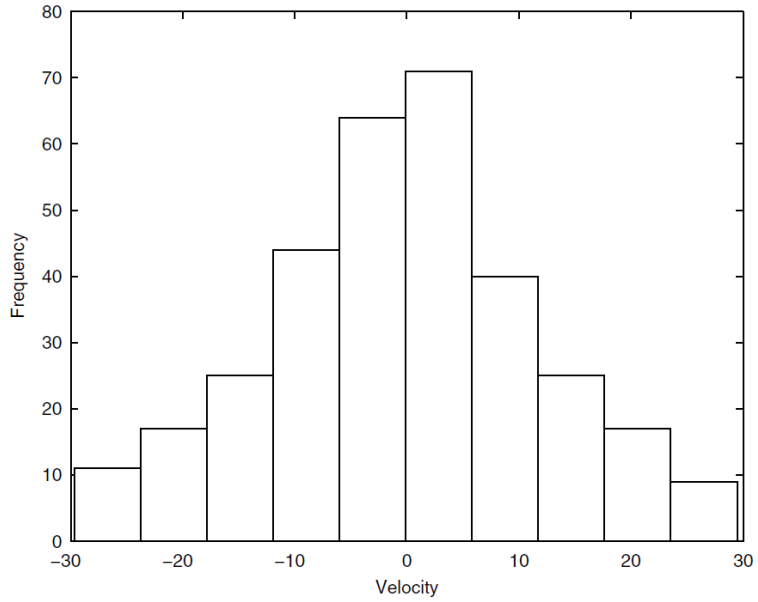


Figure 14. A histogram. (W. L. Martinez et al., 2017)

Boxplots

Through the use of boxplots, a beneficial graphical representation of the data concentration can be obtained. A boxplot reveals the symmetry, central tendency, skew and outlier data. 5 values constitute the main part of the boxplot, which are the first and third quartiles, the minimum and maximum values, and the median. Such values are used to compare the closeness of the data to them. (K. Sahoo & A. K. Samal et al., 2019)

Boxplots are also called as box and whisker diagrams and used for several years, being an effective way of visualization of the statistics summary, studying the distributions, and supplying multivariate representations with univariate information. Some properties and characteristics of boxplots have been highlighted by Benjamini (1988) in the following way:

1. Potential outlier data about the observations can be displayed by boxplots.
2. Possible alongside display of several boxplots for a better comparison of different datasets.
3. Statistics defining the data are presented in a style that gives information about the sample's skewness, spread, location, and longtailedness.
4. Ease of understanding and interpretation.
5. Ease of construct and display. (P. Bruce et al., 2020) (W. L. Martinez et al., 2017)

An example of a boxplot has been provided in the Figure 15 below.

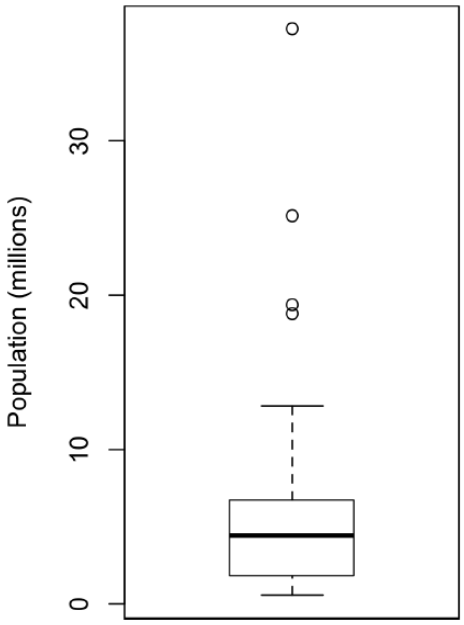


Figure 15. A boxplot. (P. Bruce et al., 2020)

A detailed representation of a boxplot with components is given in the following Figure 16.

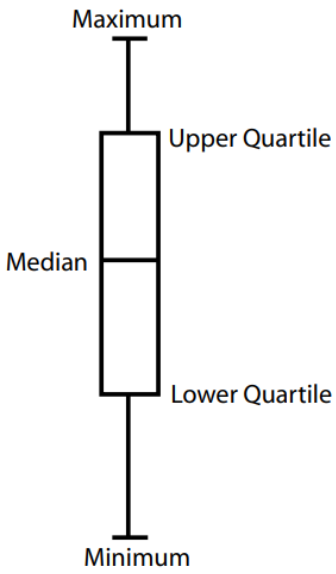


Figure 16. A boxplot with components. (K. Potter, 2006)

4.3.2 Bivariate Analysis

In order to understand the relationship between 2 parameters or between one variable and the major goal parameter and to analyse the correlation, bivariate type of EDA is being used. There

are several ways to display such an analysis. Scatterplots, boxplots, violin plots, scatterplots with distributions are some of the tools to do so. (K. Sahoo & A. K. Samal et al., 2019)

Measure of Correlation

Inspecting the relationship among input variables, and between predictors and the target is included in modelling procedures of EDA. The relationship can be positive and negative depending the correlation of high and low values. The correlation coefficient, which is also called as Pearson's correlation coefficient is a measure of to what extent the relationship between associated parameters is. This coefficient takes values in the range of -1 and 1. The equation defining the correlation coefficient has been provided in the following formulae. (P. Bruce et al., 2020)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{(n-1) * s_x * s_y}$$

Where

n is sample size

x_i and y_i are individual sample points with index *i*

$\bar{x}$ and $\bar{y}$ are means of samples expressed as: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

s_x and s_y are standard deviations expressed as: $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$; $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

n – 1 is the degree of freedom.

Other forms of correlation coefficients, such as Spearman's rho and Kendall's tau, were proposed by statisticians a long time ago. These correlation coefficients are based on data rank. These estimates are resilient to outliers and can handle certain forms of nonlinearities because they deal with ranks instead of values. For exploratory research, data scientists should typically adhere to Pearson's correlation coefficient and its robust alternatives. Rank-based estimates are particularly appealing for smaller datasets and specialized hypothesis testing. (P. Bruce et al., 2020)

Scatterplots

A scatterplot is the kind of graph in which the values of two variables in the dataset are illustrated in the Cartesian coordinate system. It is a very common way to visualize the correlation of two parameters with measured data. The data is demonstrated as aggregations of points which are the records. The scatterplot can be drawn through use of x and y axis variable values. (K. Sahoo & A. K. Samal et al., 2019) (P. Bruce et al., 2020)

A scatterplot example has been given in the Figure 17 below.

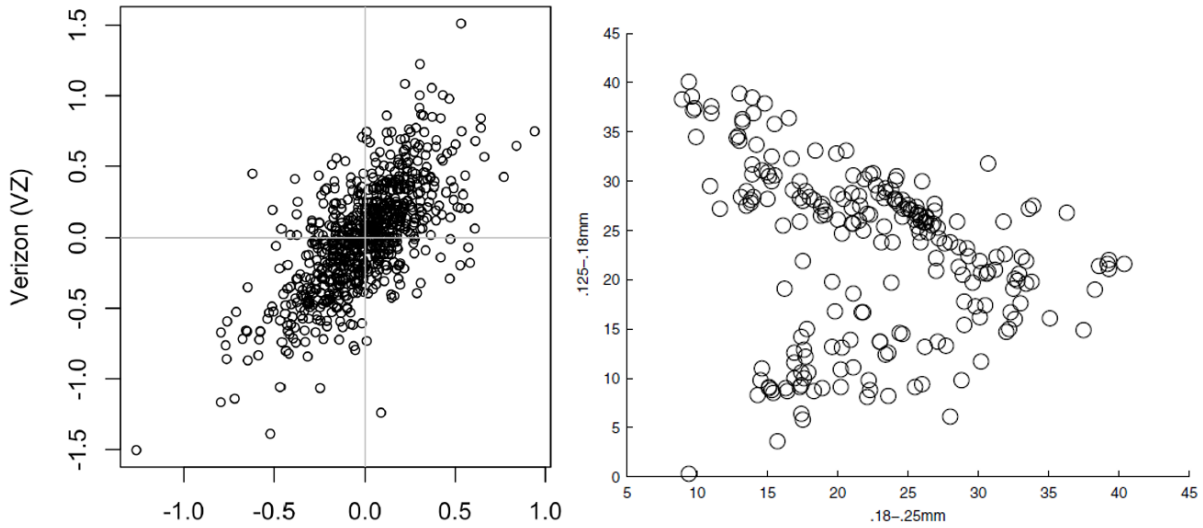


Figure 17. Examples of scatterplots. (P. Bruce et al., 2020) (W. L. Martinez et al., 2017)

Scatterplots with Histograms

In order to provide supplementary understanding and information to the bivariate analysis, there are beneficial functions in the software which add histograms to both x and y axes of the scatterplot allowing to understand the full picture with the combination of bivariate and univariate analysis. (W. L. Martinez et al., 2017)

An example of a scatterplot with marginal histograms is shown in the Figure 18.

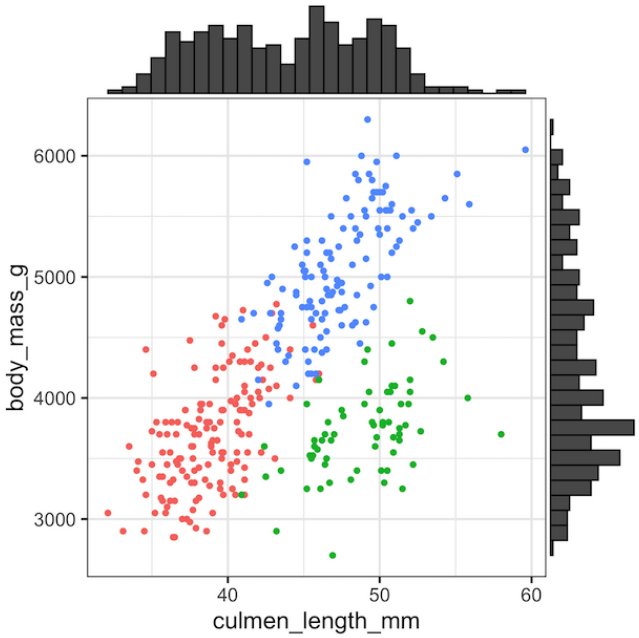


Figure 18. A scatterplot with histograms. (Datavizpyr, 2020)

4.3.3 Multivariate Data Analysis

In order to understand the correlations among several fields of the dataset and identifying the relationships between a greater number of variables, multivariate data analysis is used. Pair plots, 3-D scatterplots, heat maps and 3-D surface plots, correlation and scatterplot matrices are mostly used instances of graphical EDA. (K. Sahoo & A. K. Samal et al., 2019)

Building a correlation matrix demands the calculation of correlation coefficients for all pairs of variables. The correlation matrix is symmetrical relative to the diagonal so it would be sufficient to demonstrate whether left or right part relative to the diagonal. The correlation values on the diagonal should be equal to unity as they represent the relationship of a parameter with itself. (S. Mishra & A. Datta-Gupta, 2018)

Displaying the scatterplots of each variable pairs in a collection of multivariate data having more than two parameters is frequently a great method to start examining the data. Unfortunately, the quantity of scatterplots rapidly gets overwhelming: with 10 variables, for instance, there are 45 graphs to analyze. Arranging the pairwise scatterplots in a square grid, also referred as a scatterplot matrix or a draughtsman's plot, might aid in examining all scatterplots simultaneously. (B. S. Everitt & G. Dunn, 2001)

The examples of scatterplot and correlation matrices have been demonstrated in the following figures, Figure 19 and 20, respectively.

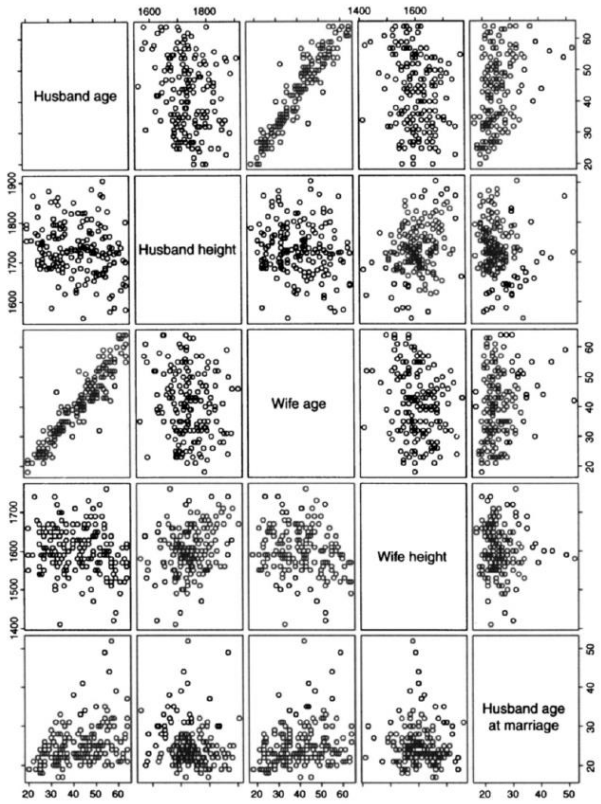


Figure 19. A scatterplot matrix. (B. S. Everitt & G. Dunn, 2001)

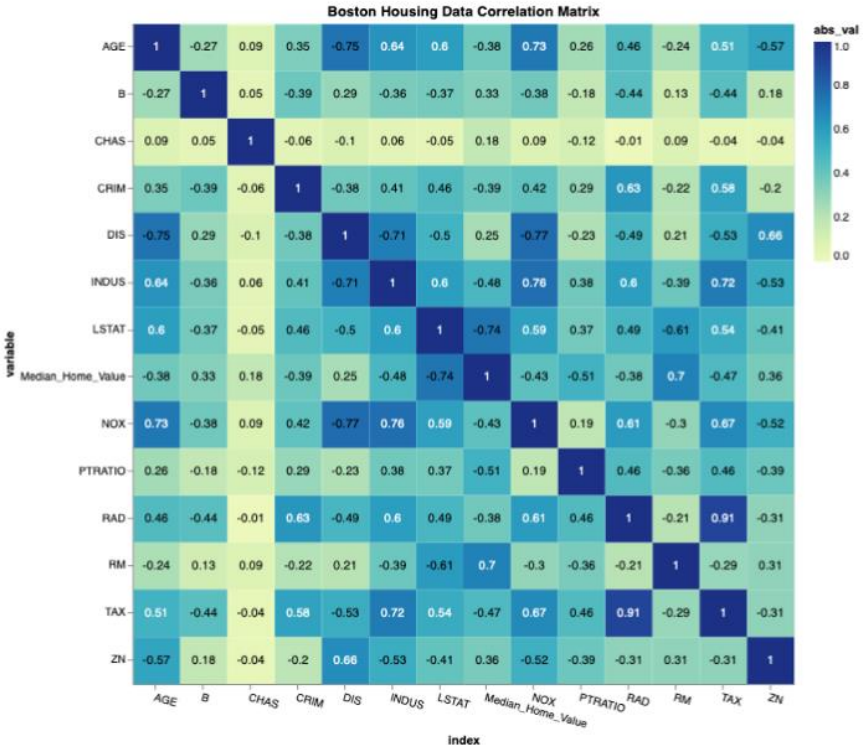


Figure 20. A correlation matrix. (M. Britton, 2020)

4.4 Building Predictive Models

After initial understanding and getting some deeper insights about the data by exploratory data analysis, building predictive models becomes the next significant step in the current study. In order to investigate the correlation between predictor/input and response/output, predictive modelling process is quite important as it also assists in forecasting the output. Such a procedure is performed through the use of machine and statistical learning techniques which this section is mainly focusing on. The application of supervised and unsupervised learning leads to understanding 'what data wants to tell' and extraction of essential trends and patterns, which is referred as 'learning from data'. Generally, statistical learning problems are categorized into two groups of supervised and unsupervised learning. Basing on the input parameters, the forecasting of the response is the target of supervised learning. However, the outcome is missing in the unsupervised learning, and the goal is to analyse and determine the patterns and associations among the predictor values. (T. Hastie et al., 2008)

In this study, the major method utilized is the supervised learning with the application of linear regression and random forest models. Firstly, linear regression and random forest models have been built to analyse the relationship between DCA constants and operational parameters. Random Forest model building involved also the modelling with and without consideration of the formation aspect. Consideration of formation aspect has been performed using dummy variables. For Linear Regression models, normalization of the data has been carried out. When it comes to the second part of the analysis, prediction of the cumulative gas production has been accomplished through building both linear regression and RF models for periods of 0.5 and 1 year.

4.4.1 Linear Regression

Regression modelling is the most widely used one when it comes to the investigation of independent and dependent variables. When the correlation between the predictor and response is defined through linear equations, linear regression is applied. This model includes only one output parameter while the number of input parameters depends on the type of linear regression model. Simple linear regression involves one predictor but the multiple regression – also referred as OLS regression (Ordinary Least Squares) – includes several independent variables. (G. James et al., 2021)

The multiple linear regression has been used in this study due to the plurality of predictors.

4.4.2 Simple Linear Regression

Like its name, simple linear regression approach is quite straightforward as it forecasts quantitative variable of response Y taking predictor X as input. A nearly linear relation between Y and X is assumed, and this relationship is mathematically given in the following equation 4. (G. James et al., 2021)

$$Y \approx \beta_0 + \beta_1 * X \tag{4}$$

where, Y is quantitative output, X is predictor, β_1 is the slope, and β_0 is the intercept. The equation 4, in other words, represents regressing Y onto X (or Y on X). In the equation 4, the two indefinite constants β_1 and β_0 are called as 'coefficients' or 'parameters' of the model. By means of the training data, these coefficients are estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ which are used to predict the estimated $\hat{y}$ using the equation 5 below. (G. James et al., 2021)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x \tag{5}$$

where, $\hat{y}$ is the prediction of Y based on that $X = x$. The hat symbol ^ denotes that the value is estimated for a coefficient, parameter or response. (G. James et al., 2021)

Practically, the model coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are not known, and the data is utilized to compute these coefficients before they can be useful in predictions. Each observation pair contains the measures of X and Y represented as (x_1,y_1), (x_2,y_2), (x_3,y_3), … , (x_n,y_n), where n is the number of observations. The target is to get estimates of coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the data available is fit well by the linear model. Fitting well means the resultant line is quite close to n points of data. Although there are several techniques to measure that 'closeness', the widely used approach is the minimization of least squares which is the method applied in the current study. The best fit is accomplished when RSS is minimized. A sample of fit in simple linear regression has been illustrated in the Figure 21, where grey vertical lines are denoting the residuals. Minimisation of least squares is performed using residuals and RSS defined through the equations 6, 7 and 8. (G. James et al., 2021)

$$e_i = y_i - \hat{y}_i \tag{6}$$

$$RSS = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 \tag{7}$$

$$RSS = \left(y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)\right)^2 + \left(y_2 - (\hat{\beta}_0 + \hat{\beta}_1 x_2)\right)^2 + \cdots + \left(y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)\right)^2 \tag{8}$$

In order to minimize the RSS, such an approach selects $\hat{\beta}_0$ and $\hat{\beta}_1$ which are determined by the equations 9 and 10 provided below. The equations 9 and 10 are also referred as 'least squares model coefficient estimates'. (G. James et al., 2021)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\hat{x})}{\sum_{i=1}^{n}(y_i-\hat{y})^2} \tag{9}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{10}$$

where $\bar{x}$ and $\bar{y}$ are the sample means expressed as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

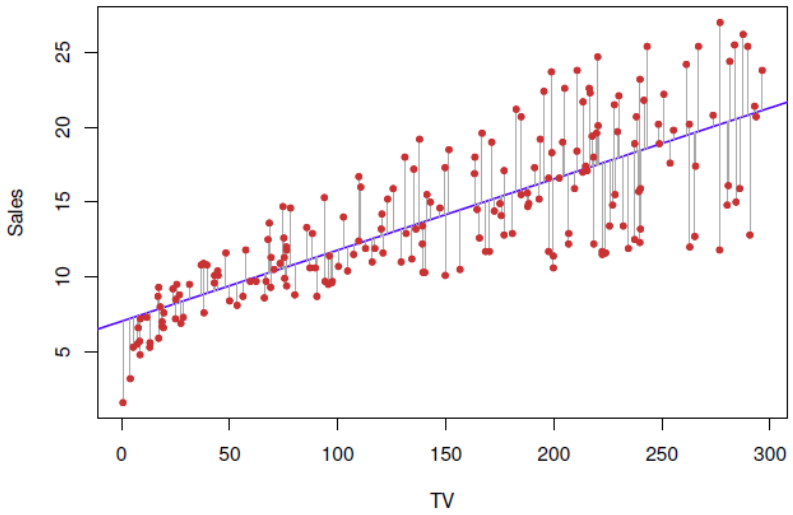


Figure 21. An example of Simple Linear Regression (SLR) fit. (G. James et al., 2021)

4.4.3 Multiple Linear Regression

To predict the response based on one input variable, the technique of simple linear regression is a beneficial approach. On the other hand, there are several more predictors in the practice.

One of the solutions can be building several simple linear regression models to each predictor variable. However, there are several problems which make this approach to be unsatisfactory. When considering separate simple linear regression models, they ignore other predictors, and estimated coefficients become inappropriate. The better method is the extension of the simple linear regression to multiple variables. This is performed by assigning different coefficients to different predictors in a single regression model. Taking the number of predictors equal to *p*, in such a case, the multiple linear regression equation becomes as the following equation 11. (G. James et al., 2021)

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \cdots + \beta_p * X_p + \epsilon \quad (11)$$

where, X_j and β_j are j-th input and the quantitative relation between that input and response, respectively. The interpretation of β_j is the mean influence on Y of a unit rise in X_j, maintaining all other inputs unchanged. Likewise in the simple regression, here the coefficients β_0, β_1, β_2, … , β_p are also unknown and should be estimated, and accordingly, the predictions are made through use of the equation 12 below. (G. James et al., 2021)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x_1 + \hat{\beta}_2 * x_2 + \hat{\beta}_3 * x_3 + \cdots + \hat{\beta}_p * x_p \quad (12)$$

The same least squares approach like in the simple linear regression is applied, and the model coefficients that minimize the RSS are chosen. The corresponding formulae of RSS has been provided in the equation 13. (G. James et al., 2021)

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \cdots + \hat{\beta}_p x_{ip}))^2 \quad (13)$$

The multiple linear regression fit with 2 predictors and one output variable represents a plane in 3-D which is chosen to minimize the RSS (vertical distance between red points and the plane). The sample of multiple linear regression fit is shown in the Figure 22. (G. James et al., 2021)

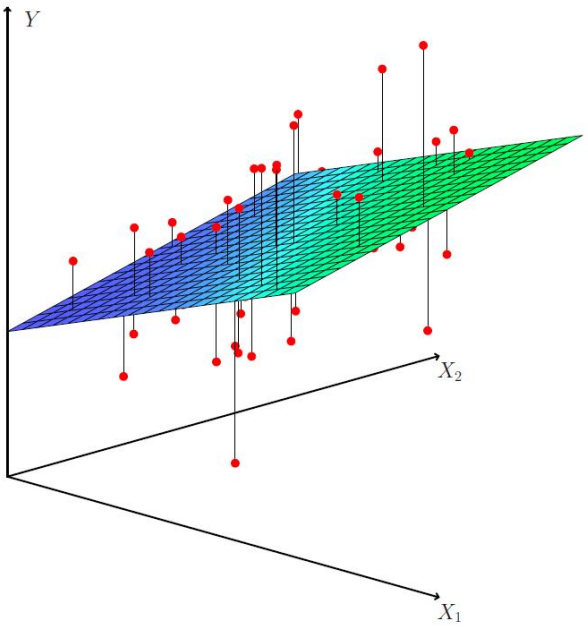


Figure 21. A sample of Multiple Linear Regression (MLR) fit. (G. James et al., 2021)

4.4.4 Ensemble Method of Random Forest

Random Forest is one of the tree-based methods used in ML and statistical learning. An ensemble approach is a method of combining a lot of 'building block' models to get one very powerful potential model. An RF model is also one of supervised ML algorithms having the main basis of decision trees. (G. James et al., 2021)

Random forest regression employs a "bagging" strategy. The model is a collection of basic regression trees, each with a number of splits based on predictor values. Based on a comparison of a given independent variable to a threshold value, each split indicates if an observation must take the right or left branch of the tree. The regression forecast is contained in the trees' terminal nodes, known as leaves. Each tree of the ensemble is being trained utilizing the training data bootstrap sample, and a random subgroup of the input variables is examined for each split in random forests. Because of this randomization, each regression tree can concentrate on subtly different parts of the predictor-response connection. In combination, the trees may transform the data into a strong tool for the prediction. (S. Mishra & M. Zhong et al., 2015)

The RF model structure has been demonstrated in the following Figure 22.

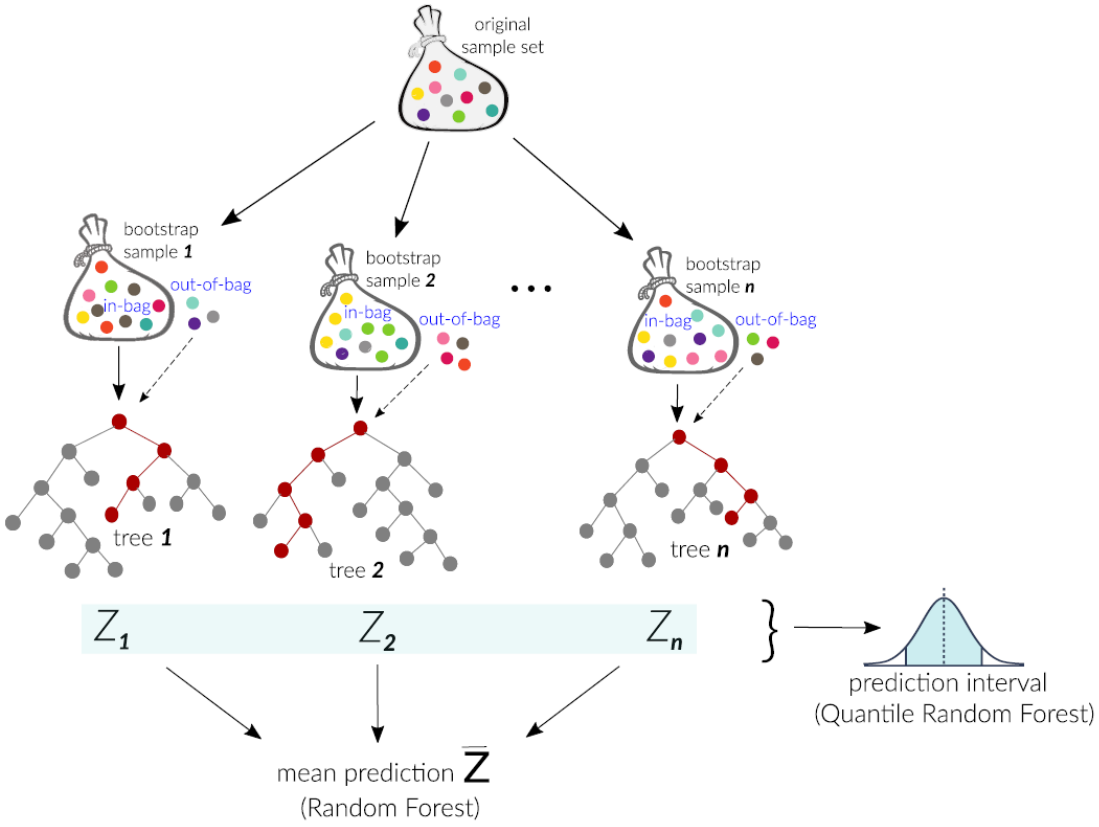


Figure 22. Structure of Random Forest model. (C. Carranza et al., 2020)

4.5 Metrics for Evaluation

In general, there are three commonly used evaluation metrics to determine the accuracy of the model, which are MAE (Mean Absolute Error), MSE (Mean Squared Error) and R-squared (R^2). MSE is the mean size of residuals, meaningly, MAE is the magnitude representing how different in average the actual and predicted outputs are. (S. Mishra et al., 2017)

The mathematical formula for MAE has been provided in the following equation 14.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{14}$$

where, $\hat{y}_i$ and y_i are predicted and actual response values, respectively.

The concept of MSE is similar to MAE, however, MSE is the mean squared difference between the actual observations and their predictions. Therefore, the units of MAE are matching the units of parameters but in case of MSE, the units are squared. The formulae for MSE calculation is given in the equation 15 below. (S. Mishra et al., 2017)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

In order to ease the interpretation of MSE, another often used metric is Root Mean Square Error (RMSE) which is easily computed by taking the square root of MSE, and has matching units with the variables in concern. The parameters mentioned above are desired to be as less as possible to obtain a good accuracy. (S. Mishra & A. Datta-Gupta, 2018)

Another important metric is R^2 which is the measure comparing residual sum of squared errors to the total sum of squared errors. The R-squared metric also provided a better understanding of the accuracy of the model because it is a proportion. It shows the variance proportion that can be explained by the model. R^2 ranges in the interval of 0 and 1, meaning that it does not depend on the scale of the *Y* variable. The better accuracy of the model is achieved when R^2 is closer to 1. The equation 16 given below shows the mathematical definition of the R-squared. (G. James et al., 2021)

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})} \tag{16}$$

where TSS is the total sum of squares defined as $TSS = \sum(y_i - \bar{y})$.

A popular strategy for model validation is to employ an independent test set in the form of wholly fresh data or a 'hold-out' section of the training dataset. In both circumstances, the

model may be fitted using the training portion of the dataset (usually 70-90% of the data) and afterwards evaluated on the independent test values (namely, the remaining 10-30% of the data) to determine the model's predictive ability for new data. K-fold cross-validation is another and a more robust option for model validation. (S. Mishra et al., 2017)

4.6 Variable Importance Analysis (VIA)

The absence of clear functional correlation between predictor and response variables as in case of complicated data driven models, to identify major input and output dependence with trivial evaluation is challenging. VIA is generally model-specific representing the uniqueness of algorithms in the model building. Different measures are being used for different models, for example, R^2-loss, Gini importance and etc. The 'R^2-loss' technique is a relatively straightforward method to determine key variables which is not stuck to any specific model. 'R^2-loss' is effective for any regression model and the reasoning behind this approach is to check how the model accuracy changes when one variable is removed. If a key parameter is absent then the accuracy should be reduced considerably. On the other hand, the removal of a relatively unimportant variable should lead to significant influence on the model accuracy. Measuring variable importance involves computation of pseudo-R^2 for the initial model with all inputs and for the one without a predictor being in check. The difference between the two models is the R^2-loss. The predictor having a greater impact, results in a bigger R^2-loss. (S. Mishra et al., 2017)

Example plots of variable importance have been given in the Figure 23.

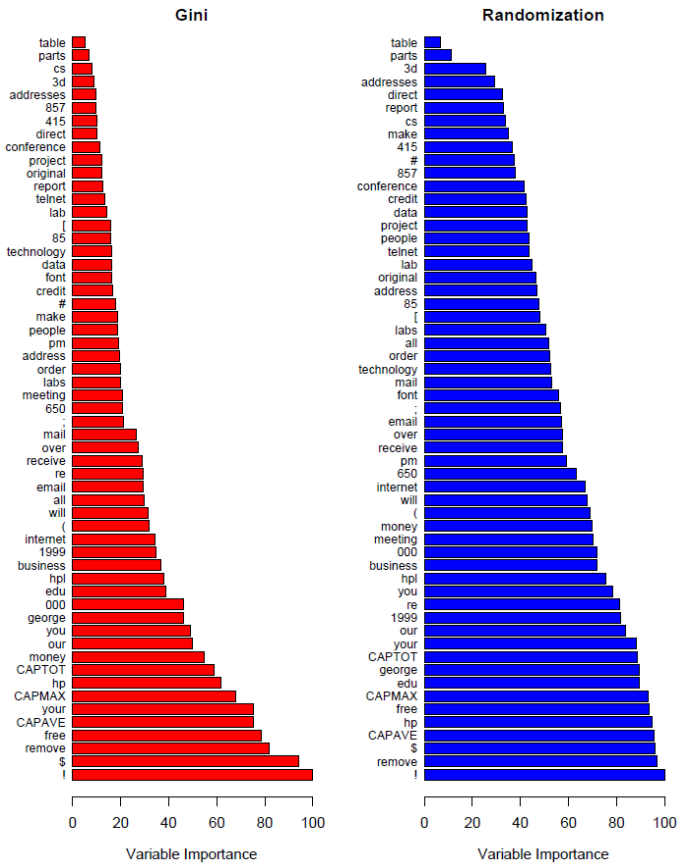


Figure 23. Examples of VIA plots. (T. Hastie et al., 2008)

Chapter 5: Results and Discussion

The results section has been divided into two parts. In the first part, the results of DCA, preparation of a dataframe for EDA, results of EDA, predictive modelling concerning DCA characteristics, and the outcome of VIA have been provided. The second part involves the results of the dataframe preparation for predictive modelling, forecasting cumulative gas production after 0.5 and 1 year, results of modelling process, and outcome of the VIA, accordingly.

5.1 Part 1: Fitting Decline Curves

In order to fit decline curves to the gas production history, there is a need for some calculation. Based on the data available in Excel spreadsheets and Arps hyperbolic decline concept (equation 1), actual and predicted production rates have been calculated for all 53 wells. The representative sample of the results for the well #1 has been provided in the Table 3 below.

Time (Days)	Gas Volume	**Predicted Prod. Rate**	**Actual Prod. Rate**	**ASE (Absolute Squared Error)**
Days	*MMscf*	*MMscf/d*	*MMscf/d*	*(MMscf/d)^2*
1	0.145	0.3086	0.145	0.0268
2	0.186	0.2980	0.186	0.0126
3	0.231	0.2882	0.231	0.0033
4	0.268	0.2791	0.268	0.0001
5	0.261	0.2706	0.261	0.0001
6	0.329	0.2627	0.329	0.0044
…	…	…	…	…
1088	0.000	0.0124	0.000	0.0002
1089	0.000	0.0124	0.000	0.0002
1090	0.000	0.0124	0.000	0.0002

Table 3. Some available and calculated data for the well #1.

In the Table 3, the headings in bold are the calculated data, and others are some of available data from the given dataset. In order to obtain the predicted production rates and therefore ASE, reasonable initial guess for decline curve constants (q_i, D_i, b) have been made. After using Excel SOLVER provided with initial guess values and an objective to minimize RSS, the machine gives a decent fit with the actual production rate observations. Based on the behavior of the fitted curve depending on the constants, some slight manual corrections have been made in order to accomplish a good fit. It is worth to note that according to the rules of DCA, the

fitting procedure covers the section or period of apparent decline. After some time periods in all wells, it can be observed from the plots that some performance boosting activities has been carried out. Such interval cannot be considered in DCA. The corresponding results of decline curve constants and SSE (or RSS) for the well #1 are given in the following Table 4.

q_i	D_i	b	SSE
MMscf/d	*1/d*	-	$(MMscf/d)^2$
0.32	0.0371	1.2	0.33

Table 4. DCA constants and RSS of the well #1.

After adjustment of manual corrections to the fitting, the final version of the fit has been fixed. The decline curve fit for the well #1 is demonstrated in the Figure 24 below.

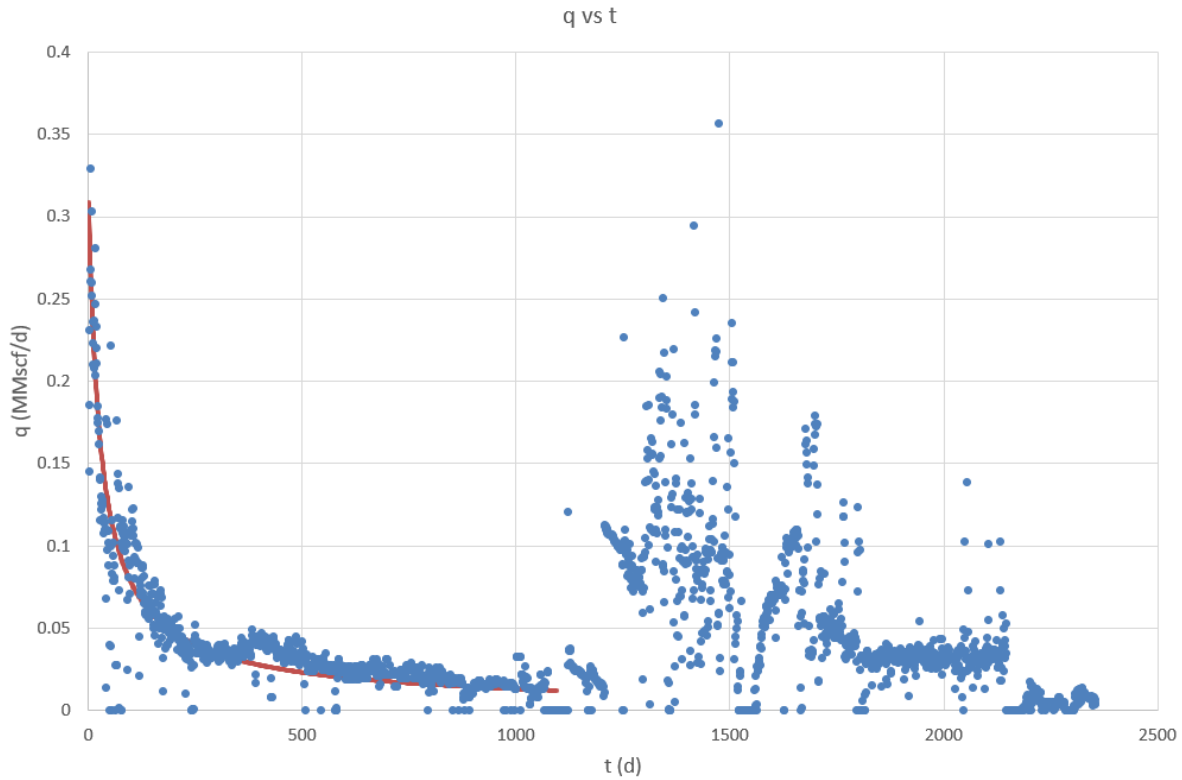


Figure 24. Well #1 production rate versus time plot with decline curve fit.

All the above procedure has been carried out for the whole dataset of 53 wells. The results of DCA characteristics obtained for all wells have been provided in the Table 5. In this table, the numbers of wells are corresponding well numbers provided in the given dataset.

The wells	q_i	D_i	b	SSE
	MMscf/d	*1/d*	-	$(MMscf/d)^2$
#1	0.32	0.0371	1.2	0.33
#3	0.79	0.03	0.1	53.67
#4	0.455	0.01	1.1	0.04
#5	0.25	0.011	0.8	0.002
#6	0.46	0.008	0.9	0.015
#7	0.46	0.009	1.2	0.026
#8	0.61	0.01	0.5	0.003
#9	0.4	0.01	1.0	0.002
#10	0.42	0.011	0.35	0.002
#11	1.1	0.037	0.95	0.079
#2	0.7	0.03	0.3	50.38
#62	11	0.031	1.18	229.9
#63	42	0.005	0.2	169.9
#64	41	0.005	0.15	7090.1
#65	20	0.006	0.55	381.2
#66	30	0.005	0.9	79.2
#67	78	0.008	0.1	151.9
#68	40	0.003	0.1	159.1
#69	180	0.009	0.1	103.5
#70	100	0.005	0.3	38.9
#71	36	0.004	0.1	2202.1
#72	22	0.0032	0.5	345.3
#73	31	0.003	1.2	445.3
#74	24	0.0024	1.1	113.1
#75	31	0.005	1.1	85.5
#76	36	0.005	1.4	870.7
#23	10.8	0.002	0.8	364.6
#24	27	0.0035	0.3	39.5
#25	79	0.013	1.3	2789.2
#26	47	0.0165	1.35	96.2

#27	32	0.033	2.0	40.8
#28	65	0.0067	0.16	12.5
#29	72	0.008	0.47	74.4
#30	43	0.01	1.4	264.3
#31	42	0.01	1.4	57.6
#32	63	0.003	0.46	377.6
#33	142	0.022	0.9	142.1
#34	30	0.011	0.8	65.9
#35	25	0.01	0.72	12.1
#36	38	0.01	1.3	37.5
#37	27	0.0065	0.4	27.1
#38	9.5	0.014	1.4	0.94
#39	7.2	0.0049	0.85	1.2
#40	8.5	0.006	1.1	0.81
#41	24	0.012	1.1	40.7
#42	40	0.022	1.5	24.4
#43	35	0.01	0.7	640.4
#44	40	0.0095	0.55	272.6
#45	62	0.012	0.37	21.9
#46	50	0.029	1.03	51.1
#47	25	0.024	1.15	19.2
#48	25	0.026	1.16	5.4
#49	10	0.011	0.99	6.2

Table 5. Decline curve constants and SSE of all 53 wells.

The plots demonstrating decline curve fits for all wells have been provided in the Appendix section.

It can be observed from the Table 5 that the values for the hyperbolic exponent b ranges from 0.1 up to 2.0, which was consistent with the concepts about the constant b for unconventional reservoirs discussed in the previous sections. It is noteworthy to mention that nominal decline rate requires high accuracy in the fitting process as it takes values of lower order of magnitude from 10^{-2} to 10^{-4}. For some wells, the SSE values obtained from fitting process are high because

of some mid-outliers which are not affecting the DCA in terms of core principles such as obtaining decline characteristics.

5.2 Part 1: Exploratory Data Analysis

Before passing to the EDA, there is a need to prepare a dataframe to analyze the correlations, data distributions, and etc. The dataframe is built based on the given dataset and the results obtained from DCA. The parameters in the dataframe are the variables in the Table 2 with the addition of three DCA parameters in the Table 5. The software being used to carry out the analysis is Python with the interactive computing platform of Jupyter Notebook. 'Amount of Total Proppant' and 'Spacing' columns in the dataframe had very few missing data. These missing points were fixed by taking the average of other values in the dataset, which is a well-known reasonable approach discussed in the previous sections.

5.2.1 Univariate Analysis

Histograms and Boxplots of Response Parameters

The univariate analysis of the response variables (q_i, D_i, b) has been performed through use of histograms, boxplots and barplots. The distribution of output variables has been provided with histograms in the Figure 25. In the Figure 25, the histograms also contain kernel density estimation (KDE) which performs kernel smoothing for density estimation.

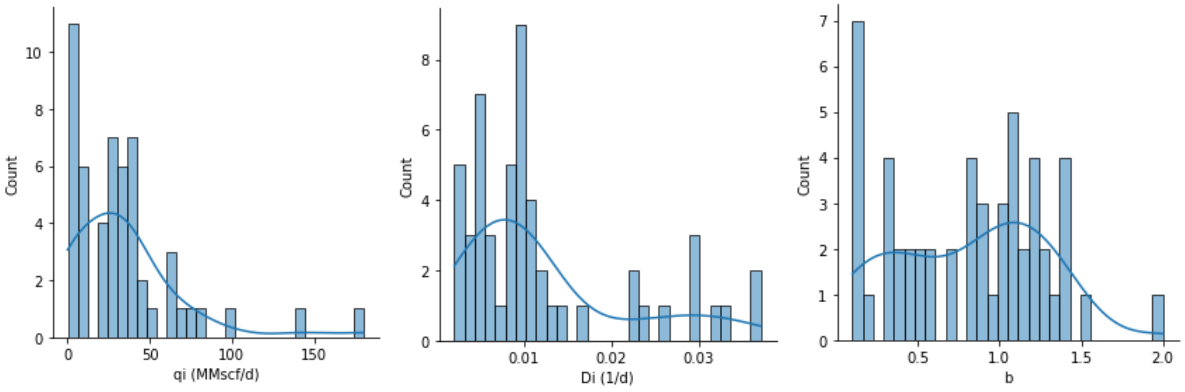


Figure 25. Histograms with KDE of response variables.

It can be observed from the Figure 25 that the distribution of q_i mainly ranges from 0 to 50 MMscf/d, and values of D_i mostly cover the interval from minimum values to nearly 0.017 1/d.

However, the distribution of the hyperbolic exponent varies involving the whole range of values from minimum to maximum. The standard deviations of q_i, D_i and b in the sample dataset are 34.8 MMscf/d, $9.6*10^{-3}$ 1/d and 0.46, respectively.

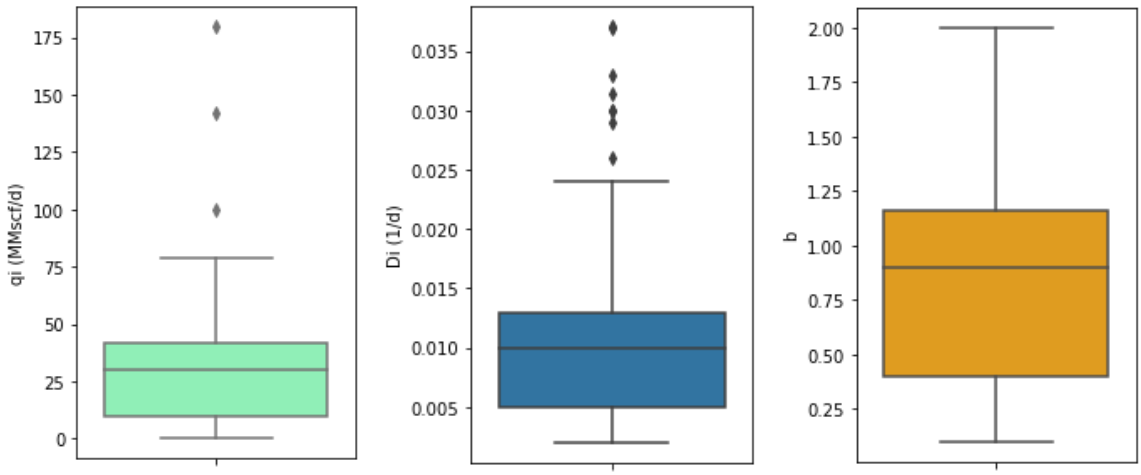
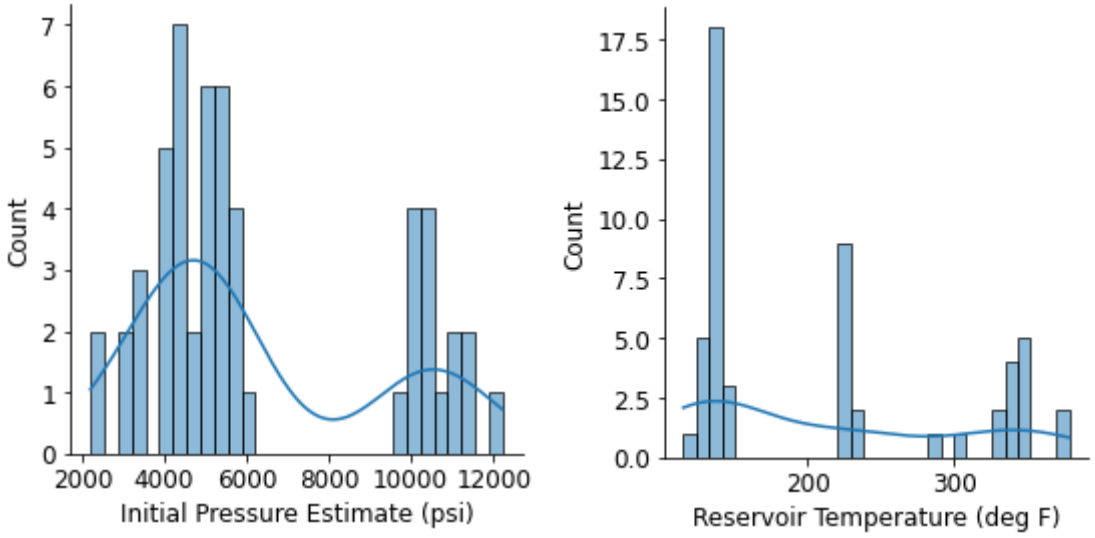


Figure 26. Boxplots for DCA constants.

Same behaviours can be observed also in the boxplot graphs in the Figure 26. There are some outliers expected via boxplots for q_i and D_i values.

Histograms and Boxplots of Input Variables

Histograms of all quantitative independent variables have been provided in the Figure 27 below.



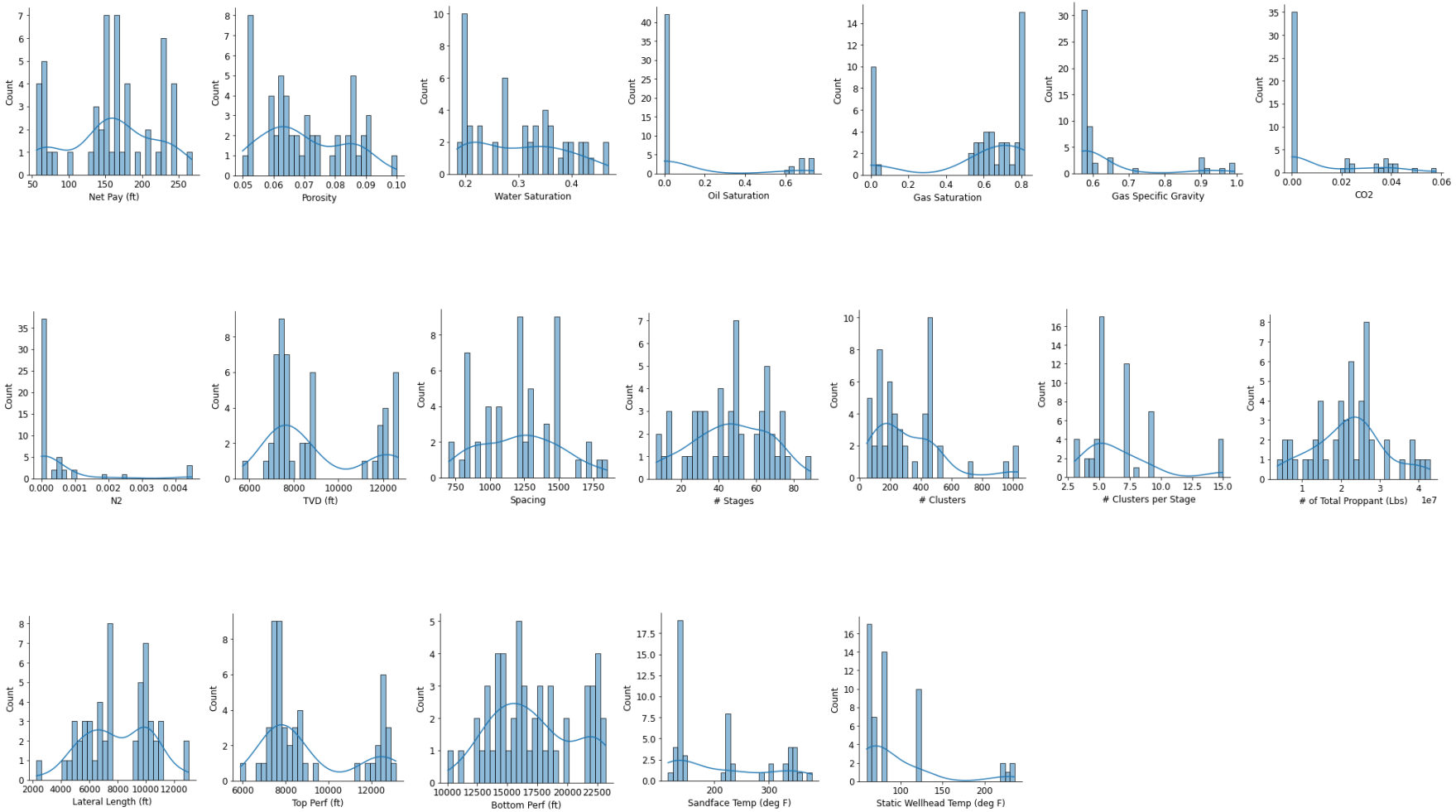
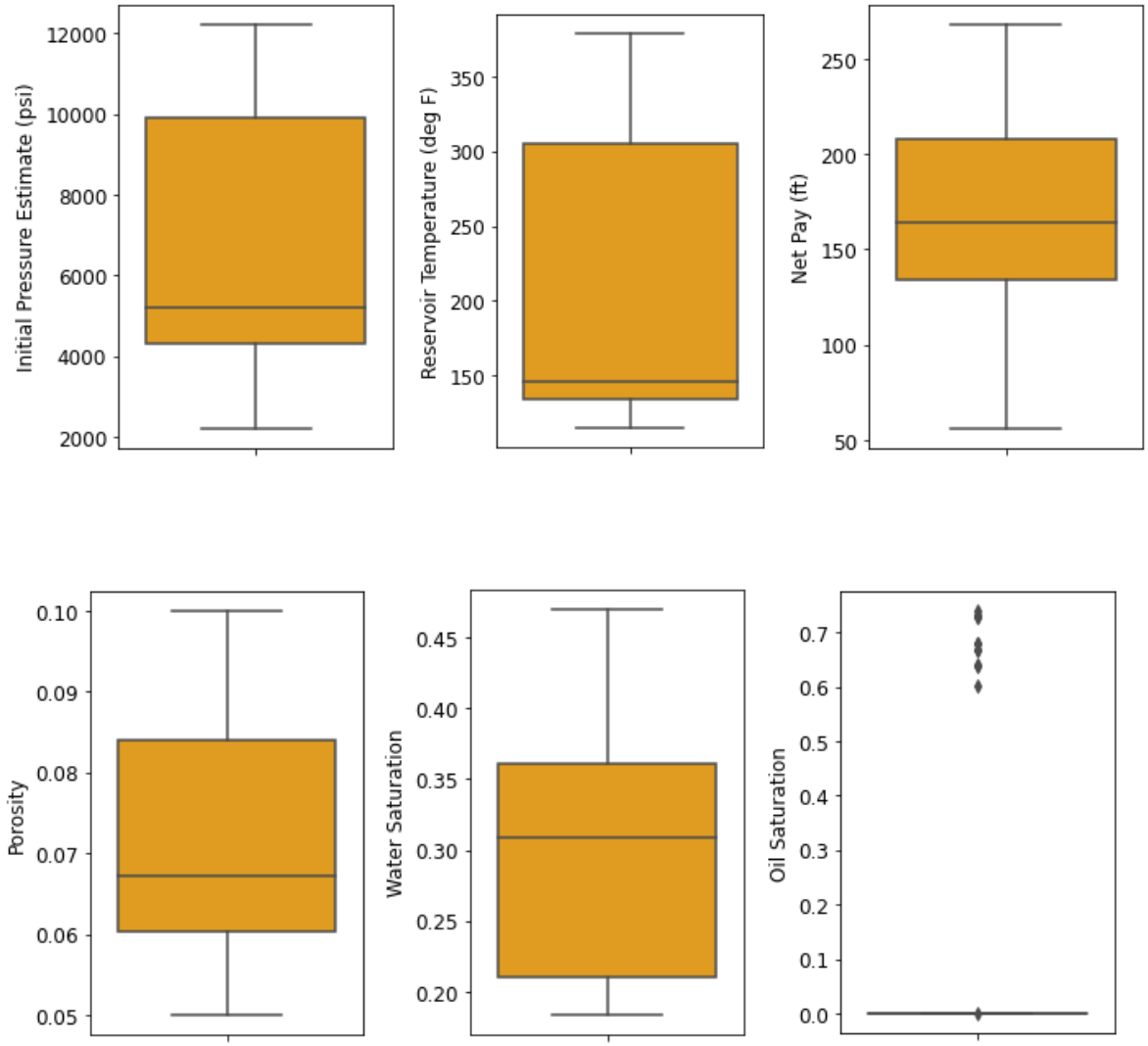


Figure 27. Histograms of independent variables.

In the Figure 27, the distributions of all input parameters have been demonstrated. The empty bars of histograms mean that there is no data point corresponding to that value. The selection of the number of bars in the histogram is up to the user, and its shape changes depending on this number. It can also be observed that the data is varying a lot for some parameters. These observations can also be revealed by boxplots which is another way of representing the results of univariate analysis. The varying distributions both in input and output variables may be a sign of both weak and strong relationship, especially in different directions. The distributions through boxplots have been shown in the following Figure 28.



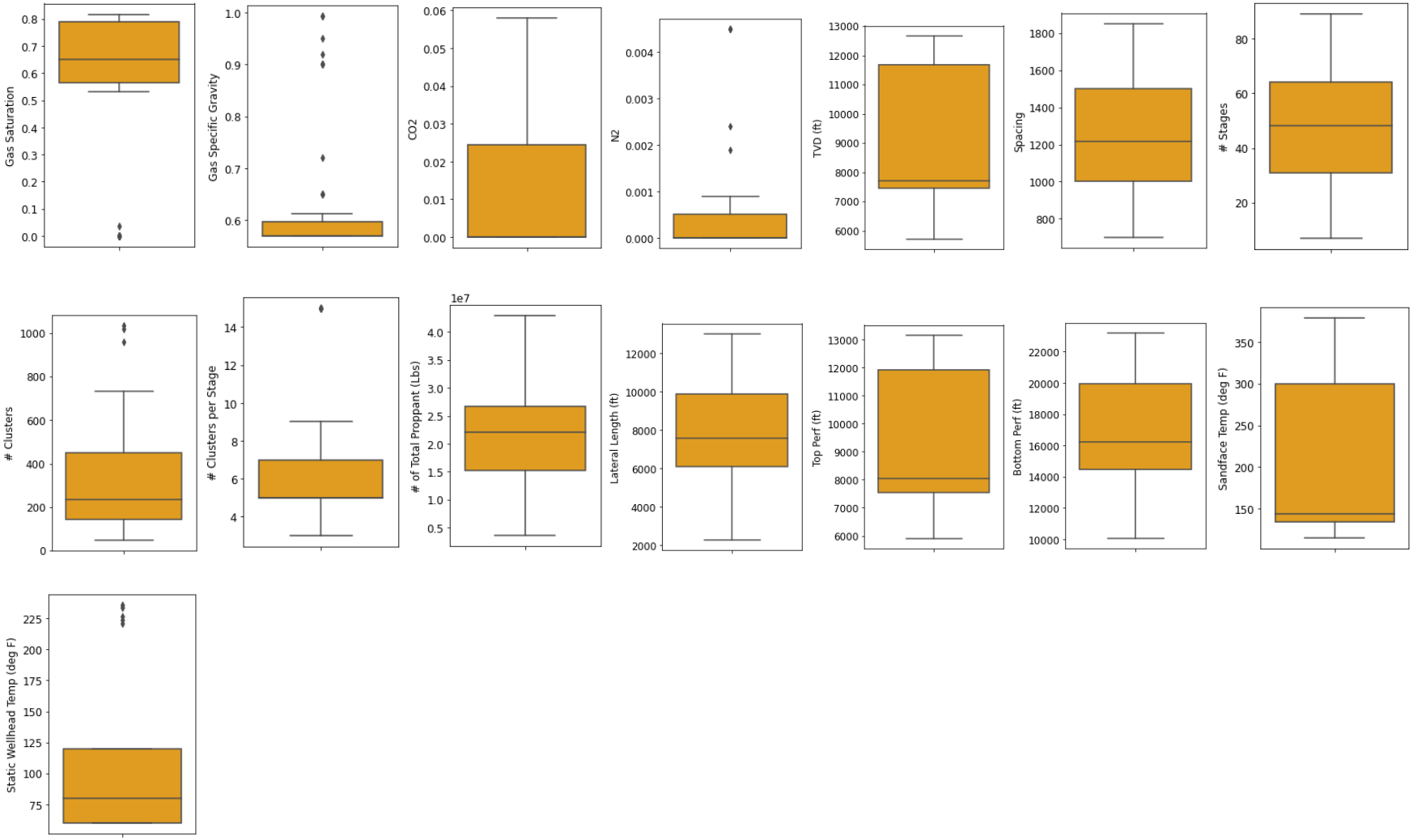


Figure 28. Boxplots of independent parameters.

5.2.2 Bivariate Analysis

In case of bivariate analysis, the correlations between different variable couples are investigated mainly using barplots and scatterplots. In the following Figure 29, the distributions of response variables depending on different formations have been plotted.

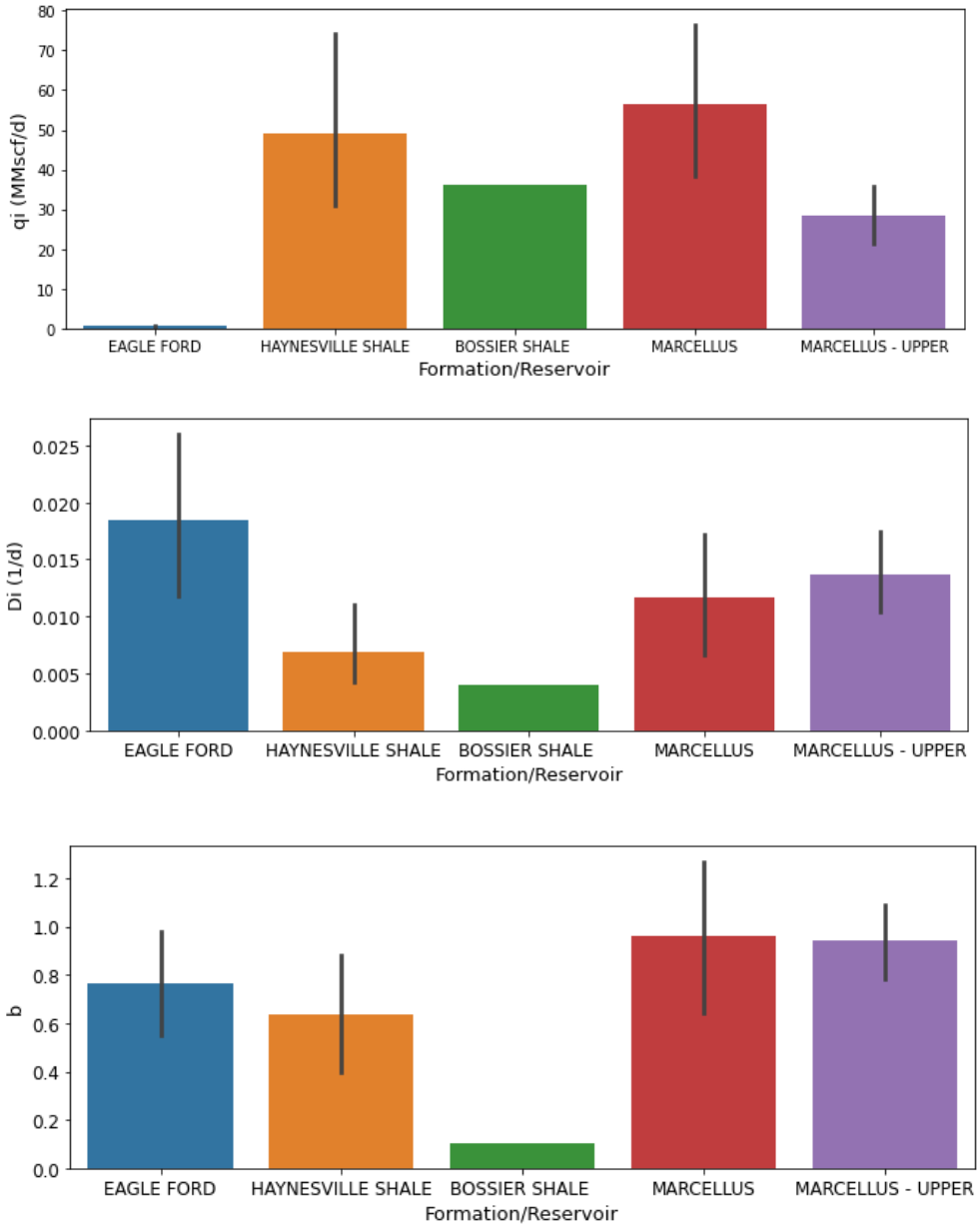


Figure 29. Barplots of output variables grouped by formations.

It can be noticed from the Figure 29 that q_i values for 'Eagle Ford' formation range in very small quantities compared to others. It can also be seen that 'Bossier Shale' formation does not have a range as it includes only one data point (only one observation), which is consistent with the information from Table 1.

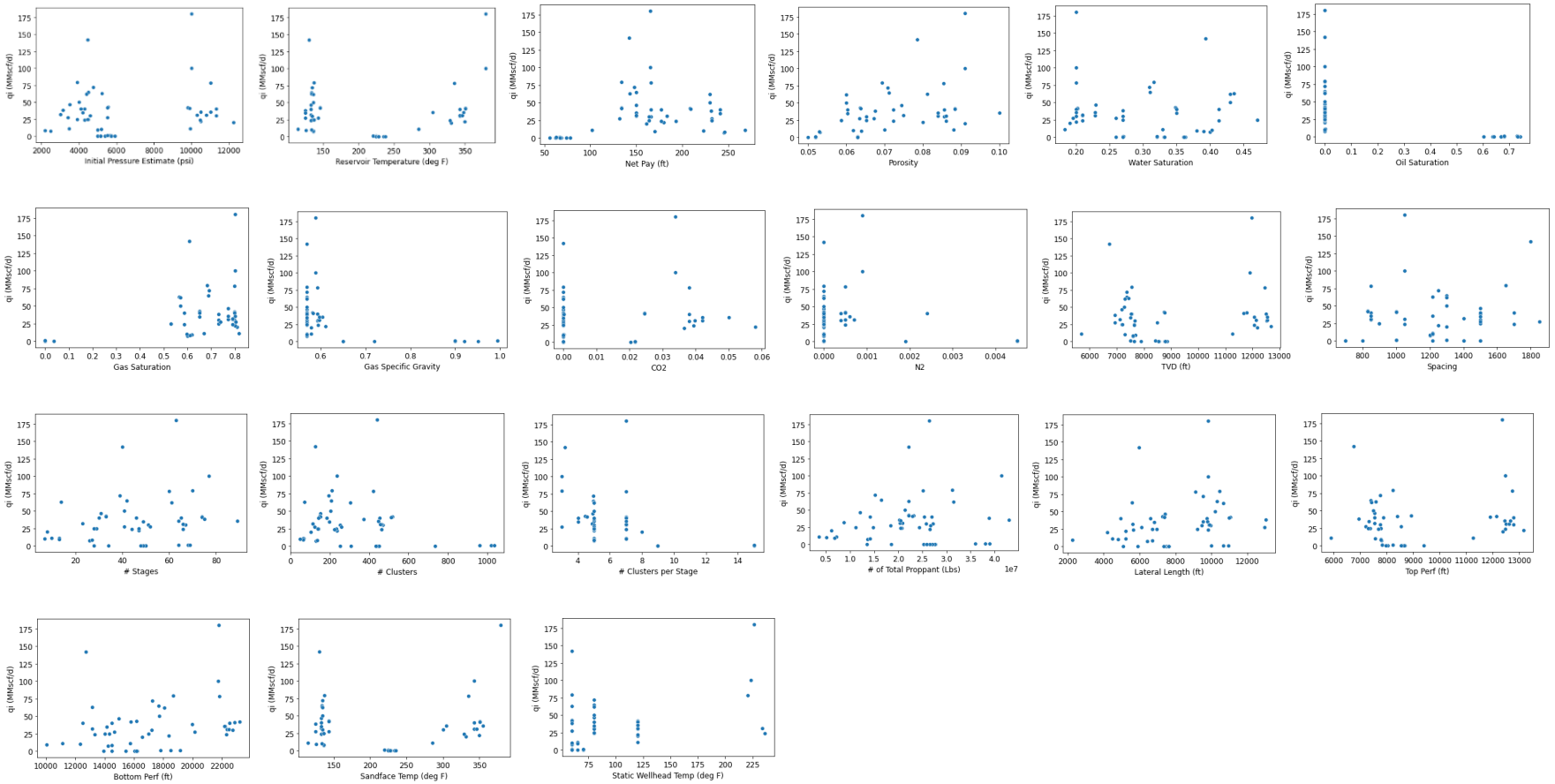


Figure 30. Scatterplots of bivariate analysis for q_i.

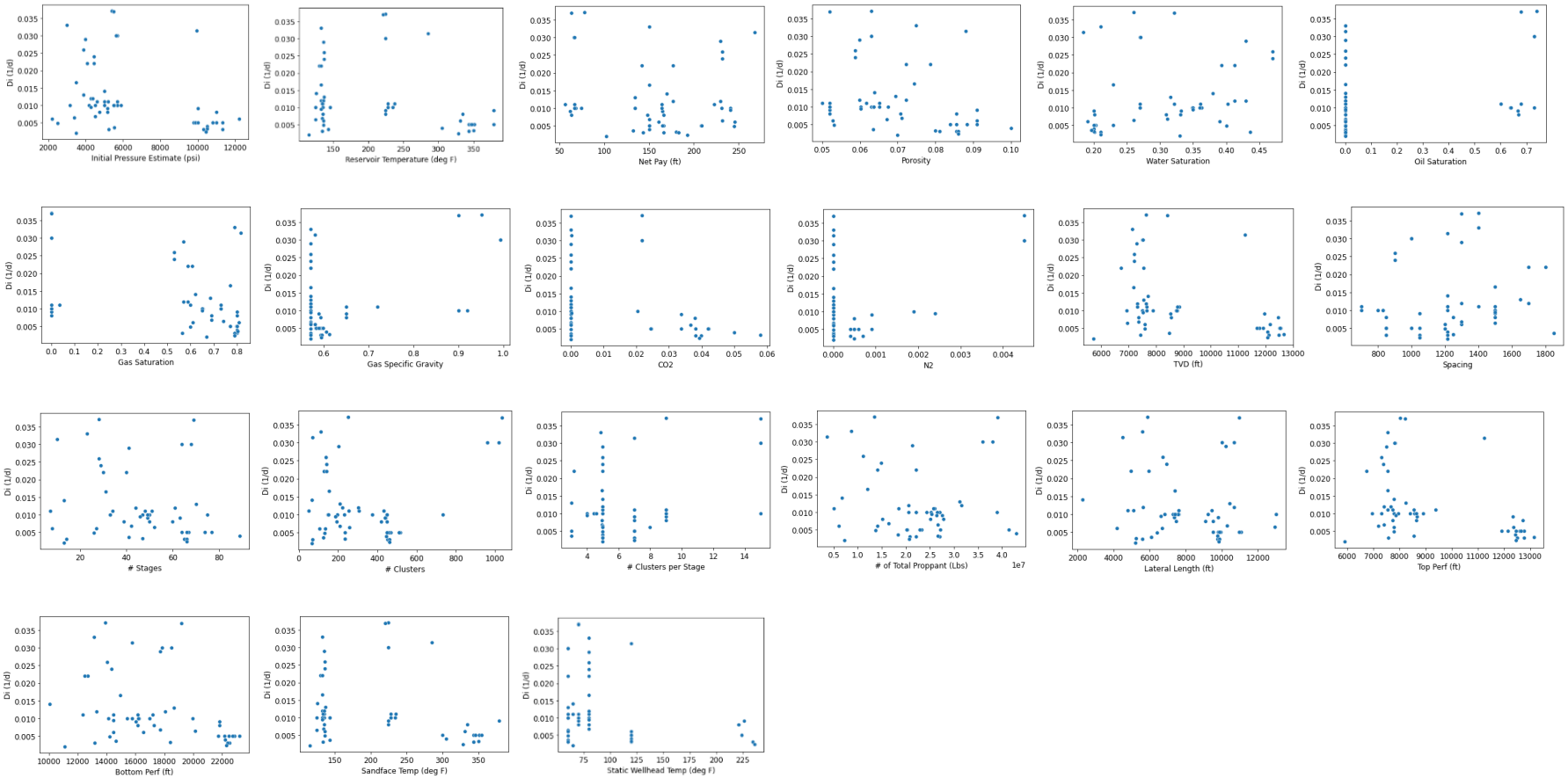


Figure 31. Scatterplots of bivariate analysis for D_i.

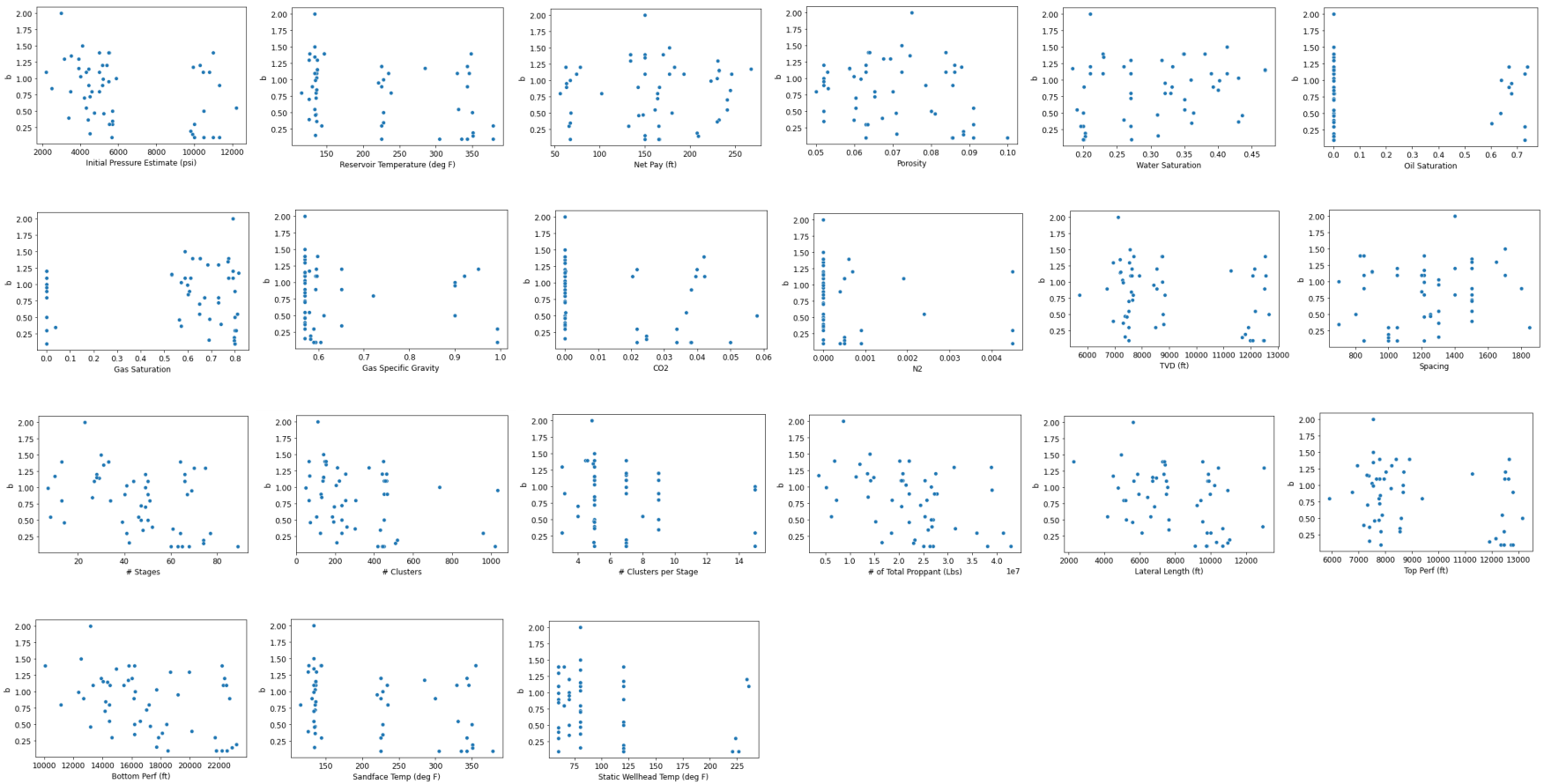


Figure 32. Scatterplots of bivariate analysis for b.

The Figures 30, 31, and 32 represent the results of bivariate analysis performed through scatterplots taking output variables (q_i, D_i, and b) on the vertical axis and independent parameters on the horizontal axis. Figures 30, 31, and 32 correspond to the dependencies having q_i, D_i, and b on the vertical axes, respectively. It is quite complicated to observe the relationship between variables from the scatterplots. It can be observed that some parameters are correlated in positive direction and some in negative. In general, there is not a variable showing a considerably significant correlation. In total, the behaviour of input parameters shows from weak to moderate correlation. Another way to demonstrate this relationship is to determine Pearson's correlation coefficients. These coefficients are provided in the Table 6 below.

	q_i	D_i	b
Initial Pressure Estimate	0.175284	-0.334467	-0.338584
Reservoir Temperature	0.114517	-0.273760	-0.344632
Net Pay	0.238378	-0.126391	0.046275
Porosity	0.494048	-0.295467	-0.238977
Water Saturation	-0.109858	0.225430	0.263085
Oil Saturation	-0.496790	0.357516	-0.051740
Gas Saturation	0.497397	-0.399709	-0.026514
Gas Specific Gravity	-0.423410	0.440442	-0.111191
CO_2	0.147134	-0.313040	-0.318732
N_2	-0.144811	0.392584	-0.198843
TVD	0.160340	-0.398895	-0.278664
Spacing	0.111654	0.134081	0.164888
Number of Stages	0.217520	-0.208061	-0.391688
Number of Clusters	-0.208726	0.177369	-0.313031
Number of Clusters per Stage	-0.458649	0.347045	-0.142994
Amount of Total Proppant	0.149398	-0.082926	-0.419993
Lateral Length	0.241357	-0.119021	-0.357321
Top Perforation	0.178937	-0.396983	-0.283788
Bottom Perforation	0.260044	-0.323025	-0.410221
Sandface Temperature	0.104161	-0.269486	-0.345918
Static Wellhead Temperature	0.421491	-0.299016	-0.234559

Table 6. Pearson's correlation coefficients for bivariate analysis.

It can be observed from the Table 6 that the results of Pearson correlation provide that there is a moderate relation between Porosity, Gas Saturation, Static Wellhead Temperature and q_i, in positive direction; and with Oil Saturation, Gas Specific Gravity, Number of Clusters per Stage in negative direction. For the other output variables, the correlations are different. There are also several weak relationships between input and output parameters. However, these analyses are subject to further investigation with predictive modelling.

5.2.3 Multivariate Analysis

The widely popular and effective tool for multivariate analysis is a heatmap with correlation coefficients. In the following Figure 33, the results of multivariate analysis have been provided in the form of a heatmap with correlation matrix.

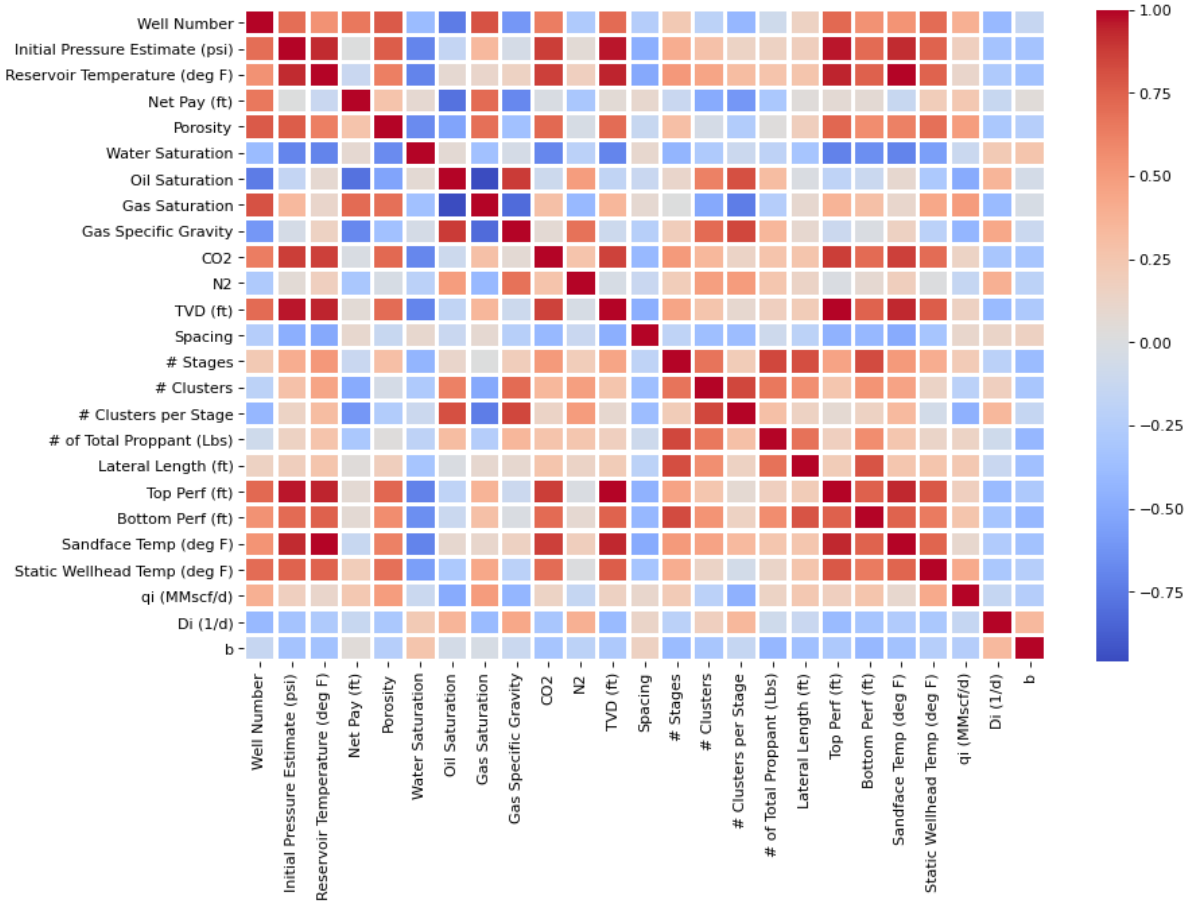


Figure 33. Heatmap with correlation matrix from multivariate analysis.

In can be seen from the Figure 33 that there are various types of correlation from weak to strong both in positive and negative directions. The diagonal in the heatmap should be disregarded because it shows the correlation of a parameter with itself which should always be equal to 1.

5.3 Part 1: Predictive Models and VIA

While building predictive models, one of the main pre-processing procedures carried out is the normalization of the dataset. Normalization is a sort of scaling which scales the values to the range from 0 to 1 assigning the minimum value of a parameter equal to 0 and the maximum value as 1. The concept of normalization is very beneficial as the variables include the values with various orders of magnitude which may influence the prediction process.

5.3.1 Multiple Linear Regression Models

As the first predictive model, Multiple Linear Regression (Ordinary Least Squares) approach has been applied. After several iterations with models, an optimal train-test split of 0.2 has been decided. Train-test split equal to 0.2 means that 80 percent of the data for each variable constitutes the training data and 20 percent for test data. Test data is so-called untouched data which is used for model validation. The OLS (Ordinary Least Squares) model performs linear regression taking the quantitative parameters mentioned in the Table 2 as input and predicting q_i, D_i, and b as response. As the OLS model can have only one output variable, three different OLS models have been built with same input but different dependent variables. The scatterplots comparing the predictions of the model with actual values of the output parameter for three different models are illustrated in the Figure 34.

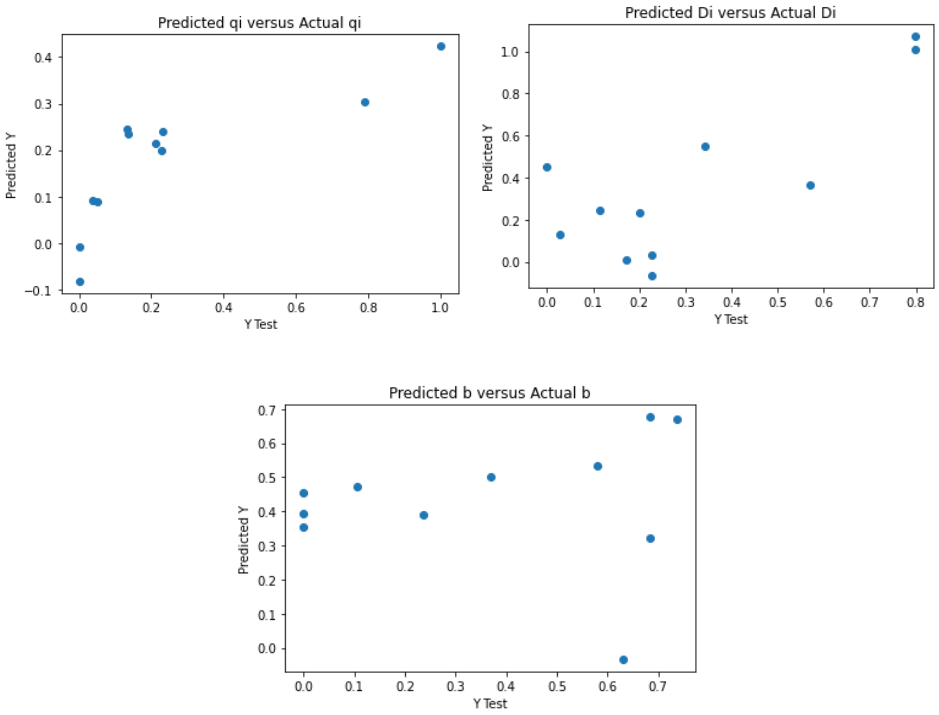


Figure 34. Predicted output vs Actual response graph for three MLR models.

In case of an ideally perfect predictive model, one-to-one correspondence should be observed from the scatterplots, meaningly that the data points should align as a straight line with the slope of unity (or 45°). It can be clearly seen from the Figure 34 that there is a very poor one-to-one correspondence between model predictions and actual values of response parameters. For the hyperbolic exponent b, it is absolutely not the case. Such a behavior is mostly related with the concept that the true relationship between the input and output variables may be strongly non-linear. The relative failure of the OLS model also suspects that the correlation is not strong so that the model fails to explain the occurred variance. Moreover, it is pivotal to mention that the lack of the data also has an influence on the results because the relationship may be so complex that just 53 observations may not be sufficient to discover it. As a supporting argument, evaluation metrics also confirm such a behavior. In the Table 7 below, evaluation metrics for each MLR model has been provided.

Evaluation Metrics	**q_i**	**D_i**	**b**
MAE	0.14	0.20	0.27
MSE	0.06	0.05	0.11
RMSE	0.23	0.23	0.34
R^2	**0.45**	**0.28**	**–0.32**

Table 7. Evaluation metrics for three MLR models.

It is obvious from the Table 7 that all three OLS models poorly predict the response variable as the R^2 values for models of q_i, D_i, and b are equal to 0.45, 0.28, and –0.32, respectively. The low values of R-squared for models of q_i and D_i reveal that the models can poorly explain the variance in the data. The R-squared result for prediction of b is even negative, showing very poor prediction. By the definition, R-squared takes a non-negative value from 0 to 1. However, mathematically, a negative R-squared is impossible only in case if the predictions are performed on the trained/fitted data. However, the R-squared is usually evaluated for test data meaning that there is a possibility of a negative R^2 when the RSS of the test data exceeds the TSS (Total Sum of Squares) of the trained data, meaning that the predictive ability of the model is very poor. The conclusions from the results of MLR application are confirming the weak correlations from bivariate and multivariate analysis. Furthermore, another reason for such results is the lack of data. For some models, the number of observations may not be enough to result in a good predictive model. In the following Figure 35, the density distribution plots of residuals for each model are given. From the Kernell density estimator, it can be seen that there is a poor normal distribution in residuals.

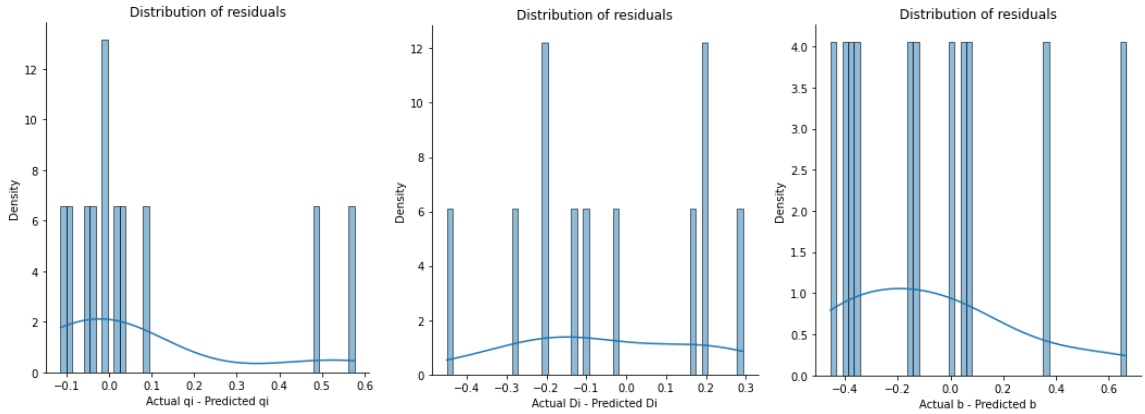


Figure 35. Distribution plots of residuals for three OLS models.

Model Coefficients for q_i	
Lateral Length	4.19
Top Perforation	3.28
Sandface Temperature	0.36
Water Saturation	0.24
Porosity	0.18
N_2	0.13
Gas Specific Gravity	0.11
Static Wellhead Temperature	0.10
Amount of Total Proppant	0.10
Spacing	0.07
Gas Saturation	0.07
Initial Pressure Estimate	0.03
Number of Clusters	- 0.00
Number of Stages	- 0.14
Oil Saturation	- 0.17
Net Pay	- 0.23
CO_2	- 0.25
Number of Clusters per Stage	- 0.25
TVD	- 0.32
Reservoir Temperature	- 0.47
Bottom Perforation	- 4.74

Model Coefficients for D_i	
Lateral Length	10.22
Top Perforation	9.71
Number of Clusters	2.73
Initial Pressure Estimate	1.07
Net Pay	0.55
Porosity	0.46
Oil Saturation	0.41
Gas Specific Gravity	0.36
Spacing	0.03
N_2	- 0.07
Water Saturation	- 0.09
Static Wellhead Temperature	- 0.10
Sandface Temperature	- 0.14
CO_2	- 0.20
Gas Saturation	- 0.34
Reservoir Temperature	- 0.59
Amount of Total Proppant	- 0.64
Number of Stages	- 0.99
Number of Clusters per Stage	- 1.66
TVD	- 3.17
Bottom Perforation	- 12.68

Table 8. Model coefficients of MLR models for q_i and D_i predictions.

Model Coefficients for b	
Top Perforation	2.76
Lateral Length	1.99
Number of Clusters	0.51
Porosity	0.32
Reservoir Temperature	0.25
Static Wellhead Temperature	0.24
CO_2	0.18
Number of Clusters per Stage	0.18
Water Saturation	0.17
Number of Stages	0.14
TVD	0.12
N_2	0.10
Spacing	0.06
Oil Saturation	0.03
Gas Specific Gravity	0.03
Gas Saturation	- 0.09
Net Pay	- 0.19
Amount of Total Proppant	- 0.85
Initial Pressure Estimate	- 1.00
Sandface Temperature	- 1.18
Bottom Perforation	- 2.69

Table 9. Model coefficients of MLR model for predictions of b.

Variable importance analysis determines the key parameters for the predictive model, and the criteria to represent it in MLR models are the model coefficients of independent variables. From the Tables 8 and 9, it can be seen that 'Lateral Length' and 'Top Perforation' appear to be most important variables for all three models in the positive direction, and 'Bottom Perforation' in the negative direction. Generally, looking at the coefficients we can confirm weak to moderate influence of input parameters on the response, which is consistent with the results of multivariate and bivariate analysis.

5.3.2 Random Forest Models

In order to further investigate and analyse the situtation, Random Forest models have been built taking the same inputs and predicting the output variables as in the case of MLR model. However, in this analysis, it was decided to include also the 'Formation/Reservoir' parameter to the list of independent variables based on the fact that RF is tree-based method. Therefore, six RF models are built in order to explore the effect of the formation, meaningly two models for each response parameter with and without consideration of the formation. The

'Formation/Reservoir' parameter consists of the names of the formations from which the wells are producing, so it is a categorical data. As we are dealing with quantitative analysis, there is a need to change them from categorical to numerical. To do so, the concept of dummy variables has been applied. In the Tables 10 and 11, the results of evaluation metrics of RF models without and with consideration of 'Formation' parameter are demonstrated respectively.

Evaluation Metrics of RF	q_i	D_i	b
MAE	7.65	0.00	0.15
MSE	124.26	$2.20*10^{-5}$	0.04
RMSE	11.15	0.00	0.21
R^2	**0.86**	**0.56**	**0.77**

Table 10. Evaluation metrics for three RF models without 'Formation'.

Evaluation Metrics of RF	q_i	D_i	b
MAE	7.00	0.00	0.17
MSE	107.45	$2.01*10^{-5}$	0.05
RMSE	10.37	0.00	0.22
R^2	**0.88**	**0.60**	**0.74**

Table 11. Evaluation metrics for three RF models with 'Formation'.

It can be obviously observed from the Tables 10 and 11 that the results have significantly improved when RF approach is applied. The interpretation of the comparison should include the fact that the data in the case of MLR is normalized. The R^2 has significantly increased meaning that the model can explain the majority of the variance. When it comes to the comparison between results of RF models with and without 'Formation' variable, it is worth to mention that inclusion of formation has improved the model in case of q_i and D_i. Nevertheless, this addition worsened the model for the prediction of b. The improvements in the results can also be concluded from the plots in the Figures 36 and 37 provided below. The data points in the scatterplots are more nearly located to the unity slope straight line than the messy distributions in the results of OLS models.

Figures 38 and 39 provide the results of variable importance analysis. VIA in Random Forest approach is performed through use of Mean Decrease in Impurity (MDI) which is also called as Gini Importance.

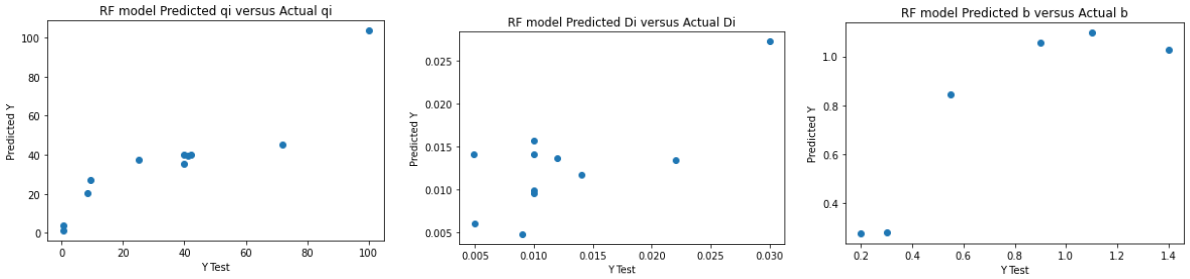


Figure 36. Plots of Predicted response vs Actual output for three RF models without 'Formation' consideration.

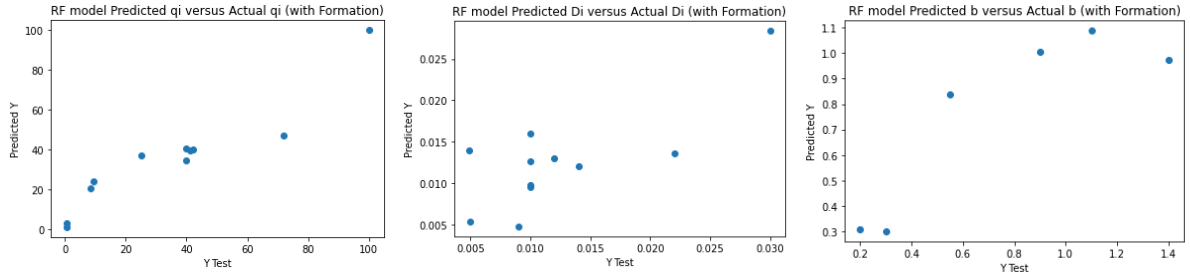


Figure 37. Plots of Predicted response vs Actual output for three RF models with 'Formation' consideration.

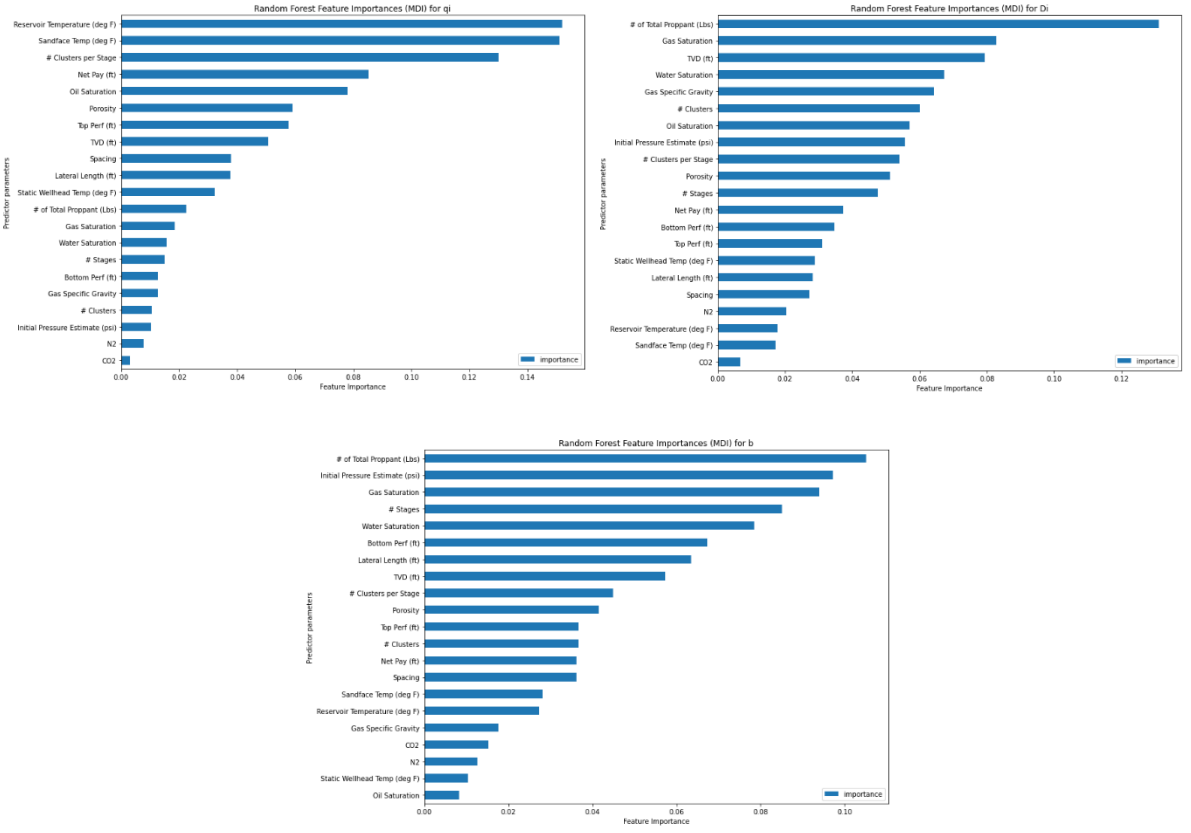


Figure 38. Variable Importance for output parameters without 'Formation'.

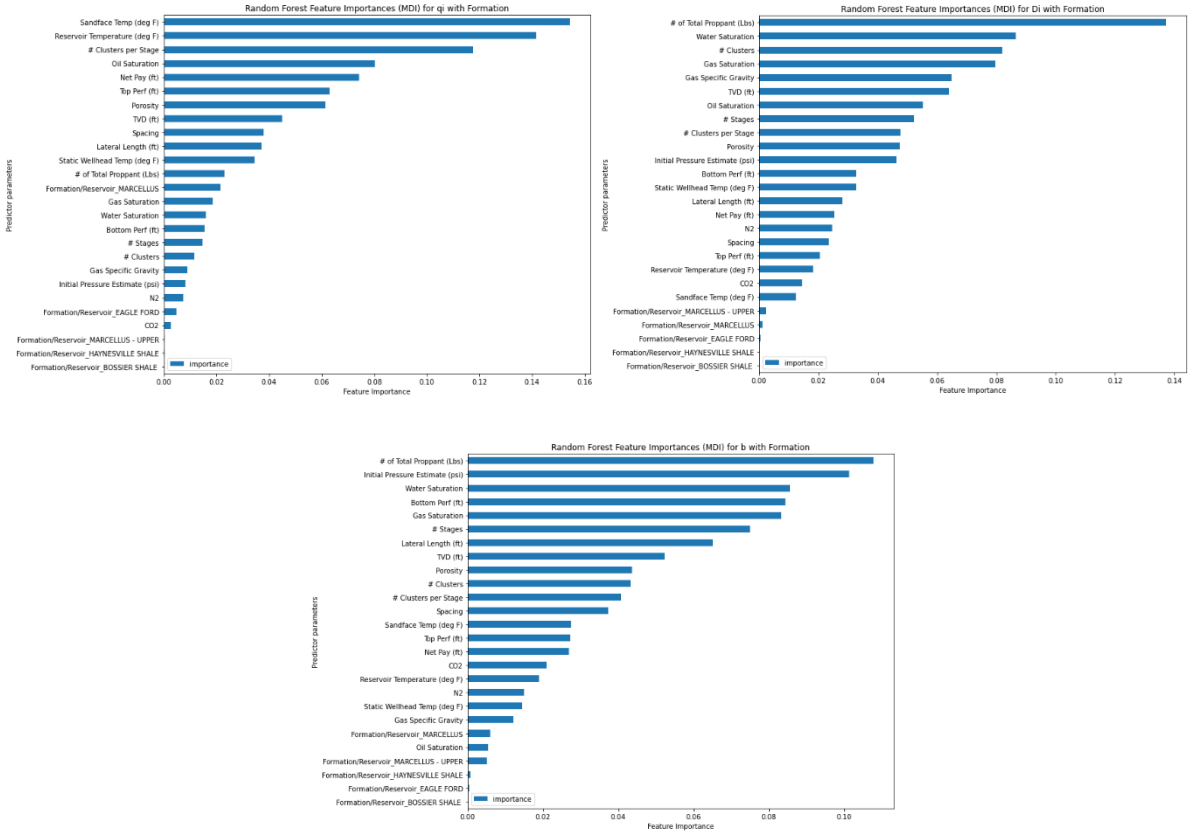


Figure 39. Variable Importance for output parameters with 'Formation'.

According to VIA of RF models without taking 'Formation' parameter into account (Figure 38), three most important parameters for the models are different. Reservoir Temperature, Sandface Temperature and Number of Clusters per Stage are the main independent variables for the model of q_i. However, for D_i and b two of three major features coincide which are Amount of Total Proppant and Gas Saturation. The VIA results of RF and OLS models differ from each other having only minor similarities. The results of RF models with the consideration of 'Formation' variable are approximately same with the case of without it. The results of both OLS and RF models reveal that there are also unnecessary variables almost not contributing to the predictions.

5.4 Part 2: Preparation of the Dataset

In the second part of the analysis, the cumulative gas production after 0.5 and 1 year have been predicted using same dataset and modelling techniques. The dataframe created for the predictive models consists of the same input parameters as in the part 1 but with different response variable being cumulative gas production after a certain period. The periods of 0.5 and 1 year have been selected according to the availability of the data in the original SPE dataset. In the case of 0.5 year, all wells contain production data throughout the period of 183

days. However, for 1 year predictions, the well #3 had production data until 249th day. In order to fill the value in the dataframe prepared, the extrapolation principle have been used for this well. To do so, a second decline curve was fit to the last decline, and the value for G_p after 365 days has been computed using the decline characteristics corresponding to that last decline. The equation utilized for this extrapolation is the Arps hyperbolic decline formula demonstrated in the equation 1 (see Methodology section). Fitting the second decline curve for the well #3 is performed in the same way likewise in the part 1 for the other wells. The decline curve fit for the well #3 is provided in the Figure 40 below.

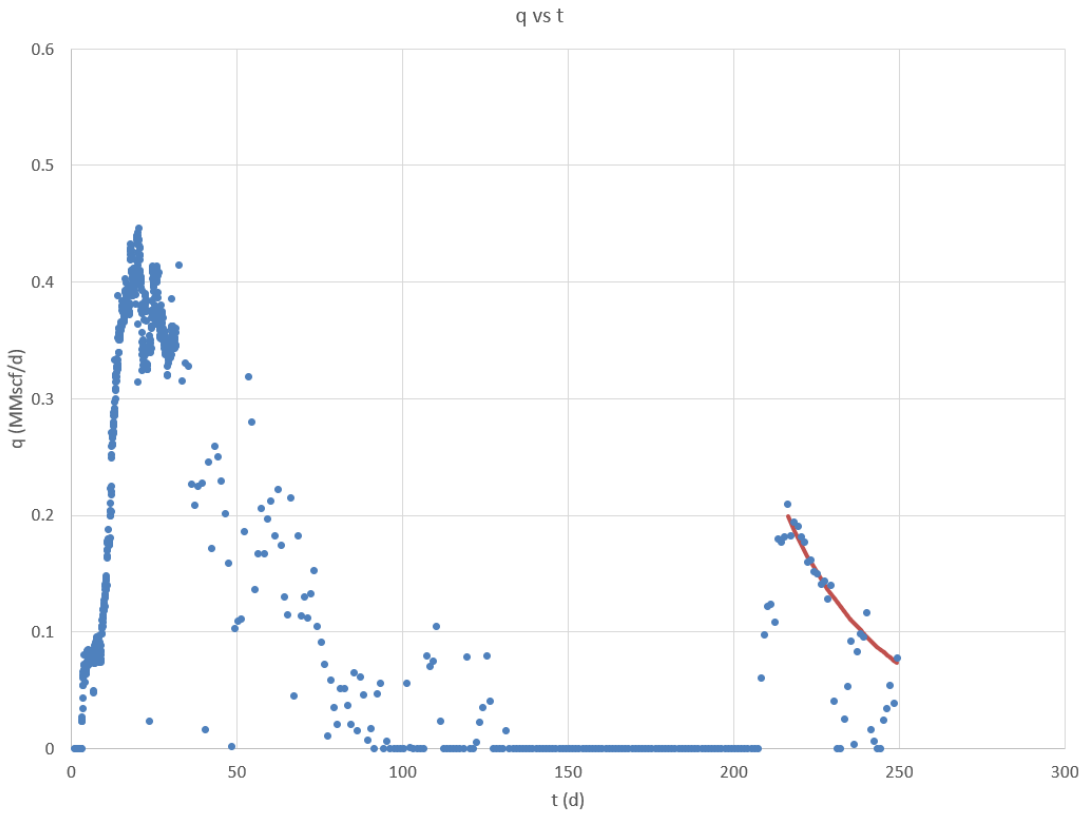


Figure 40. Well #3 production rate versus time plot with second decline curve fit.

The DCA parameters obtained from the second decline curve fitting of the well #3 are provided in the Table 12. The results of extrapolation procedure led G_p after 1 year to be equal to 22.96 MMscf which was very reasonable considering the behavior of the gas production.

q_i	**D_i**	**b**	**SSE**
MMscf/d	*1/d*	-	*(MMscf/d)^2*
10200	0.086	0.092	0.1

Table 12. DCA constants and RSS of the well #3 after second decline curve fitting.

5.5 Part 2: Exploratory Data Analysis

EDA performed for the second part is analogous in the concepts and methods to the EDA in the part 1.

5.5.1 Univariate Analysis

Histograms and Boxplots of Response Parameters

The distributions of response variables through histograms and boxplots have been given in the Figures 41 and 42.

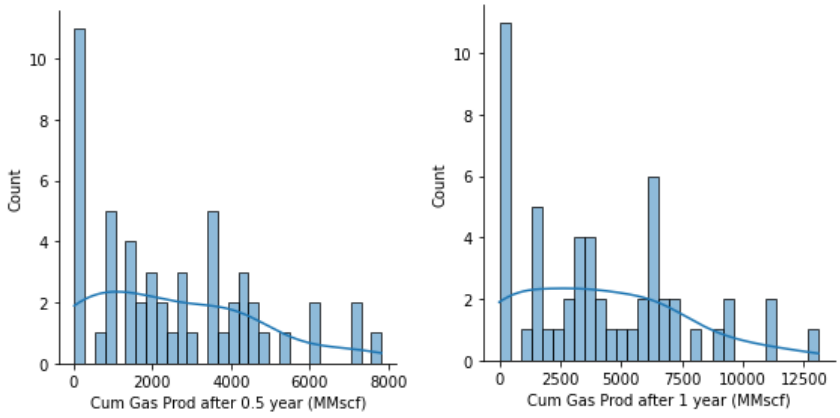


Figure 41. Histograms with KDE of response variables for part 2.

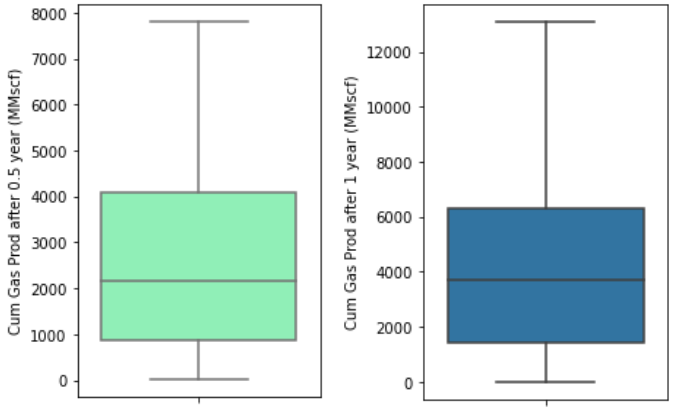


Figure 42. Boxplots of output parameters for part 2.

From the Figure 41 and 42, it can be observed that the range of dependent variables is quite large, with means of G_p (cumulative gas production) at 0.5 year and G_p at 1 year being around 2000 and 4000 MMscf, respectively.

5.5.2 Bivariate Analysis

In the following Figure 43, the distributions of output variables depending on different formations have been plotted using barplots.

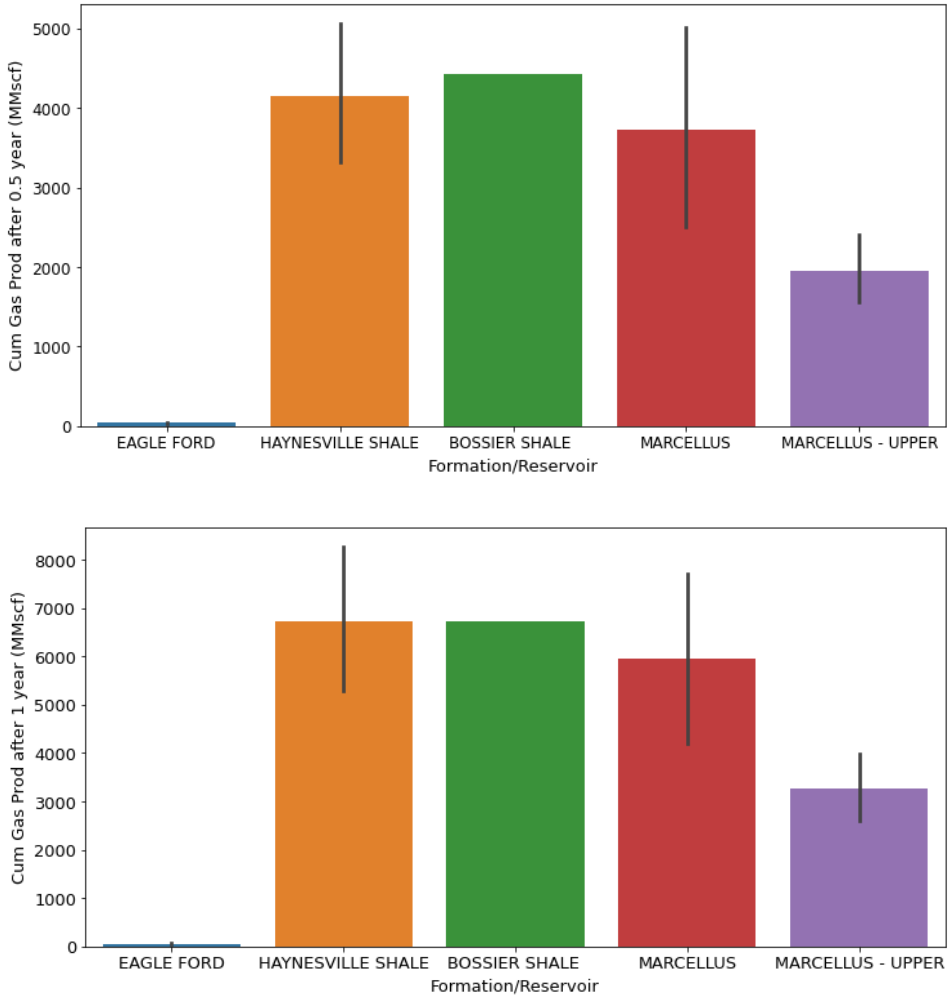


Figure 43. Barplots of output parameters for part 2 grouped by formations.

It can be noticed from the Figure 43 that throughout the periods of time, 'Eagle Ford' formation has not been so productive as its values range in very low quantities. It can also be mentioned that 'Bossier Shale' formation does not have a range as it includes only one observation (only one data point), which is consistent with the information provided in the Table 1.

In the Figures 44 and 45, the scatterplots of dependent variables versus input parameters are illustrated. The results shown in these graphs are very similar to the scatterplots generated in the part 1.

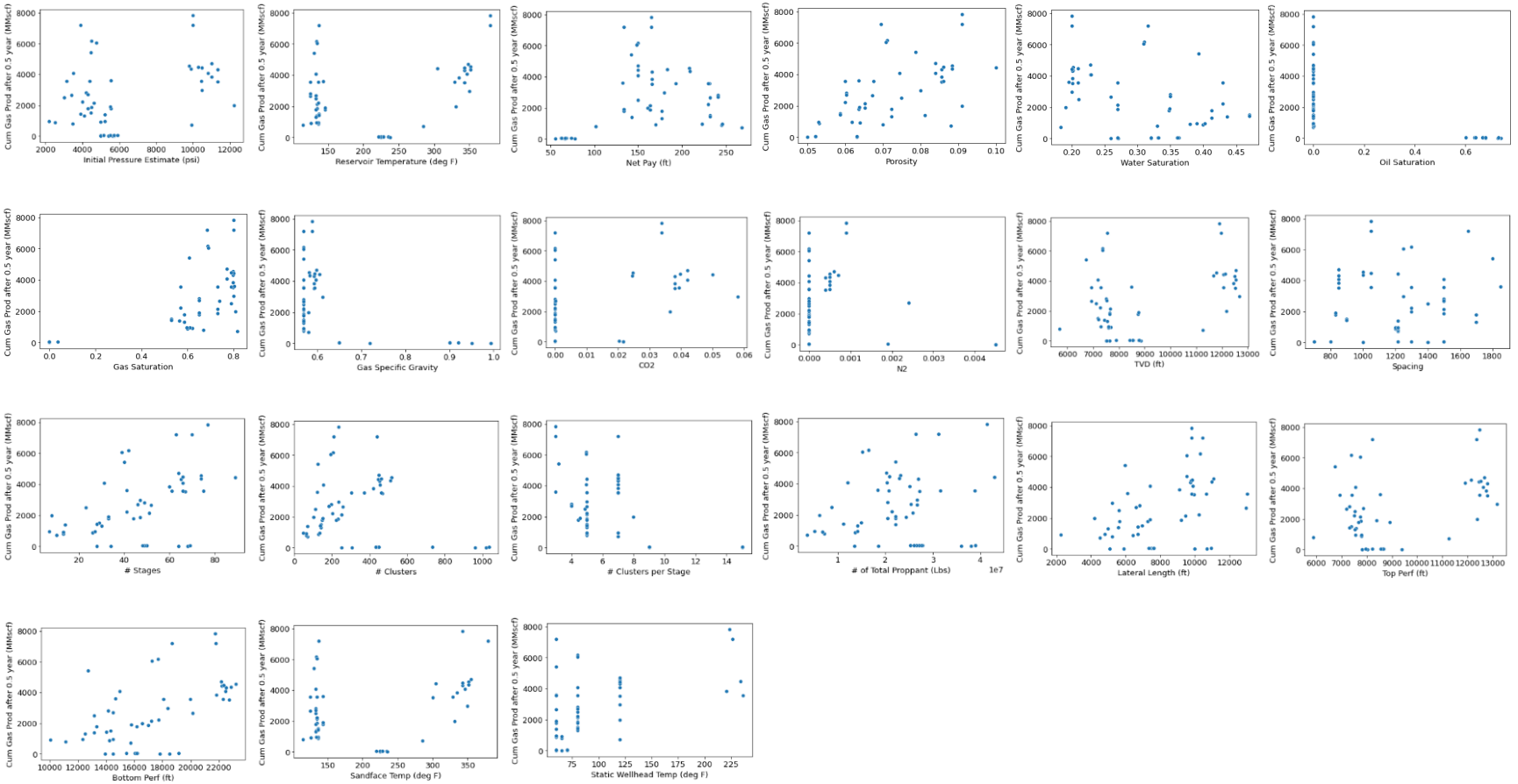


Figure 44. Scatterplots of bivariate analysis for G_p after 0.5 year.

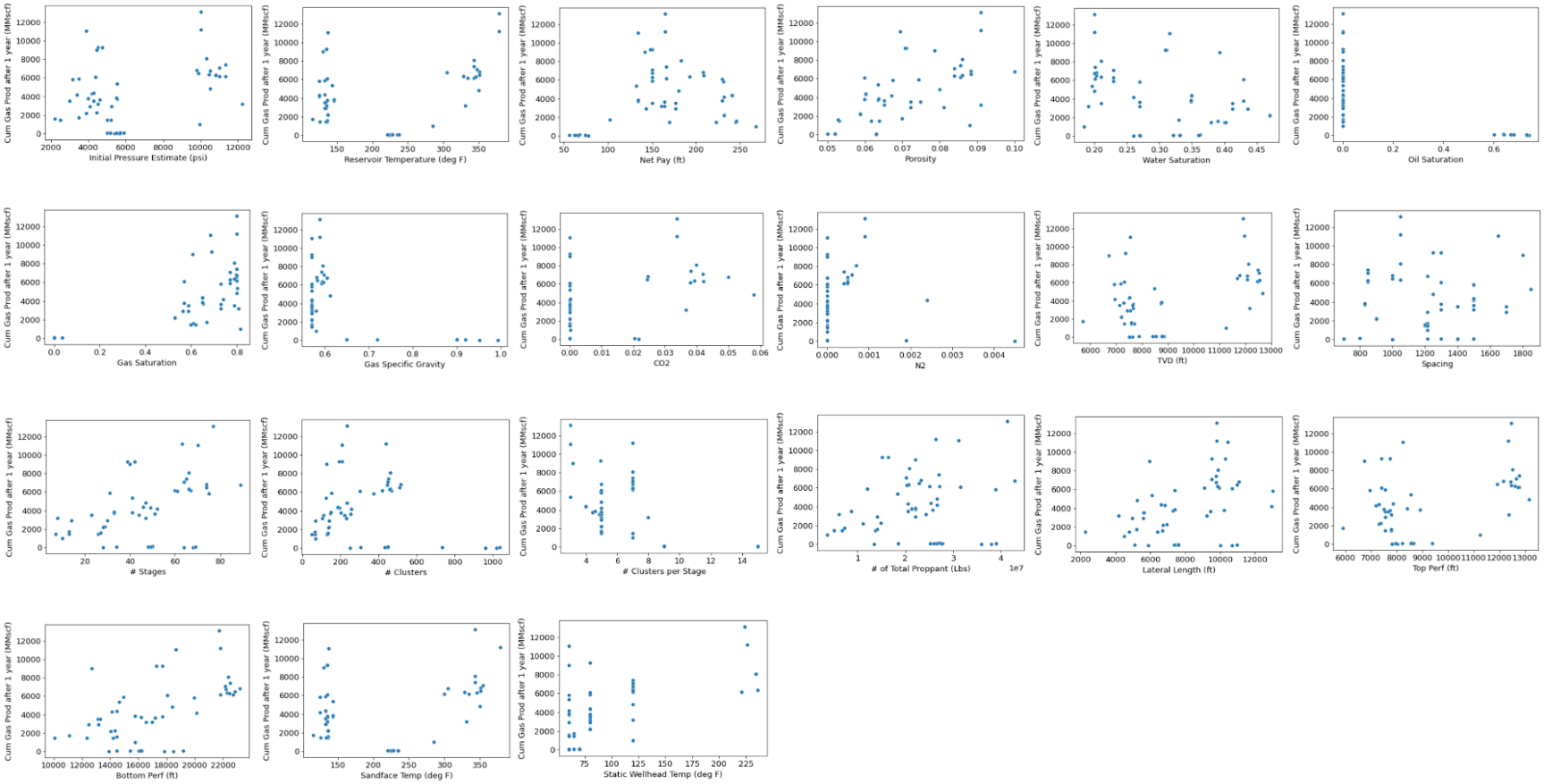


Figure 45. Scatterplots of bivariate analysis for G_p after 1 year.

	G_p after 0.5 year	G_p after 1 year
Initial Pressure Estimate	0.372175	0.382886
Reservoir Temperature	0.293218	0.294997
Net Pay	0.320618	0.327015
Porosity	0.653933	0.664648
Water Saturation	-0.406605	-0.383847
Oil Saturation	-0.611986	-0.629827
Gas Saturation	0.690122	0.700367
Gas Specific Gravity	-0.503962	-0.519237
CO_2	0.378362	0.387273
N_2	-0.182302	-0.188080
TVD	0.385093	0.394408
Spacing	0.064261	0.054388
Number of Stages	0.455430	0.455516
Number of Clusters	-0.144434	-0.155400
Number of Clusters per Stage	-0.518263	-0.535324
Amount of Total Proppant	0.195632	0.212880
Lateral Length	0.447223	0.441113
Top Perforation	0.414922	0.422815
Bottom Perforation	0.546680	0.547035
Sandface Temperature	0.279740	0.278131
Static Wellhead Temperature	0.533500	0.559797

Table 13. Pearson's correlation coefficients for bivariate analysis for part 2.

Pearson's correlation coefficients have been used for bivariate analysis and are shown in the Table 13. In contrast to the analogous analysis performed in the first part, here the majority of input variables are significantly correlated to the cumulative gas production. It can be obviously observed that Gas Saturation, Porosity and Oil Saturation are the most closely correlated to both outputs considering both positively and negatively directed relationships.

5.5.3 Multivariate Analysis

In the following Figure 46, the results of multivariate analysis have been provided through a heatmap with correlation matrix. The diagonal in the heatmap should be neglected because it displays the correlation of a parameter with itself which should always be equal to unity. The observations from the multivariate analysis confirm the judgements made throughout the bivariate analysis.

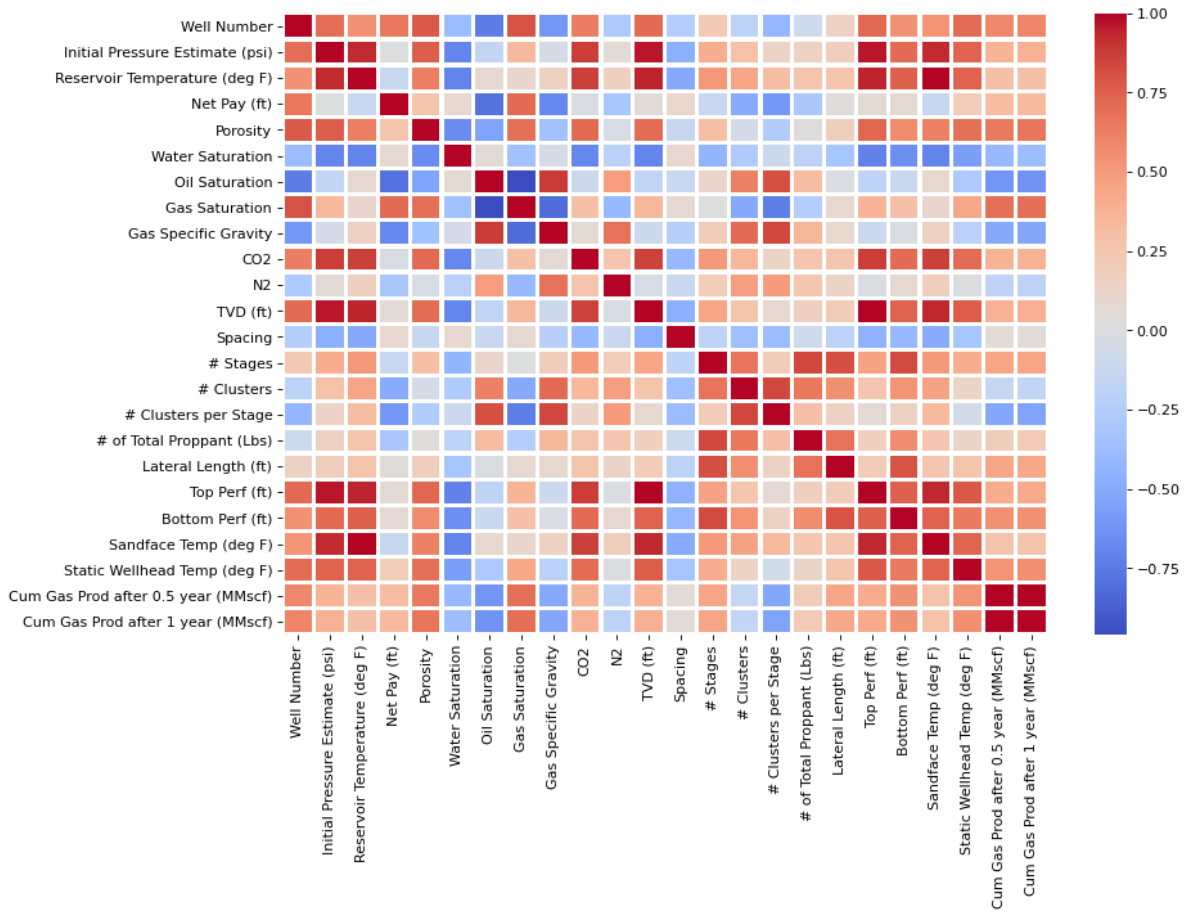


Figure 46. Heatmap with correlation matrix from multivariate analysis in part 2.

5.6 Part 2: Predictive Models and VIA

In the process of building a predictive model, normalization is an essential procedure and it concerns, especially, multiple linear regression models. In this section, the results of predictive models such as OLS and RF are discussed. The models take the same input variable as for the case in the part 1 but the output variables are different here. Thus, the overall process is conducted analogously with slight amendments.

5.6.1 Multiple Linear Regression Models

Similar to the modelling performed in the first part, MLR models have been built with a train-test split of 0.2, meaning that the test data comprises the 20 percent of the total data. Nevertheless, the OLS models in this part try to predict cumulative gas production for half- and one-year periods. Taking into account that such type of models can deal with one output variable at a time, two MLR models are built and corresponding results of predicted versus actual data for both models using scatterplots are demonstrated in the Figure 47.

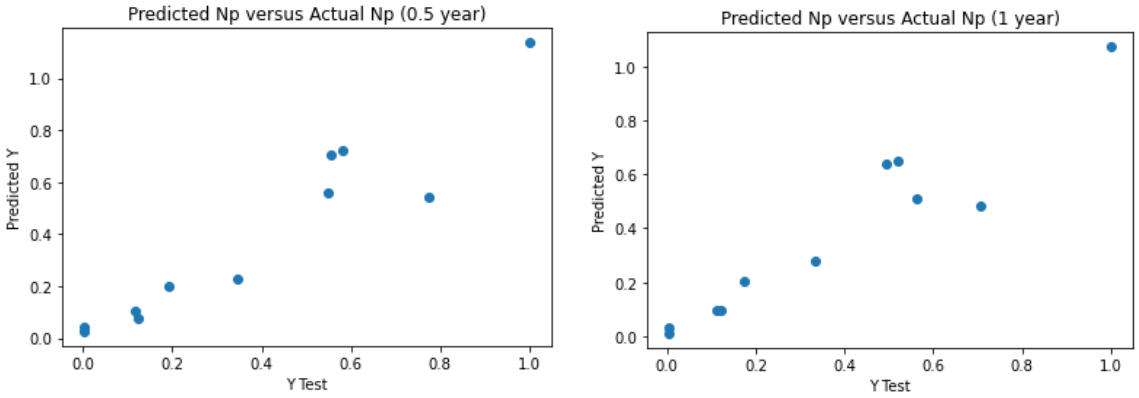


Figure 47. Predicted data vs Actual data plot for two MLR models in part 2.

It can be seen from the Figure 47 that the one-to-one correspondence is not ideal but the data points are aligned quite good with the straight line of unity slope. The results can also be confirmed by the evaluation metrics provided in the Table 14. It is noteworthy to mention that the interpretation of results should also consider the normalization of the dataset as a pre-processing step.

Evaluation Metrics	**G_p after 0.5 year**	**G_p after 1 year**
MAE	0.08	0.07
MSE	0.01	0.01
RMSE	0.11	0.10
R^2	**0.88**	**0.90**

Table 14. Evaluation metrics for two MLR models in part 2.

The results of model evaluation given in the Table 14 reveal that both models have performed very good considering the R^2 values as high as 0.88 and 0.90. This means that the models are able to explain the variance up to 90 percent.

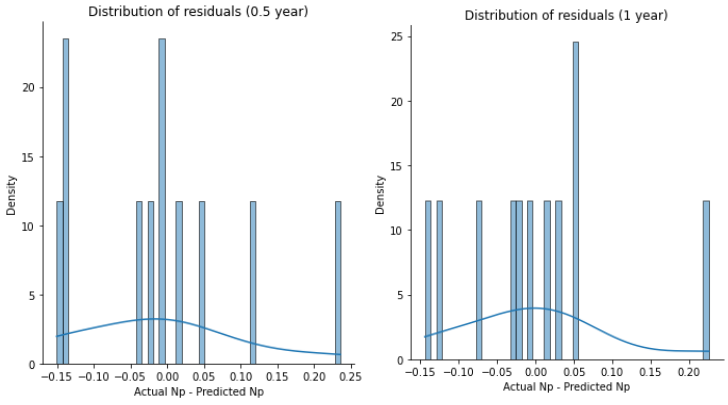


Figure 48. Residuals distribution plots for both OLS models in part 2.

In the Figure 48, the density distribution graphs of residuals for both models are provided. The density distribution has the tendency to the shape of normal distribution but not formed completely which may be due to the small amount of test data on which the models are evaluated.

Model Coefficients for G_p after 0.5 year		*Model Coefficients for G_p after 1 year*	
Lateral Length	12.56	Lateral Length	7.80
Top Perforation	9.55	Top Perforation	5.87
Sandface Temperature	0.86	Number of Stages	0.65
Number of Stages	0.79	Sandface Temperature	0.40
Initial Pressure Estimate	0.33	Water Saturation	0.27
Gas Specific Gravity	0.25	Gas Saturation	0.27
Gas Saturation	0.22	Gas Specific Gravity	0.23
Water Saturation	0.20	Initial Pressure Estimate	0.18
Spacing	0.13	Spacing	0.16
Porosity	0.12	Porosity	0.07
N2	-0.04	N_2	0.06
Static Wellhead Temperature	-0.15	Reservoir Temperature	0.03
Number of Clusters	-0.17	Static Wellhead Temperature	-0.01
CO2	-0.24	Number of Clusters per Stage	-0.14
Number of Clusters per Stage	-0.27	Amount of Total Proppant	-0.27
Oil Saturation	-0.32	CO_2	-0.28
Net Pay	-0.34	Number of Clusters	-0.31
Amount of Total Proppant	-0.49	Net Pay	-0.40
Reservoir Temperature	-0.57	Oil Saturation	-0.40
TVD	-1.47	TVD	-0.97
Bottom Perforation	-14.96	Bottom Perforation	-9.16

Table 15. Model coefficients of two MLR models.

VIA determines the major variables for predictive models, and the criteria to represent it in MLR models are the model coefficients of predictors. It can be observed From the Table 15 that ‘Lateral Length’ and ‘Top Perforation’ appear to be most important parameters for both models in the positive direction, and ‘Bottom Perforation’ in the negative direction.

5.6.2 Random Forest Models

To further investigate and analyse the situtation, RF models are built which takes the same inputs and predicts the response variables as in the case of MLR models. In the Table 16, the evaluation metrics of Random Forest models have been given.

Evaluation Metrics of RF	**G_p after 0.5 year**	**G_p after 1 year**
MAE	548	1058.9
MSE	734405	$2.6*10^6$
RMSE	857	1612
R^2	**0.88**	**0.84**

Table 16. Evaluation metrics for both RF models in part 2.

Based on the results of RF models shown in the Table 16, the models are performing almost as good as OLS models. In general, Random Forest approach is considered as a more sophisticated technique compared to Linear Regression, meaningly, in the majority of cases former performs better than latter. Such results were observed in the part 1, however, in this case, the opposite took place. MLR models have better results in terms of performance which may be due to the possible linear relation between input and output. Nevertheless, both methods show good results close to 0.9, meaning that the 90 percent of variance can be explained by the models. The similarity of performances also can be deducted from the scatterplots of predicted values versus actual data provided in the Figure 49 below.

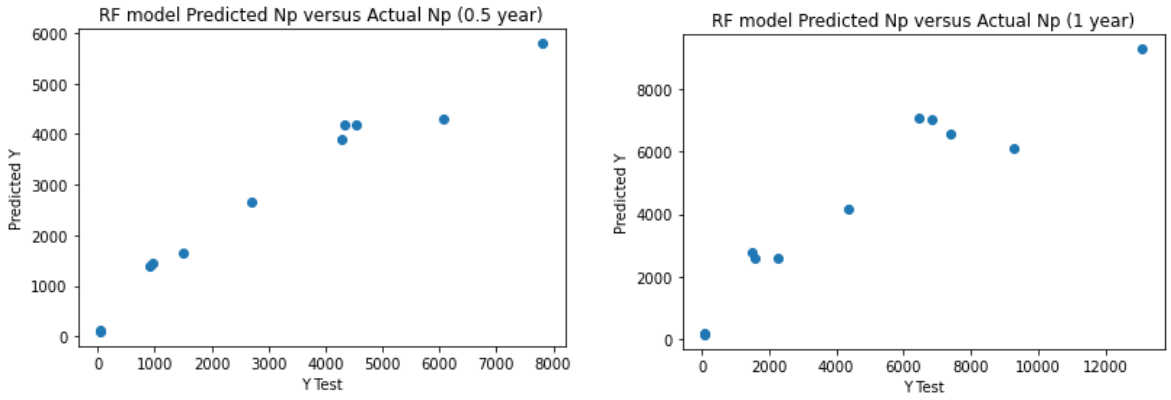


Figure 49. Graphs of Predicted values vs Actual data for both RF models in part 2.

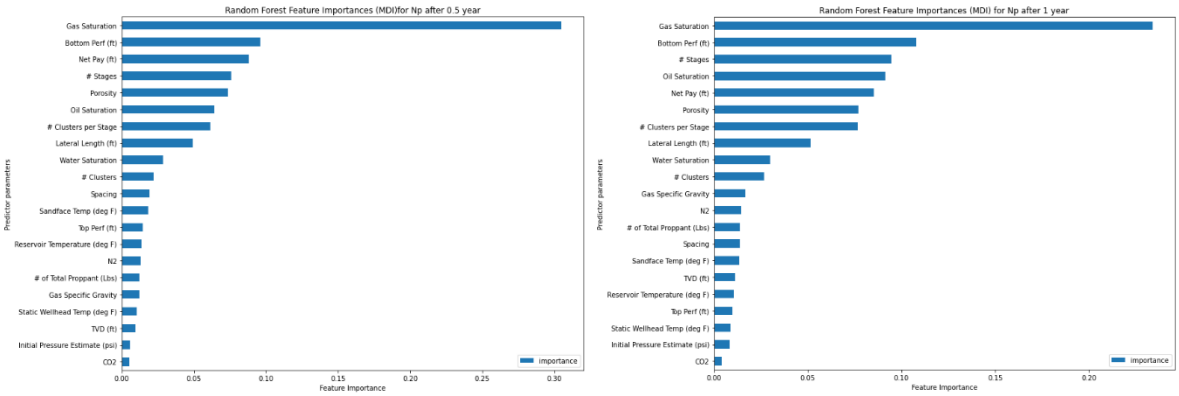


Figure 50. Feature Importance for response parameters in part 2.

Figure 50 shows the results of feature importance analysis. VIA in Random Forest approach is carried out using of Mean Decrease in Impurity (MDI) which is also called as Gini Importance. According to the feature importance analysis of RF models (Figure 50), the most important parameter for both models is ‘Gas Saturation’. The VIA results of RF and OLS models differ from each other having also some similarities such as the importance ‘Bottom Perforation’ parameter. The results of both RF and OLS models disclose that there are also unnecessary variables almost not contributing to the predictions.

Chapter 6: Conclusion

In conclusion, this research study comprises the investigation of the correlation between operational/reservoir parameters and decline curve characteristics with additional predictions on cumulative gas production. The focus of the study is shale gas wells from different unconventional reservoirs. Exploratory techniques have been performed on the dataset of SPE to form an understanding of the patterns and features. After EDA has been carried out, statistical learning and different ML algorithms were applied to build predictive models for the estimation of the relationship and forecasting. The performance efficiency of the models was evaluated to observe the precision of the predictions and feature importance analysis were conducted to determine driving variables.

The principal conclusions drawn from this research have been provided in the following way.

1. The results of univariate, bivariate and multivariate analysis demonstrated that there was a correlation to some extent between some independent variables and output parameters.
2. The metrics used to evaluate the models determined that in the most cases, the predictions of Random Forest models were much more accurate compared to OLS regression as the proportion of the explained variance was as high as up to almost 90 percent for some models. The only exception is the modelling for predicting cumulative gas production after 1 year in which the OLS model outperforms the RF model. These results confirm the well-known power of random forest models in machine learning. In addition, MLR models for prediction of nominal decline rate and hyperbolic exponents showed very poor results.
3. The driving parameters determined by VIA were generally different for different models with some similarities which confirms the feature importance analysis being model-specific.
4. Random_state feature (selects the data points randomly for train-test splitting) required in the model building process was generating different results of prediction accuracy of models. Another feature of models changing the results was the parameter ‘n’ in RF models which is deciding the number of trees used in the method.
5. An important issue in the results is the lack of data which is significant in data analytics. Only 53 observation points made the whole procedure very challenging because

generally, more data is used in statistical learning and having as more data as possible is always better for predictions.

Recommendations for further research studies on these topics are as follows:

- Studies including much more data points can be involved to enhance the analysis for deeper investigations.
- The model features can be tuned for better results and their influence can be inspected.
- Different decline curve models such as Duong, SEDM and PLE can be used for fitting decline curves and the influence on the results can be investigated.
- Changes can be made to the list of operational and reservoir parameters to discover different correlations.
- Other machine learning techniques, for example deep learning (Neural Network), can be used for deeper comparisons on the best performances.

References

Alarifi, S. A. (2021) 'Production Data Analysis of Hydraulically Fractured Horizontal Wells from Different Shale Formations', *Appl. Sci.*, https://doi.org/10.3390/app11052165

Arps, J. J. (1945) 'Chapter II. Petroleum Economics', *PETROLEUM TECHNOLOGY*.

Britton, M. (2020) *Escape the Correlation Matrix into… Feature Space*. Available at: https://towardsdatascience.com/escape-the-correlation-matrix-into-feature-space-4d71c51f25e5

Bruce, P., Bruce, A. & Gedeck, P. (2020) *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python* (2nd ed.). O'Reilly Media.

Carranza, C. et al. (2020) 'Root zone soil moisture estimation with Random Forest', *Journal of Hydrology*, https://doi.org/10.1016/j.jhydrol.2020.125840.

Chen, Z. et al. (2021) 'Application of statistical machine learning clustering algorithms to improve EUR predictions using decline curve analysis in shale-gas reservoirs', *Journal of Petroleum Science and Engineering*, https://doi.org/10.1016/j.petrol.2021.109216

Datavizpyr (2020) *How To Make Scatterplot with Marginal Histograms in R?* Available at: https://datavizpyr.com/how-to-make-scatterplot-with-marginal-histograms-in-r/

Duong, A. N. (2011) 'Comparison of Decline Curve Analysis Methods with Analytical Models in Unconventional Plays', *SPE*.

Everitt, B. S. & Dunn, G. (2001) *Applied Multivariate Data Analysis* (1st ed.). Wiley.

Gong, X. et al. (2010) 'Bayesian Probabilistic Decline Curve Analysis Quantifies Shale Gas Reserves Uncertainty', *CSUG/SPE*.

Han, D. et al. (2019) 'Production Forecasting for Shale Gas Well in Transient Flow Using Machine Learning and Decline Curve Analysis', *URTeC*.

Hui, G. et al. (2021) 'Machine learning-based production forecast for shale gas in unconventional reservoirs via integration of geological and operational factors', *Journal of Natural Gas Science and Engineering*, https://doi.org/10.1016/j.jngse.2021.104045

Jebb, A. T., Parrigon, S. & Woo, S. E. (2016) 'Exploratory data analysis as a foundation of inductive research', *Human Resource Management Review*, http://dx.doi.org/10.1016/j.hrmr.2016.08.003

Jobson, J. D. (1992) *Applied Multivariate Data Analysis. Vol. II: Categorical and Multivariate Methods*. Edmonton: University of Alberta.

Li, Y. & Han, Y. (2017) 'Decline Curve Analysis for Production Forecasting Based on Machine Learning', *SPE*.

Martinez, W. L., Martinez, A. R. & Solka, J. L. (2017) *Exploratory Data Analysis with MATLAB (Chapman & Hall/CRC Computer Science & Data Analysis)* (3rd ed.). Chapman & Hall/CRC.

Meyet, M., Dutta, R. & Burns, C. (2013) 'Comparison of Decline Curve Analysis Methods with Analytical Models in Unconventional Plays', *SPE*.

Milo, T. & Somech, A. (2020) 'Automating Exploratory Data Analysis via Machine Learning: An Overview', *SIGMOD'20*, https://doi.org/10.1145/3318464.3383126

Mishra, S. (2012) 'A New Approach to Reserves Estimation in Shale Gas Reservoirs Using Decline Curve Analysis Models', *SPE*.

Mishra, S. & Lin, L. (2017) 'Application of Data Analytics for Production Optimization in Unconventional Reservoirs: A Critical Review', *URTeC*, doi:10.15530/urtec-2017-2670157

Mohammadpoor, M. & Torabi, F. (2018) 'Big Data analytics in oil and gas industry: An emerging trend', *Petroleum*, https://doi.org/10.1016/j.petlm.2018.11.001

Morgan, E. (2018) 'Accounting for Serial Autocorrelation in Decline Curve Analysis of Marcellus Shale Gas Wells', *SPE*.

Nelson, B. et al. (2014) 'Predicting Long-term Production Behavior of the Marcellus Shale', *SPE*.

Pearson, R. K. (2018) *Exploratory Data Analysis Using R (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)* (1st ed.). Chapman & Hall/CRC.

Potter, K. (2006) 'Methods for Presenting Statistical Information: The Box Plot', *University of Utah*.

Sahoo, K. et al. (2019) 'Exploratory Data Analysis using Python', *IJITEE*, doi: 10.35940/ijitee.L3591.1081219

Satter, A. & Iqbal, G. M. (2015) *Reservoir Engineering: The Fundamentals, Simulation, and Management of Conventional and Unconventional Recoveries* (1st ed.). Gulf Professional Publishing.

Schuetter, J., Mishra, S. & Zhong, M. (2018) 'A Data-Analytics Tutorial: Building Predictive Models for Oil Production in an Unconventional Shale Reservoir', *SPE*.

Schuetter, J., Mishra, S. & Zhong, M. (2015) 'Data Analytics for Production Optimization in Unconventional Reservoirs', *URTeC*, doi: 10.15530/urtec-2015-2167005

Sircar, A. et al. (2021) 'Application of machine learning and artificial intelligence in oil and gas industry', *Petroleum Research*, https://doi.org/10.1016/j.ptlrs.2021.05.009

Tadjer, A., Hong, A. & Bratvold, B. R. (2021) 'Machine learning based decline curve analysis for short-term oil production forecast', *Energy Exploration & Exploitation*, doi: 10.1177/01445987211011784

Taji O. & Alp D. (2021) 'Comparison of Type Well Generation Methods for Unconventional Reservoirs', *SPE Reservoir Evaluation & Engineering*.

Tan, L., Zuo, L. & Wang, B. (2018) 'Methods of Decline Curve Analysis for Shale Gas Reservoirs', *Energies*, doi:10.3390/en11030552

Tanriverdi, I. (2021) *Beginners Guide to Explanatory Data Analysis*. Available at: https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-explanatory-data-analysis/

Vikara, D., Remson, D. & Khanna, V. (2020) 'Machine learning-informed ensemble framework for evaluating shale gas production potential: Case study in the Marcellus Shale', *Journal of Natural Gas Science and Engineering*, https://doi.org/10.1016/j.jngse.2020.103679

Vyas, A., Datta-Gupta, A. & Mishra, S. (2017) 'Modeling Early Time Rate Decline in Unconventional Reservoirs Using Machine Learning Techniques', *SPE*.

Wei, T. N. (2022) *Explaining negative R-squared*. Available at: https://towardsdatascience.com/explaining-negative-r-squared-17894ca26321

Yuan, Y. et al. (2020) ‘Production decline analysis of shale gas based on a probability density distribution function’, *Journal of Geophysics and Engineering*, doi:10.1093/jge/gxz122.

Zhenke, X. & Morgan, E. (2019) ‘Combining Decline-Curve Analysis and Geostatistics To Forecast Gas Production in the Marcellus Shale’, *SPE*.

Zhong, M. et al. (2015) ‘Do Data Mining Matter? : A Wolfcamp “Shale” Case Study’, *SPE*.

Appendix

In this part, decline curves fit to other 52 wells have been given.

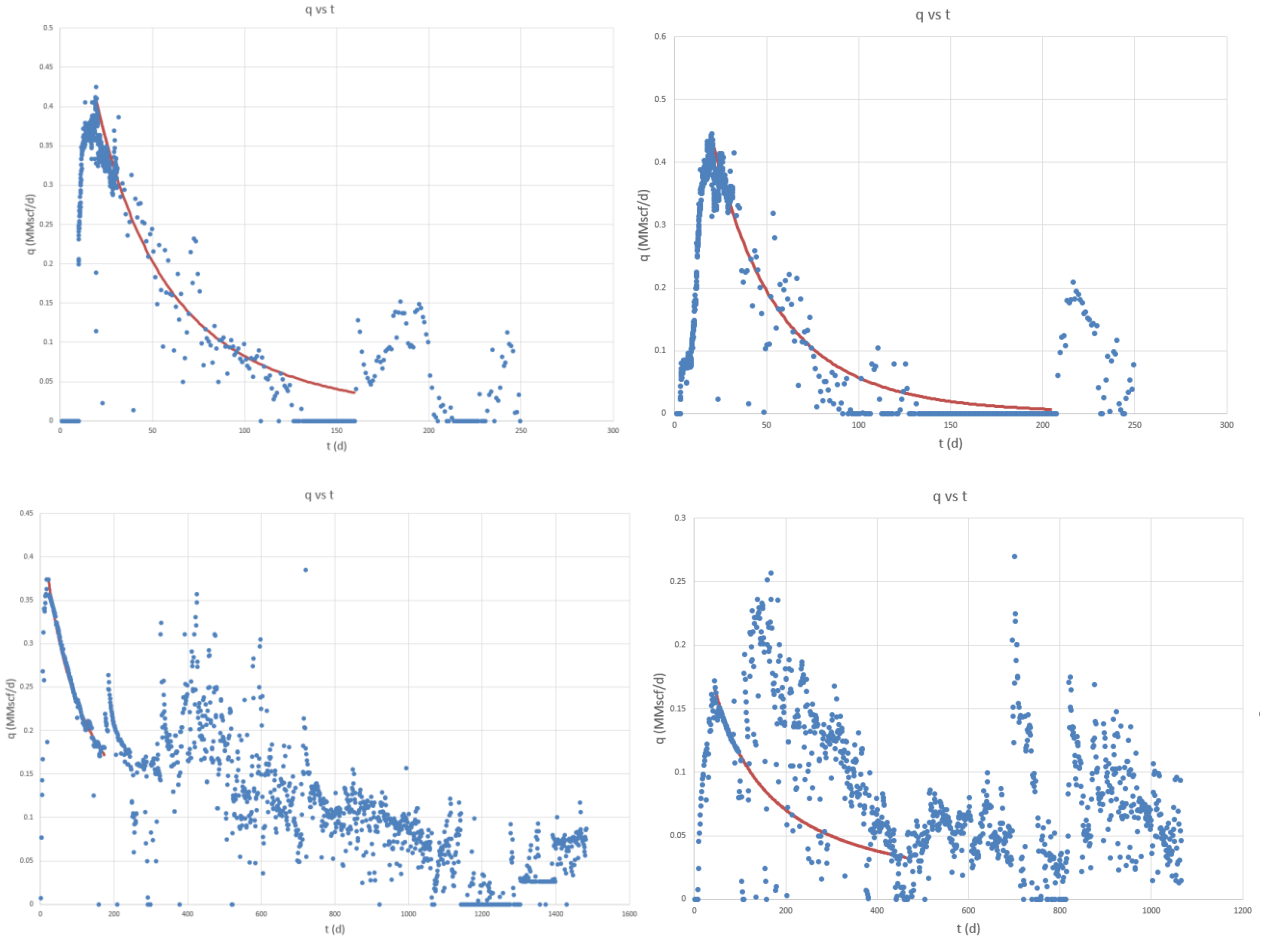


Figure A-1. Decline curve fit for the wells #2 (upper left), #3 (upper right), #4 (bottom left), #5 (bottom right).

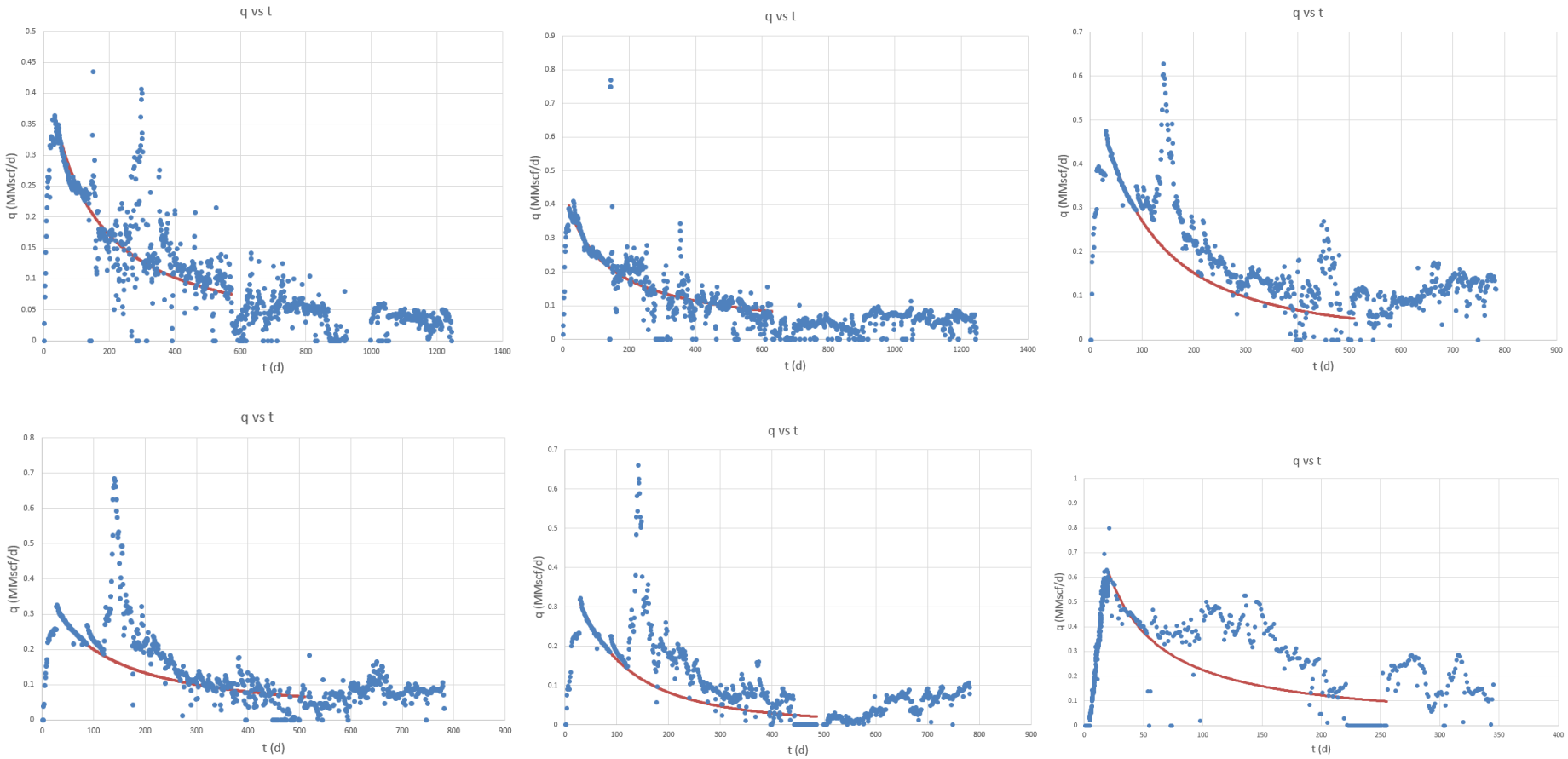


Figure A-2. Decline curve fit for the wells #6 (upper left), #7 (upper middle), #8 (upper right), #9 (bottom left), #10 (bottom middle), #11 (bottom right).

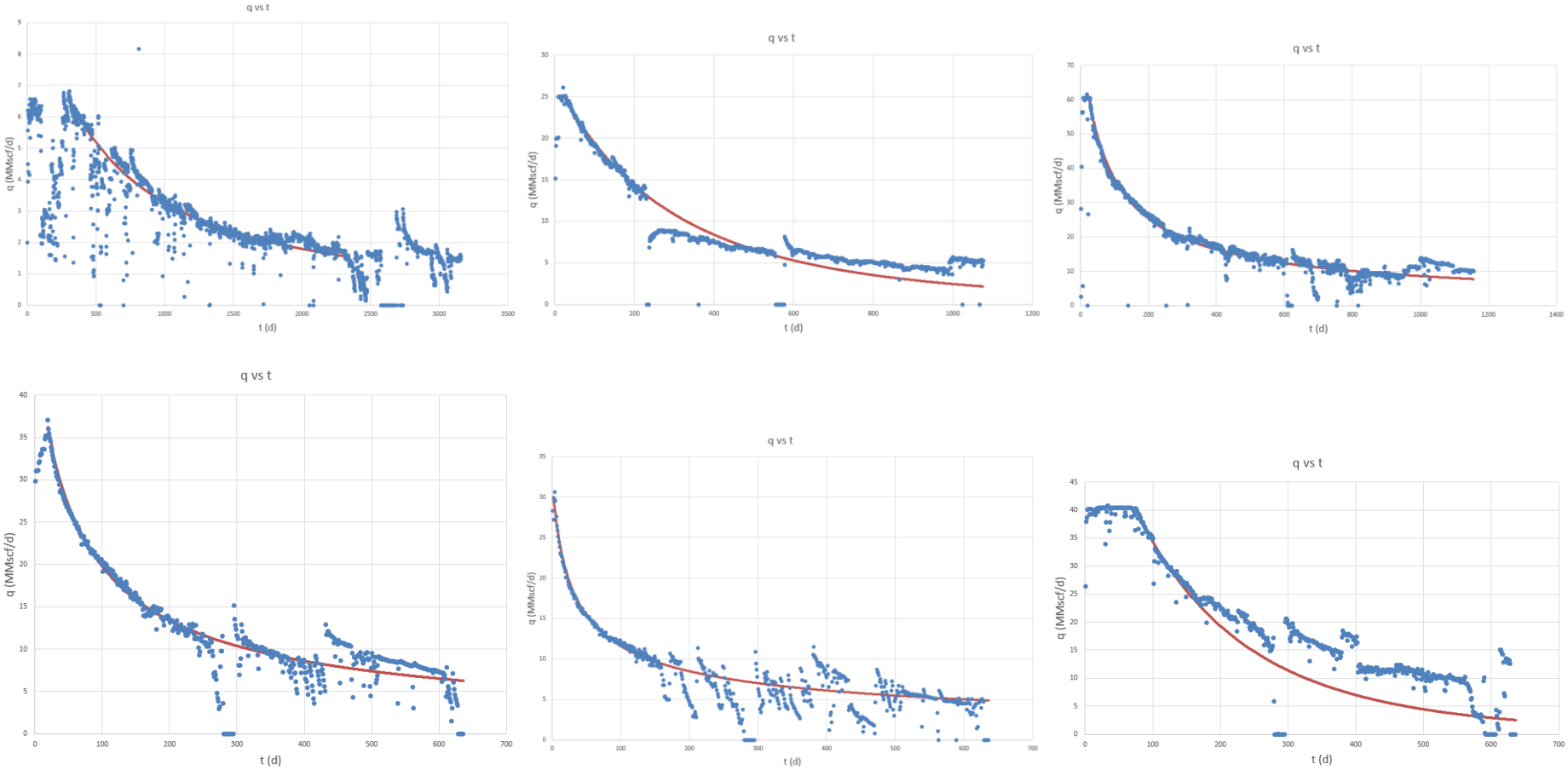


Figure A-3. Decline curve fit for the wells #23 (upper left), #24 (upper middle), #25 (upper right), #26 (bottom left), #27 (bottom middle), #28 (bottom right).

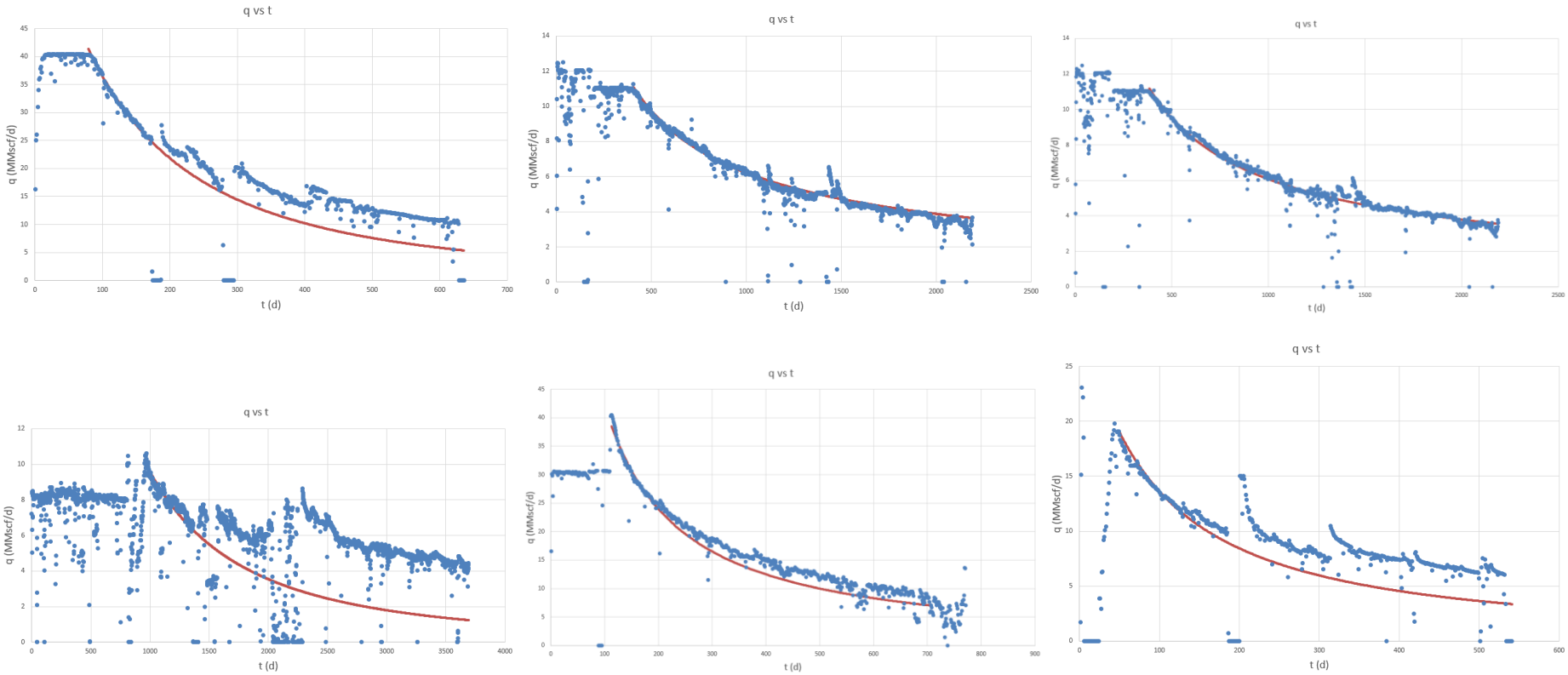


Figure A-4. Decline curve fit for the wells #29 (upper left), #30 (upper middle), #31 (upper right), #32 (bottom left), #33 (bottom middle), #34 (bottom right).

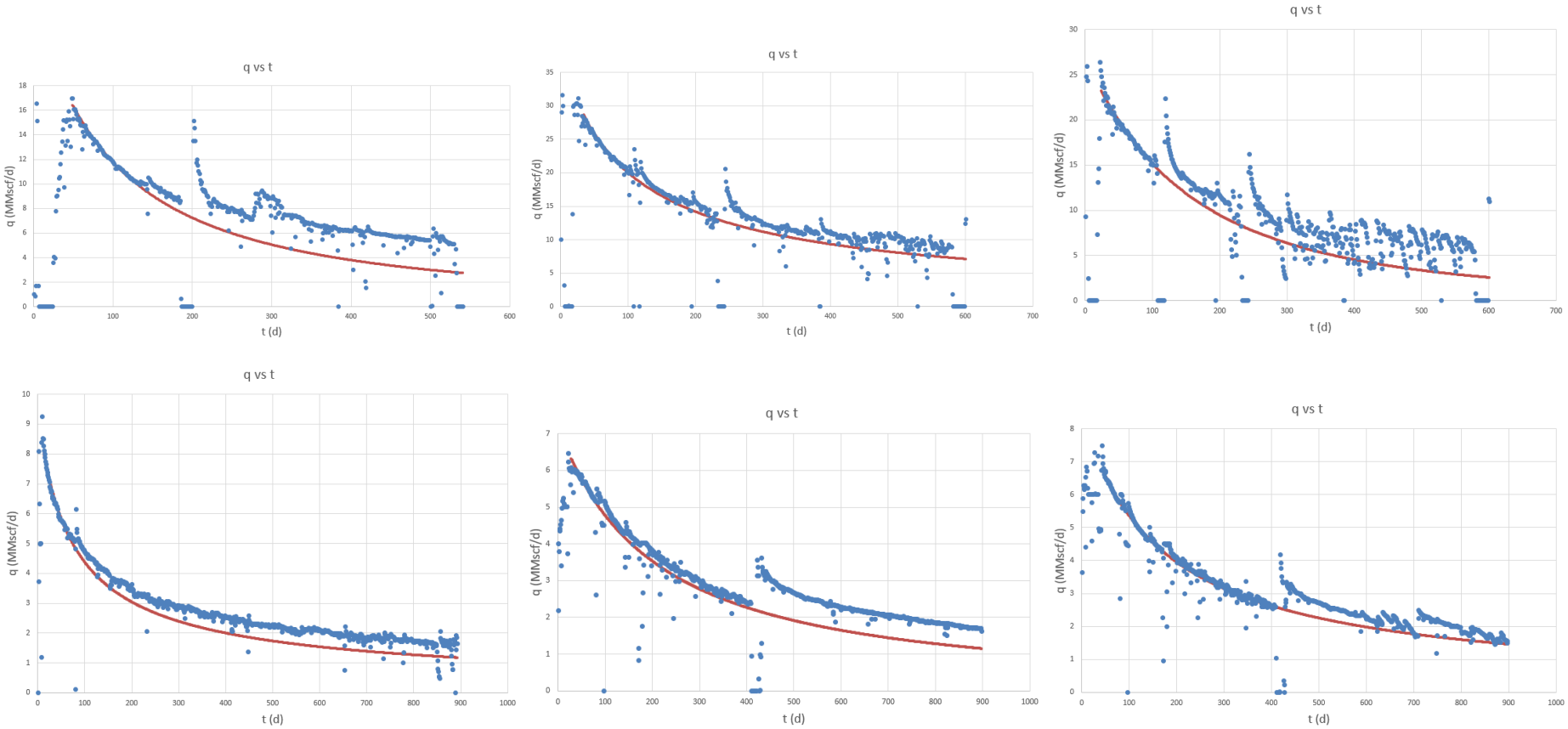


Figure A-5. Decline curve fit for the wells #35 (upper left), #36 (upper middle), #37 (upper right), #38 (bottom left), #39 (bottom middle), #40 (bottom right).

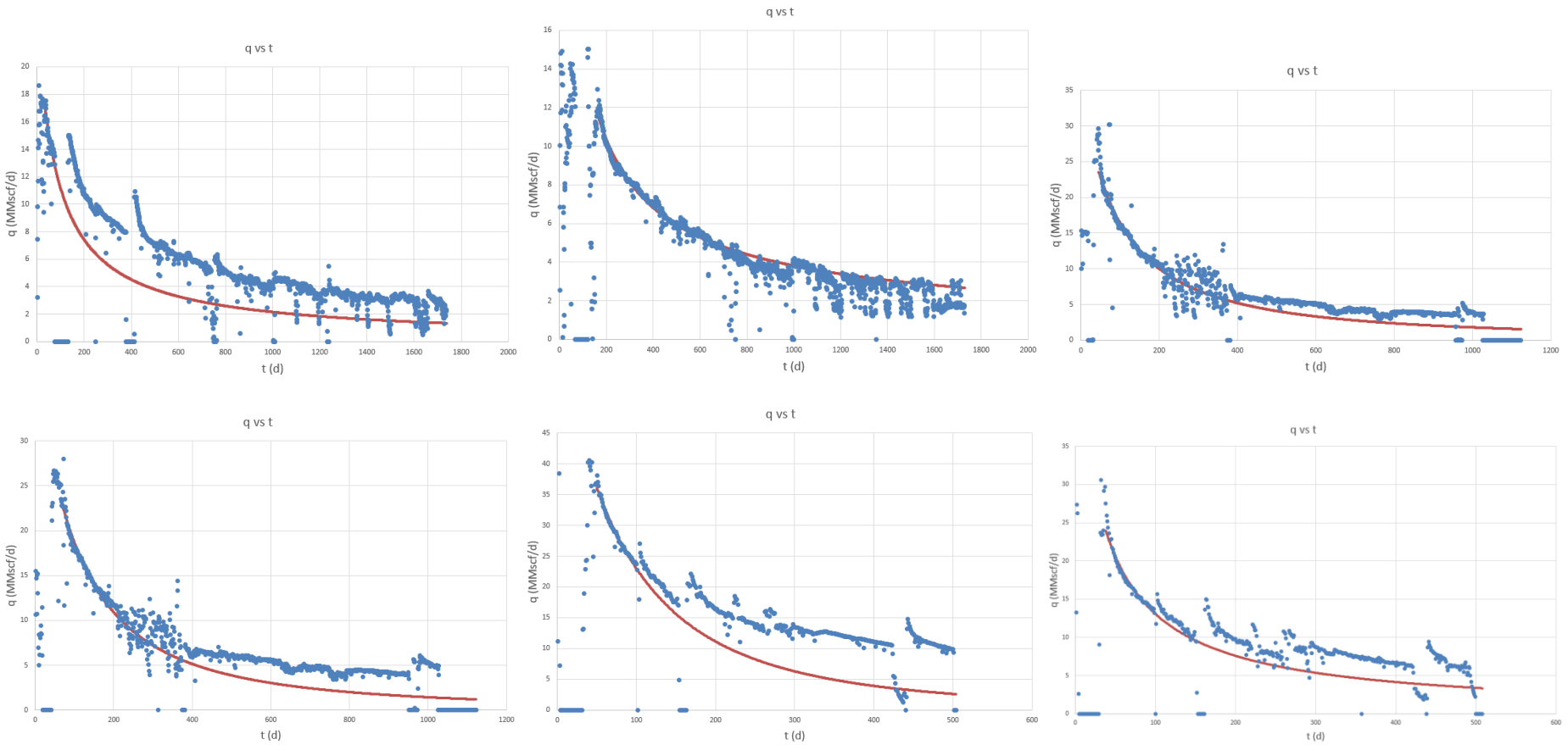


Figure A-6. Decline curve fit for the wells #41 (upper left), #42 (upper middle), #43 (upper right), #44 (bottom left), #45 (bottom middle), #46 (bottom right).

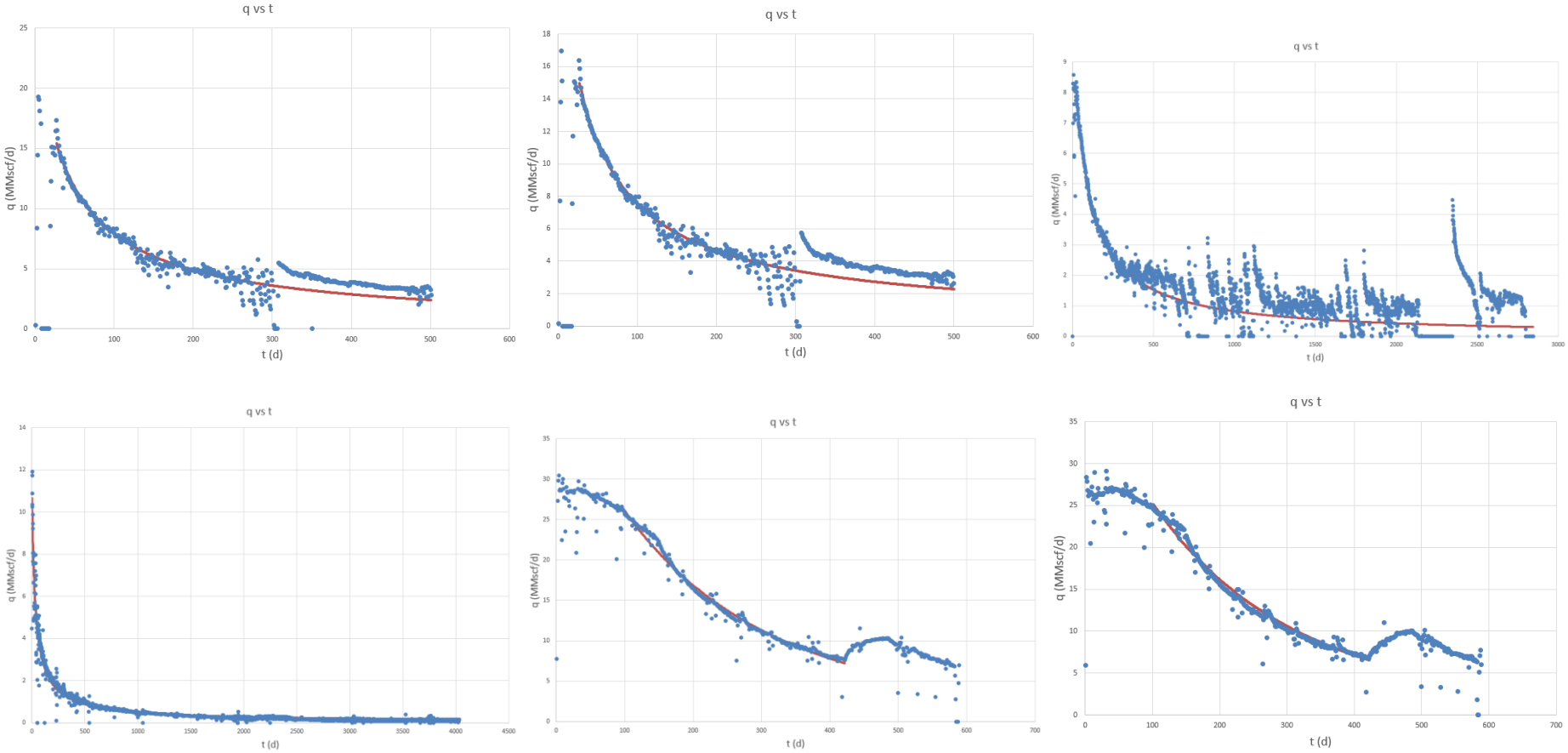


Figure A-7. Decline curve fit for the wells #47 (upper left), #48 (upper middle), #49 (upper right), #62 (bottom left), #63 (bottom middle), #64 (bottom right).

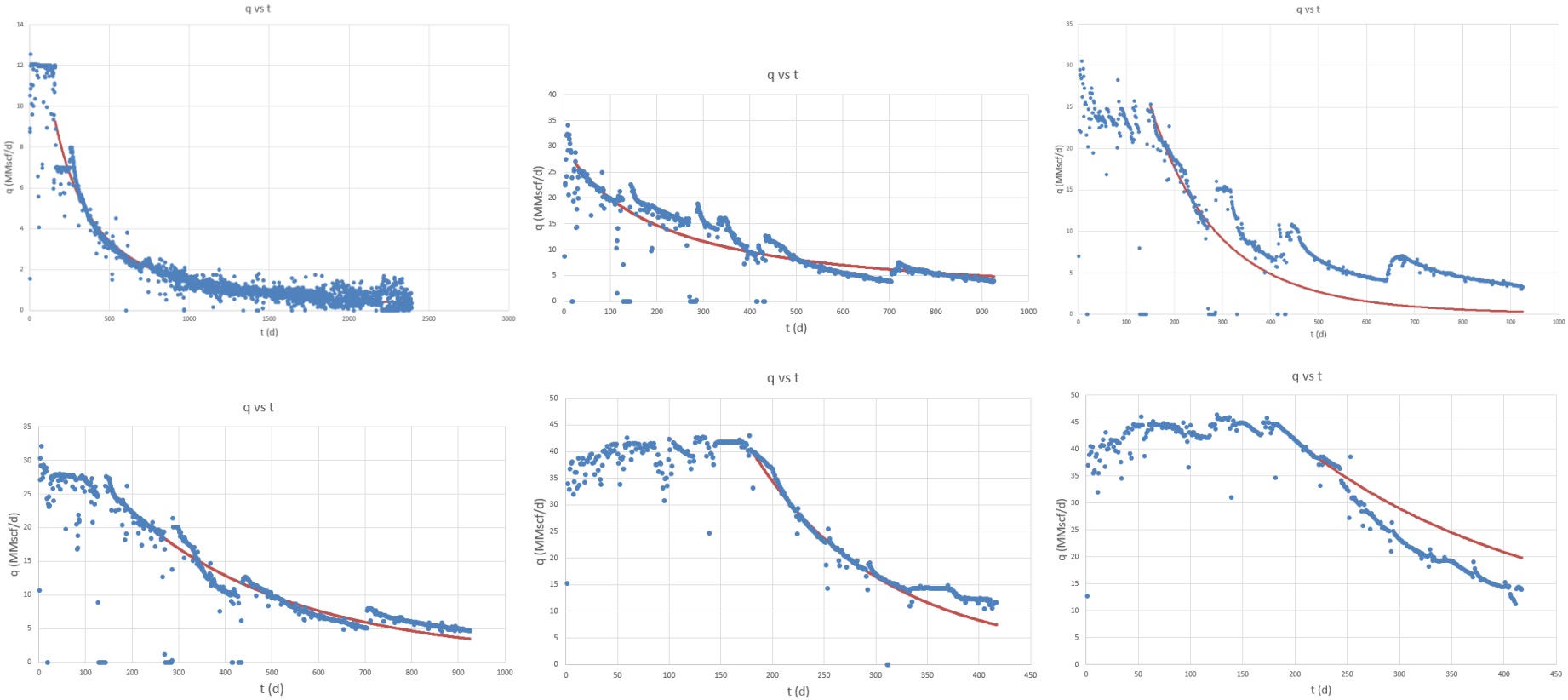


Figure A-8. Decline curve fit for the wells #65 (upper left), #66 (upper middle), #67 (upper right), #68 (bottom left), #69 (bottom middle), #70 (bottom right).

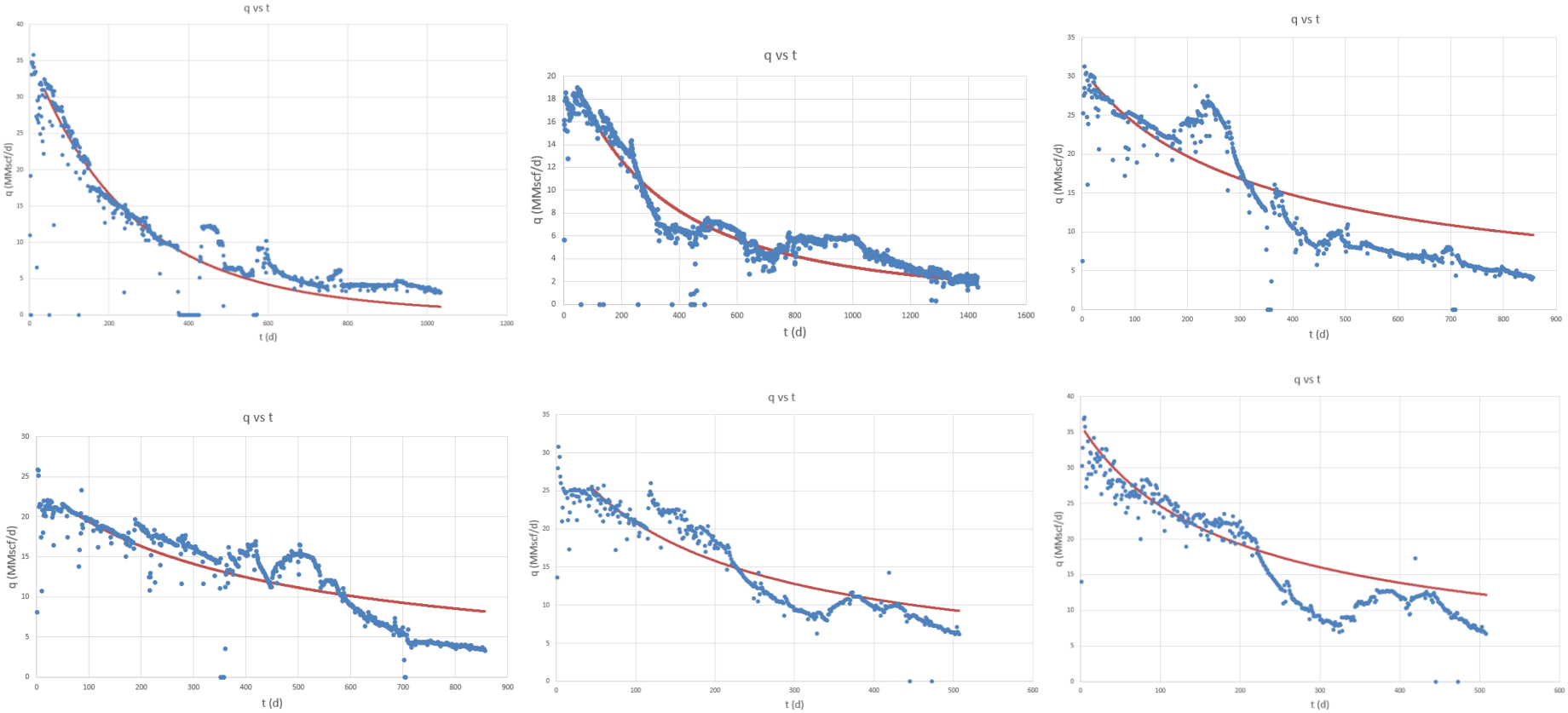


Figure A-9. Decline curve fit for the wells #71 (upper left), #72 (upper middle), #73 (upper right), #74 (bottom left), #75 (bottom middle), #76 (bottom right).