UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

# PERFORMANCE PREDICTION FOR DEEPWATER GULF OF MEXICO USING DATA MINING

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

PRIYANK SRIVASTAVA
Norman, Oklahoma
2016

PERFORMANCE PREDICTION FOR DEEPWATER GULF OF MEXICO USING DATA MINING

A THESIS APPROVED FOR THE
MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

BY

_____
Dr. Xingru Wu, Chair

_____
Dr. Deepak Devegowda

_____
Dr. Matthew Pranter

Dedicated to

*"Human Mind and its constantly changing nature"*

## Acknowledgements

I would like to thank Dr. Xingru Wu for his guidance and motivation. In spite of being busy, he was always there to answer my weird emails, no matter where in world he was. I will always be inspired to work hard like him. He does not say much, but whatever he says has so much meaning than you realize at first glance. Thanks Dr. Wu, for your motivation, for being so nice and patient. I would also like to thank Dr. Devegowda, Dr. Amin, Dr. Laxmivarahan and Dr. Pranter for their technical inputs. Special thanks to Dr. Pournik, Dr. Escobar and Dr. Rai for their support and giving me privilege for being their teaching assistant. I feel blessed having so many professor's encouragement and motivation in every semester. Acknowledgement to Dr. John Grace for providing $GOM^3$ software to be used for some of the work done in this thesis. Special thanks to Dr. Melissa and Matthew cook for helping with virtual reality concept.

I would have not survived this place without help of my friends, whom I was lucky enough to know Abhishek, Ishank, Shashwat, Beam, Son, Ankita, Sassan, Laura, Ruben Huy, Yiwen, Mohamed, Felipe, Abhinav and my only coffee friend Laura Lozano. Above all, I thank my little brother Parag, my parents and family for their constant support and encouragement.

# Table of Contents

Acknowledgements ..... iv

List of Tables ..... viii

List of Figures ..... ix

Abstract ..... xi

Chapter 1: Introduction ..... 1

1.1 Objective ..... 1

1.2 Organization of Chapters ..... 2

Chapter 2: Data Mining Methods and Applications to Petroleum Engineering ..... 4

2.1 Methods for Data Mining ..... 4

2.1.1 Steps for Data Mining ..... 4

2.1.2 Classification Algorithms ..... 7

2.1.3 Prediction Algorithms ..... 10

2.2 Application of Data Mining to Petroleum Engineering ..... 14

2.2.1 Determination of recovery factor ..... 14

2.2.2 Optimization of data acquisition/field development plans ..... 15

2.2.3 Well and Field Key Performance Drivers ..... 16

2.2.4 Improving drilling and hole conditions ..... 17

2.2.5 Production data analysis ..... 17

Chapter 3: Reservoir Forces and Dimensionless Numbers ..... 19

3.1 Factors affecting reservoir performance ..... 19

3.2 Forces operating inside reservoir ..... 19

3.3 Dimensionally Scaled Reservoir Models ..... 22
3.4 Tarner's Method for Performance Prediction ..... 27
Chapter 4: Deepwater Gulf of Mexico Exploratory Data Analysis ..... 29
4.1 Geological Background for Gulf of Mexico (GOM) ..... 29
4.1.1 Introduction ..... 29
4.1.2 Stratigraphy for GOM ..... 30
4.1.3 Petroleum System & Depositional Model for Deepwater GOM ..... 33
4.1.4 Protraction area and Leasing ..... 36
4.2 Deepwater GOM Oilfield Dataset Description ..... 37
4.2.1 Description of Key Attributes ..... 37
4.2.2 Statistics for Qualitative Attributes ..... 38
4.2.3 Statistics for Quantitative Attributes ..... 42
4.2.4 Statistics for Dimensionless numbers ..... 49
Chapter 5: Results and Conclusions ..... 51
5.1 Classification of oilfields Using Original Attributes ..... 52
5.1.1 Hierarchical Clustering (HC) ..... 52
5.1.2 K-means Clustering ..... 53
5.1.3 Self-Organizing Maps (SOM) ..... 56
5.1.4 Principal Component Analysis (PCA) ..... 59
5.1.5 Partial Least Square Regression (PLS) ..... 61
5.2 Classification of Water Drive Oilfields Using Dimensionless Numbers ..... 62
5.2.1 Clustering ..... 62
5.2.2 Self Organizing Maps ..... 65

5.2.3 PCA and PLS ..... 67

Chapter 6 Conclusions and future work ..... 73

6.1 Summary of work ..... 73

6.2 Conclusions ..... 73

6.3 Suggested future work ..... 74

References ..... 76

Appendix A: Nomenclature ..... 79

Appendix B: K-Means on log-transformed attributes ..... 82

# List of Tables

Table 1 : Data types for different attributes..... 5

Table 2 : Advantages and Disadvantages of different data mining techniques..... 7

Table 3 : Classification of factors affecting reservoir performance ..... 19

Table 4 : Fractional flow equation for different drive mechanism..... 21

Table 5 : list of Dimensionless Numbers ..... 23

Table 6 : Relation of dip angle with Field structure..... 25

Table 7 : Total reserves form sands of different geological ages (D. Nixon 1999) ..... 32

Table 8 : Description of architectural elements found in GOM..... 34

Table 9 : Biochoronostratigraphy for dGOM oilfields..... 39

Table 10 : Classification of field class based on BOEM data ..... 40

Table 11 : Classification of Field Structure (FSTRUC)..... 40

Table 12 : Classification of Primary and secondary trap type..... 41

Table 13 : Classification for drive mechanism..... 42

# List of Figures

Figure 1 : Generalized process of data mining ..... 4

Figure 2 : Workflow for SOM models ..... 10

Figure 3 : Graphical representation of PCA operation ..... 11

Figure 4 : Graphical schematic for PLS algorithm (Source: camo.com) ..... 13

Figure 5 : Workflow for application of Tarner's method of performance prediction ..... 28

Figure 6 : USGS Map of Gulf of Mexico ..... 30

Figure 7 : Local Stratigraphic and N-S Deposition X-section for GOM sands ..... 31

Figure 8 : Three-Dimensional description of Gulf of Mexico ..... 33

Figure 9 : X-section for Lower Tertiary Wilcox Trend ..... 36

Figure 10 : Filtered list of attributes used for data mining operation in dGOM ..... 38

Figure 11 : Distribution of Recovery Factor and Permeability ..... 43

Figure 12 : Distribution of Porosity and Water Saturation ..... 44

Figure 13 : Distribution of Pressure gradient (SDPG) and Solution GOR (RSI) ..... 45

Figure 14 : Distribution of Oil Thickness and Oil Area. ..... 46

Figure 15 : Distribution of Oil in-place and Cumulative produced oil ..... 47

Figure 16 : Over-estimated OIP as compare to physical production ..... 48

Figure 17 : Distribution of Capillary Number ($N_{pc}$) and Gravity Number ($N_g$) ..... 49

Figure 18 : Distribution of Density number ($D_n$) and Aspect ratio ($R_l$) ..... 50

Figure 19 : Workflow for Prediction of Recovery Factor in dGOM oilfield dataset ..... 51

Figure 20 : Workflow for Classification of Recovery factor ..... 52

Figure 21 : Dendograms of testing dataset ..... 53

Figure 22 : Elbow Plot for determination of number of optimum clusters. ..... 54

Figure 23 : K-means clustering on all 395 oilfields in dGOM ..... 55

Figure 24 : Distribution of original attributes in individual clusters ..... 56

Figure 25 : Validation of convergence in iterations ..... 57

Figure 26 : The U-matrix plot to identify clusters within the SOM group ..... 58

Figure 27 : Distribution of each attribute across the latent space ..... 59

Figure 28 : Bi-plot for K-means clusters in a principal component (PC) space ..... 60

Figure 29 : PLS match and circle of correlation for Cluster 1, 2 and 3 ..... 62

Figure 30 : Distance based dendogram on normalized dimensionless numbers ..... 63

Figure 31 : Optimum number of clusters for K-means clustering ..... 64

Figure 32 : Scatter Plot for Dimensionless numbers ..... 65

Figure 33 : U-Matrix plot for clusters obtained using SOM ..... 66

Figure 34 : SOM iterations reaches convergence at after 70 cycles ..... 66

Figure 35 : Heat map for dimensionless numbers in SOM latent space ..... 67

Figure 36 : Clusters of dGOM oilfields having water drive in PC space ..... 68

Figure 37 : PLS result for Cluster-1. $R^2$ 0.92 ..... 69

Figure 38 : PLS result for Cluster-2. $R^2$ 0.61 ..... 69

Figure 39 : PLS result for Cluster-3. $R^2$ 0.1 ..... 70

Figure 40: PLS result for Cluster-4. $R^2$ 0.4 ..... 70

Figure 41 : Distribution of Normalized dimensionless numbers ..... 71

Figure 42: Distribution of Normalized recovery factor ..... 72

## Abstract

The estimation of recovery factor is important in every stage of hydrocarbon development, and multiple traditional techniques are available for engineering application for particular fields. However, the estimated recovery factor can be very different using these methods. This is particular true for deepwater development as the parameters associated with the recovery factor estimation have significant uncertainties. The objective of this study is to apply the data mining technologies based on the data from the developed fields in Gulf of Mexico.

Using database of 395 Deepwater Gulf of Mexico (dGOM) oilfields with 84 attributes, set of dimensionless variables are calculated; and these dimensionless variables are used as the input for data mining with an aim of obtaining the recovery factors. A subset of 59 oilfields that have water drive mechanism are selected for discovering the generalized correlation for recovery factor using data mining techniques. In the study, a variety of data mining techniques such as K-means and principal component analysis are used for classifying oilfields into four categories. Subsequently, partial least square (PLS) regression is used to relate the dimensionless variables to the recovery factor from sparse data in dGOM. However, not all clusters show very high coefficient of correlation, hence limiting the applicability of this method. This study shows that dimensionless numbers, together with data mining techniques, can be very useful to predict field behavior in terms of recovery factor for sparse datasets with widely scattered reservoir properties. This information can be further used for the preparation of data acquisition and risk assessment plans to set up a framework for decision-making on risks and uncertainty for optimizing reservoir management and production forecast.

# Chapter 1: Introduction

*"He who has a why to live, can bear any how" (Friedrich Nietzsche)*

## 1.1 Objective

**Problem statement**

Recovery factor is the critical parameter to justify reservoir development and economics. At any stage in reservoir development, reservoir management team requires a realistic estimate of recovery factor that is controlled by many geological and engineering factors. Typically, the recovery factor of a field in exploratory stage is estimated using analogs or material balance. While for fields in appraisal or development phase, it is estimated using numerical simulation or decline curve analysis. In addition, comprehensive simulations can be very costly and time consuming for large databases of reservoirs. All the above approaches are deterministic that rely on accuracy of input dataset. In addition, results obtained from these models are subjective, demanding good history matching by tuning a large number of known and unknown parameters, which leads to non-unique solutions. This study uses data analytics for development of easy to implement generalized recovery factor correlations for reservoirs in terms of dimensionless numbers, these correlations can be quickly used in absence of information and time, required for comprehensive simulation.

**The new approach**

A new approach of integrating dimensionless groups with data mining techniques is applied in this thesis for the estimation of recovery factor and performance prediction of reservoirs. From a database of more than 1300 deepwater reservoirs, models are first scaled using dimensionless groups developed for reservoirs with immiscible

displacement of oil by water. Subsequently, these dimensionless groups along with data mining techniques of K-means, PCA and PLS are used to classify and develop a generalized correlation for recovery factor.

## 1.2 Organization of Chapters

This thesis is organized in six chapters. Chapter 2 starts with methods used for pattern recognition from big datasets and its applications in petroleum engineering. This chapter is divided into two sections. First section describes mathematics and fundamentals behind various methodologies for classification. Second section illustrates the use of data mining in petroleum engineering viz. prediction of recovery factor, optimization of data acquisition and field development plans, improving drilling and well performance and production data analysis.

Chapter 3 contains the summary of various reservoir forces and energy on which performance of a typical immiscible displacement is dependent. First section describes grouping of factors affecting reservoir performance into inherent reservoir factors and external controllable factors. Second section describes details of inherent reservoir factors by describing physical laws to capture these factors using fractional theory (Leverett, 1941). Third section introduces dimensionless numbers and their importance in scaling of reservoir models. Last section in this chapter defines theory of Tarner's method for performance prediction of reservoir under depletion drive mechanism.

Chapter 4 starts with geological description of Gulf of Mexico (GOM) basin along with stratigraphy, depositional model and petroleum system formation. It explains basis of Bureau of Ocean and Energy Management (BOEM) dataset in terms of different

protraction areas and leasing blocks. Second section explains GOM dataset with statistics for various quantitative and qualitative attributes used in data mining process.

Chapter 5 gives detail of the results obtained by applying various data mining algorithms to water drive reservoirs in dGOM oilfields. Data mining algorithms (Clustering, K means, SOM, PCA and PLS) are applied to original attributes in 395 dGOM oilfields and subsequently to 59 reservoirs having water drive. Latter case study uses dimensionless numbers for development of correlation to predict recovery factor in water drive reservoirs of dGOM oilfields. This thesis ends in chapter 6 that describes contribution and conclusions of the present work, with future research directions.

# Chapter 2: Data Mining Methods and Applications to Petroleum Engineering

*"All models are Wrong, Some are useful" (George Box)*

## 2.1 Methods for Data Mining

### *2.1.1 Steps for Data Mining*

Data mining is a soft computing knowledge discovery process which is tolerant of uncertainty and imprecise information (Nikravesh 2004) and thus, can be used with noisy data collected in reservoir exploration and production. A meaningful data mining workflow relies on three qualities (1) Data cleaning (2) Well-defined target to predict (3) Good validation to avoid over fitting. The general process of data mining is divided into four steps as shown in **Figure 1**

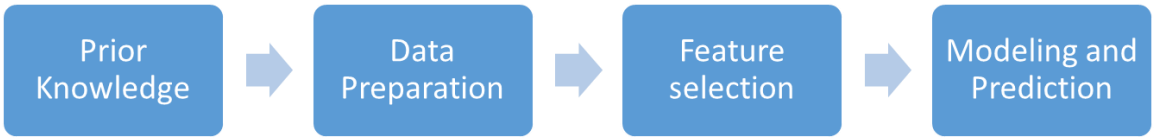



**Figure 1 : Generalized process of data mining**

Prior knowledge refers to the information known apriori about the system. This aims to define the scope and objective of the problem, understanding the journey of data from its acquisition to reporting is a critical part of data mining process. For this thesis, the apriori knowledge suggests field performance to be dependent on geology, reservoir properties, and well performance. However, it is difficult to delineate specific factors under each category that affects the recovery factor and by how much. Based on data mining algorithms this thesis gives the quantitative contribution of each of the factors in recovery factor estimation for Deepwater Gulf of Mexico oilfields.

Second step is data preparation, where data needs is checked for its quality, outliers, redundant, un-useful and missing values. Data preparation requires conversion of different categorical and character data types to numerical quantities that be can be analyzed using required algorithms. All attributes are scaled and normalized before application of any model. This insures that any one particular attribute does not dominate the results. For Deepwater Gulf of Mexico (dGOM) oilfields a set of 84 attributes is divided into 5 categories viz. (1) Identification tags & dates (2) Geological attributes (3) Reserves and Production (4) PVT parameters (5) Completions. **Table 1** gives a classification of each variable into categorical, numerical and string data types. Chapter#4 describes this data in detail.

Table 1 : Data types for different attributes

| **Categorical** | **Numeric** | **String/Dates** |
|---|---|---|
| DRIVE | All reserves & production category | SANDNAME |
| FTRAP1 | All petro-physical properties | POOLNAME |
| FTRAP2 | All PVT parameters | FDYEAR |
| CHRONOZONE | | WELLAPI |
| FCLASS | | EIAID |
| SD_TYPE | | SDYEARH |
| FSTRU | | PLAYNAME |
| RES_TYPE | | SAND |

In third step, data is reduced to its essential characteristics or features by removing all unnecessary and redundant attributes. More data may increase the confidence in the derived model, but it makes it complicated by "curse of dimensionality" thus degrading performance (Vazirgiannis, Halkidi, & Gunopulos, 2012). Based on the domain knowledge and processed data, parameters needs to be configured to suite appropriate data mining algorithms (Holdaway 2014). Partial contributions of attributes affecting the outcome needs to be accounted. In this thesis, feature selection is done by use of dimensionless numbers. Since dimensionless numbers are based on geometry (aspect ratio), dynamic forces (density number) and kinetic forces (fluid flow velocity). The use of dimensionless numbers provides a set of physical attributes on which performance of any general reservoir model should be dependent.

Last step is the Modeling, where suitable algorithms are selected for new knowledge discovery. A model is an abstract representation of the data and its relationships in a given data set (Kikani, 2013). Data mining methods are based on the idea of identifying and describing interesting patterns of extracted knowledge. Selection of data mining algorithm is dependent on the objective question asked, domain expertise and data availability. **Table 2** provides details of various data mining methods used in this thesis with advantages and disadvantages for each of them.

Table 2 : Advantages and Disadvantages of different data mining techniques

| Mining Technique | Pros | Cons |
|---|---|---|
| K- Means Clustering | Easy to implement. Saves computational time. | Initialization of cluster centroid can get stuck in local optima. Visualization may be difficult for high-dimensional dataset |
| Principal component analysis | Robust tool for dimensionality reduction of the multivariate dataset. | No physical meaning can be assigned to principal components. Applicable only for correlated attributes. Non-parametric method. So cannot incorporate prior knowledge. Leads to loss of information. |
| Identification Trees | Requires minimum effort in data preparation step, can handle wide data types, not sensitive to outliers. | Over fitting can be a major issue without pruning. Sensitive to statistical irregularities of the training set. |
| Self-Organizing maps | Unsupervised learning algorithm. Non-linear generalization of PCA. | Not suitable for a very large dataset. They are just a dimensionality reduction tool. Classification must be done using other algorithms. Difficult to interpret. |
| Partial Least regression (PLS) | Best suited for multi collinearity correlated datasets. | Assumes linearity of correlation. It is not necessary that maximum dynamics occur in direction of maximum variance. |

### *2.1.2 Classification Algorithms*

Classification techniques predicts target variables based on grouping the selected features. A well-defined set of classes and a training set of pre-classified examples characterizes the classification (Kotu and Deshpande 2014). Some common techniques for classification task are hierarchical clustering, K-means clustering, Linear Discriminant, Naïve Bayes and Logistic regression. Clustering is the method to organize big dataset into sensible groupings. One of the main limitation of clustering is uncertainty in number, shape and size of clusters. In addition, visualization of clusters for a high dimensional dataset could be a difficult task depending on the resolution of visualization method.

**Agglomerative Hierarchical Clustering (HC)**

This technique organizes data in pairs and groups based on dissimilarity matrix, which are shown in form of dendograms. An advantage of this method is it does not need a-priori information about number of clusters. It is necessary to scale and center all attributes before applying this method to ensure unbiased clustering of original dataset. This algorithm is computationally expensive for large datasets (Kantardzic 2011).

**K-Means clustering algorithm**

It is a decision boundaries formation algorithm that obtains decision boundaries between the dataset instead of clustering structure. The purpose of K-means clustering is to optimize the following objective function

$$\mathbf{E} = \sum_{\mathbf{i=1}}^{\mathbf{c}} \sum_{\mathbf{x} \in \mathbf{c_i}} \mathbf{d}(\mathbf{x}, \mathbf{m_i}) \qquad \textbf{Eqn. 1}$$

$m_i$ = cluster centroid for cluster $C_i$

d (x, $m_i$) = Euclidean distance between a point x and centroid ($m_i$)

The optimum number of clusters is reached when there is no relative reduction in 'SSW' defined as

$$\mathbf{SSW} = \sum \sum_{\mathbf{j=1}}^{\mathbf{n}} \left(\mathbf{X_{1j}} - \boldsymbol{\mu}_{\mathbf{x}}\right)^{\mathbf{2}} \qquad \textbf{Eqn. 2}$$

'n' is the number of data points in a cluster. $X_{1j}$ is dissimilarity between cluster centroid and each data point and $\mu_x$ mean of all data points in a cluster.

Limitation of K-means is that it may require multiple random realizations, if centroids are stuck in local minima. In addition, all attributes need to be quantitative in order to obtain proper decision boundary for partition. Other complex clustering algorithms such as PAM (partitioning around meloids), Fuzzy clustering, density based

clustering and grid-based clustering are also available in literature but is not used in this thesis.

**Self-Organizing Maps (SOM)**

SOM are unsupervised competitive learning algorithm, which is a simpler form of neural nets useful in visualization of multidimensional dataset in 2-D space. The main objective of SOM is to preserve the topology of multidimensional data; i.e. to get a new set of data from the input data such that the new set preserves the structure (clusters, relationships etc.) of the input data. Steps for a SOM algorithm is illustrated in **Figure 2**

consider a random vector V to be tested in SOM

$$\begin{bmatrix} \boldsymbol{\alpha_1} \\ \boldsymbol{\alpha_2} \\ \boldsymbol{\alpha_3} \end{bmatrix} = \overrightarrow{\boldsymbol{V}} \qquad \textbf{Eqn. 3}$$

The best matching unit (BMU) is selected by minimization of following function

$$\boldsymbol{BMU} = \sum(\boldsymbol{R_i} - \boldsymbol{\beta_i})^2 \qquad \textbf{Eqn. 4}$$

$R_i$ = node weight vector $\beta_i$ = BMU vector

Once BMU is matched, modification in weights of neighboring nodes of BMU is done based on shrinking radius of the neighborhood.

$$\begin{bmatrix} \boldsymbol{\alpha_1} \\ \boldsymbol{\alpha_2} \\ \boldsymbol{\alpha_3} \end{bmatrix}_{\boldsymbol{New}} = \begin{bmatrix} \boldsymbol{\alpha_1} \\ \boldsymbol{\alpha_2} \\ \boldsymbol{\alpha_3} \end{bmatrix}_{\boldsymbol{Old}} + \boldsymbol{\lambda_1} \left[ \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \boldsymbol{\beta_3} \end{bmatrix}_{\boldsymbol{BMU}} - \begin{bmatrix} \boldsymbol{\alpha_1} \\ \boldsymbol{\alpha_2} \\ \boldsymbol{\alpha_3} \end{bmatrix}_{\boldsymbol{Old}} \right] \qquad \textbf{Eqn. 5}$$

Where,

$\lambda_1 = Weight\ of\ the\ node\ , this\ quantity\ changes\ with\ distance\ of\ block\ from\ BMU$

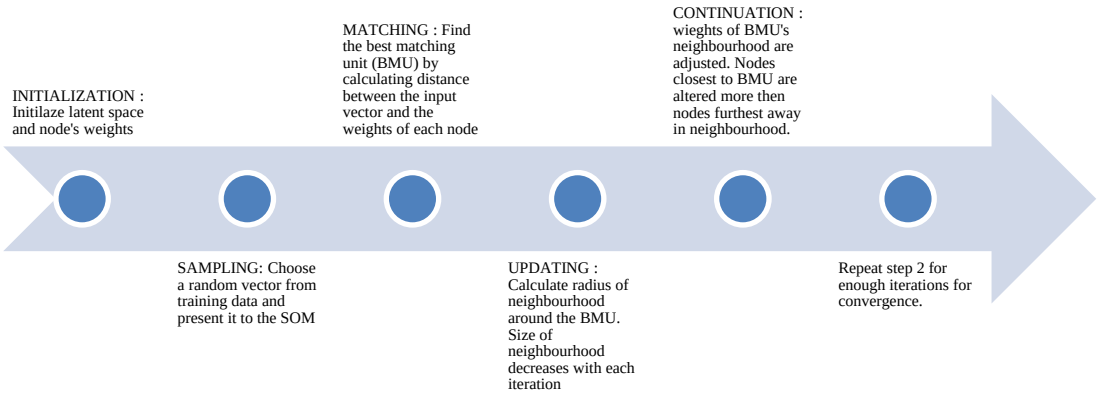



**Figure 2 : Workflow for SOM models**

The iteration is continued until stabilization of $E_{QE}$ that is defined as

$$\boldsymbol{E_{QE} = \frac{1}{M}\sum_{P=1}^{M}\left\|X_P - M_{C(P)}\right\|} \qquad \textbf{Eqn. 6}$$

The most popular scheme for visualization of SOM is unified distance matrix (U-matrix). U-matrix is the difference of similarity between adjacent blocks in latent space as the SOM is colored by the values of U-matrix elements. This visualization is used for obtaining number of clusters in dataset, as data points with similar adjacent nodes in latent space will have same color in U-matrix plot.

### *2.1.3 Prediction Algorithms*

Prediction aims at estimating the future values of some interesting variables based on other variables correlated to it (Vazirgiannis, Halkidi, and Gunopulos 2012). Predictive analytics tries to use different validated models for predicting future probabilities and trends. These algorithms rely on the discovery of patterns obtained during classification

task. The following algorithms were used in this study to predict the recovery factor in dGOM oilfield dataset.

**Principal Component Analysis**

It is a method to reduce the dimensionality of a dataset with a large number of correlated attributes. Graphically, PCA rotates original axes to the direction of maximum variability in dataset as shown in **Figure 3**. All the attributes present in dataset are transformed to a new set of variables based on scaled covariance matrix. The principal components (PCs) are uncorrelated and ordered. So that first few contain most of the variance of the original data set. Mathematically, the principal components are the eigenvectors of the covariance matrix of the original attributes.

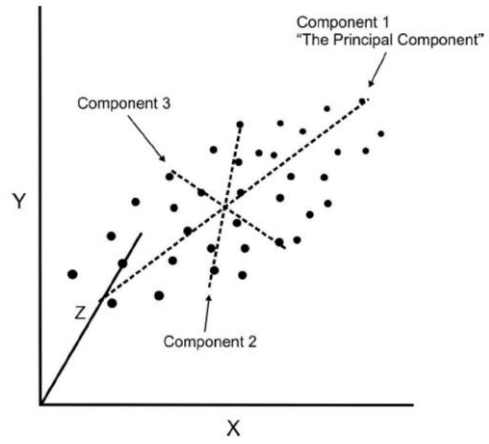



**Figure 3 : Graphical representation of PCA operation**

Consider a matrix $X_{mxn}$, its covariance matrix is represented as:

$$C_x = \frac{1}{n} XX^T \qquad \textbf{Eqn. 7}$$

For a symmetric non-singular matrix, this is written:

$C_x = EDE^T$ , E being eigenvector for the covariance matrix; and principal component can be given by,

$$\boldsymbol{P = E^T X} \qquad \textbf{Eqn. 8}$$

PCA algorithms results in output of loadings (eigenvalues) for each individual attributes. Higher loading implies higher importance of that component. Component scores calculation is done by multiplication of scaled value of original component with component loading, summed over all variables. Visualization of bi-plots of component scores with original variables used in PCA on a 2-D or 3-D graph revels inherent clusters present in the dataset. Some of the main limitations of PCA is its assumption of linearity in dataset. In addition, PCA will lead to higher loading for components with large variance as compare to lower variance components. Also, it is not necessary that direction of maximum variance will contain good features for classification.

**Partial Least Square (PLS)**

PLS is a technique that reduces the predictors (X) to a smaller set of uncorrelated components and performs least square regression on these components. This algorithm tries to identify the latent factors (combination of variables) which account for most of the variation in the response. In constructing the principal components of X, the PLS algorithm iteratively maximizes the strength of the relation of successive pairs of X and Y component scores by maximizing the covariance of each X-score with the Y variables. Graphically, PLS algorithm is explained in **Figure 4**

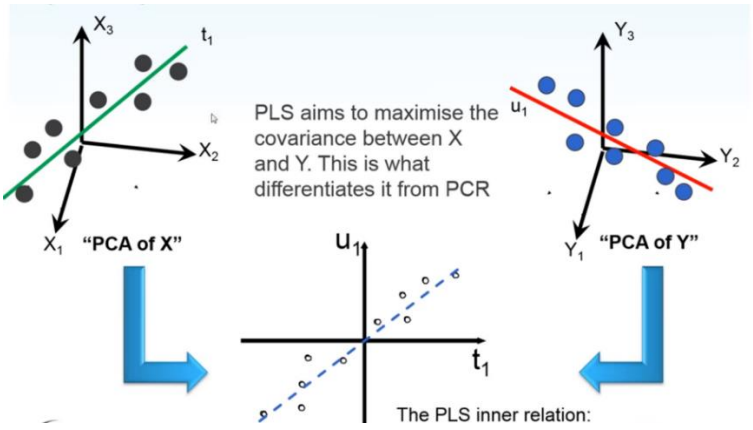



**Figure 4 : Graphical schematic for PLS algorithm (Source: camo.com)**

Mathematically, consider a model on which PLS is applied:

$$\boldsymbol{Y = B * X + E} \qquad \textbf{Eqn. 9}$$

This model is written in form of latent variables (principal components)

$$\boldsymbol{X = T * P' + E_o} \qquad \textbf{Eqn. 10}$$
$$\boldsymbol{Y = U * Q' + F_o} \qquad \textbf{Eqn. 11}$$

'P' & 'Q' are orthogonal matrices and 'T' & 'U' are loading matrix. The multivariate linear regression is performed on 'T' and 'U' as follows:

$$\boldsymbol{U = T * B + F1} \qquad \textbf{Eqn. 12}$$
$$\boldsymbol{Y = X * (P * B * Q')} + \boldsymbol{F} \qquad \textbf{Eqn. 13}$$

Overall regression coefficient is P*B*Q′.

PLS is a good algorithm to describe multicollinearity in dataset. While, if data is very high dimensional (p (no of features)>n (no. of observation))) it suffers from overfitting and bias-variance trade-off.

## 2.2 Application of Data Mining to Petroleum Engineering

### *2.2.1 Determination of recovery factor*

Classic reservoir engineering assumes that recovery factor is dependent on rock properties, fluid properties, geological structures and mode of production (Craft, 1959). Currently recovery factor is estimated by following five techniques (1) Using analogous reservoir (2) Using Volumetric calculations (3) Material balance calculations (4) Decline curve analysis (5) Numerical reservoir simulation. Approaches like analogy and reserve volumetric are highly susceptible to errors. Decline curve analysis is a mathematical technique that does not take physics of flow into account. Sophisticated computer models requires accurate descriptions of subsurface for recovery factor estimation. However, all of these are deterministic approaches where accurate information is needed for building a geological model. This dilemma calls for an alternative approach in handling the inaccurate data.

In 1936, Schilthuis was able to calculate recovery factor based on tank model material balance equations. Buckley & Leverett (1942), Tarner (1944) and Muskat (1945), who calculated oil recovery to be expected for a particular rock and fluid characteristic at any stage of depletion under different drive mechanisms, improvised the method. These methods along with estimating recoveries, provided causes of low recovery in terms of forces operating inside the reservoir. It became possible to develop correlations to recovery factor for many fields based on data specific models. Using datasets and multivariate statistics, many researchers (Craze & Buckley (1945), Vietti et. al (1945), Muskat & Taylor (1946), Guthrie & Greenberger (1955)) have developed

different linear correlations to estimate recovery factor as a function of rock and fluid properties.

Data analytics describe the physical system from set of observations. Since use of data analytics does not require apriori knowledge of reservoir models, correlations of field performance is developed with aids of trends and patterns revealed by data. Data mining integrates and visualizes all the available structured and unstructured data from a static geological description to dynamic fluid flow behavior.

### *2.2.2 Optimization of data acquisition/field development plans*

Another vital information for field development planning is, correlating use of data with its necessity in decision-making process. Information is only valuable if it allows us to improve decision-making (Kikani, 2013). Kikani (2013) describes Value of Information (VOI) as a way to quantify the use and subsequent benefit of data collected tasks. Often the decision to develop newly discovered hydrocarbon resources must be made with insufficient or analogous information. This uncertainty can be reduced by using data analytics to minimize collection of irrelevant data while obtaining maximum information about the reservoir. Since data acquisition in upstream industry is a very expensive process, especially in deepwater environment, one of the most important uses of data science is optimization of data acquisition plans and removal of redundant acquisition techniques used in past ventures.

Data mining methods provides a novel way to use all the data that geoscientists and engineers generate for finding and producing hydrocarbons efficiently. This helps to automate simple decisions & guide harder ones, ultimately reducing the risk and resulting in finding and producing more oil & gas. This thesis demonstrates application of data

mining techniques to reservoir engineering for prediction of field performance especially recovery factor based on sparse dataset recorded by various operators in Deepwater Gulf of Mexico (dGOM) over number of years.

### *2.2.3 Well and Field Key Performance Drivers*

Reservoir characterization is defined as a use and integration of huge amount of data obtained in exploratory and appraisal phase of reservoir development in order to delineate and obtain realistic production forecast from reservoir models (Teh, 2012). Industry standard approach is assimilation and interpretation of representative knowledge from reservoir rock and fluid data acquisition programs into a reservoir simulation model.

Modern technologies and high resolution sensors has made it possible to monitor wells in real time. With advent of newer computationally intensive technologies like Distributed Acoustic Sensing (DAS), Distributed Temperature System (DTS); the use of analytics for proper utilization of all the collected information has become imperative (Holdaway 2014). These high resolution sensors collect data of order of terabytes every day for example: permanent downhole gauges can monitor pressure, temperature or flow rate changes with sampling rate as low as 2 seconds leading to more than 16 million discrete time and pressure values per year (Chorneyko 2006). These huge datasets when integrated across the field have power to generate hidden trends and patterns based on proper data mining model. The transformed model is subsequently used for classifying conditions in individual wells or exploratory fields where sufficient information is not available.

Bob Shelley (2007) applied data mining technologies like artificial neural networks (ANN) and self-organizing maps to identify production drivers for 32 high

temperature wells. They were able to observed significant previously unknown patterns for generation of reasonable prediction model while using ANN models, thus was able to identify key parameters on production performance. Data mining use this integrated dataset to generate new knowledge of unknown drivers for well and field performance, develop new methodology for prediction of well productivity and recovery factor, defines best practices in completion and reservoir management. Thus, enabling development of performance metrics for field to be a successful producer (Abou-Sayed 2012).

### *2.2.4 Improving drilling and hole conditions*

Drilling usually involve collection of huge amount of surface data (stand pipe pressure, mud static and circulating density etc.) and downhole data from measuring while drilling (MWD) and logging while drilling (LWD) units. The acquired data can be used in real-time to seek trends and correlations between various parameters to obtain and improve signs of hole conditions and drilling efficiency. Johnston and Guichard (2015) used data analytics approach to link drilling parameters (weight of bit, Rate of penetration, torque and caliper) as an indicator of hole condition and flagging of bad holes.

With availability of suffient wells both spatially across field and vertically through formations, predictive data mining models provide drilling engineers with apriori indications of difficult areas to improve drilling efficiency thus reducing nonproductive time. Other application of data analytics in drilling engineering include Smart kick detection (Brakel et al. 2015).

### *2.2.5 Production data analysis*

Similarly, data analytics can help us to determine cause of pump failure in artificial lift and can also help in predicting when the next pump failure will occur (Liu et al. 2010).

Other data analytics techniques like artificial neural networks are being used for training production data and using the trained model to predict production rates for newer wells (Elmabrouk, Shirif, and Mayorga 2014).

# Chapter 3: Reservoir Forces and Dimensionless Numbers

*"Where oil is first found, is in the minds of men"  (Wallace Pratt)*

## 3.1 Factors affecting reservoir performance

There are many factors influencing the hydrocarbon production and reservoir performance. Wyckoff (1940) divided factors affecting reservoir performance into following two parts:

- Inherent reservoir factors (over which humans have no control)
- Controlled factors (under partial or complete human control)

While, the controlled factors are always dependent on Inherent reservoir factors. **Table 3** lists the factors that can affect reservoir performance (Wyckoff 1940).

Table 3 : Classification of factors affecting reservoir performance

| Inherent Reservoir factors | Controlled factors |
|---|---|
| Structural characteristics: Type of trap (Dual, Faulted, lenticular or stratigraphic) | Well BHP: Controlled by chokes, tubing size etc. |
| Amount of Closure or Steepness of the Dip within the reservoir | Location of well and perforation. |
| Pay zone characteristics: sand type ; porosity ; pay thickness ; no. of pay zones | The diameter of well or quality of completion. |
| fluid PVT behavior ; fluid content (size of gas cap) ; boundaries of gas oil contact | Use of newer technologies like frac. pack completions , multilateral etc. |

## 3.2 Forces operating inside reservoir

Pirson (1977) gave a comprehensive review of types of forces and energies acting inside the reservoir that drives hydrocarbons to producing wells or retain it within reservoir. These forces and energies are a function of reservoir rock/fluid properties, reservoir structure, producing histories and processes. He described four fundamental forces that

control fluid flow in the reservoir as body (gravity) forces, static pore pressure forces, viscous forces and static interfacial tension due to fluid-fluid and fluid-rock interactions.

Gravity is the macroscopic body force that occurs and dictates the movement of reservoir fluid inside the reservoir on the large scale. Differential gravitational forces generally have a negligible effect on the performance of high-pressure fields or fields with high vertical permeability (Pirson 1977, Calhoun Jr 1976). However, gravity plays a critical role in recovery from depleted fields and fields with sufficient relief in geological structures (Pirson 1977). Gravity number is the dimensionless quantity that quantifies ratio of gravity to inertial (viscous) forces.

Viscous forces is a friction forces that acts within the moving fluid or provide resistance to movement of static fluid. In liquids, viscous force is due to attraction between molecules while in gases, viscosity is a result of collisions between the molecules. They are retarding force and work is needed against them to produce fluid from the reservoir. Viscous forces become more dominant when the rate of fluid flow is high.

Capillary forces arise due to rock and fluid interactions and are a function of interfacial liquid tensions, pore size/shape, and wetting properties of the rock. The capillary pressure difference existing in a porous medium between two immiscible phases is a function of interfacial tension, the average size of capillaries (pore size distribution) and saturation distribution which controls the curvature of the interface (Pirson 1977). Capillary forces manifest themselves in the subsurface as the capillary pressure that affects saturation distributions. Capillary pressure is magnitude of pressure that is applied on a non-wetting fluid in order to reach a certain saturation in that fluid.

During the performance of most of the reservoir combination of above forces are active but at times one group of force may dominate others. The forces and other lithological and petrophysical properties dictate the distribution, simultaneous movement, and displacement of fluids within the reservoir. Driving mechanisms can be captured by use of fractional flow equation described (Leverett 1941) by:

$$f_d = \frac{1 - \frac{K_o}{\mu_o * q_t}\left\langle \frac{dP_c}{du} + g * \Delta\rho \sin\alpha \right\rangle}{1 + \frac{k_o \mu_D}{k_D \mu_o}} \qquad \textbf{Eqn. 14}$$

Here, subscript 'D' is used for displacing fluid. While more than one mechanism is simultaneously active in the reservoir, it is desirable to know if any particular mechanism is more dominant in a reservoir. **Table 4** lists the modification of above equation to describe contribution from each mechanism.

Fractional flow formula provides a description of fundamental production processes occurring during immiscible displacement. The main factors which influence recovery according to fractional flow equation can be summed up as (1) Relative permeability ratio (2) Fluid viscosity ratio (3) Presence of connate water phase (4) Presence of gas phase (5) Presence of formation dip (6) Effect of capillary-pressure gradient in the direction of flow.

Table 4 : Fractional flow equation for different drive mechanism

| Mechanism Type | Contribution factor |
|---|---|
| Frontal water or gas drive | $f_w = \frac{1}{1 + \frac{K_o}{K_w}\frac{\mu_w}{\mu_o}}$ |
| Gravity drive | $f_g = \frac{1 - \frac{K_o}{\mu_o * q_t}\langle g\Delta\rho \, sin\alpha \rangle}{1 + \frac{k_o}{k_D}\frac{\mu_g}{\mu_o}}$ |

| | |
|---|---|
| Capillary drive | $f_d = \frac{1 - \frac{K_o}{\mu_o * q_t}\langle\frac{d\,P_c}{du}\rangle}{1 + \frac{k_o}{k_D}\frac{\mu_D}{\mu_o}}$ |
| Combination drive | $f_d = \frac{1 - \frac{K_o}{\mu_o * q_t}\langle\frac{d\,P_c}{du} + g * \Delta\rho \sin\alpha\rangle}{1 + \frac{k_o}{k_D}\frac{\mu_D}{\mu_o}}$ |

### 3.3 Dimensionally Scaled Reservoir Models

The performance of a reservoir is a interplay of various variables, and comparison between performance of different reservoirs can be done if the variables are "properly scaled" (Geertsma, Croes, and Schwarz 1956). Different variables affecting reservoir performance are combined as similarity groups using dimensional analysis. If the values of similarity groups is same for a model and a prototype, then the model is properly scaled and can be used for comparison in terms of field performance (Geertsma, Croes, and Schwarz 1956). Dimensionless analysis of the governing equations for fluid flow in a reservoir provides insight into the relative importance of driving forces such as viscous, gravity, and capillary forces on the displacement mechanisms (Wu et al. 2008). Dimensional analysis uses primary variables affecting the physical system to determine minimum number and form of similarity groups (Fox, McDonald, and Pritchard 1985).

The use of dimensionless numbers makes it possible to report the results in a manner that makes them applicable to systems other than the one used to acquire the data (Peters, Afzal, and Gharbi 1993). For example, pressure transient testing uses dimensionless pressure and rate terms for universal application of diffusivity equation. The geometrically similar system will result in same dimensionless variables irrespective

of the scale of measurement. For petroleum engineering applications, dimensionless numbers relate the effects of various forces involved in the system under consideration. Other advantages of using scaled models are, it leads to a reduced number of dimensions required to describe the system and eliminates the need for unit conversion (Shook, Li, and Lake 1992).

There are many different definitions of dimensionless variables reported in the literature (Shook et al., 1992; Rapoport, 1955). Dimensionless numbers generated by Shook, Li et al. (1992) are based on forces (gravity, viscous, capillary & dispersion) controlling the displacement process are described in **Table 5.** This dimensionless group, combines the variables that govern the immiscible displacement of oil by liquid (Geertsma, Croes, and Schwarz 1956). The flow process considered for generation of these numbers is two-phase immiscible displacement with constant viscosity and residual saturation of each phase.

Table 5 : list of Dimensionless Numbers

| Dimensionless number | Formula |
|---|---|
| Capillary Number ($N_{pc}$) | $N_{Pc} = \frac{\lambda_{r2}^{o}\sigma}{LU_t}\sqrt{\phi K_x}$ |
| Gravity Number ($N_g$) | $N_g = \frac{K_z\lambda_{r2}\Delta\rho g cos\alpha}{u_t}\frac{H}{L}$ |
| Aspect Ratio ($R_l$) | $R_l = \frac{L}{H}\sqrt{\frac{K_z}{K_x}}$ |
| Density Number ($D_n$) | $N_\rho = \frac{\rho_o}{\Delta\rho}$ |

This similarity group assume system geometrically analogous with similar petro-physical properties. Novakovic (2002) articulated that gravity number scales viscous and capillary forces while aspect ratio scales spatial interplay. Therefore, these numbers if used jointly with mobility ratio will result in the system's behavior that is properly scaled. For dGOM oilfield dataset. These numbers are calculated by use of assumptions given below. Proper scaling of models may require additional numbers if any of these assumptions is broken (Rapoport, 1954).

1.) Homogenous reservoirs having immiscible displacement of oil by water.
2.) Relative permeability function and contact angle for all the reservoirs are considered same.
3.) Viscosity of water is taken constant as 0.5 cp; Constant Permeability anisotropy of 0.1 is assumed for all fields.
4.) Relative permeability is obtained from Corey's relationship (BrooksRH 1964)

$$\boldsymbol{K_{rw} = \left(\frac{1-S_o}{1-S_w}\right)^2 * \left(\frac{1-S_o-S_w}{1-S_w}\right)^2} \qquad \textbf{Eqn. 15}$$

5.) Interfacial tension between hydrocarbon and water system is calculated using (Firoozabadi and Ramey Jr 1988) :

$$\boldsymbol{\sigma_{hw} = \left(\frac{1.58*(1-\rho_o)+1.76}{t_r^{0.03125}}\right)^4} \qquad \textbf{Eqn. 16}$$

$$\boldsymbol{t_r = \frac{reservoir\ temperature}{T_{co}}} \ ; \ \boldsymbol{T_{co} = 24.28 * K_w^{1.77} * \rho_o^{2.12504}} \qquad \textbf{Eqn. 17}$$

$K_w =$ 11.7 $(watson\ characterization\ factor)$

Reservoir temperature is calculated using the static temperature gradient reported in the database.

6.) ‘L’ is calculated by assuming well is centered in a cylindrical reservoir of given area. Sand is assumed to be homogenous with only one well per sand.

7.) Average daily production rate is defined as the ratio of total cumulative production to number of days field has produced. Fluid flow velocity is estimated by dividing the average daily production rate reported for the well block by reported surface area of that sand.

**8.)** Dip angles are calculated by assuming structure of field to be related to dip angle as per Table-6

Table 6 : Relation of dip angle with Field structure

| FSTRU | Dip Angle |
|---|---|
| A / Anticline | 10 |
| B / Fault | 11 |
| C / Sallow Salt diapir | 12 |
| D / Intermediate Salt diapir | 13 |
| E / Deep Salt dome | 14 |
| F / Salt Ridge | 15 |
| G/ Shale diapir | 16 |
| H/ Unconformity | 17 |
| I/ Stratigraphic | 18 |
| J/ Reef | 19 |
| K/ Rollover growth fault | 20 |
| L/ Rotational Slump block | 21 |
| M/ Non-piercement Louann Salt | 22 |
| N/ Thrust Fault | 23 |
| U / Unknown | 0 |

### *3.3.1 Global Capillary Number* ($N_{Pc}$)

This dimensionless number is defined by the ratio of capillary to inertial forces. Several authors have provided many empirical models to determine capillary number from small-scale laboratory measurement to large-scale field measurements (Moore and Slobod 1956, Melrose , Slattery, Sagis, and Oh 2007, Chatzis, Morrow, and Lim 1983, Rapoport and Leas , Geertsma, Croes, and Schwarz 1956, Craig et al. 1957). Capillary forces are important in pore scale modeling stage where fundamental issue is to determine residual saturations and scaling groups that control them (Novakovic 2002). High capillary number is a result of either high capillary forces (small pore throat, high interfacial tension) or reduced flow velocity (low inertial forces). Capillary forces results in dispersion of the immiscible displacement.

### *3.3.2 Gravity Number* ($N_g$)

Gravity number is ratio of gravity forces to inertial forces, while this number can also be defined by using ratio of time required for a fluid to be moved across reservoir as a result of gravity (density difference) to time required to move the fluid across reservoir due to viscous forces (Novakovic 2002). This number is sensitive to relief in geological structure. Hence, requires dip angle measurement for it's calculation.

### *3.3.3 Density Number* ($D_N$)

This number relates contrast of densities in displacing to displaced fluid in the immiscible displacement process. For dGOM oilfields dataset this number is calculated by assuming constant in-situ water density of 62.4 $lb/ft^3$ while density of oil is deduced from its API gravity.

### *3.3.4 Effective aspect ratio* ($R_l$)

This dimensionless number combines the anisotropy ratio ($K_z/K_x$) and the aspect ratio (L/H). It is a measure of relative flow capacity of a medium in vertical and horizontal direction or it is a viscous force ratio in horizontal and vertical direction when flow area and the rate is same. Physically the effective aspect ratio, $R_l$ is a ratio of characteristic time require fluid to travel in a horizontal direction to that of in vertical direction. Thus, if $R_l$ is large this implies pressure and saturation variations in the vertical direction are much less than those in horizontal direction hence flow in the vertical direction can be ignored (Novakovic 2002). $R_l$ plays a critical role in determining capillary-gravity vertical equilibrium condition. For calculating this number in dGOM dataset constant permeability anisotropy ratio of 0.1 is assumed.

## 3.4 Tarner's Method for Performance Prediction

Tarner (1944) proposed a solution to forecast reservoir performance for depletion drive by means of material balance and instantaneous gas-oil ratio equations. Tarner's method work by applying these two equations to find field pressure attained for each assumed increment of stock-tank oil production.

$$\boldsymbol{N = \frac{n[B_o + B_g(GOR - R_s) - (W - w)]}{mB_{gi}\left(\frac{B_g}{B_{gi}} - 1\right) + B_g(R_{si} - R_S) - (B_{oi} - B_o)}} \qquad \textbf{Eqn. 18}$$

$$\boldsymbol{R = R_s + \frac{B_o K_g U_o 1}{K_o U_g B_g}} \qquad \textbf{Eqn. 19}$$

Assuming,

No water drive (W =0) and no produced water (w =0) and no original gas cap (m =0).

The gas production increment ($G_2 - G_1$) between two cumulative oil productions, $n_2$ and $n_1$ can be written as

$$\boldsymbol{G_2 - G_1 = N\left[R_1 - R_2 - \frac{(B_{oi} - B_2)}{B_{g2}} - \frac{(B_{oi} - B_1)}{B_{g1}}\right] - n_2\left(\frac{B_2}{B_{g2}} - R_2\right) + n_1\left(\frac{B_2}{B_{g2}} - R_1\right)} \qquad \textbf{Eqn. 20}$$

$$G_2 - G_1 = \frac{(R_1+R_2)}{2}(n_2 - n_1) \qquad \textbf{Eqn. 21}$$

The calculation of instantaneous GOR requires knowledge of fluid saturation that can be obtained by:

$$S_0 = (1 - S_w)\frac{(N-n)B_o}{NB_{oi}} \qquad \textbf{Eqn. 22}$$

Once fluid saturation is known relative permeability can be calculated using relative permeability curves. **Figure 5** shows the workflow to be used for calculation of field performance using Tarner's method.

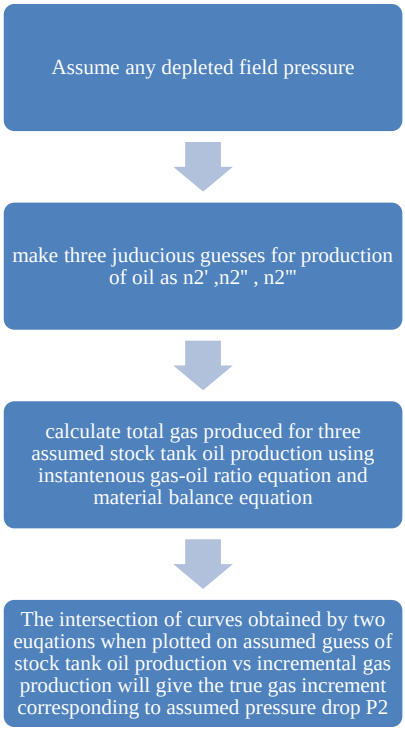



**Figure 5 : Workflow for application of Tarner's method of performance prediction**

# Chapter 4: Deepwater Gulf of Mexico Exploratory Data Analysis

*"We usually find gas in new places with old ideas. Sometimes, also, we find gas in an old place with a new idea, but we seldom find much gas in an old place with an old idea" (Parke Dickey)*

## 4.1 Geological Background for Gulf of Mexico (GOM)

### *4.1.1 Introduction*

Offshore oil & gas exploration in GOM began in 1940s, which was gradually extended to deepwater and ultra-deepwater continental shelf and slopes. The first discovery in deepwater GOM (dGOM) was made in 1975 by shell in Mississippi canyon. Since 1975 there have been over 300 deepwater discoveries in the GOM (Post et al. 2012). **Figure 6** describes the bathymetric map of GOM basin. The abyssal plain (Sigsbee) is separated from continental shelf by broad slope region which contains numerous salt structures and is affected by downslope salt movement like prominent submarine cliff called Sigsbee escarpment. The slope regions is fed by several canyons from east to west. These canyons provided great supply of reservoir quality sand for storage of hydrocarbons produced from organic rich marine sediments. While the salt structures provides efficient seals for trapping the hydrocarbons.

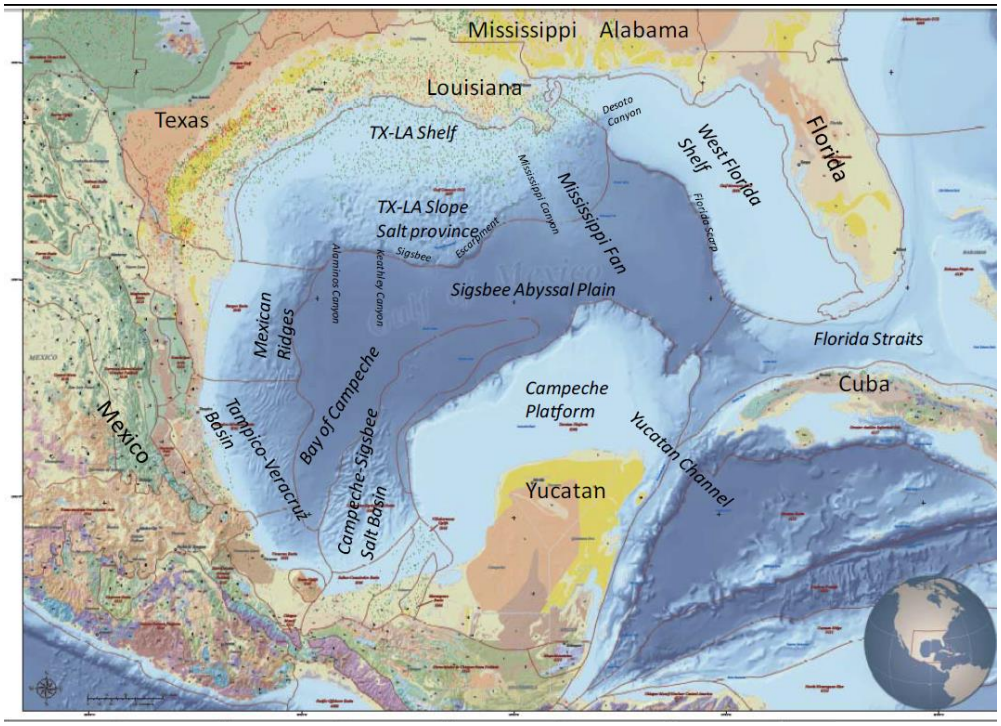



**Figure 6 : USGS Map of Gulf of Mexico. It contains three continental selves, West Florida Shelf, Texas-Louisiana Shelf and the Campeche Platform. The continental slope bathymetry contains numerous salt mini-basins, salt canopies and salt diapers. Sigsbee escarp is a prominent cliff that trends East-West on southern Texas-Louisiana slope and turns northeast towards Mississippi delta. Canyons cutting shelf from east to west are Desoto canyon, Mississippi canyon, Keathley Canyon and Alaminos Canyon.**

### *4.1.2 Stratigraphy for GOM*

**Figure 7** represents the generalized stratigraphic section of GOM shelf region along with generalized north-south cross sectional depositional system as described by Galloway (2008). Stratigraphy of deepwater GOM is divided into three main groups: Neogene (Pliocene & Miocene), Paleogene (Oligocene, Eocene, and Paleocene) and Mesozoic (Cretaceous, Jurassic and Triassic). Paleogene includes Frio sandstone (Alaminos canyon), Upper & Lower Wilcox (Alaminos canyon, Keathley Canyon, Walker ridge blocks).

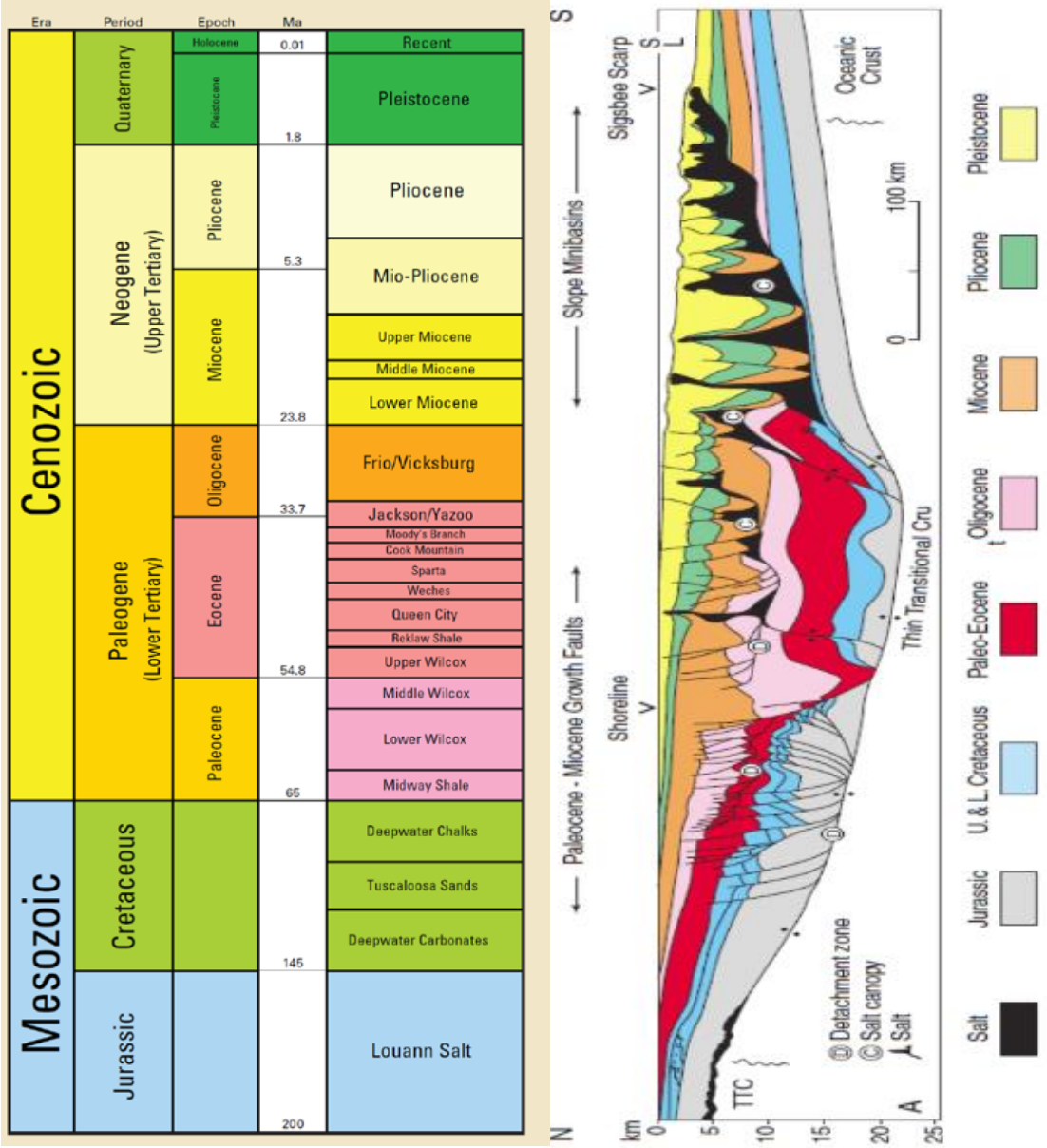



**Figure 7 : Local Stratigraphic and N-S Deposition X-section for GOM sands (Slatt and Zou (2014), Galloway (2008))**

Lach (2010) divided major resource potential for GOM into Neogene and Paleogene reservoirs. While Neogene reservoirs are mature in deep water production experience. The Paleogene age reservoirs are in exploration and appraisal stage. GOM deposits are

characterized by Louann salt, which is sandwiched between organic rich Miocene sands and Plio- Pleistocene deposits. **Table 7** gives hydrocarbon accumulation and production from each plays as given by D. Nixon (1999).

Table 7 : Total reserves form sands of different geological ages (D. Nixon 1999)

| Chronostratigraphic unit | Original proved reserve | Cumulative Produced |
|---|---|---|
| Miocene | 41.9% | 43.5% |
| Pleistocene | 36.2% | 36.5% |
| Pliocene | 18.6% | 19.1% |
| Mesozoic | 2.9% | 0.4% |
| Oligocene | 0.4% | 0.5% |

Most of the discovered original oil in-place (OOIP) in Paleogene age reservoirs is in Keathley canyon and Walker ridge. **Figure 8** illustrates location of various geological features along with prevalent drive mechanisms in GOM basin. Many reservoirs with depletion drive are located on eastern part of GOM. The Paleogene play has the highest risk in reservoir quality, water depth, drilling costs, reservoir complexity and infrastructure (Lach 2010). The Mesozoic section comprises of main source rocks (Jurassic & mid-cretaceous) as well as salt (Jurassic) that dictates the trapping mechanism for the GOM fields.

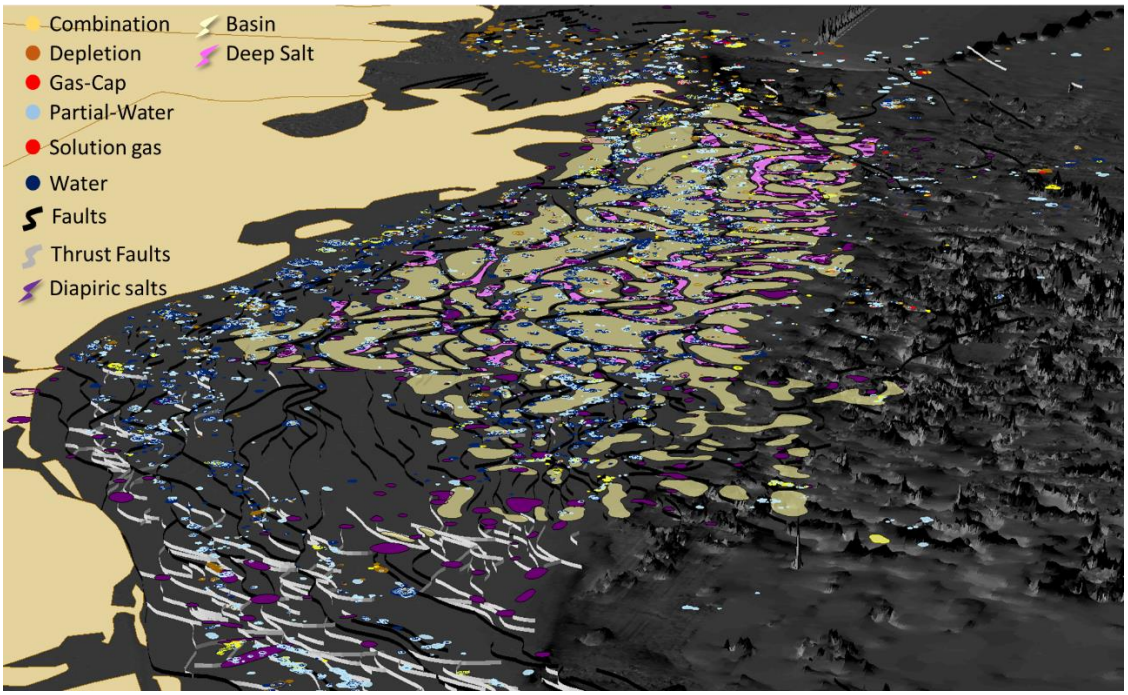



**Figure 8 : Three-Dimensional description of Gulf of Mexico geology and drive mechanisms using GOM3 (Courtesy: Earth Science Associates)**

### *4.1.3 Petroleum System & Depositional Model for Deepwater GOM*

Conventional hydrocarbon production is becoming increasingly focused on deep and ultra-deep water GOM (Disenhof, Mark-Moser, and Rose 2014). Optimum conditions of rate of sedimentation and presence of abundant organic matter supported formation of GOM hydrocarbons during Mesozoic era, while impermeable salt and shale deposited during Jurassic period provides the seal and complex topographical features. Majority of depositional history of the basin is concentrated to period after middle Jurassic (Galloway (2008), Woods, Salvador, and Miles (1991)). The complex interaction of GOM's depositional history with its structure and salt tectonics has resulted in a heterogeneous system that requires varied techniques to evaluate the risks of hydrocarbon extraction (Disenhof, Mark-Moser, and Rose 2014).

According to Galloway (2008), GOM basin reached present structural configuration in the early cretaceous period. Early cretaceous deposition was dominated by carbonates and evaporites on the shelves and shallow marine clastics on northern basin rims (McFarlan Jr and Menes 1991). While large volume of clastics were deposited in northern GOM during Cenozoic period (Salvador 1991). Major architectural elements and their characteristics are given in **Table 8** .

Table 8 : Description of architectural elements found in GOM

| Architectural Element | Fields | Characteristics |
|---|---|---|
| Confined channels & mini-basins | Eastern Mississippi canyon ; Northern green canyon ; Garden banks (Mars-Ursa) | Mid Slope; Debris flow; Low length to thickness ratio; Discontinuous & high heterogeneity. |
| Distributary channel fills | Central green canyon ; Mississippi canyon (Thunder horse & Tahiti) | Continuous geometry; lateral continuity; ratio of length to thickness >100. |
| Distributary lobes | -- | Sheet sandstone good lateral connectivity; 4-way turtle structure |
| Mini-basins | Green canyon (Condor and Droshky); Garden banks | Shallow amplitude plays in upper Miocene, Pliocene & Pleistocene |
| Conventional Miocene subsalt deposits | -- | Channelized sheet sandstones |

**Neogene Depositional Model**

Neogene deposition in deepwater GOM is connected with salt tectonics. The field is classified into following categories based on structural style: (1) Structural traps (faults/turtle structure/anticline) (2) Stratigraphic traps (Pinch outs/unconformity) (3) Combinational traps (salt flanks/diapers). Purely stratigraphic traps are common for deepwater sands older than cretaceous in northeastern margin. But discovery of purely stratigraphic trap is difficult and there occurrence is very rare (Lach 2010). Neogene

deposition are characterized by rapid sedimentation in mini-basins resulting in high porosity over-pressured reservoirs.

**Paleogene (Wilcox sands) deposition model**

Three main reservoir target for Paleogene sands are (1) Frio (Oligocene) (2) Upper Wilcox (Eocene) (3) Lower Wilcox (Eocene to Paleocene). While lower Wilcox is dominant in western and central GOM. Upper Wilcox and Frio occurs only in western GOM. Wilcox sands is correlated from Alaminos canyon in west to walker ridge in east that is approximately 400 miles (**Figure 9**). This lateral continuity is evidence of good quality reservoirs occurring in these areas resulting in discovery of Jack and St. Malo fields (Meyer et al. 2005). Paleogene reservoirs have a complex digenetic history and several factors control the quality of reservoir in Paleogene sands.

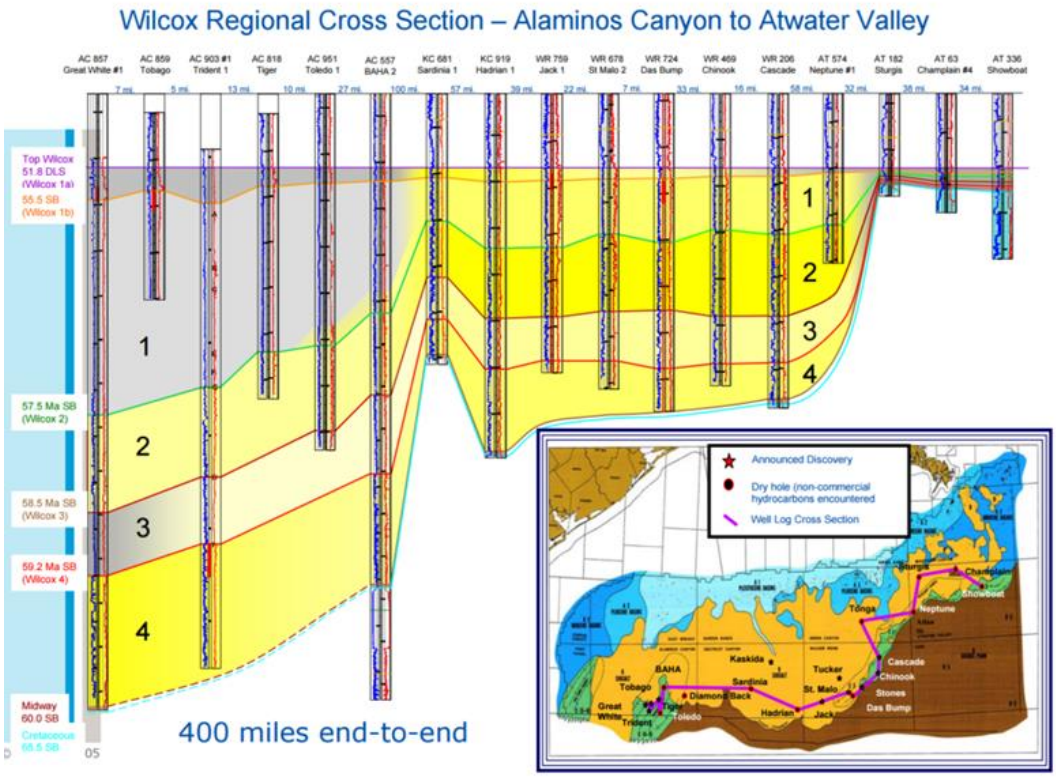



**Figure 9 : X-section for Lower Tertiary Wilcox Trend in Deepwater Gulf of Mexico (Rains, Zarra, and Meyer 2006)**

### *4.1.4 Protraction area and Leasing*

Gulf of Mexico Outer continental shelf (OCS) planning area is divided into three planning area viz. western, central & eastern. Each zone is subdivided into number of protraction area containing different sands and blocks. While extensive exploration is done in western and central part. Eastern part remains a very good exploration prospect for future development (Lach 2010). Fields in GOM are identified by protraction area and block number. For example, MC 807 stands for field in Mississippi canyon in block 807. Each block may contain many lease areas and different fields. BOEM has published reservoir data for 1300 fields based on initial geological and engineering analyses. Of these 1300 fields, 633 are expired (depleted) fields. After converting gas volumes (bcf) and oil

volumes (bbl) to Barrels oil equivalent (BOE), there are 395 oil fields and 905 gas fields (fields with GOR < 9,700 SCF/STB are classified as oil producers). As of 2013, BOEM reports estimated original reserves of 22.19 billion bbl. Oil and 193 trillion SCF gas from 1300 fields. All the reserves estimates reported by BOEM are based on volumetric and performance methods. Quick look Analysis of the 1,300 oil and gas fields indicates that the GOM is a gas-prone basin.

## 4.2 Deepwater GOM Oilfield Dataset Description

### *4.2.1 Description of Key Attributes*

In order to understand the GOM dataset. Deep-water oilfields (Water depth >1000 ft., SAND_TYPE = "O") is selected. This gave 395 wells that are used for applying data mining algorithms. In order to use this dataset for analysis it is necessary to understand the sources from which data was procured. **Figure 10** gives the list of attributes is used for analysis. For geological attributes, most of the data comes from seismic & geologist's interpretation. Production data form individual well is a result of flow rate allocation based on number of wells flowing to particular platform/gathering stations. GOR is derived from cumulative production of oil and gas. Petro-physical and PVT parameters are as reported by various operators to BOEM. Therefore, most of the data is result of individual interpretation of engineers that can vary significantly from operator to operator. The goal of data mining is to use this uncertain dataset to figure out unknown correlation and patterns present in data.

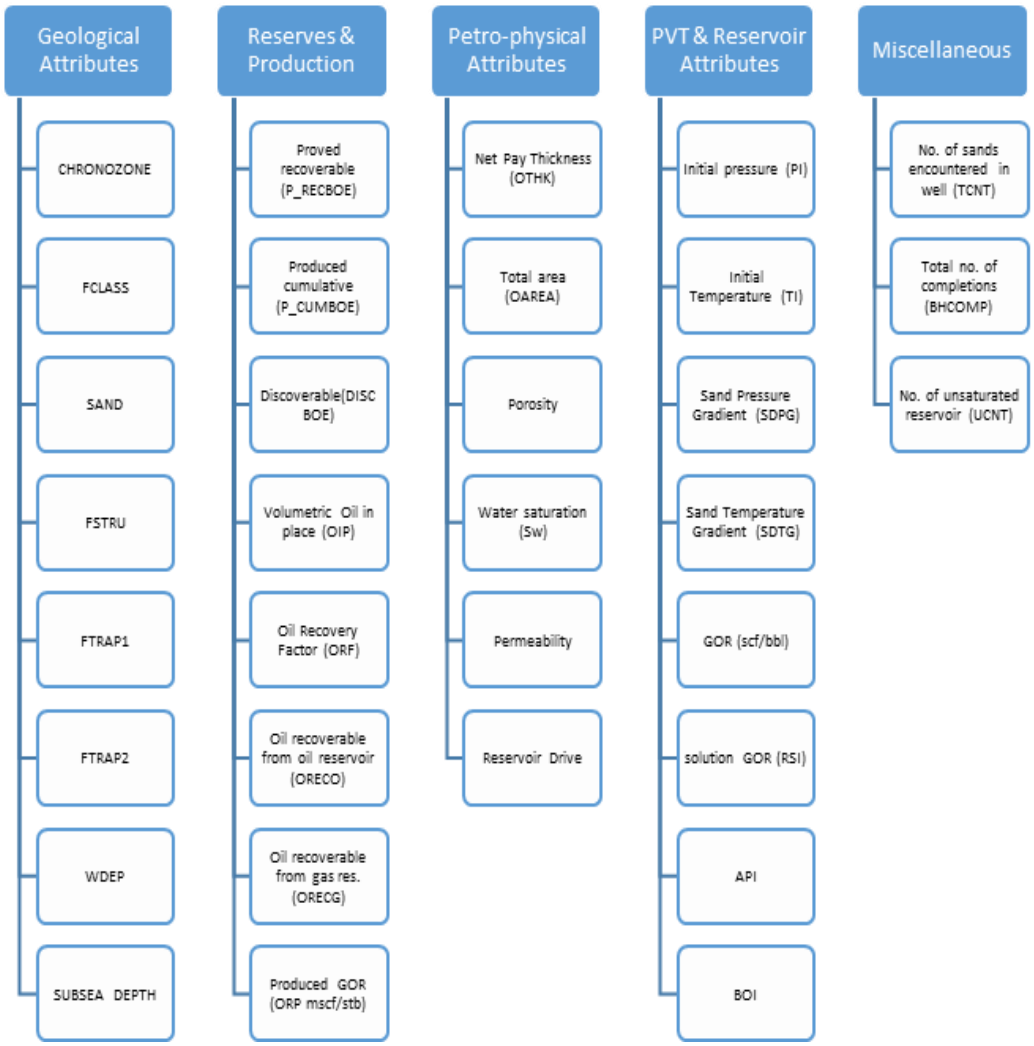



**Figure 10 : Filtered list of attributes used for data mining operation in dGOM. Refer to Nomenclature for definitions for used keywords.**

### *4.2.2 Statistics for Qualitative Attributes*

Based on the sequence biostratigraphy discovery, wells in deepwater oilfields is classified in following chornozone units (**Table 9**). As observed in Table 7 most of deep-water oil production has come from Miocene, Pliocene and Pleistocene reservoirs while some gas fields are present in Paleocene period is not discussed in this thesis. **Table 9-Table** 13 gives distribution of various qualitative attributes in dGOM oilfields. “QQ code” signifies

the numerical values used in data mining algorithms when qualitative variable is converted to quantitative.

**Table 9 : Biochoronostratigraphy for dGOM oilfields dataset sorted chronologically. Absence of paleogene period indicates no oilfields were present in dataset for this period.**

| Period | CHRONOZONE NAME | CODE | QQ code | Wells |
|---|---|---|---|---|
| Quaternary | Pleistocene-Upper | PLU-LL | 19 | 42 |
| Quaternary | Pleistocene-Middle | PLM | 10 | 6 |
| Quaternary | Pleistocene-Lower | PLL | 16 | 29 |
| Neogene | Pliocene-Upper | PU | 20 | 64 |
| Neogene | Pliocene-Lower | PL | 15 | 111 |
| Neogene | Miocene-Upper | MUU (Younger than MLU) | 12 | 72 |
| Neogene | Miocene-Upper | MLU | 6 | 15 |
| Neogene | Miocene-Middle | MUM (Younger than MMM) | 17 | 6 |
| Neogene | Miocene-Middle | MMM | 8 | 24 |
| Neogene | Miocene-Lower | MLM | 5 | 10 |
| Neogene | Miocene-Lower | MUL | 9 | 2 |
| Neogene | Miocene-Lower | MML | 7 | 4 |

**Table 10 : Classification of field class based on BOEM data. When the leases make formal commitment to develop and produce it is classified as RJD. During the period of infrastructure set-up, the accumulation is classed as PDN. After production of accumulation began it is assigned as PDP.**

| Field Class | Code | Wells |
|---|---|---|
| Proved Undeveloped Reserves | PDN | 7 |
| Proved Developed Producing Reserves | PDP | 365 |
| Reserves Justified for Development | RJD | 20 |

**Table 11 : Classification of Field Structure (FSTRUC)**

| Field Structure | FSTRU code | QQ code | Wells |
|---|---|---|---|
| Anticline | A | 1 | 102 |
| Fault | B | 2 | 28 |
| Shallow Salt Diapers | C | 3 | 24 |
| Intermediate Salt Diapers | D | 4 | 65 |
| Deep Salt Dome | E | 5 | 62 |
| Salt Ridge | F | 6 | 76 |
| Unconformity | H | 8 | 2 |
| Stratigraphic | I | 9 | 19 |
| Thrust Fault | N | 14 | 7 |

**Table 12 : Classification of Primary and secondary trap type for dGOM oilfields. It is interesting to note that 197 fields have missing FTRAP2. While majority of fields is dominated by salt type primary traps.**

| Field Trap Type | Trap Code | QQ code | Wells with Primary trap type (FTRAP1) | Wells with Secondary trap type (FTRAP2) |
|---|---|---|---|---|
| Missing | ? | -- | 6 | 197 |
| Anticlines | A | 1 | 23 | 16 |
| Faulted Anticlines | B | 2 | 56 | 11 |
| Rollover anticlines into growth fault | C | 3 | 1 | 6 |
| Normal Fault | D | 4 | 38 | 34 |
| Reverse Fault | E | 5 | 5 | 12 |
| Turtle Structure | F | 6 | 20 | 10 |
| Flank traps (salt or shale diapirs) | G | 7 | 181 | 6 |
| Sediments overlying domes | H | 8 | 5 | 15 |
| Up dip facies change | J | 9 | 13 | 9 |
| Up dip pinch out | K | 10 | 18 | 14 |
| On lap sands | M | 12 | 6 | 9 |
| Subsalt trap | Q | 15 | 19 | 23 |

**Table 13 : Classification for drive mechanism reported to BOEM by operators. Most of the dGOM oilfields are operating under varying water drive mechanism.**

| Drive Mechanism | Code | QQ code | No. of fields |
|---|---|---|---|
| Missing | ? | -- | 4 |
| Combination | COM | 1 | 29 |
| Depletion | DEP | 2 | 45 |
| Partial | PAR | 4 | 196 |
| Solution Gas Drive | SLG | 5 | 31 |
| Unknown | UNK | 6 | 8 |
| Water Drive | WTR | 7 | 76 |

### *4.2.3 Statistics for Quantitative Attributes*

**Figure 11-Figure 15** illustrates histograms and box plots for various quantitative reservoir and production attributes. It can be observed that Oil thickness (OTHK), oil area (OAREA), Solution GOR (RSI), Water saturation (SW) and Permeability have log normal distribution leading to similar distribution for oil reserves (OIP) and cumulative production (P_CUMCOIL).

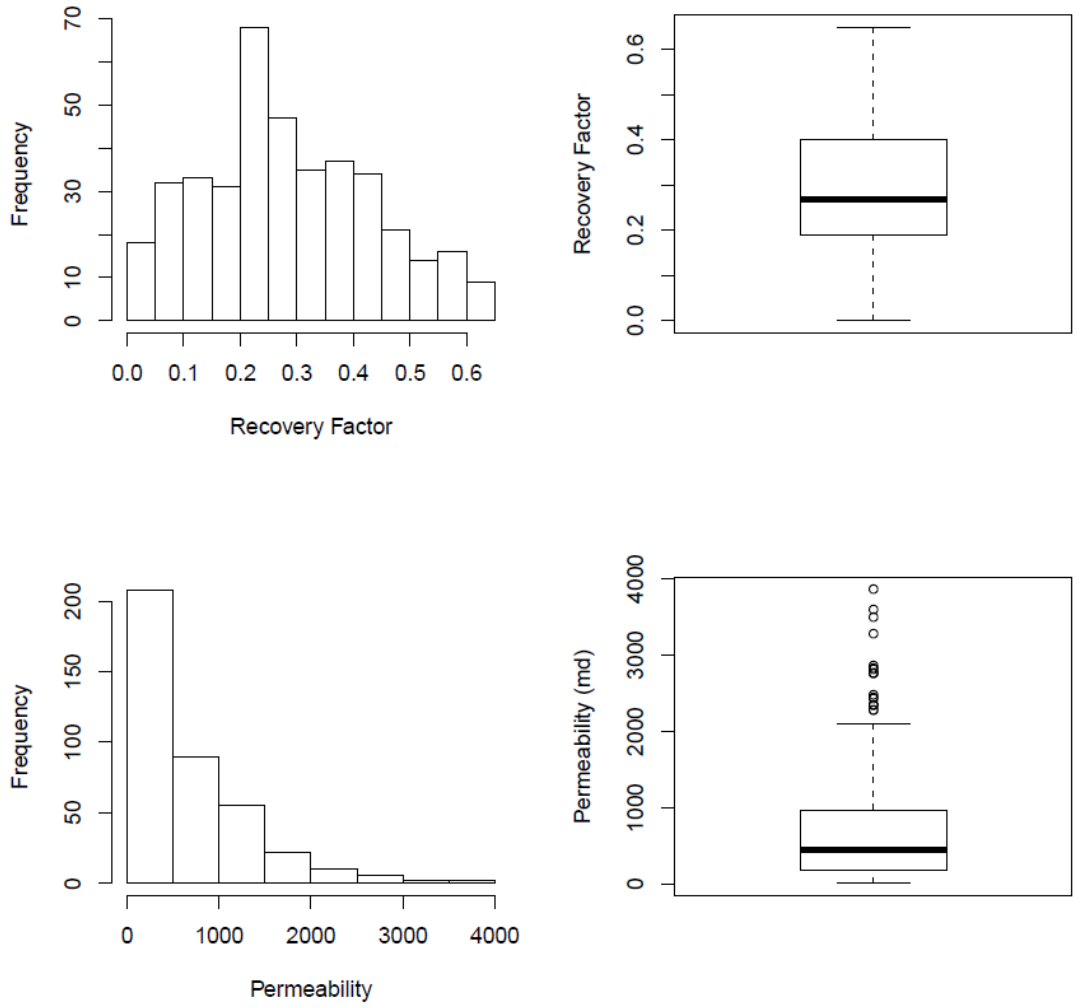



**Figure 11 : Distribution of Recovery Factor and Permeability**

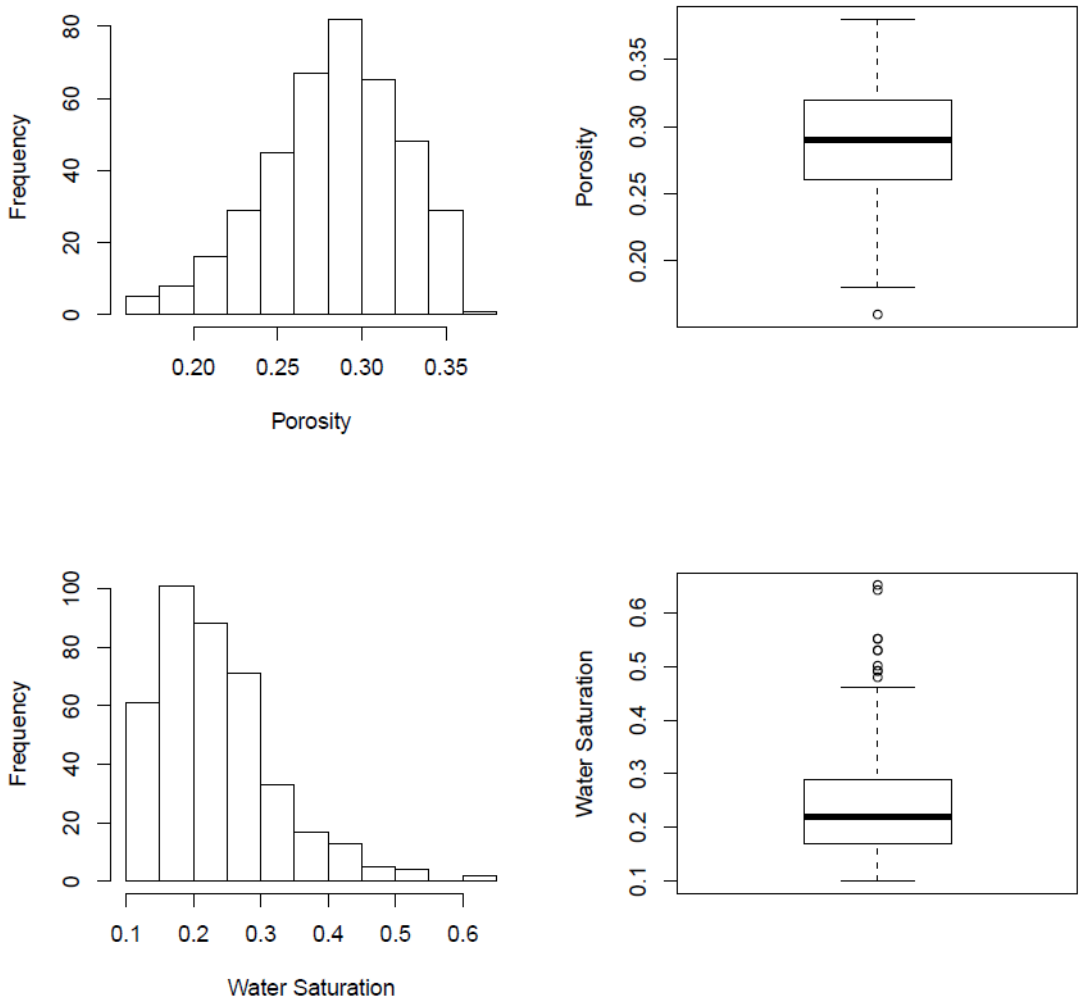



**Figure 12 : Distribution of Porosity and Water Saturation**

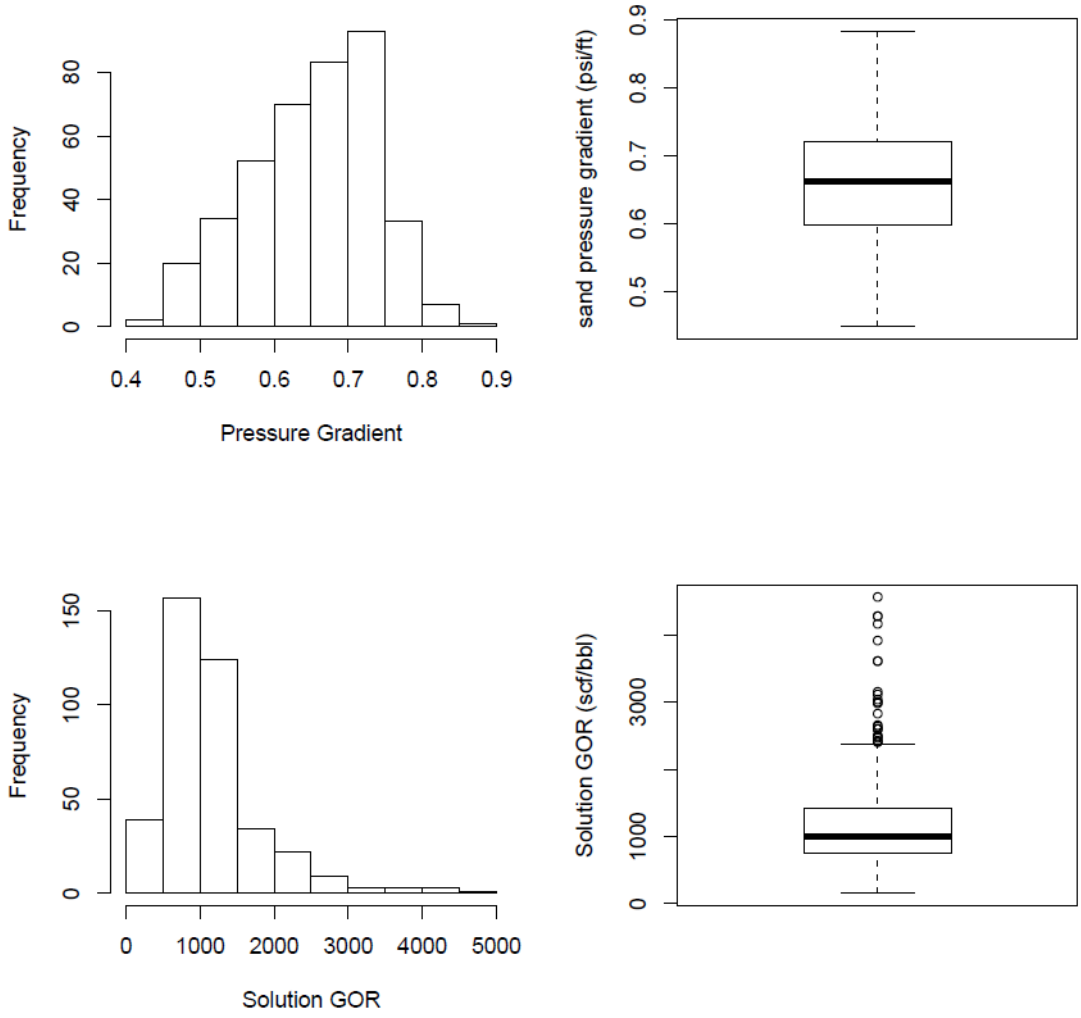



**Figure 13 : Distribution of Pressure gradient (SDPG) and Solution GOR (RSI)**

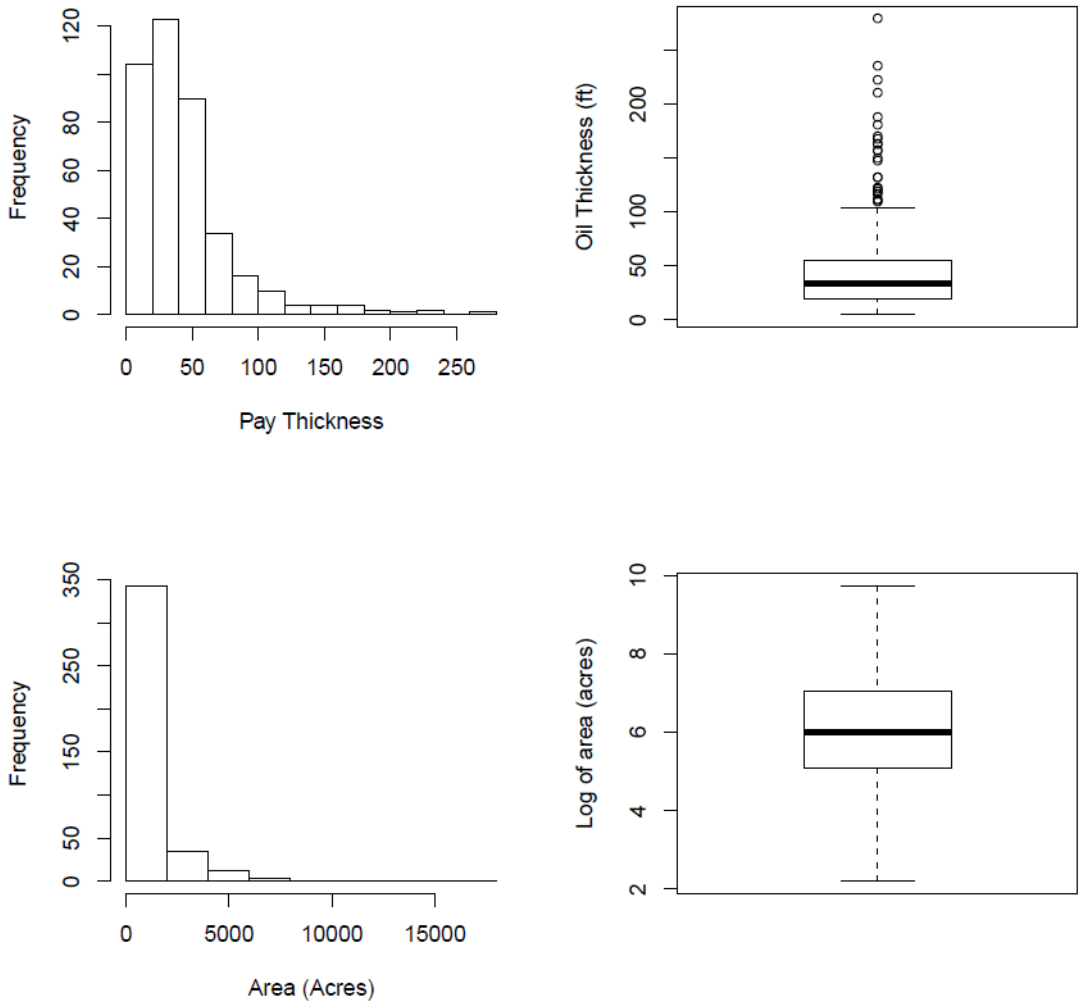



**Figure 14 : Distribution of Oil Thickness and Oil Area.**

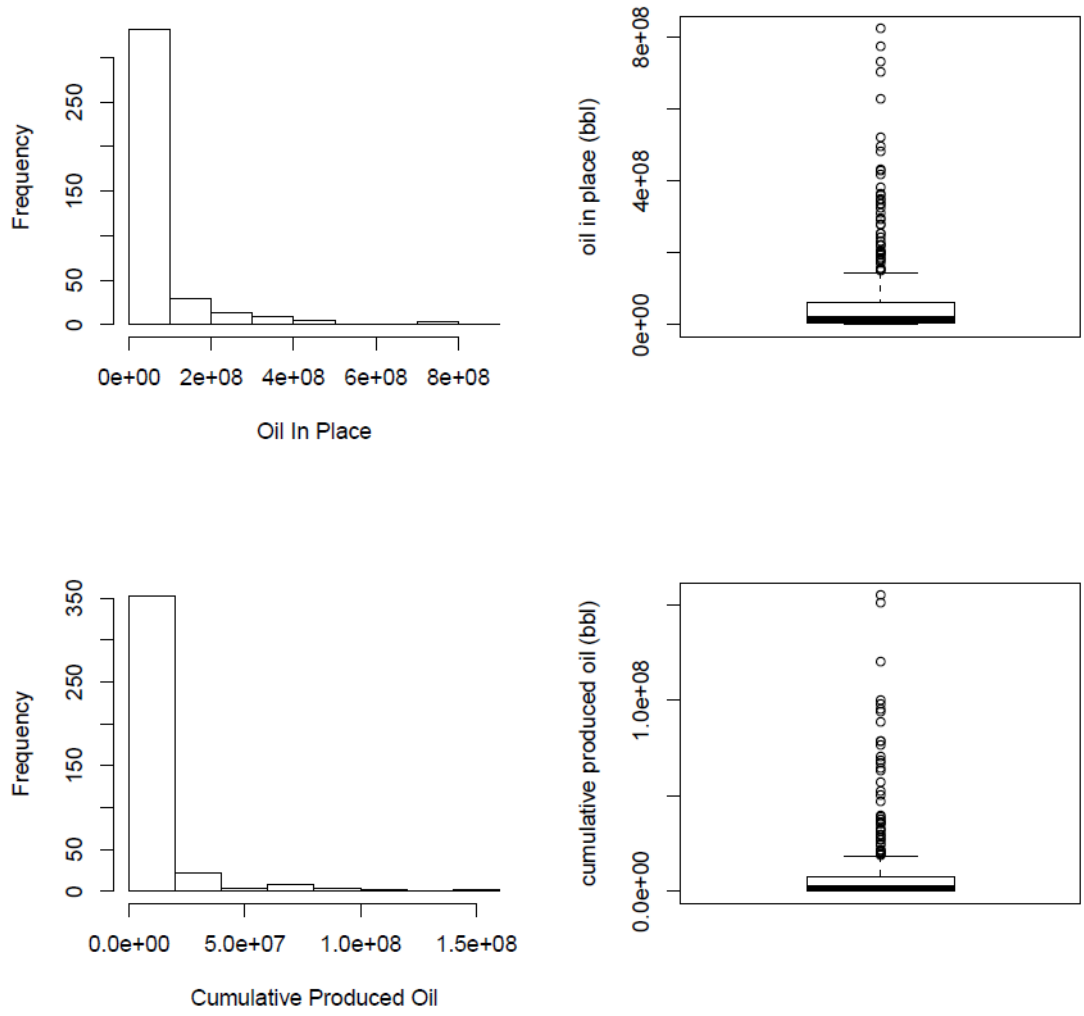



**Figure 15 : Distribution of Oil in-place and Cumulative produced oil. It can be deduced that most of the reservoir are small in size.**

**Figure 16** displays the position of various reservoirs as compare to 100 % recovery line.

Points lying furthermost relative to this line have lower values of recovery factors.

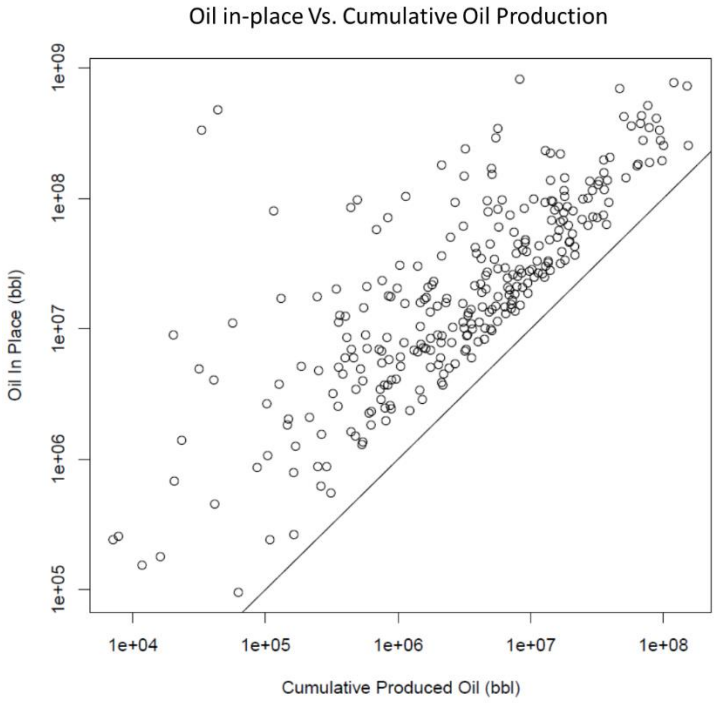



**Figure 16 : Over-estimated OIP as compare to physical production. Unit slope line represents 100% recovery.**

As of December 31, 2013, Minerals Management Services (MMS) estimated oil reserves in deep-water GOM is 3.67 bbl of oil and 9 trillion scf of gas from 667 active fields, but many of these fields have recovery factors ranging 6%-25% (Beshears 2013). The remaining oil is significant and provides the incentive to develop the methodology for prediction of recovery factor to guide for full field development of these fields. Since lower tertiary trend is still in exploration and appraisal stage, the reported volumes by MMS excludes most paleogene discoveries. The average recovery factor determined for Neogene/Pleistocene reservoir using a volume-weighted average is 28.9%. This average is based on statistical data for 392 oilfields in deepwater GOM.

### *4.2.4 Statistics for Dimensionless numbers*

The set of dimensionless numbers is calculated for fifty-nine oil reservoirs under water drive mechanism. The statistical distribution for these numbers is given in **Figure 17- Figure 18** .

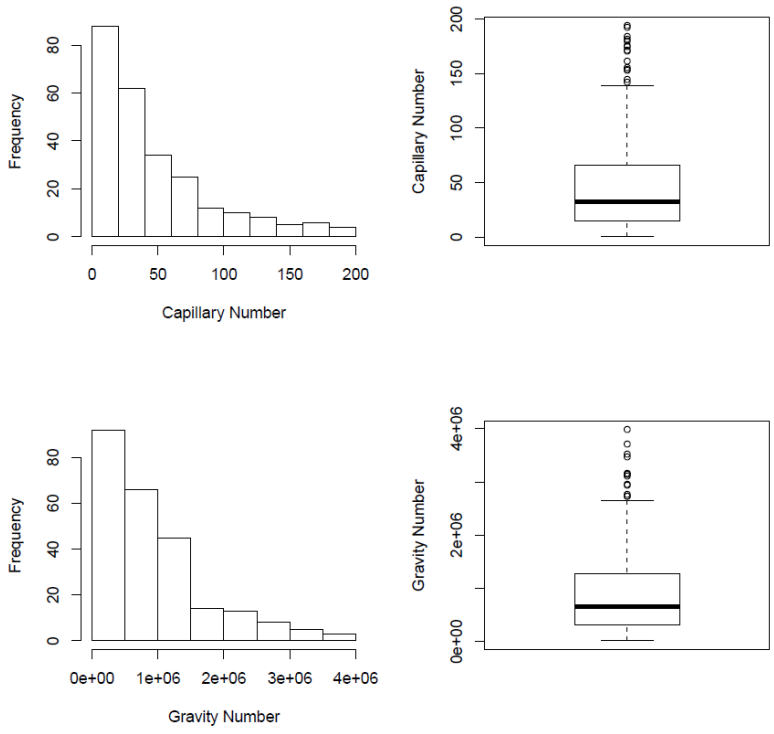



**Figure 17 : Log- normal Distribution of Capillary Number ($N_{pc}$) and Gravity Number ($N_g$)**

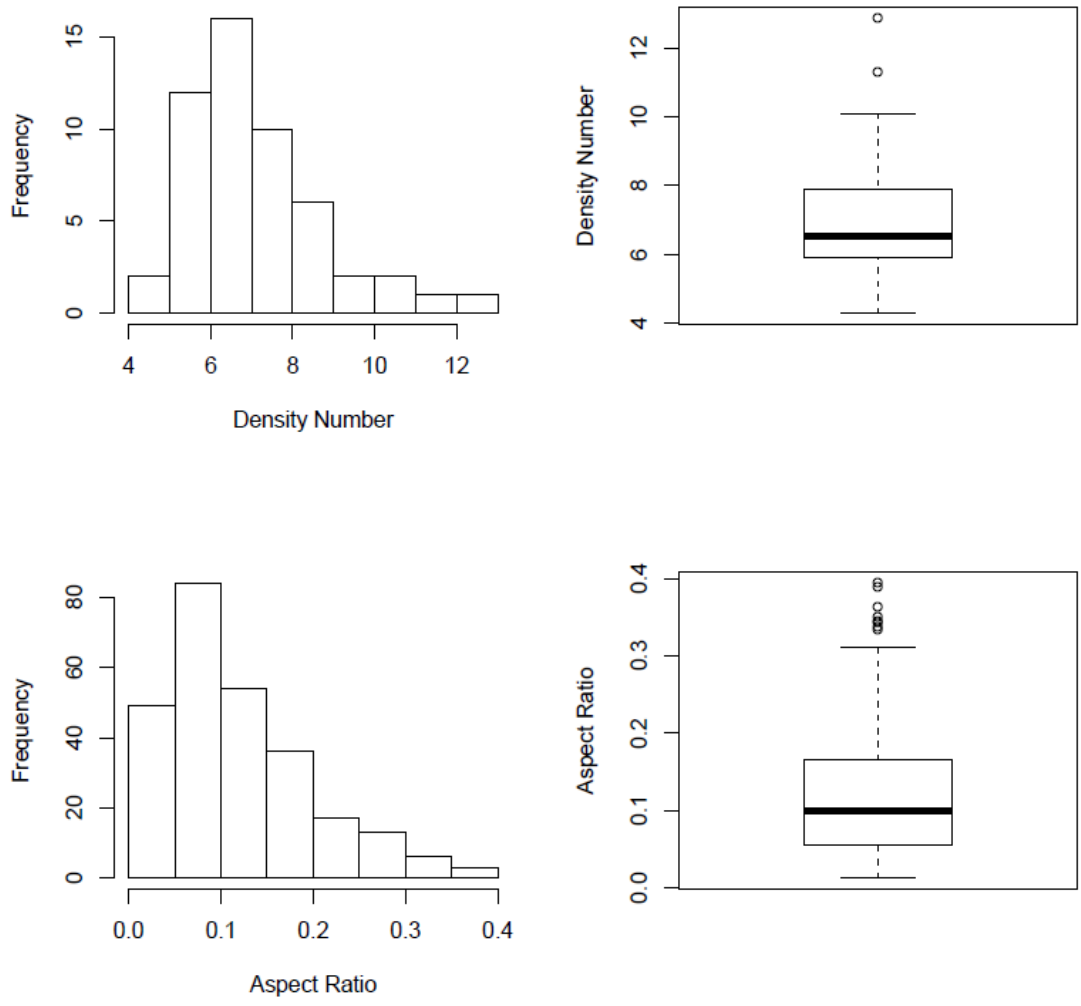



**Figure 18 : Distribution of Density number ($D_n$) and Aspect ratio ($R_l$)**

# Chapter 5: Results and Conclusions

*"Reservoirs pay little heed to either wishful thinking or libelous misinterpretation. They continually unfold a past which inevitably defies all "man-made" laws. To predict this past while it is still the futures is the business of the reservoir engineers. But whether the engineer is clever or stupid, honest or dishonest, right or wrong, the reservoir is always "right"" (Muskat 1947)*

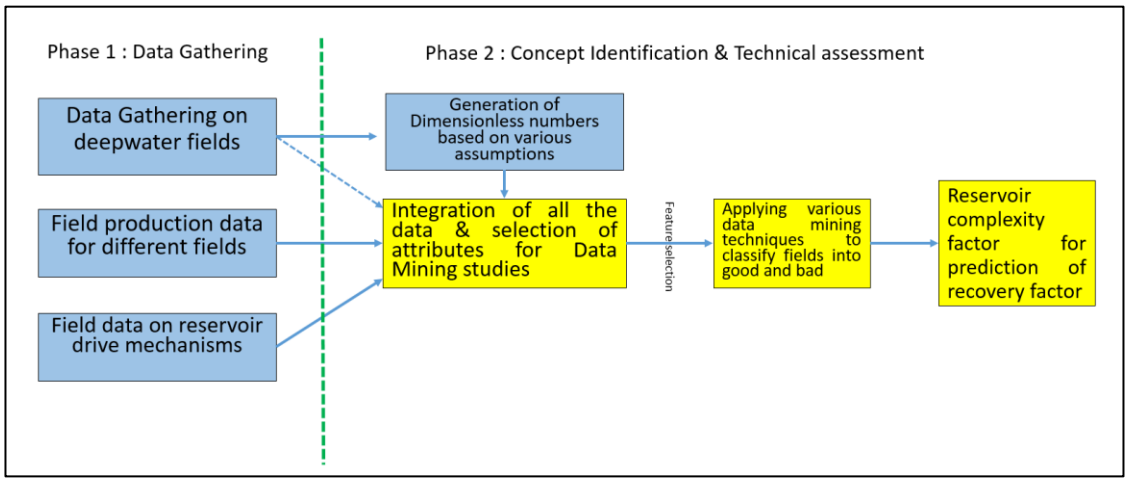



**Figure 19 : Workflow for Prediction of Recovery Factor in dGOM oilfield dataset**

**Figure 19** describes the Workflow used in the integration of all the measured and calculated data for prediction of recovery factor on dimensionally scaled reservoir models. Feature selection of important parameters is done based on domain knowledge attributing to recovery factor prediction. Following features are selected for application of data mining algorithms: FSTRUC, FTRAP1, CHRONOZONE, OTHK, OAREA, POROSITY, PERMEABILITY, SW, SDPG, SDTG, RSI, BOI, WDEP, ORF, P_CUMCOIL, P_RECOIL (refer to nomenclature section for definition of each feature).

## 5.1 Classification of oilfields Using Original Attributes

General workflow adopted for clustering is shown in **Figure 20**. Initially, classification algorithms are run on 282 deepwater oilfields obtained after removing outliers and fields with no data. In addition, since drive mechanisms are chief factor in final recovery factor calculation (Arps and Roberts 1955). Hence, fifty-nine reservoirs having water drive mechanism are scaled using dimensionless numbers. They are subsequently used in clustering and regression algorithms.

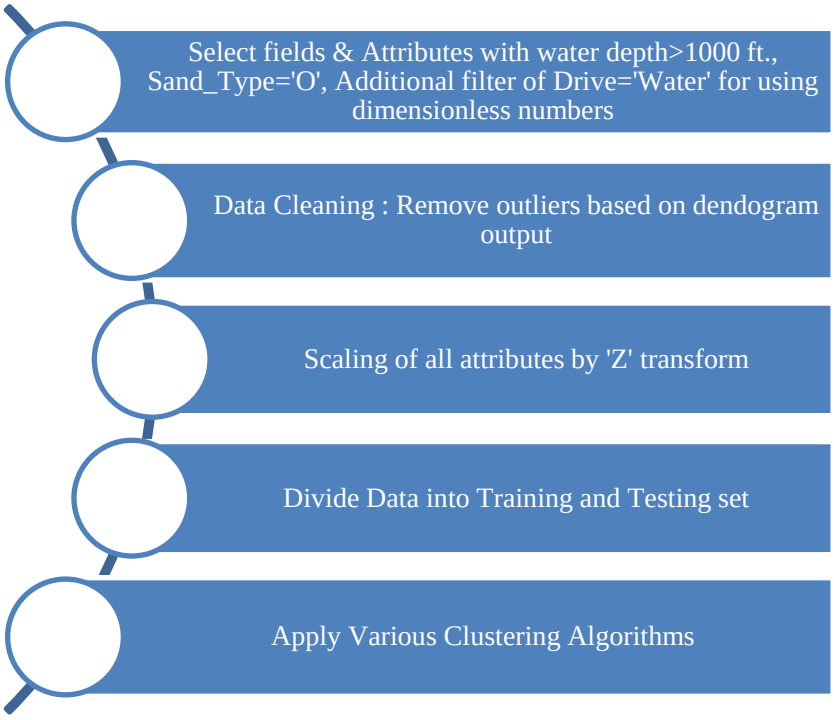



**Figure 20 : Workflow for Classification of Recovery factor**

### *5.1.1 Hierarchical Clustering (HC)*

First forty fields from set of fifty-nine fields having water drive are selected to train the hierarchical clustering model. Applying a distance based clustering techniques with 'complete' linkage algorithm results in **Figure 21.** Three sets of clusters are obtained as

illustrated by horizontal dotted line. This algorithm gave accuracy of 32% based on application of trained model to testing dataset.

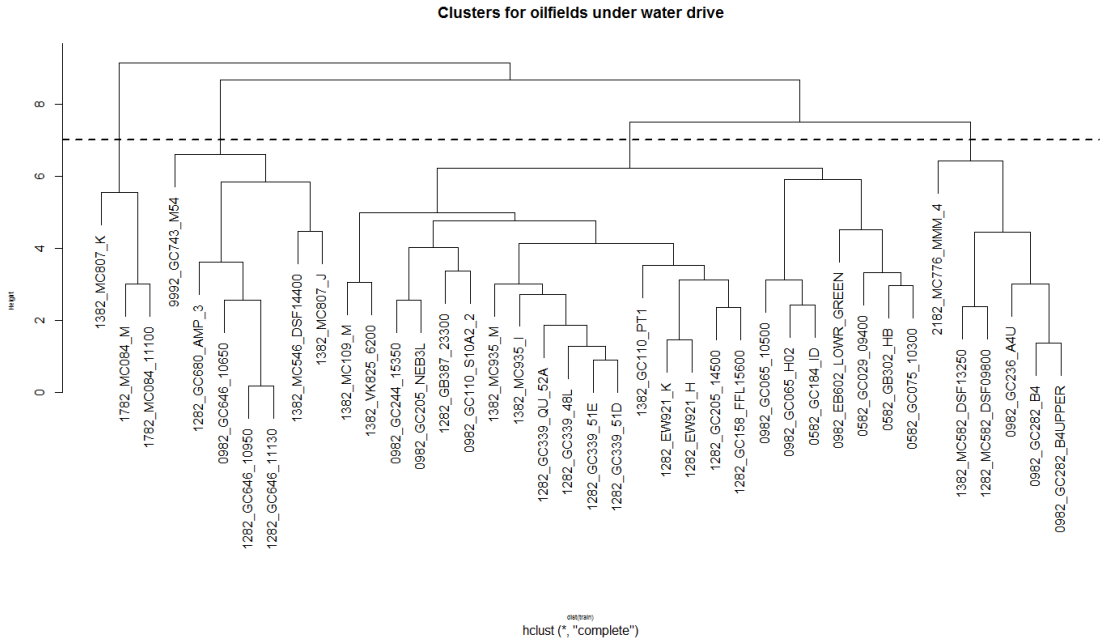



**Figure 21 : Dendograms of testing dataset**

### *5.1.2 K-means Clustering*

After removing outliers, Two eighty-two oilfields are used for classification based on original attributes. This centroid based clustering scheme leads to three clusters as determined from elbow plot **Figure 22** . Cluster '1' have 118 members. Cluster '2' have 144 members and cluster '3' have 20 members. **Figure 23** illustrates behavior of different clusters as a function of original attributes used in K-means clustering. Due to wide scattering of data, clusters overlie each other and interpretation is difficult using this figure. **Figure 24** reveals distribution of original attributes among different clusters. Clearly, higher recovery factor is associated with lower water depth, higher solution GOR, higher-pressure gradient. However, lower field area and pay thickness contribute

towards cluster with higher recovery factor. The reason for this anomaly is not completely known.

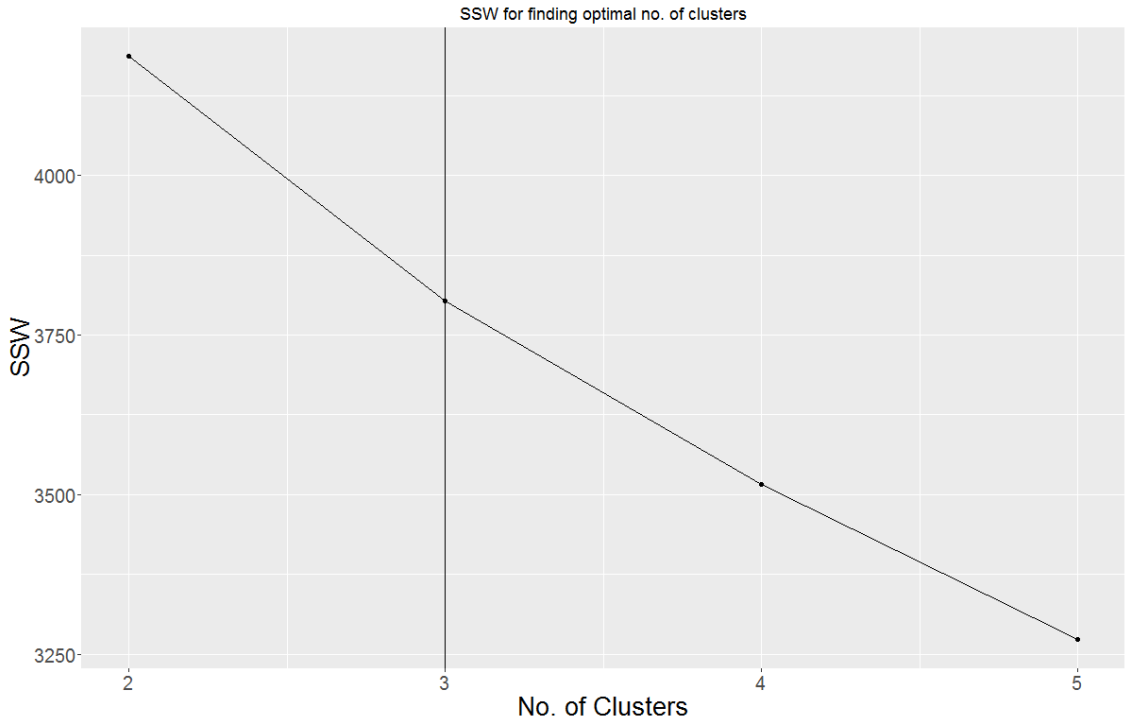



**Figure 22 : Elbow Plot for determination of number of optimum clusters.**

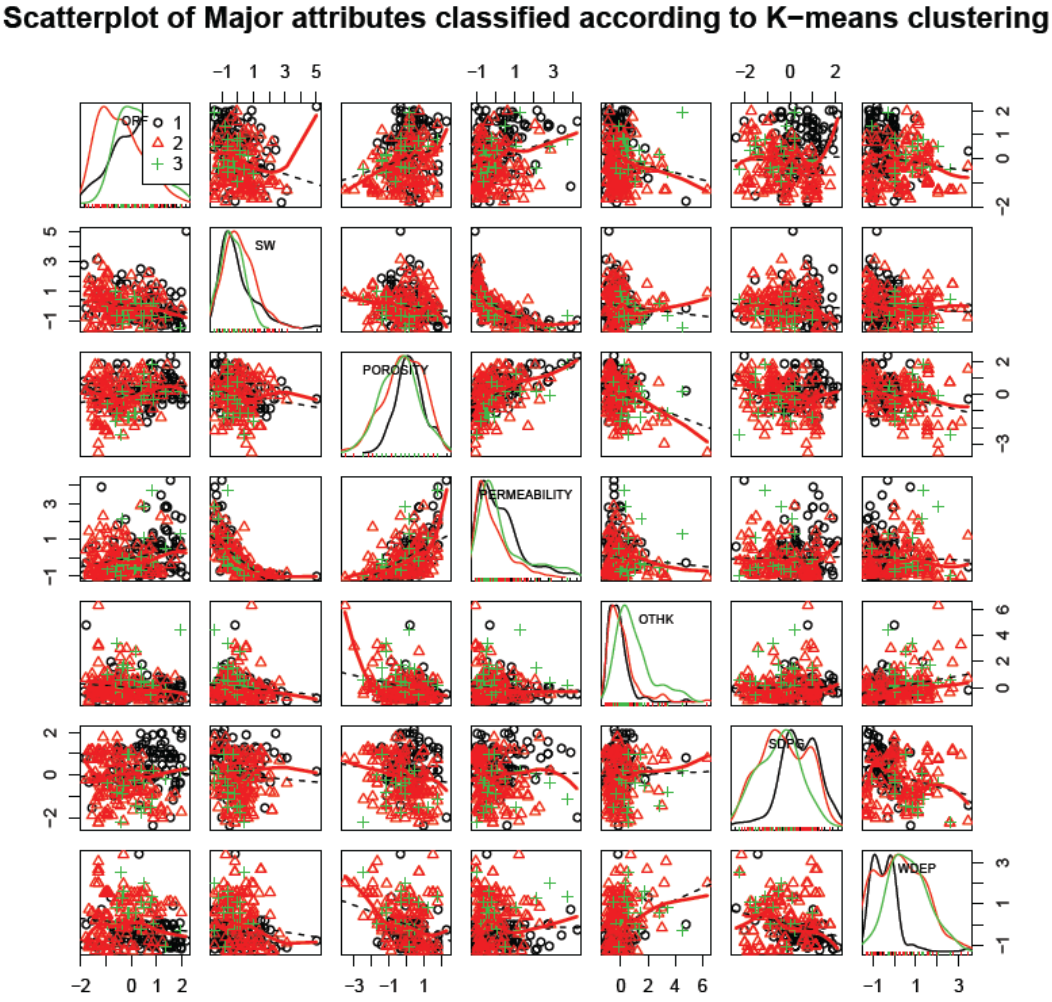



**Figure 23 : K-means clustering on all 395 oilfields in dGOM**

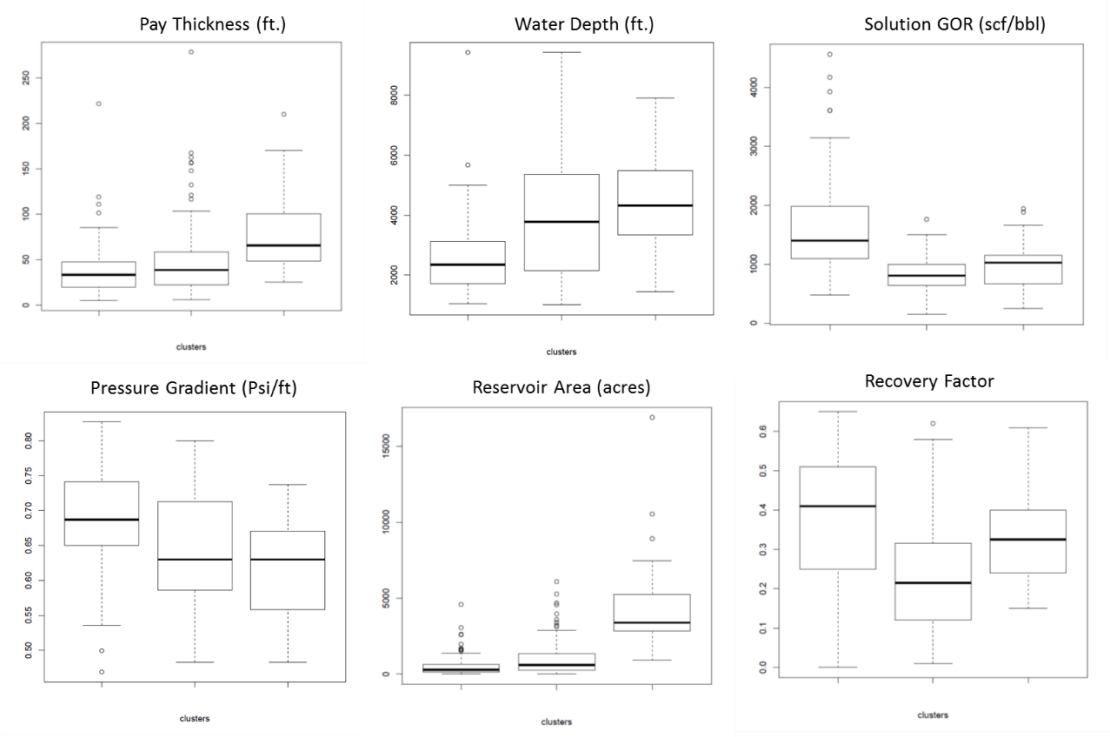



**Figure 24 : Distribution of original attributes in individual clusters**

### *5.1.3 K-Means after transformation of Non-normal distributed attributes to Gaussian distribution.*

Based on the initial distribution of attributes. Attributes which are log-normally distributed (OTHK, OAREA, PERMEABILITY, RSI, BOI, P_CUMCOIL, P_RECOIL) are converted to Gaussian distribution using log transformation. Subsequently, K-means is run on this transformed set of variables. The results obtained are described in Appendix-B.

### *5.1.4 Self-Organizing Maps (SOM)*

These algorithms provides a robust technique for visualization of inherent clusters in the dataset. **Figure 25** shows SOM training iterations progress, y-axis represents distance

from each node. The attainment of plateau at approximately 80$^{th}$ iteration is the criterion for convergence.

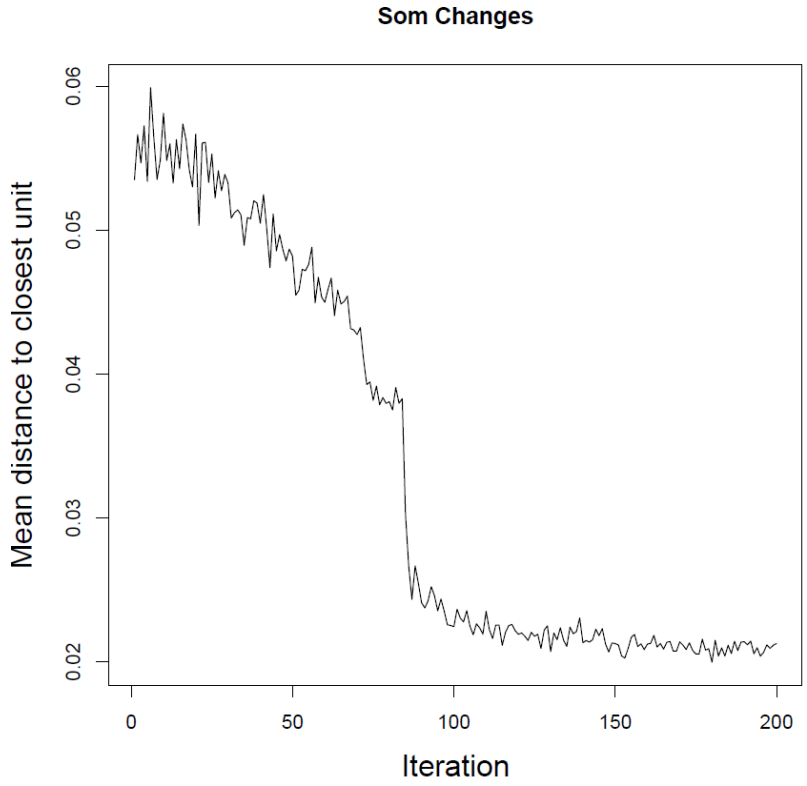



**Figure 25 : Validation of convergence in iterations**

**Figure 26** shows a U-matrix neighbor distance plot. This plot is used to define number of clusters in a dataset. Areas of low neighboring distances indicates groups of nodes that are similar. Thus, areas with large distances is visualized as natural boundaries between clusters.

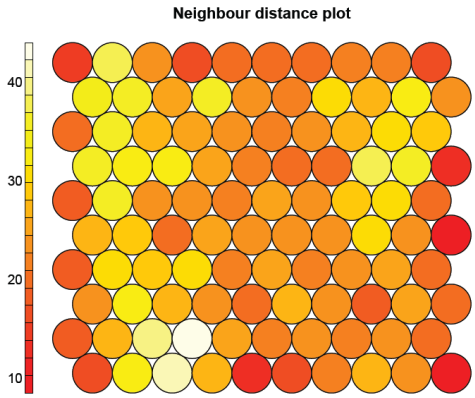



**Figure 26 : The U-matrix plot to identify clusters within the SOM group**

**Figure 27** is the heat map plot for different attributes used in SOM. It can be inferred that top left nodes with higher recovery factor are associated with high permeability, low water saturation, intermediate GOR, lower water depth, higher choronozone and FSTRUC values. However, interpretation of these maps are highly subjective therefore, caution and validation is require before using them as a classifying tool.

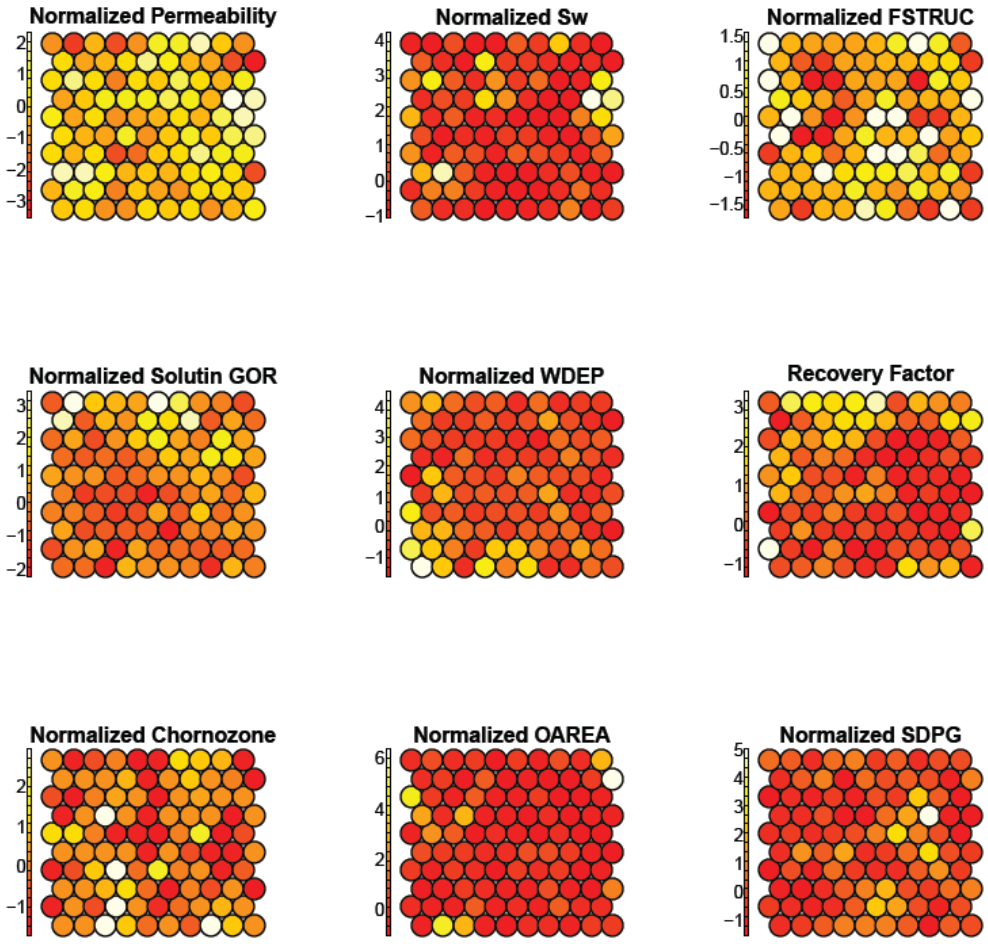



**Figure 27 : Distribution of each attribute across the latent space.**

### *5.1.5 Principal Component Analysis (PCA)*

First two principal components (PC's) were able to explain approximately 50% of variance in the training dataset. Interestingly, clusters obtained by K-means were distinctly seen in PC space but first two PC's that limits its application capture only 30% of variability. These cluster points when overlain on bi-plot of original attributes used in analysis are shown in **Figure 28**. Three distinct clusters is easily inferred from this figure,

it also gives importance of original attributes in the given clusters. Cluster '3' (blue points) is influenced by P_RECOIL, P_CUMCOIL. Cluster '2' (green points) are inclined towards WDEP, SW, OTHK. Cluster '3' (pink points) are dominated by DRIVE, Permeability, Porosity and reservoir fluid properties.

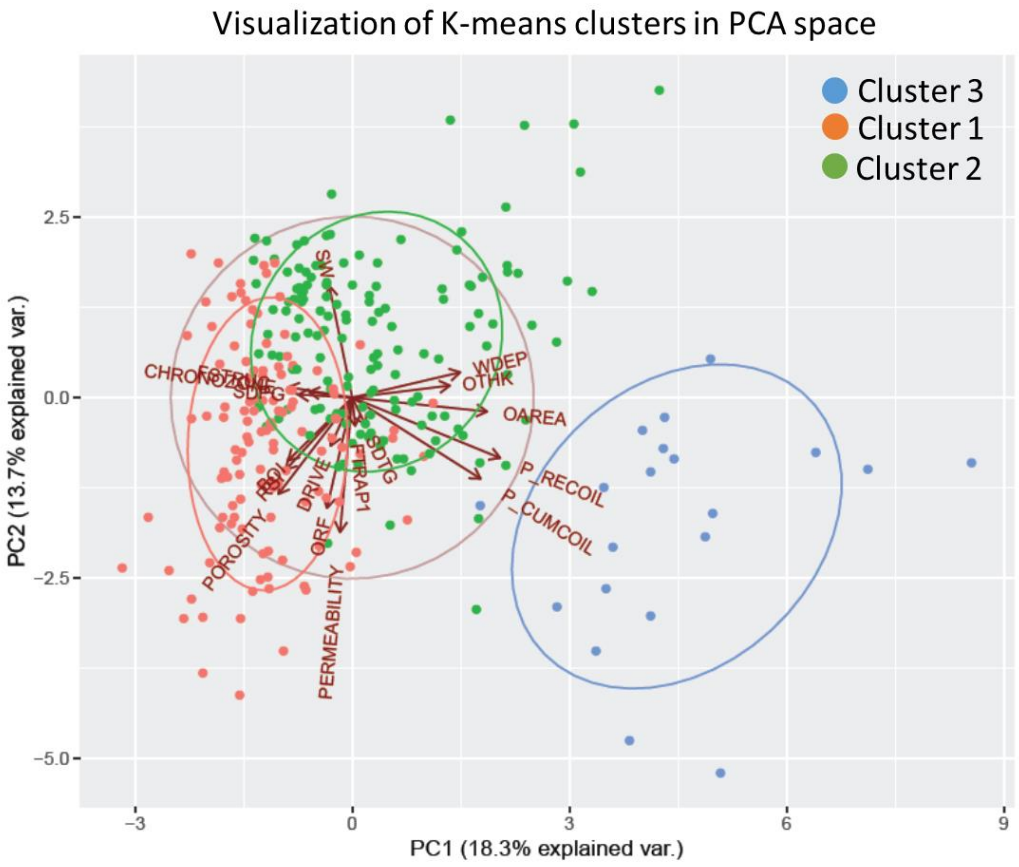
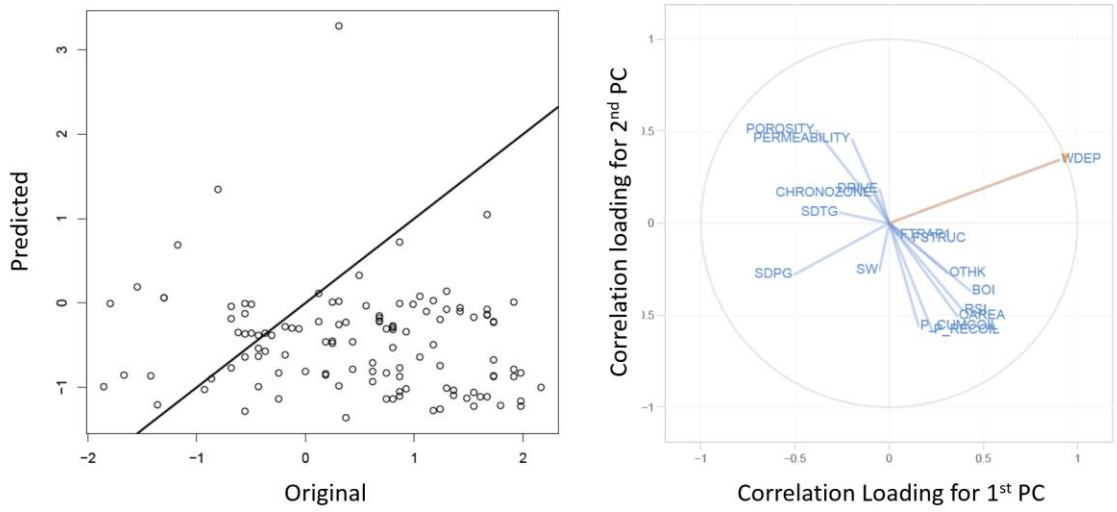



**Figure 28 : Bi-plot for K-means clusters in a principal component (PC) space. Arrows show correlation between the used attributes. See nomenclature for definition of each attribute.**

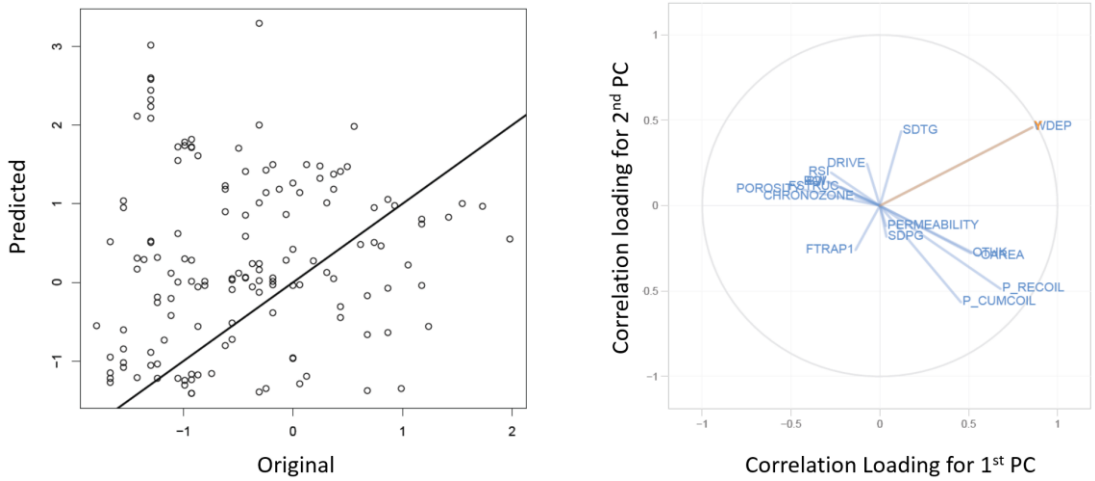
### *5.1.6 Partial Least Square Regression (PLS)*

Using training dataset for PLS regression of recovery factor results in very low $R^2$ as given in **Figure 29** below. Circle of correlation depicts how the target variable 'Y' (recovery factor) is related to various original attributes used in calculations. Due to high scattering of dataset, univariate PLS have a very low $R^2$ for all the three clusters. Highest $R^2$ is 0.2 for cluster 3. Thus, in conclusion PLS is not the appropriate technique for regression in this dataset.

**Univariate PLS on Cluster 1. 'Y' is recovery factor**




**Cluster 2**



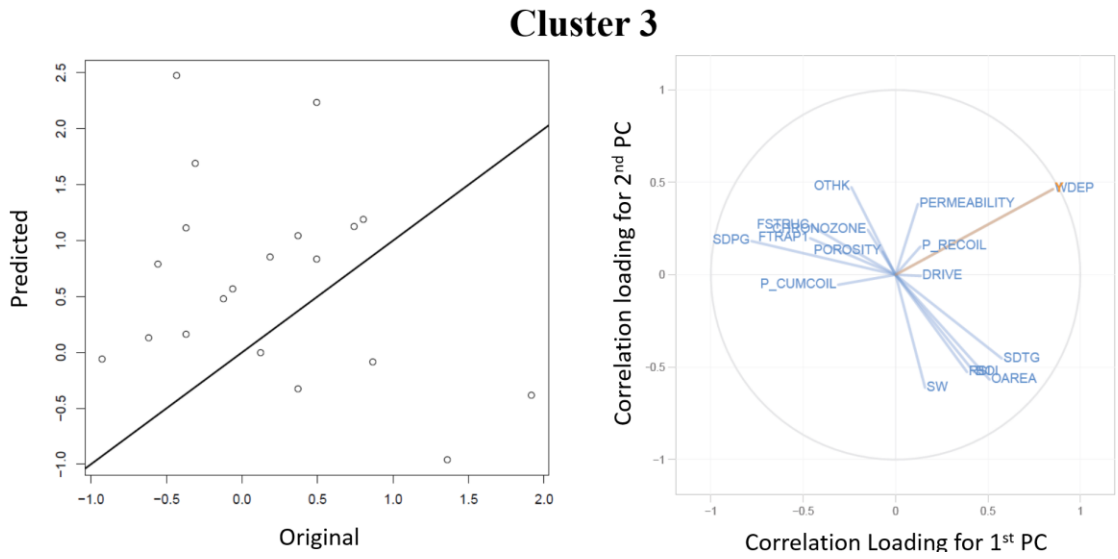



**Figure 29 : PLS match and circle of correlation for Cluster 1, 2 and 3**

## 5.2 Classification of Water Drive Oilfields Using Dimensionless Numbers

Using same dataset as in section 5.1.1, Dimensionless numbers are calculated for each field. Dip angles are assumed as per **Table-6**

**Table 6**. Due to lognormal distribution of dimensionless numbers, they are converted to their corresponding logarithmic values before application of any classification technique. In addition, five fields in block '1582_MC' and '0582_GB/GC' have to be removed as outliers based on dendogram output.

### *5.2.1 Clustering*

**Figure 30** illustrate dendogram for oilfields under water drive based on euclidean distance between dimensionless numbers. **Figure 31** displays elbow plot to optimize number of clusters in k-means clustering. '4' clusters are chosen based on this sum of square within clusters (SSW) method. The use of K-means and hierarchical clustering

leads to four clusters. Cluster ‘1’ have eight members, cluster ‘2’ have seven members, cluster ‘3’ has nine members and cluster ‘4’ have sixteen members.

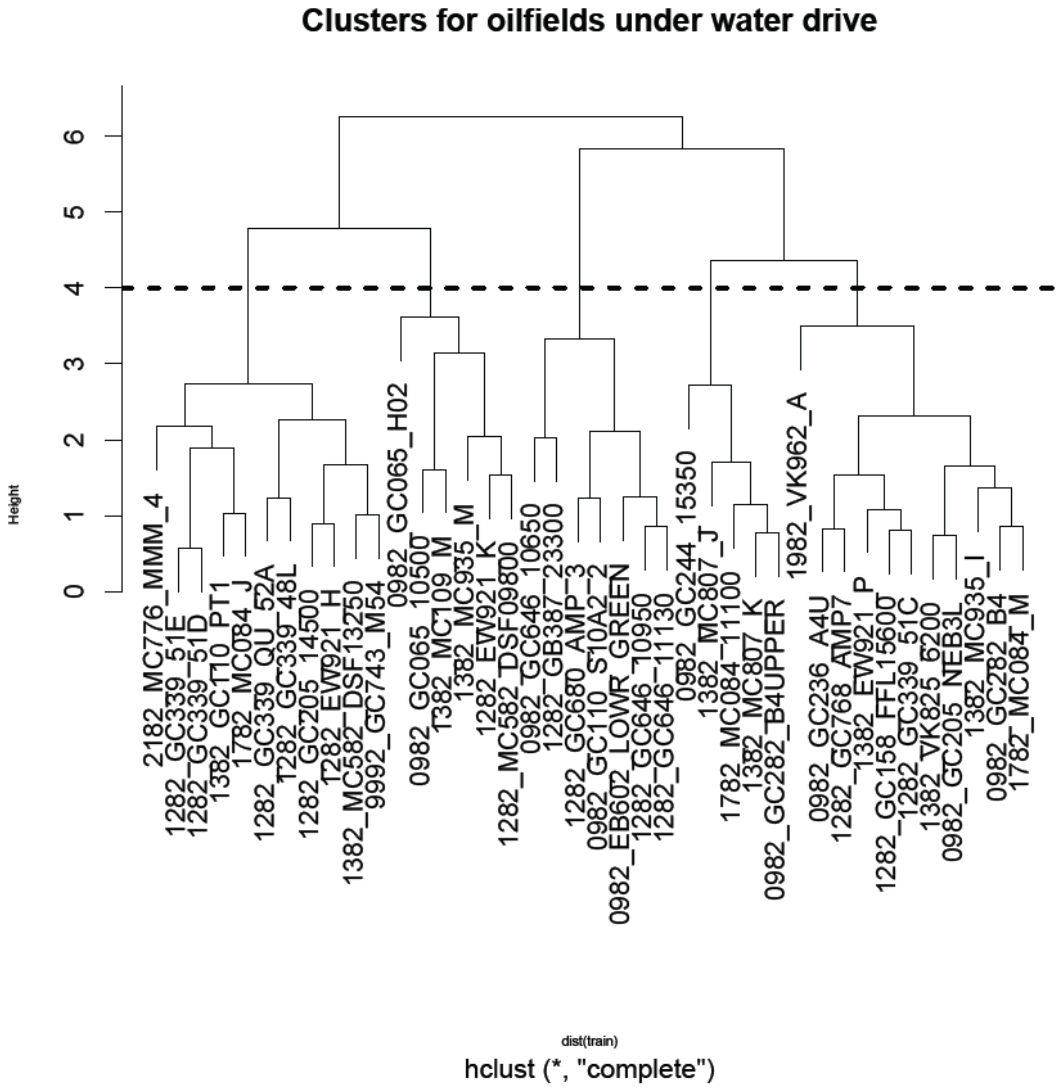



**Figure 30 : Distance based dendogram on normalized dimensionless numbers. X-axis is defined by sand names.**

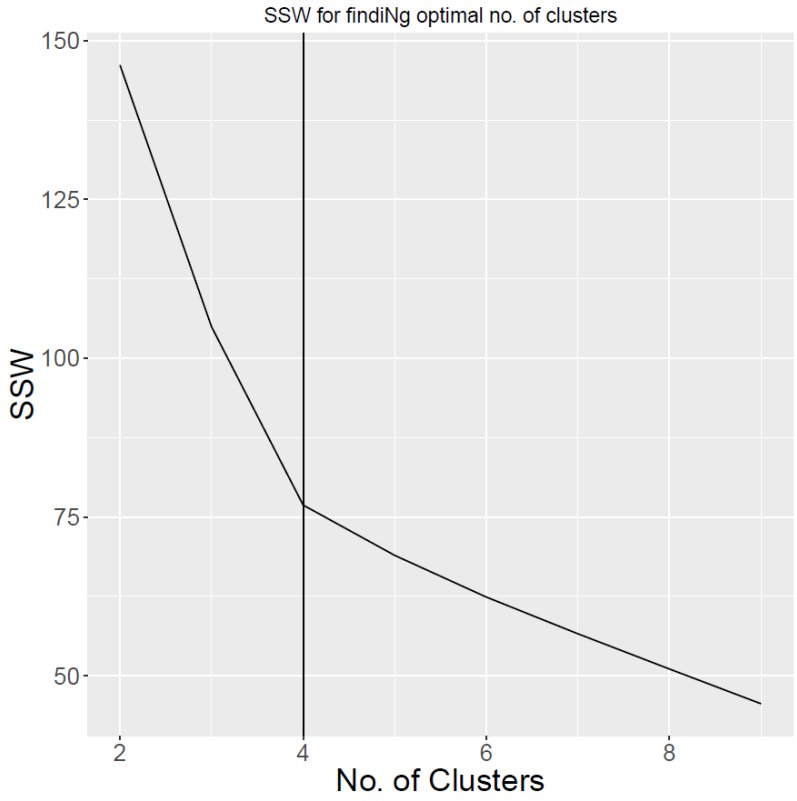



**Figure 31 : Optimum number of clusters for K-means clustering**

**Figure 32** describes the scatter plot for various clusters in dimensionless variable. This figure shows cluster '2' (Red triangles) as having low recovery, this corresponds to lower aspect ratio ($R_l$) and high capillary (Npc) and gravity (Ng) numbers. Cluster '1' (Black circles) have intermediate recovery factor. While cluster '3' (Green plus) and cluster '4' (Blue crosses) have high to intermediate recovery factor range. These clusters correspondingly relates with dimensionless numbers. Cluster '4' with smaller $N_g$ values displays higher magnitude of density number and aspect ratio implying these reservoirs to be dominated by magnitude of residual oil saturations and relative permeability. Cluster '1' have high recovery factor owning to lower magnitude of capillary number, density number and aspect ratio. The figure also shows that at intermediate values of Npc

clusters overlap each other in recovery factor. The reason for this behavior is not completely known, but it can be reasoned out due to different petro-physical characteristics of reservoirs, which present dimensionless groups does not capture.

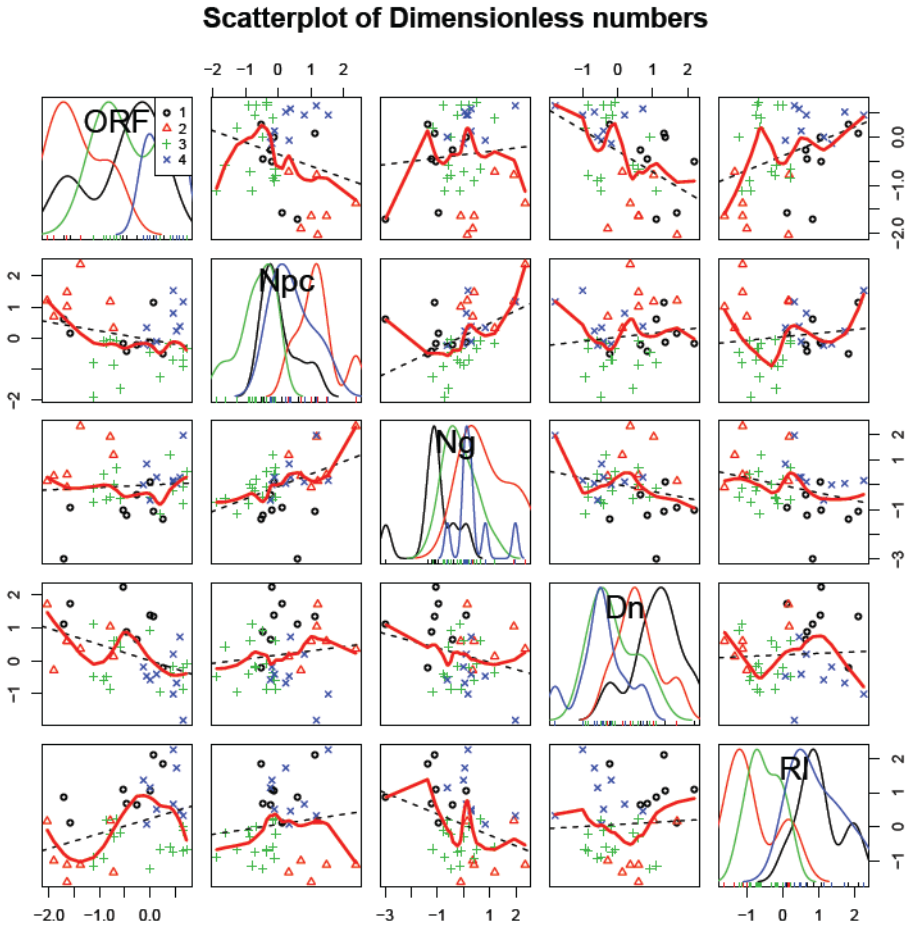



**Figure 32 : Scatter Plot for Dimensionless numbers**

### *5.2.2 Self Organizing Maps*

**Figure 33** is a neighbor distance plot. The nodes with low distance indicate groups of nodes similar to each other while areas with high distance indicate dissimilar nodes. **Figure 34** shows successful convergence of SOM algorithm iterations for finding BMU. **Figure 35** displays relationship between normalized values of dimensionless numbers

and recovery factor. So, based on this figure higher recovery factor (top right nodes in ORF) is related to higher capillary no. (Npc), intermediate gravity number (Ng), lower density no. (Dn) and higher aspect ratio ($R_l$).

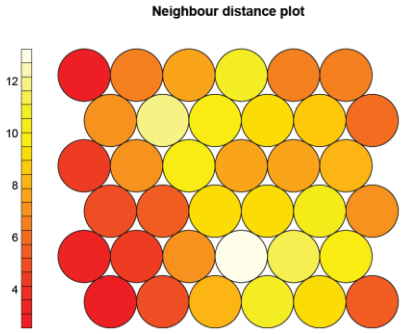



**Figure 33 : U-Matrix plot for clusters obtained using SOM**

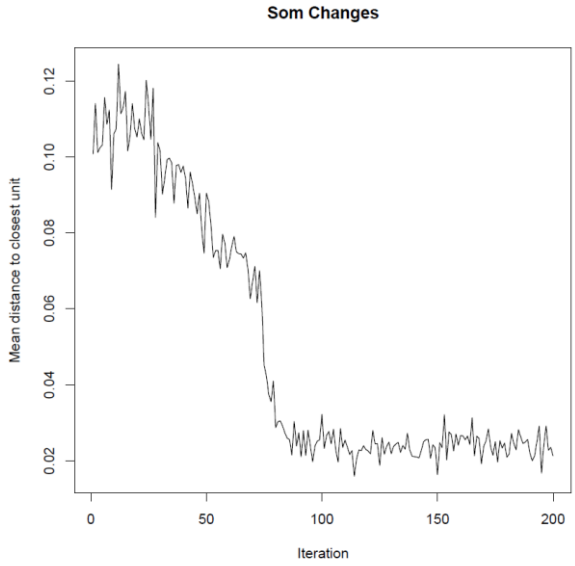



**Figure 34 : SOM iterations reaches convergence at after 70 cycles**

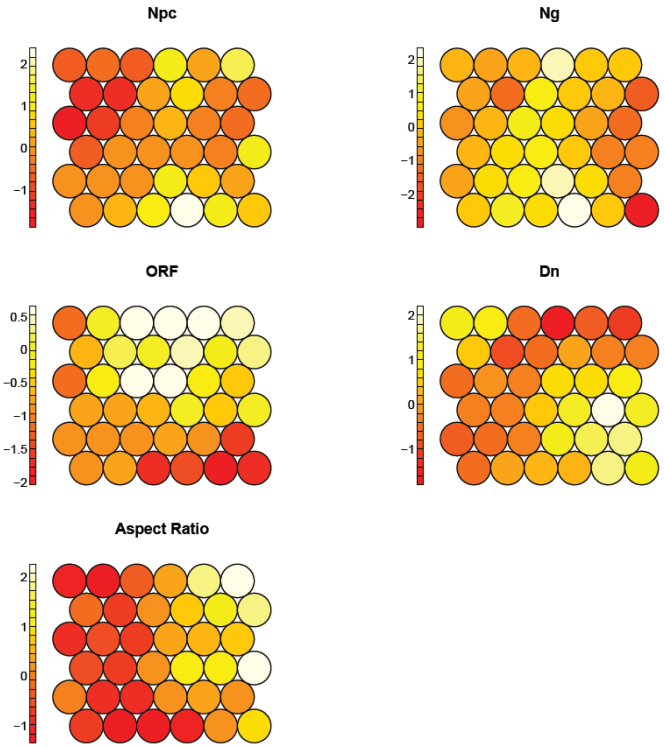



**Figure 35 : Heat map for dimensionless numbers in SOM latent space**

### *5.2.3 PCA and PLS*

**Figure 36** illustrates behavior of clusters using K-means in PCA space. Two PC's are able to capture 63% of variation in data. The figure also illustrates how dimensionless numbers affect grouping of clusters. For example, capillary number (Npc) dominates cluster 2. While cluster '4' is affecting by density number and aspect ratio. Although, there is no clear distinction between cluster '4' and cluster '3' in this plot.

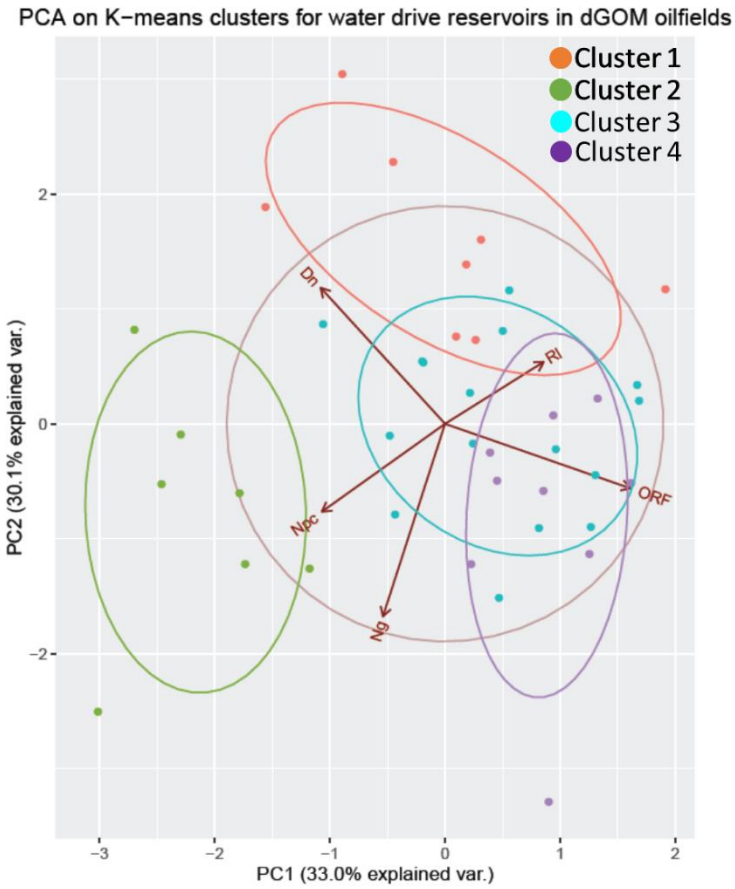



**Figure 36 : Clusters of dGOM oilfields having water drive in PC space. Arrows represents correlation between different dimensionless numbers and recovery factor (ORF).**

**Figure 37** to **Figure 40** shows PLS match for different clusters. Clusters '1' and '2' shows good match due to linearity of data points. Circle of correlation shows how target variable 'Y' (recovery factor) is related with individual dimensionless numbers.

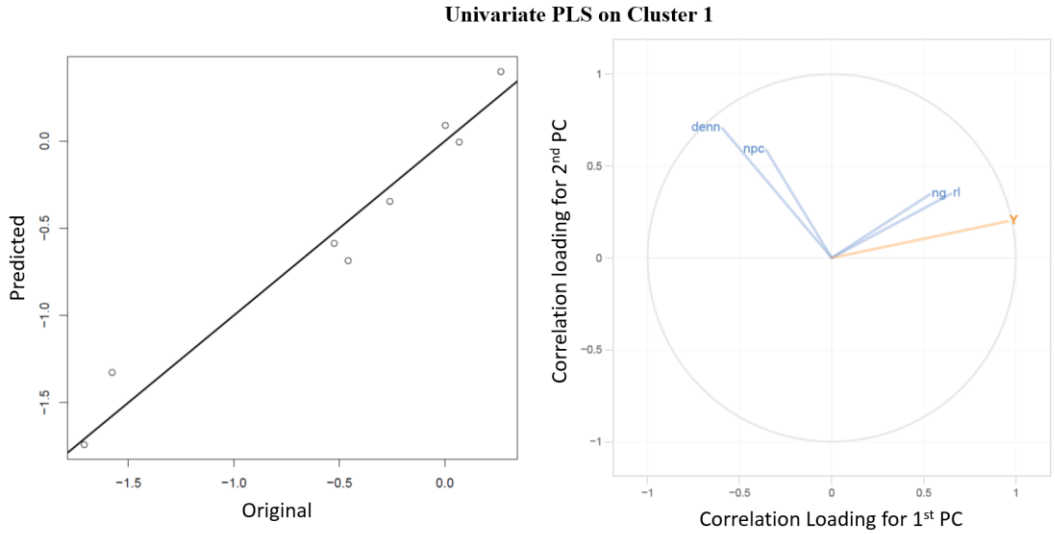



**Figure 37 : PLS result for Cluster-1. $R^2$ 0.92**

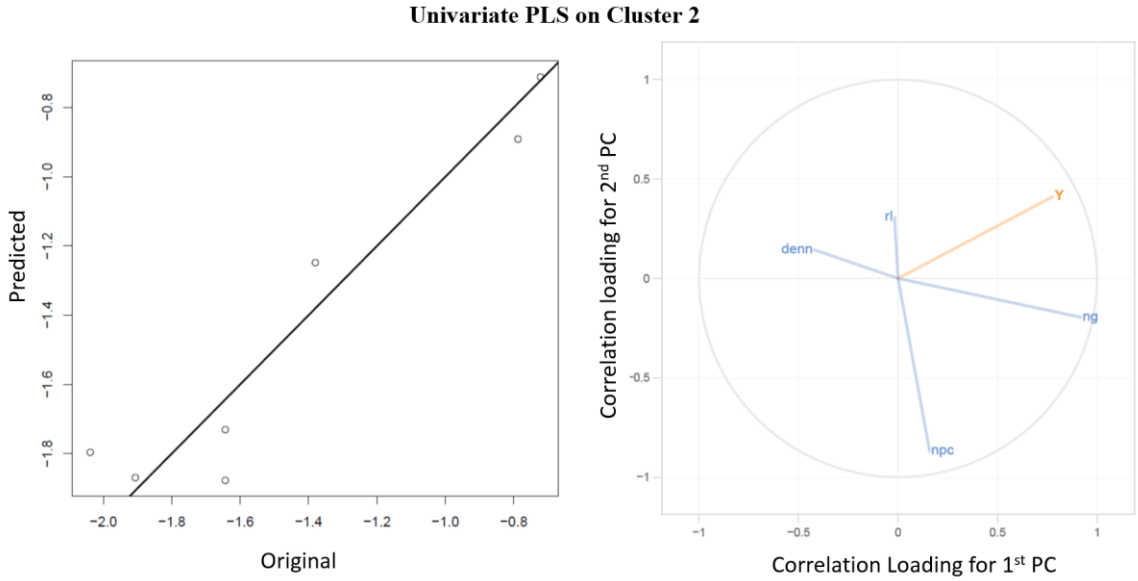



**Figure 38 : PLS result for Cluster-2. $R^2$ 0.61**

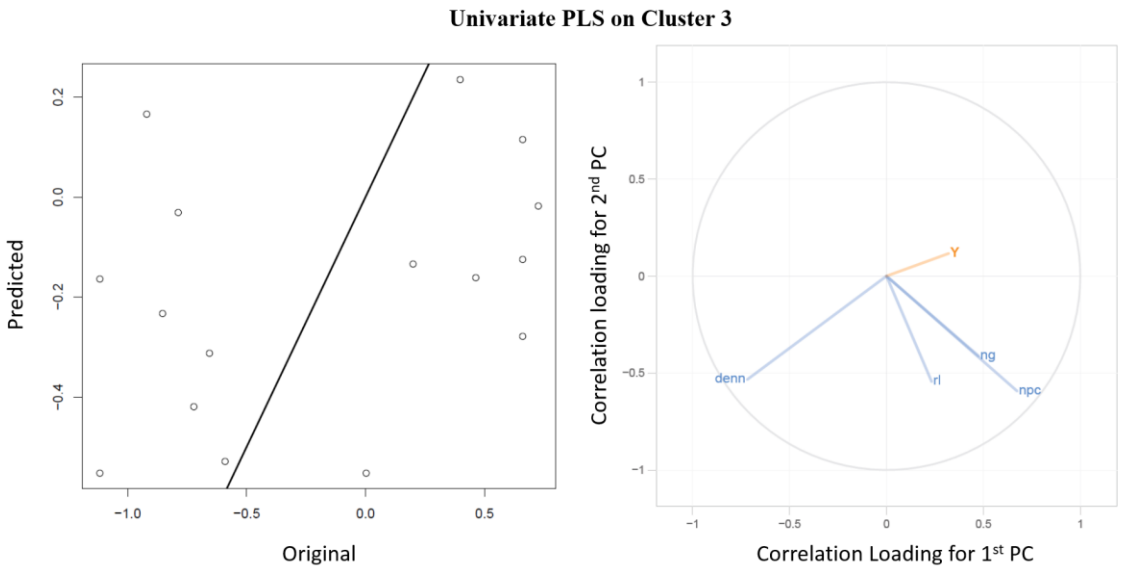



**Figure 39 : PLS result for Cluster-3. $R^2$ 0.1**

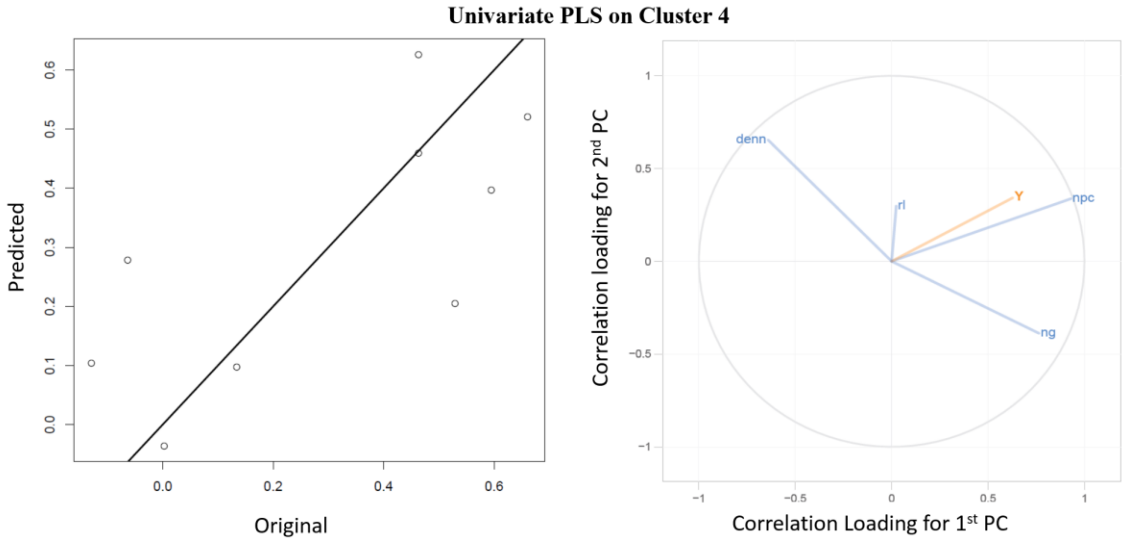



**Figure 40: PLS result for Cluster-4. $R^2$ 0.4**

Very low $R^2$ for cluster '3' makes it interesting to see why we are having the anomaly for this cluster. Visualizing distribution of dimensionless numbers in individual clusters using box plots **Figure 41** cluster '3' have lowest capillary number while its recovery factor cover ranges for cluster '1' and cluster '4' (**Figure 42**)

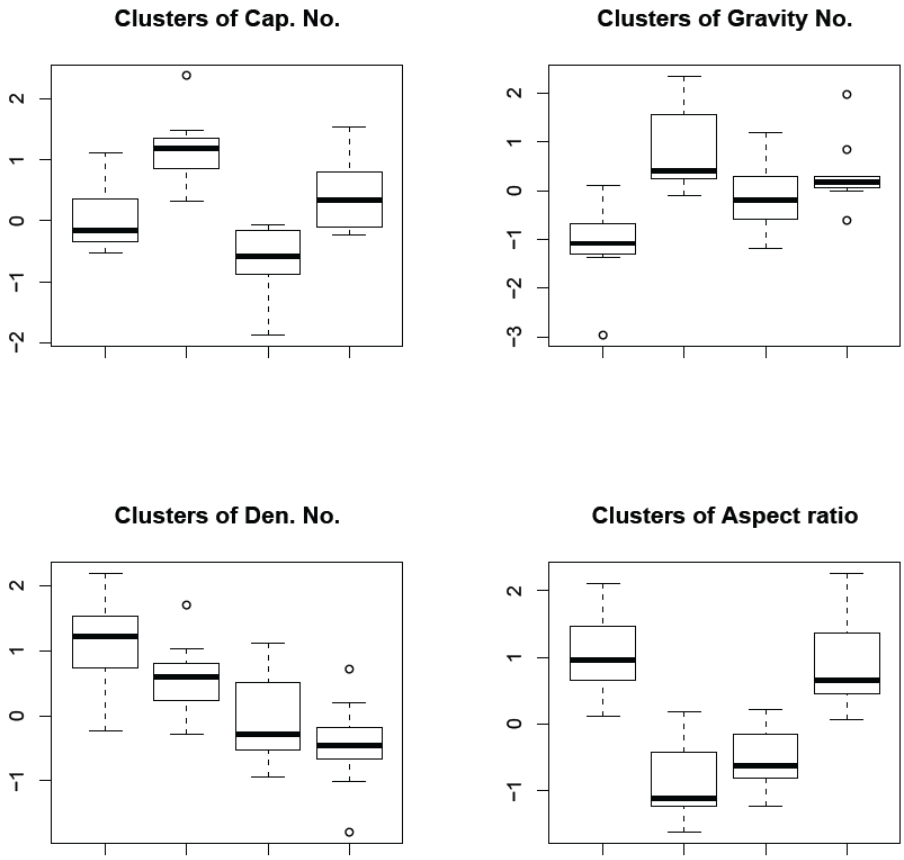



**Figure 41 : Distribution of Normalized dimensionless numbers in individual clusters.**

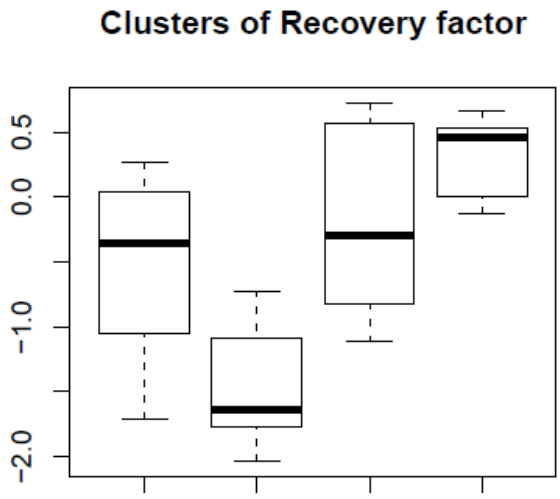



**Figure 42: Distribution of Normalized recovery factor values for different clusters**

**Cluster '1' PLS equation**

$$RF = -0.63 - 0.29N_{pc} + 0.48N_g - 0.175D_n + 0.813R_l$$

**Cluster '2' PLS equation**

$$RF = -1.87 - 0.67N_{pc} + 0.69N_g + 0.5D_n - 0.37R_l$$

**Cluster '3' PLS equation**

$$RF = -0.23 - 0.01N_{pc} + 0.24N_g - 0.35D_n - 0.11R_l$$

**Cluster '4' PLS equation**

$$RF = 0.131 + 0.33N_{pc} + 0.11N_g + 0.13D_n + 0.043R_l$$

# Chapter 6: Conclusions and Future work

*"Information is Powerful, but it is how we use it that will define us" (Larry Page)*

## 6.1 Summary of work

In this study, classification and regression algorithms are applied to deepwater oilfields dataset from Gulf of Mexico. Even though it is possible to classify fields in three categories using original attributes. It becomes necessary to scale reservoir models before using regression algorithms. This thesis successfully applies dimensionless variables for immiscible displacement of oil by water to dGOM oilfields under water drive mechanism. The result after application of PCA, K-means and PLS algorithm is a set of generalized correlation for prediction of recovery factor. However, only some of the clusters shows good regression coefficient limiting the applicability of PLS technique to particular class of reservoir under water drive.

## 6.2 Conclusions

From this work, the following conclusions can be drawn:

1. K-means was able to classify fields according to recovery factor by using original attributes; three clusters are distinctly visualized in PCA space. However, since cluster analysis is an exploratory tool; the outputs of clustering algorithms only suggests some truth in hypotheses; they cannot be used to prove any hypothesis about natural organization of data.
2. Regression is the self-assured technique in mathematics that can be utilized for prediction. PLS is used in this thesis for prediction of recovery factor, the maximum coefficient of correlation by using PLS on original attributes is 0.2 for cluster 3. Whereas, by using k-means on dimensionally scaled reservoir models

i.e. the set of dimensionless numbers, the maximum coefficient of correlation is 0.92 in cluster 1.

3. On the contrary, some clusters show no correlation with very low $R^2$. This can be partially attributed to limitations of PLS methodology which works best on multicolliear correlated dataset. Other reasons may include inability of present dimensionless groups to scale the reservoir properly as these groups assumes all fields to have similar geometrical and petro-physical systems.
4. This work successfully uses dimensionless numbers derived by previous researchers along with data mining technology to generate generalized correlations for oil reservoirs. Use of K-means, PCA and PLS provided techniques that can quickly estimate the recovery factor from limited data, when comprehensive simulations on large number of reservoirs is too costly and time consuming.
5. The findings from data mining should be supported by a plausible theory. Beguiling story can disguise weaknesses in the data. In words of Patrick Whitson MIT professor on artificial intelligence "Given enough time, enough attempts, and enough imagination, almost any set of data can be teased out of any conclusion". Just having a large amount of data is no guarantee of the success of a data-mining project.

### 6.3 Suggested future work

This work is a starting point for generating estimates of recovery factor at any location in Gulf of Mexico using historical data of fields based on data mining approach. This information can be further used for the preparation of data acquisition and risk assessment

plans to set up a framework for decision-making on risks and uncertainty for optimizing reservoir management and production forecast. However, further investigations are needed to describe reasons for low regression coefficients in some of the clusters. Similar workflow with other set of dimensionless groups is used in reservoir having solution gas drive or depletion drive mechanism. In addition, research is needed for studying interaction between coefficients for dimensionless numbers in developed correlations. Present work may be extended to predict reservoir performance as a function time using Tarner's and Muskat's theories of material balance.

# References


Abou-Sayed, Ahmed. 2012. "Data mining applications in the oil and gas industry." *Journal of Petroleum Technology* 64 (10):88-95.

Arps, JJ, and TG Roberts. 1955. "The effect of the relative permeability ratio, the oil gravity, and the solution gas-oil ratio on the primary recovery from a depletion type reservoir." *Trans. AIME* 204:120-127.

Ayatollahi, Shahab, Abdolhossein Hemmati-Sarapardeh, Moahammad Roham, and Sassan Hajirezaie. 2016. "A rigorous approach for determining interfacial tension and minimum miscibility pressure in paraffin-CO 2 systems: Application to gas injection processes." *Journal of the Taiwan Institute of Chemical Engineers* 63:107-115.

Beshears, John. 2013. "The performance of corporate alliances: Evidence from oil and gas drilling in the Gulf of Mexico." *Journal of Financial Economics* 110 (2):324-346.

Brakel, Jan, Brian Tarr, William Cox, Fredrik Jorgensen, and Haakon Vidar Straume. 2015. "SMART Kick Detection: First Step on the Well-Control Automation Journey." *SPE Drilling & Completion*.

BrooksRH, CoreyAT. 1964. "Hydraulicpropertiesofporous media." *Colorado State University, Hydro Paper* 3:27.

Calhoun Jr, John C. 1976. "Fundamentals of reservoir engineering."

Chatzis, Ioannis, Norman R Morrow, and Hau T Lim. 1983. "Magnitude and detailed structure of residual oil saturation." *Society of Petroleum Engineers Journal* 23 (02):311-326.

Chorneyko, David M. 2006. "Real-Time Reservoir Surveillance Utilizing Permanent Downhole Pressures-An Operator's Experience." SPE Annual Technical Conference and Exhibition.

Craig, FF, JL Sanderlin, DW Moore, and TM Geffen. 1957. "A laboratory study of gravity segregation in frontal drives." *Trans. AIME* 210 (275):1957.

D. Nixon, Barbara J. Bascle, David A. Marin, Lesley. 1999. "Offshore Atlas project: Methodology and results." *Marine Georesources and Geotechnology* 17 (2-3):211-212.

Disenhof, Corinne, MacKenzie Mark-Moser, and Kelly Rose. 2014. "The Gulf of Mexico Petroleum System–Foundation for Science-Based Decision Making." *Journal of Sustainable Energy Engineering* 2 (3):225-236.

Elmabrouk, S, E Shirif, and R Mayorga. 2014. "Artificial Neural Network Modeling for the Prediction of Oil Production." *Petroleum Science and Technology* 32 (9):1123-1130.

Firoozabadi, Abbas, and Henry J Ramey Jr. 1988. "Surface tension of water-hydrocarbon systems at reservoir conditions." *Journal of Canadian Petroleum Technology* 27 (03).

Fox, Robert W, Alan T McDonald, and Philip J Pritchard. 1985. *Introduction to fluid mechanics*. Vol. 7: John Wiley & Sons New York.

Galloway, William E. 2008. "Depositional evolution of the Gulf of Mexico sedimentary basin." *Sedimentary basins of the world* 5:505-549.

Geertsma, J, Go A Croes, and N Schwarz. 1956. "Theory of dimensionally scaled models of petroleum reservoirs." *Trans. AIME* 207:118-127.
Holdaway, Keith. 2014. *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data Driven Models*: John Wiley & Sons.
Johnston, J, and A Guichard. 2015. "New Findings in Drilling and Wells using Big Data Analytics." Offshore Technology Conference.
Kantardzic, Mehmed. 2011. *Data mining: concepts, models, methods, and algorithms*: John Wiley & Sons.
Kotu, Vijay, and Bala Deshpande. 2014. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*: Morgan Kaufmann.
Lach, Joseph. 2010. "IOR for Deepwater Gulf of Mexico." *Knowledge Reservoir, December*.
Leverett, MoC. 1941. "Capillary behavior in porous solids." *Transactions of the AIME* 142 (01):152-169.
Liu, Yintao, Ke-Thia Yao, Shuping Liu, Cauligi Srinivasa Raghavendra, Tracy Lynn Lenz, Lanre Olabinjo, F Burcu Seren, Sanaz Seddighrad, and CG Dinesh Babu. 2010. "Failure Prediction for Artificial Lift Systems." SPE Western Regional Meeting.
McFarlan Jr, Edward, and L Silvio Menes. 1991. "Lower Cretaceous." *The Gulf of Mexico Basin: Boulder, Colorado, Geological Society of America, The Geology of North America, v. J*:181-204.
Melrose, J. C. "Role of Capillary Forces In Detennining Microscopic Displacement Efficiency For Oil Recovery By Waterflooding." doi: 10.2118/74-04-05.
Meyer, Dave, Larry Zarra, David Rains, Bob Meltz, and Tom Hall. 2005. "Emergence of the Lower Tertiary Wilcox trend in the deepwater Gulf of Mexico." *World Oil* 226 (5):72-77.
Moore, TF, and RL Slobod. 1956. "The effect of viscosity and capillarity on the displacement of oil by water." *Producers Monthly* 20 (10):20-30.
Nikravesh, Masoud. 2004. "Soft computing-based computational intelligent for reservoir characterization." *Expert Systems with Applications* 26 (1):19-38.
Novakovic, Djuro. 2002. "Numerical reservoir characterization using dimensionless scale numbers with application in upscaling." Louisiana State University.
Peters, Ekwere J, Nadeem Afzal, and Ridha Gharbi. 1993. "On scaling immiscible displacements in permeable media." *Journal of Petroleum Science and Engineering* 9 (3):183-205.
Pirson, Sylvain Joseph. 1977. *Oil reservoir engineering*: RE Krieger Publishing Company.
Post, Paul, Ralph Klazynski, Elizabeth Klocek, Thierry DeCort, Tommy Riches, Kun Li, Erin Elliott, and Rex Poling. 2012. "Assessment of undiscovered technically recoverable oil and gas resources of the Atlantic Outer Continental Shelf 2011 as of January 1, 2009." *Bureau of Ocean Energy Management BOEM* 16:39.
Rains, David B, Larry Zarra, and David Meyer. 2006. "AVThe Lower Tertiary Wilcox Trend in the Deepwater Gulf of Mexico."
Rapoport, LA, and W Leas. "J.: 1953, Properties of linear waterfloods." *Trans. AIME* 198:139-148.

Salvador, Amos. 1991. "Origin and development of the Gulf of Mexico basin." *The gulf of Mexico basin*:389-444.
Shook, Mike, Dachang Li, and Larry W Lake. 1992. "Scaling immiscible flow through permeable media by inspectional analysis." *In Situ;(United States)* 16 (4).
Slatt, Roger M, and Fuge Zou. 2014. "Turbidite Petroleum Geology in the Deepwater/Subsalt Gulf of Mexico."
Slattery, John C, Leonard Sagis, and Eun-Suok Oh. 2007. *Interfacial transport phenomena*: Springer Science & Business Media.
Tarner, Jack. 1944. "How different size gas caps and pressure maintenance programs affect amount of recoverable oil." *Oil Weekly* 144 (2):32-34.
Vazirgiannis, Michalis, Maria Halkidi, and Dimitrious Gunopulos. 2012. *Uncertainty handling and quality assessment in data mining*: Springer Science & Business Media.
Woods, RD, Amos Salvador, and AE Miles. 1991. "pre-Triassic." *The Gulf of Mexico Basin: Geological Society of America, The geology of North America, v. J*:109-130.
Wu, Xingru, Gary A Pope, G Michael Shook, and Sanjay Srinivasan. 2008. "Prediction of enthalpy production from fractured geothermal reservoirs using partitioning tracers." *International Journal of Heat and Mass Transfer* 51 (5):1453-1466.
Wyckoff, RD. 1940. "Factors affecting reservoir performance." Drilling and Production Practice.

# Appendix A: Nomenclature

| | |
|---|---|
| API | Oil API degree |
| BHCOMP | Total number of bottom-hole completions in well |
| BOI | Initial formation volume factor |
| dGOM | Deepwater gulf of mexico |
| DISCOIL/BOE | Discovered oil/ bbl. oil equivalent (bbls.) |
| FCLASS | Field class |
| FSTRU | Field structure code |
| FTRAP1 | Field primary trap type |
| FTRAP2 | Field secondary trap type |
| GOR | Gas-Oil ratio (scf/bbl) |
| H | Net pay thickness |
| OAREA | Oil zone area (acre-ft.) |
| OIP | Oil initially in-place (bbls.) |
| ORECG | Oil recoverable from gas reservoir |
| ORECO | Oil recoverable from oil reservoir |
| ORECO_AF | Recoverable oil/acre. ft |
| ORF | Oil recovery factor |
| ORP | Produced GOR from oil reservoir (Mcf/stb) |
| OTHK | Oil zone thickness (ft.) |
| P_CUMOIL/BOE | Proved cumulative oil/ bbl. oil equivalent (bbls.) |
| P_RECOIL/BOE | Proved recoverable oil/ bbl. oil equivalent |
| PC1 | First Principal component |

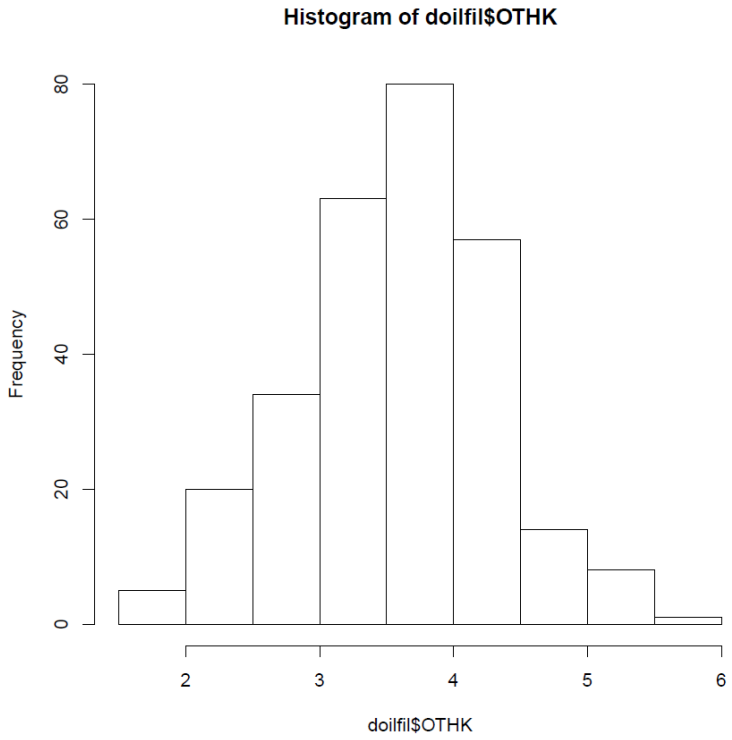
| | |
|---|---|
| PI | Initial Pressure (Psi) |
| $R_l$ | Dimensionless aspect ratio |
| RSI | Solution Gas-Oil ratio |
| SDPG | Static Pressure Gradient |
| SDTG | Static Temperature Gradient |
| SS | Subsea depth (ft) |
| SW | Water Saturation |
| TCNT | Total number of sand intersected by well |
| TI | Initial Temperature (Deg F) |
| UCNT | Number of under-saturated sand encountered in the well |
| WDEP | Water depth (ft.) |

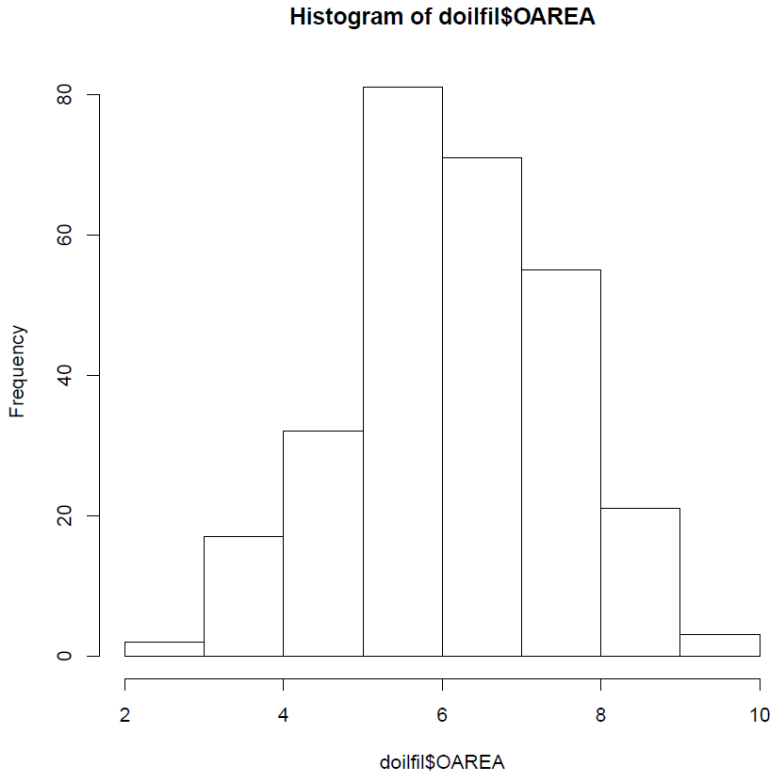
| **Symbol** | **Description** |
|---|---|
| $K_x$ | Average horizontal permeability of reservoir , md |
| $K_Z$ | Vertical permeability of reservoir , md |
| $U_t$ | Total fluid velocity (oil+water) , ft./day |
| $\lambda_{r2}^{o}$ | Relative mobility of residual phase-2 |
| $\rho_l$ | Density of non-wetting liquid phase |
| Dn | Dimensionless density number |
| $K_{rw}$ | Relative permeability to water |
| l | Length of reservoir |
| Ng | Dimensionless gravity number |
| Npc | Dimensionless capillary number |

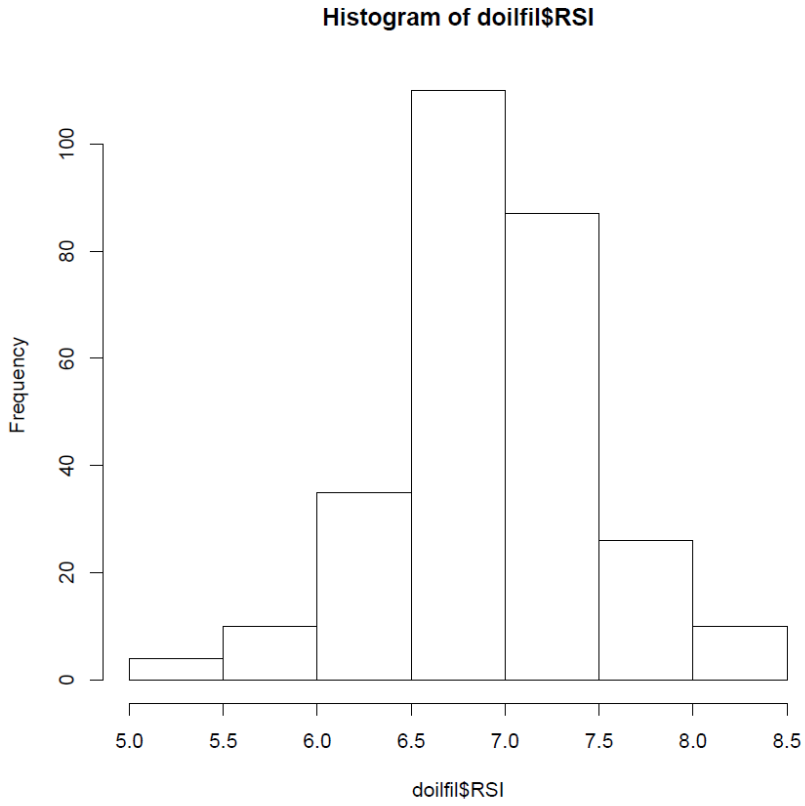
| | |
|---|---|
| Rl | Dimensionless aspect ratio |
| $\sigma_{hw}$ | Interfacial tension of hydrocarbon-water system |
| H | Net pay thickness |
| $\Delta\rho$ | Density difference between oil-water densities |
| $\alpha$ | Dip angle |
| $\phi$ | Porosity |

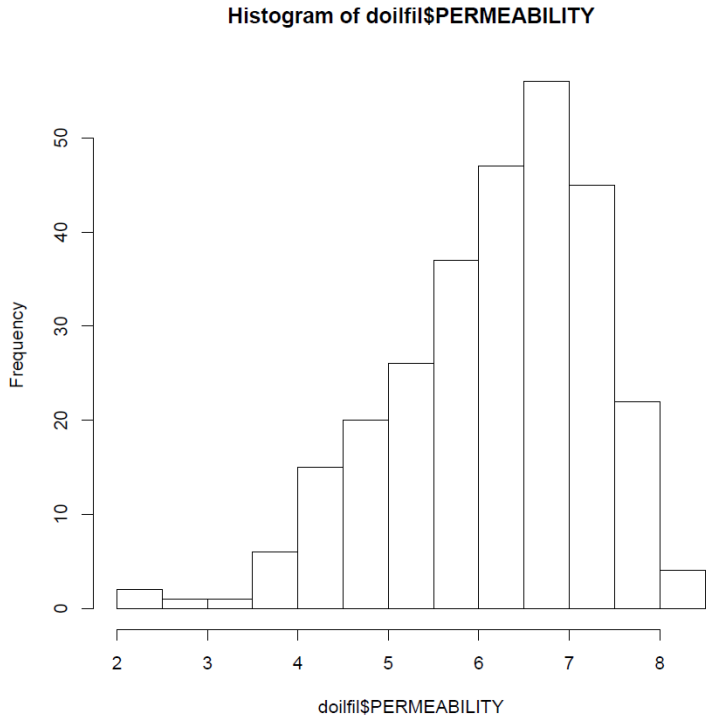
# Appendix-B: K-Means on log-transformed attributes

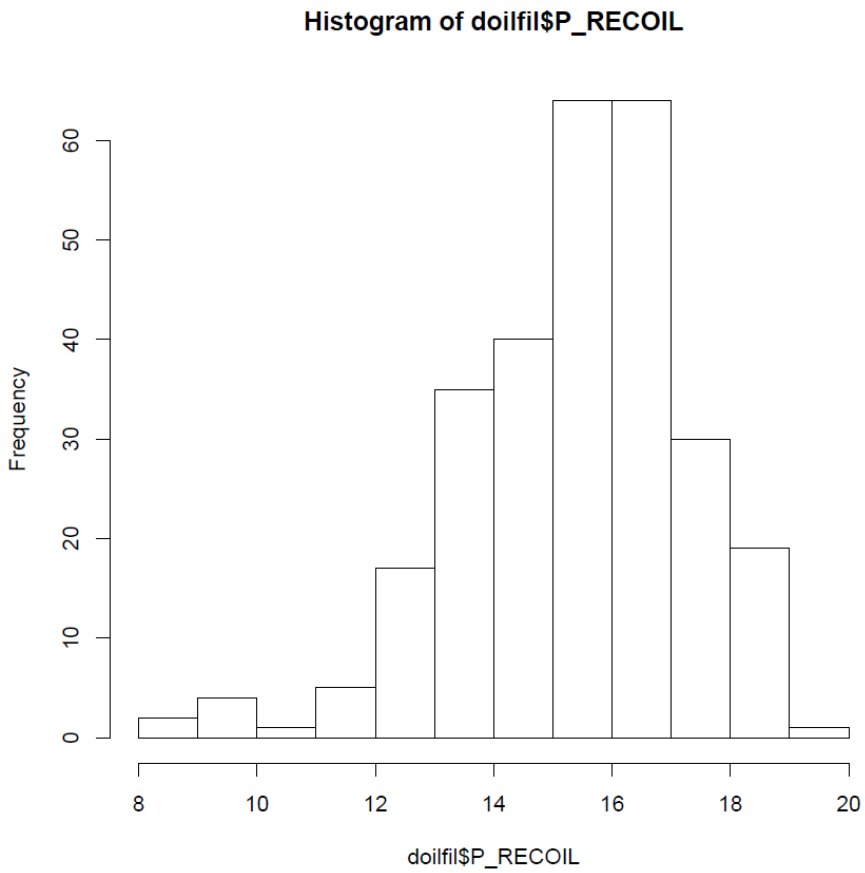
## Histogram of transformed attributes



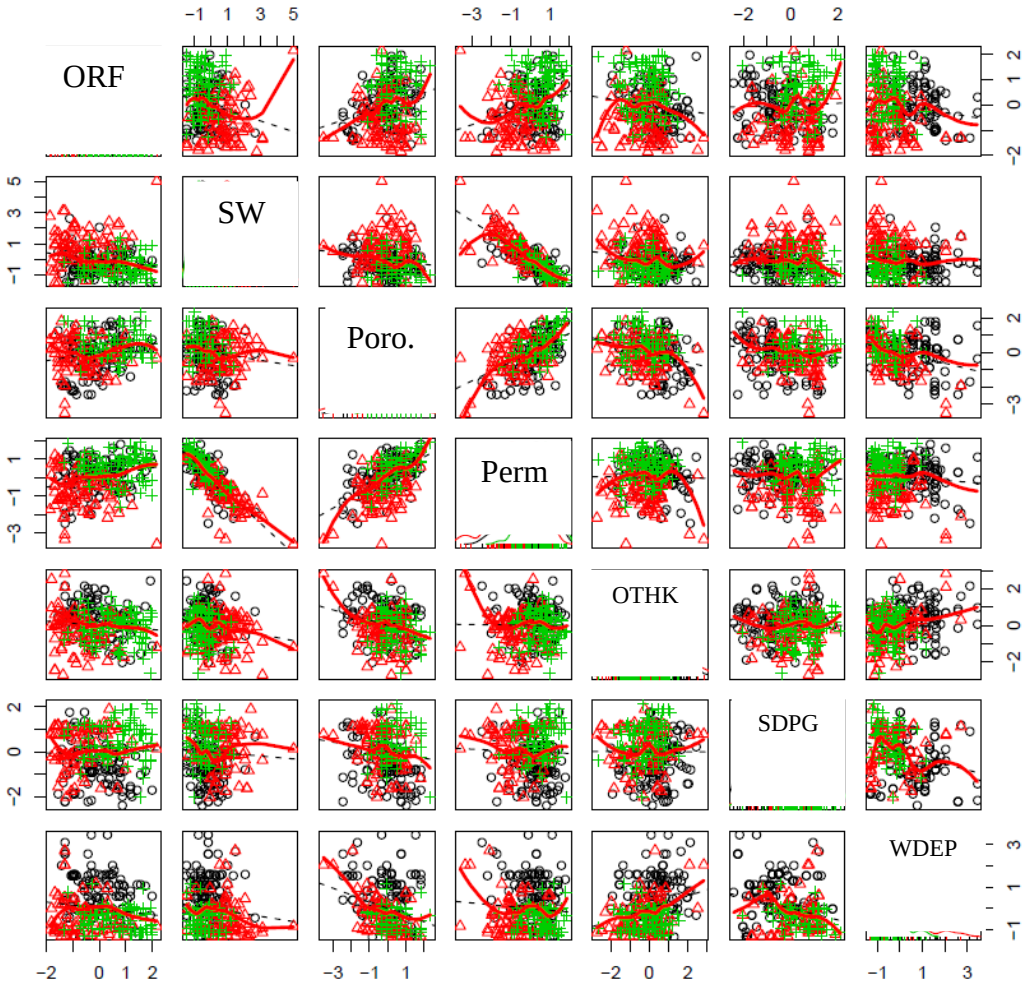

Histogram of doilfil$OAREA
Frequency
0
20
40
60
80
2
4
6
8
10
doilfil$OAREA


Histogram of doilfil$RSI
Frequency
0
20
40
60
80
100
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
doilfil$RSI


Histogram of doilfil$PERMEABILITY
Frequency
0
10
20
30
40
50
2
3
4
5
6
7
8
doilfil$PERMEABILITY


Histogram of doilfil$P_RECOIL
Frequency
0
10
20
30
40
50
60
8
10
12
14
16
18
20
doilfil$P_RECOIL

## Scatterplot after K-means Clustering



## Distribution of recovery factor for K-means clusters