NTNU
Norwegian University of
Science and Technology

Automatic Lithology Prediction from Well Logging Using Kernel Density Estimation

Anisa Noor Corina

Petroleum Engineering
Submission date: June 2016
Supervisor: Sigve Hovda, IPT

Norwegian University of Science and Technology
Department of Petroleum Engineering and Applied Geophysics

Preface

This thesis is a one of my fruitful result of a keen work for the past one semester. This thesis is also part of the requirement for a master degree in Petroleum Engineering, Department of Petroleum Technology and Applied Geophysics, Norwegian University of Science and Technology (NTNU). The study described herein began in spring 2016 to the extent of 30 educational points. Apart from the efforts of myself, the success of this study depends on the guideline of many others. I take this opportunity to thank these people.

First, I express my deep gratitude to my supervisor, Sigve Hovda, for the time and constructive response to the work. His enthusiasm working in this project has preserved the optimistic attitude and made this project possible. I would also like to thank Erik Skogen for the supportive discussion of well logging which I found very valuable for this work. I would also like to thank AGR company for providing access to iQx software, including the access to the data. And, I would also like to thank MATLAB © 2015 The MathWorks, Inc.,Natick, Massachusetts, United States, for supporting the computation within the study. Among all of the benefits provided by Institute Petroleum of Technology (IPT) NTNU, my curiosity of interdisciplinary working on this project together with these people have benefited myself more than most.

Despite the academic support, I would thank my all of my friends from Petroleum Engineering/Petroleum Geosciences, especially cohort 2014, who became my loyal partners during the long journey of 2 years of the master degree. I would also thank my friends from Indonesia who bring the joy during the rough hour working on my thesis.

Most importantly, none of this could have happened without my family. To my mom, the person that I adore the most, I sent my deepest love and thank for being there all the times for me. I would also to thank my dad, sister, and brother who never stop giving me comfort even though we are miles away and I am forever grateful.

Trondheim, 2016-06-09

Anisa Noor Corina

Abstract

This thesis presents an automatic real-time analysis of lithology interpretation through a method of statistical analysis: kernel probability density method. The goal of this thesis is to develop a method for interpreting and predicting lithology from the borehole geophysical data in real time. Prior to the development, the data is explored to check the data quality and the requirement of data correction. In addition, from exploratory data analysis, the data characteristics can be observed thus the best-fit classification method can be selected. The study focuses on the univariate analysis of gamma-ray data in classifying shale and non-shale lithology. In addition to univariate analysis, a preliminary study of bivariate analysis is also provided in this thesis. The bivariate analysis combines the gamma-ray and the neutron data.

Within the study, the models of probability density are constructed by using kernel estimator. The data source for the models are extracted from 3 wells in the North Sea, they are Well 15/5-7 A, Well 15/6-11 S, and Well 15/6-9 S. The application of kernel method on gamma-ray data returns a good estimation and appropriates the non-parametric distribution of the data. There are two different types of model constructed based on the type of classification rule. The first model is constructed solely using gamma-ray data. While, the second model is constructed by combining gamma-ray data and geological description which is represented as prior probability value in the classification rule. Once the models are ready, the models are validated and tested with a set of testing data in order to assess the misclassification rate.

There are three different experiments performed based on the source of the testing data set. These experiments are executed in order to assess how precise is the model classifying lithology by using testing data from different well locations. One of the experiment tests the models with a dataset taken from the same source of the model. Meanwhile, the other two experiments test the models with a dataset taken from the different source of the models. The validation shows promising result and proves that gamma-ray is a representative variable in classifying lithology.

The evaluation techniques within this study can be applied in the practice to interpret and predict the lithology. By applying the technique, it is expected that the reading from the logging tools can be processed and the result of lithology type with the prediction can be automatically returned in the surface. The application is also beneficial to reduce the time required for lithology interpretation during drilling operation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Subsurface lithology within drilling operation is interpreted with the usage of different techniques, involving seismic records, mud logging, and well logging. In the petroleum industry, well logs are the main source of information concerning the subsurface formations. And, the variation in borehole geophysical data usually is used to relate any change in lithology and geological properties. In addition to mud logging, real-time drilling measurement nowadays is improving through the development of measurement-while-drilling. Moreover, logging within drilling phase is even possible through logging-while-drilling (LWD). Continuous transmission of the information from MWD and LWD has shown the benefits in helping the drilling decisions and real-time formation evaluation (Bonner et al., 1992).

Despite of well logging, mud logging also provides information for lithology interpretation. Within mud logging, measurement of the progress of the drilling operation and the contents of the formation are recorded. The drilling parameters which are recorded include weight on bit (WOB), hook load, and mud properties. Meanwhile, the cuttings from the drilled formation inside borehole is circulated to the surface. By visualizing the cutting sample, the lithology can be approximated. Generally, cuttings samples require an amount of time to be circulated to the surface which lead to a delayed interpretation. Restoring the information from cuttings also can be difficult due to the requirement of correlating the cutting origin, fluid loss in the borehole, flushed rock fragment. These factors cause cutting visualization alone does not provide an accurate lithology interpretation and requires a combination with information from well logs and mud logs.

A wide span of methods in lithology interpretation has been proposed through combinations of various measurements both in the qualitative and quantitative evaluation. Some of the qualitative evaluations from mud logging include ROP interpretation (such as identification of drilling break and drill-off trend), drilling force, and bit evaluation, and specific energy deflection (Provost (1987), Ziaja and Roegiers (1998), Laosripaiboon et al. (2015)). Meanwhile, qualitative evaluations from logging measurement include visualizations of multiple logs , photoelectric (P_e) factor interpretation, and gamma-ray evaluation for shale identification (Gardner and Dumanoir (1980), Serra et al. (1985), Dewan (1986)). Qualitative methods itself are inadequate for lithology interpretation, especially in formation with complex lithologies which require a large set of logging information.

Within time, lithology interpretation has expanded and starts to consider the usage of quantitative methods, such as crossplot, statistical analysis, and neural network.

Crossplot is one of the basic method of quantitative evaluation which is performed by plotting data points of two or more than two different log data. The geophysical data which are widely used for crossplotting are density, neutron, sonic, and P_e. The types and the applications of crossplot method has been studied by Burke et al. (1969) and Clavier and Rust (1976). However, these methods still require manual analyses and can not be applied for automatic interpretation.

There are a plethora of statistical classification methods, such as discrimination analysis, linear regression, kernel estimation, etc. The selection between these methods is greatly dependent on the character of the data, so that the method can provide an interpretation that is important for model building. Probabilistic classification is a common classification method which can predict the belonging of a member by returning the probability. An early research by Delfiner et al. (1987) has shown how statistical analysis can be applied for lithology determination and prediction. This research proposed a procedure which combines modern wireline measurement in order to produce automatic lithologic description. Within the procedure, lithology classification by discriminant analysis and probability calculation by Bayesian rule was introduced. The application of this research was shown in a case study by Busch et al. (1987). This research showed that the statistical discriminant analysis is possible to predict lithology formation. However, the proposed method in these researches is limited for geophysical data with normal (Gaussian) distribution thus, this method is not flexible to be applied in non-parametric distribution.

A statistical method called kernel density estimator provides an estimation of the probability density function for a non-parametric distribution and examines the multimodality of data. A research by Silverman (1986) showed that kernel density estimation is the excellent tool for estimating the univariate, bivariate, or trivariate data. The application of kernel density estimation in borehole geophysical data was performed by Mwenifumbo (1993). Within this research, the kernel density method was applied for analysis of univariate and bivariate data to identify lithology and sulfide mineralization. However, the assessment of the statistical significance was not performed.

Until now, there has been no study of lithology prediction based on borehole geophysical data which applied statistical analysis. Within this study, a method of lithology classification and prediction from borehole geophysical data is developed by applying kernel density method as the statistical analysis. The objective of this study is that the developed method can be applied to give interpretation and prediction in a real-time operation. Specifically, the main focus of this study is the application of univariate kernel density which is not extensively used. The models from gamma-ray data constructed by kernel estimator are assessed by the used of the confusion matrix to understand the model accuracy in classifying shale and non-shale lithology. The results are presented in term of misclassification rate. In addition to that, a brief insight of bivariate analysis to improve lithology classification is also provided.

The models are tested with two different classification rules to assess the effect from adding prior probability value which value is taken from the geological description. The models are also tested in different experiments with testing datasets which are taken from different well locations. This study shows that gamma-ray is a good variable for lithology classification. But, the accuracy is dependent on the source of data which is tested towards the models. Moreover, the method proposed in this study can be applicable for lithology prediction, even though the application is still limited because it processes the data from the current depth of logging tool, not beyond the logging tool. However, this application can be beneficial to improve lithology interpretation during the drilling operation and provides an automatic lithology prediction.

The rest of report is structured as follows. Chapter 2 introduces the theory of the well logging and several methods of statistical analysis, they are exploratory data analysis, kernel density estimation, and classification. Chapter 3 introduces the source of the data which is used within this study and the data management. Chapter 4 introduces the data exploration of gamma-ray data and hypothesis testing on the data. Chapter 5 introduces the application of kernel density method into gamma-ray data and the validation of models for classification purpose.

Chapter 2

Background Theory

This purpose of this chapter is to discuss the theories which relevant to this study. The first theory discusses the well logging in petroleum industry, including the description of gamma ray and neutron logging which are relevant to the geophysical data which is used in this study. The next theories discussed are relevant to the methodology adapted in this study, they are exploratory data analysis, hypothesis testing, and kernel density estimation.

2.1 Well Logging

In the petroleum industry, well logging plays a crucial role as a tool to interpret downhole conditions. Well logging is divided into two types, surface and downhole logging. Surface logging records all information during drilling operation through sensors located at the surface. Meanwhile, downhole logging records information from sensors located at the downhole tools. The information recorded in surface logging are (i) drilling parameters, such as hookload, torque, and Rate of Penetration (ROP), (ii) mud returns, and (iii) cuttings from downhole. By closely monitoring the surface measurements, events occurring in the borehole can be identified by looking at values outside of the normal ranges. In addition, cuttings from downhole can be used to describe the geological properties and detect hydrocarbon traces (Wilson, 1955).

Downhole logging, also called as wireline logging, measures the properties of rocks surrounds the borehole, such as rock radioactivity, resistivity, etc. The tool is suspended on a cable or wire and can be run between drilling operations and at the end of drilling. The recent development allows wireline logging to be run during a drilling operation, which is called as Logging while Drilling (LWD). By combining LWD with the Measurement while Drilling (MWD) system, information can be transmitted from downhole to surface almost continuously during the drilling operation.

Some of the wireline tools measure properties that give a direct result and do not require to be interpreted, while some of the others require interpretation. Most of the times, the interpretation requires a collaboration of results from different wireline tools because each tool has a limited measurement and the results can be masked by the rock or fluid properties. As an example, resistivity measurement is affected by formation temperature because the resistivity tools are not sensitive to temperature. This event can lead to misinterpretation of reservoir fluids in the formation. Therefore, a combination of resistivity and temperature data will give a better interpretation.

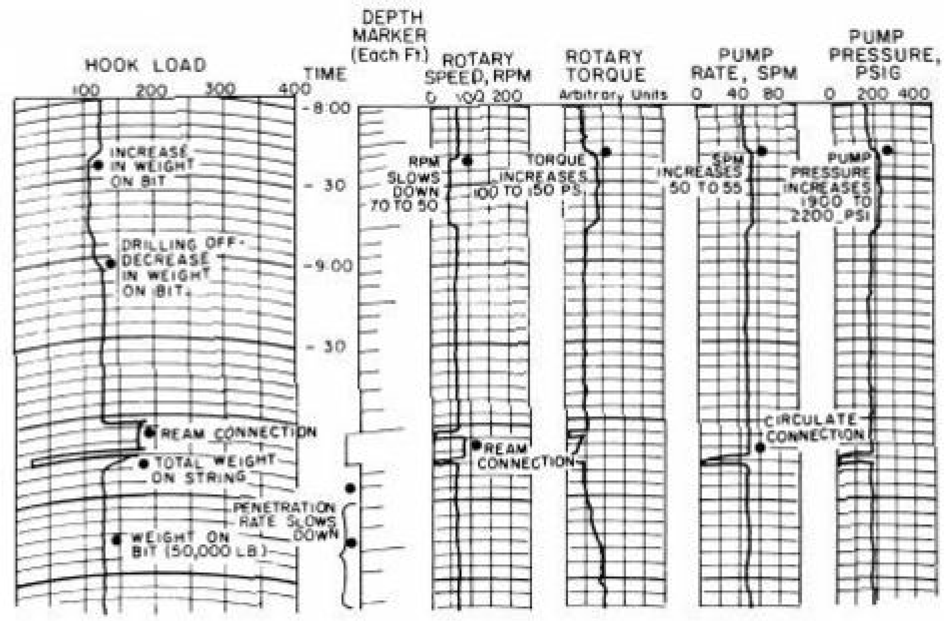


Figure 2.1: Example of time-based surface logging (Bourgoyne et al., 1985)

2.1.1 Gamma Ray Logging

Gamma Ray (GR) tool measures the natural radioactivity of minerals contained in the rocks. Most of the rocks contain natural occurring radioactive elements, such as potassium, uranium, and thorium in different amounts, and all of these emit gamma rays (Schlumberger Educational Services, 1989).

GR log is useful for correlating zones from one well to others and indicating shale in the formation, due to high content of radioactive minerals in shale. A rough estimation of clay volumes (V_{cl}) can be calculated using GR reading. By setting *sand point*, minimum GR reading (γ_{min}) which indicates 100 % sand content, and *shale point*, maximum GR reading (γ_{max}) which indicates 100 % shale content, GR index (I_{GR}) can be calculated by linear scaling (Ellis and Singer, 2010) :

$$I_{GR} = \frac{\gamma_{log} - \gamma_{min}}{\gamma_{max} - \gamma_{min}} \tag{2.1}$$

Poupon and Gaymard (1970) proposed that shale volume is equal with I_{GR}. Beside linear scaling, there are several different approaches that consider the effect of clay distribution in the reservoir rock, clay mineral, and clay bound. These methods are summarized in Table 2.1 and visualized in Figure 2.3.

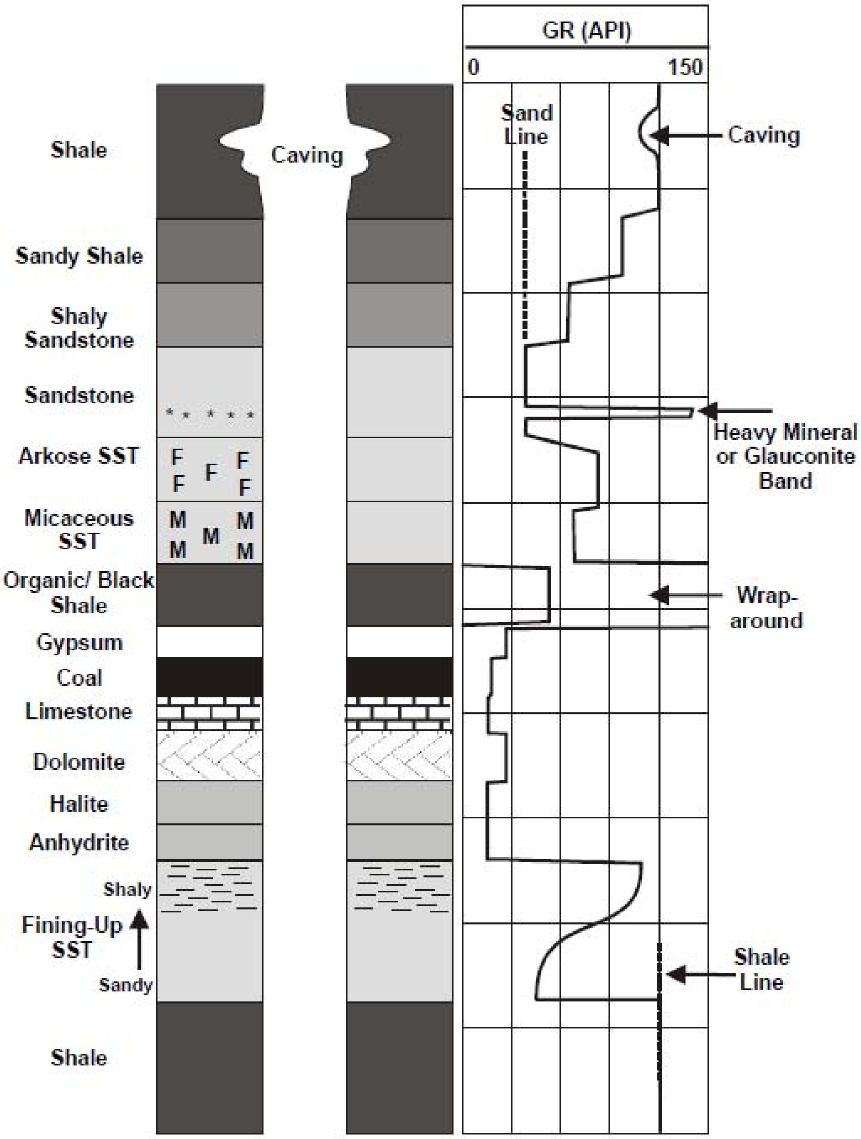


Figure 2.2: Gamma ray reading for various lithology (Glover, 2001)

Method	Equation
Clavier et al. (1971)	$V_{sh} = 1.7 - [3.38 - (I_{GR} + 0.7)^2]^{0.5}$
Larionov (1969), for tertiary rock	$V_{sh} = 0.083 \times [2^{(3.7058 \times I_{GR})} - 1]$
Larionov (1969), for older rock	$V_{sh} = 0.33 \times (2^{I_{GR}} - 1)$
Stieber (1970)	$V_{sh} = \dfrac{3I_{GR}}{(1+2I_{GR})}$

Table 2.1: Methods of calculating shale volume

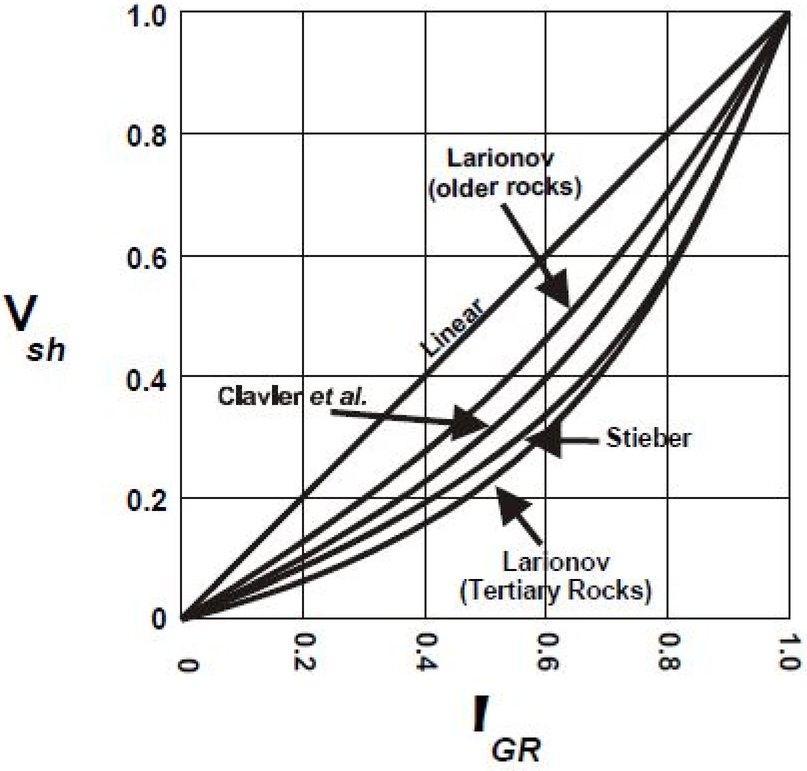


Figure 2.3: Comparison of shale volume by using different methods (Glover, 2001)

However, determining lithology shaliness only by using GR index can cause misinterpretations, such as in cases of uranium-rich formations, sandstone containing mica, and nonradioactive clays. These misinterpretations can be prevented by the use of spectral gamma ray which measures not only the total radioactivity, but also the concentration of potassium (K), thorium (Th), and uranium (U).

GR tools are sensitive to a number of factors (Bateman, 2012):

1. Eccentricity of gamma ray tool
2. Hole size
3. Mud weight
4. Casing weight and size
5. Cement thickness

Modern logs usually have automatic corrections applied to GR readings. However, a set of correction charts is available to correct GR manually to environmental conditions such as hole size and mud weight (Schlumberger Wireline & Testing, 1998). The corrections are considered to be crucial so that the logging data can be representative.

2.1.2 Neutron Log

In general, neutron log measures the amount of hydrogen in formations. The neutron tool releases high energy neutron into the formation which will scatter elastically with nuclei. The energy will be reduced to the thermal energy level (≈ 0.025eV) and then the neutron will be absorbed by the nucleus while emitting γ-rays.

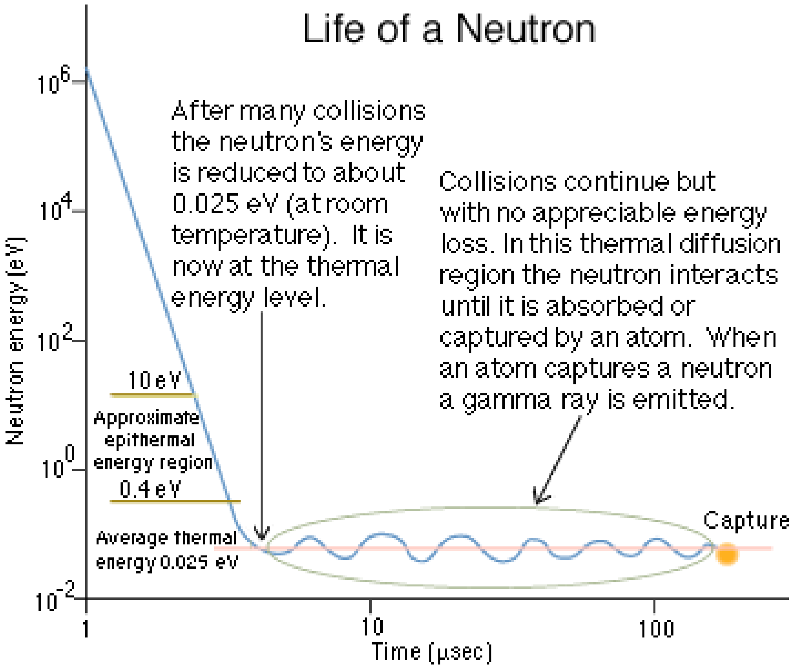


Figure 2.4: The graph of neutron life after neutron is emitted by the neutron tool

The energy loss due to elastic scattering is maximum when a neutron collides nucleus with the same mass (i.e. hydrogen). Therefore, the count rate to slow down the neutron and the distance traveled by neutron depend on the amount of hydrogen. Because hydrogen is mostly found in pores (composed in water or hydrocarbon), the neutron log is related to porosity function. The count rate in high porosity rocks is slower than in low porosity rocks.

Mainly, there are three different types of neutron tools available, they are:

1. The gamma ray/neutron tool (GNT)
2. The sidewall neutron porosity tool (SNP)
3. The compensated neutron log (CNL)

The detailed explanation of each tool can be found in a textbook written by Bateman (2012). The data measured by neutron tools are shown in the unit of porosity or hydrogen index. The neutron tools are calibrated in limestone filled with fresh water. Therefore, the results often presented in equivalent limestone porosity units.

The neutron logs are used for porosity calculation which assumes that the contribution of elements other than hydrogen is negligible. The second use of neutron logs is lithology determination. The source of a hydrogen atom is not only from fluids occupying the pore space but it can be originated from bound water molecules in shales, crystallized water in evaporites, or hydrated minerals in igneous and metamorphic rocks (Glover, 2001).

The apparent porosity of shale is varying, but it is usually higher than the apparent porosity identified in carbonate and sandstone rocks. This high porosity reading by neutron tool is caused by the effect of hydrogen contained in the bound water in shale. However, shale identification by using neutron log requires extra concern due to the effect of hydrocarbon gas which may present and disturbs the log.

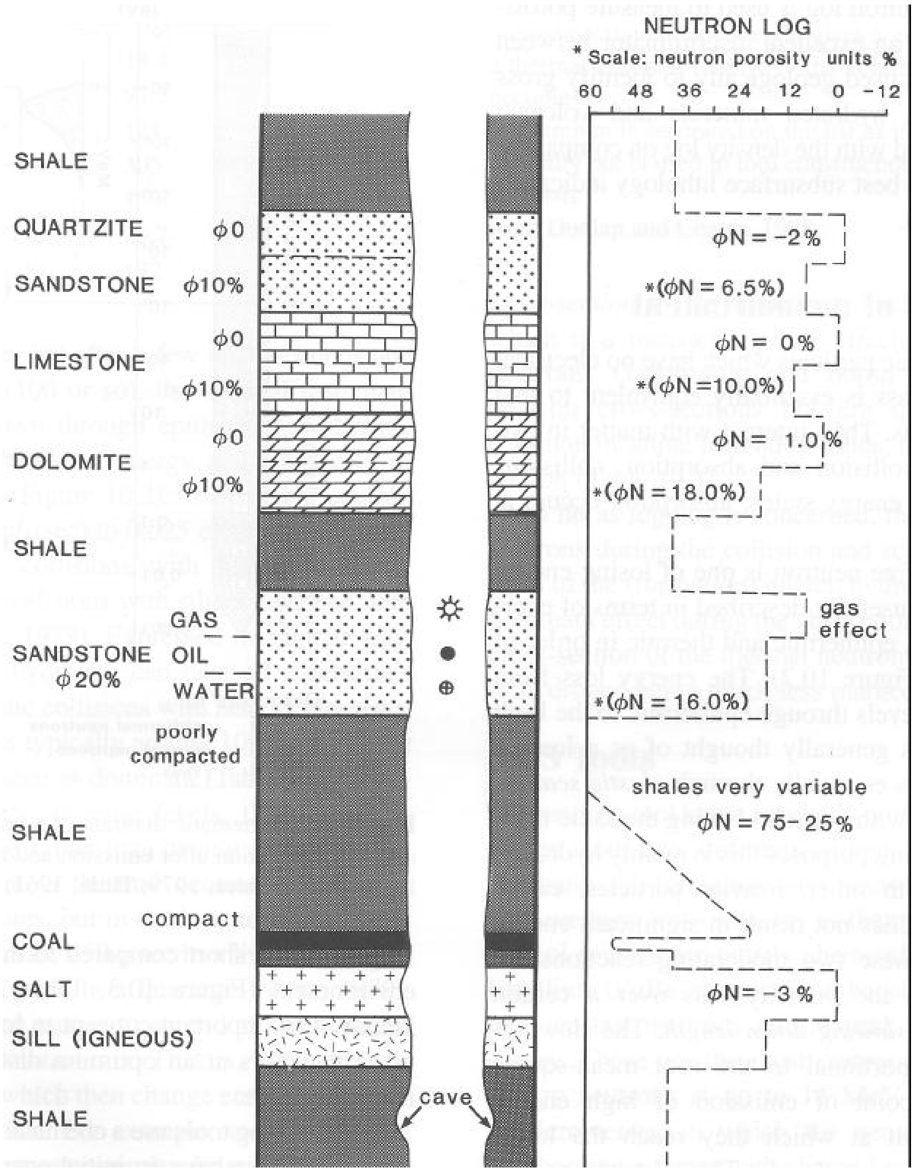


Figure 2.5: Typical apparent porosity from neutron log for varies lithologies

2.2 Exploratory data analysis

The motivation of data exploration is to extract any important information of the data and understand the data behavior. This step is found to be crucial because we would set assumptions and test hypotheses within this process (Tukey, 1977). The method adapted in data exploration is called as Exploratory Data Analysis (EDA). This method provides the most appropriate way to explore, summarize data, and also create a visualization of the data. By implementing EDA, it is expected to gain some confidences of the data. Most of the EDA use graphical techniques to get data visualization, such as histograms, box plots, and steam-leaf plots.

2.2.1 Histogram

Histogram is one of graphical techniques which serves the data by using bars to show data distribution. During the construction, the data values are divided into series of intervals which illustrated in bars. The number of intervals are often referred as *bin*, which each interval contains a certain range of values. Each bar usually has a consistent and equal ranges with others, and its interval is not overlapping with others and adjacent.

Histograms are often constructed as frequency or density histogram. The description of each type of histogram is explained below:

Class frequency histogram measures the number of occurrences which values are falling within a given class interval.

Relative frequency histogram calculates the frequency of each interval divided by the total number of measurements. This histogram has a total height of bars equal to 1 or 100%.

Density histogram calculates the relative frequency of each interval divided by bin width. In other words, density histogram shows relative frequency as area of each bar. In density histogram, the total area of bars is normalized to 1.

Histogram is very well suited to illustrate a large set of data and continuous data. With the advantage of its simplicity of construction, histogram is a graphical technique which is commonly used. However, describing data using histogram may encounter some difficulties and requires some understandings. A histogram is very sensitive to bin width because of its effect to the graph smoothness. It is evidenced that larger bin width (fewer bins) reduces noises and makes the graph oversmoothed, while smaller bin width (more bins) makes the graph undersmoothed.

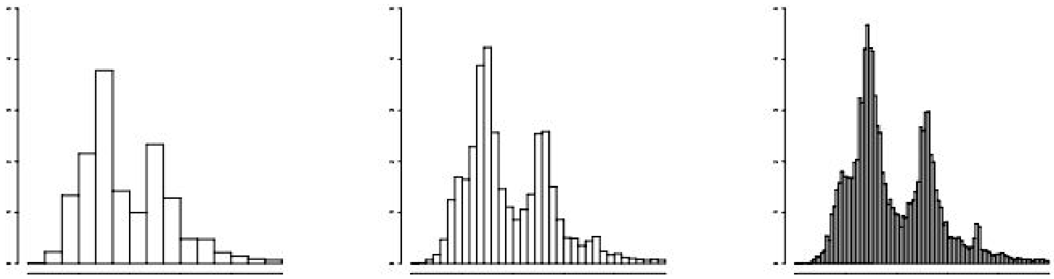


Figure 2.6: Histograms of a data set with 15, 35, and 100 bins (Scott, 2004)

An example is taken from a paper written by Scott (2004) to give an illustration of the effect of the bin width. A set of data with 21,640 data points are constructed using histograms with 3 different bins Figure 2.6). Undersmoothed histogram contains high variability in value even though it has a smaller bias, while oversmoothed histogram has the opposite effect. Thus, choosing the most optimal bin width is crucial to avoid misinterpretation (Simonoff, 1996).

There are many theories have been developed in determining the optimum bin width. However, the application of these methods depends on the data distribution and the goal of analysis. Some of the methods proposed calculation of bin numbers, k, and some proposed calculation of bin width, h. The relationship of bin numbers and bin width is:

$$k = \frac{x_{max} - x_{min}}{h} \tag{2.2}$$

where x is random variable. The most common methods on calculating optimum bin width are explained below.

Sturges method. This method suggests to calculate bin width with formula (Scott, 1992),

$$k = 1 + log_2 n \tag{2.3}$$

where n is the number of data points. Sturges derived the formula above based on binomial distribution with normal distribution data. This method is popular due to its simplicity in the calculation. However, this method may give poor result if the data is not normal.

Scott method. This method is derived by minimizing the integrated means squared error of the density estimate for normally distributed data (Scott, 1992). The calculation requires standard deviation [1] , σ.

$$h = \frac{3.491\sigma}{n^{1/3}} \tag{2.4}$$

Freedman-Diaconis. This method is suitable for data containing a large number of outliers or heavy-tailed distribution. The method is a modification of Scott method, replacing σ with parameter interquartile range (IQR), the distance between the lower and upper quartiles (Scott, 1992). Description of quartile can be found in Chapter 2.2.2.

$$h = \frac{2(IQR)}{n^{1/3}} \tag{2.5}$$

Histogram is also referred as simple univariate density estimator to approach probability density function. To improve the smoothness in estimate density function, the application of histogram often combined with kernel density estimator. The detailed explanation of kernel estimator will be covered under 2.4.

2.2.2 Boxplot

Boxplot is a graphical technique to examine the shape of data distribution by using parameters called quartiles. In addition, boxplot is also used to study the variability of values in a set of data (Ott and Longnecker, 2010).

Quartiles are three points dividing a dataset which is arranged from the lowest to the highest value equally into 4 groups. The first quartile (Q_1), which is called lower quartile, has a value between the smallest value and median of a dataset. The second quartile (Q_2) is called as the median of a dataset. The median value is taken from a

[1] Standard deviation measures the variation of data set ,$\sigma = \sqrt{\frac{\sum_i (x-\bar{x})^2}{n-1}}$ (Ott and Longnecker, 2010)

data point located in the middle of a dataset which is arranged from the lowest to the highest value. In other words, a median is the center of data distribution. Last, third quartile (Q_3), which is called upper quartile, has a value between the median and the highest value of a dataset. The difference between upper and lower quartiles value is defined as an interquartile range.

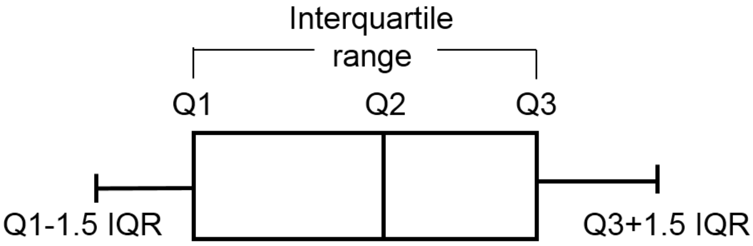


Figure 2.7: An example of boxplot with whiskers of 1.5 IQR of upper and lower quartiles

The boxplot is constructed by creating a box with two sides which values are equal to the upper and lower quartiles. Within the box, a line is drawn to indicate median (see figure 2.7). The boxplot is often constructed with whiskers to indicate data variability. There are several ways to plot the whiskers depend on the information to be represented. In most cases, the whiskers represent the minimum and maximum values of the data. The whisker can also set at a value equal to 1.5 IQR of the upper and lower quartiles, and this boxplot is often referred as Tukey boxplot (Frigge et al., 1989).

2.3 Hypothesis testing

2.3.1 Motivation of hypothesis testing

Hypothesis testing is a method for testing hypothesis about a group within a population (Privitera, 2015). The hypothesis testing is started by defining the null hypothesis (H_0), a statement of a population parameter that is assumed to be true. Hypothesis testing tests the null hypothesis in order to check whether the statement is likely to be true or not. The statement which opposes the null hypothesis is called alternative hypothesis (H_1).

The methods to compute the test statistic are varied depends on the data characteristic. The hypothesis test applied in this study was Kruskal-Wallis test. Kruskal-Wallis test is a nonparametric test (distribution free) for assessing differences in a continuous dependent variable which is presumed containing independent variables (3 or more groups) (Kruskal and Wallis, 1952). Kruskal-Wallis test is a rank-based test which is an extension of Mann-Whitney test. However, the test does not reveal which group of independent variables that is significantly different from each other. In other words, the test is only limited to inform that at least two groups are different. The assumptions required for Kruskal-Wallis test, are:

1. The dependent variable is a continuous variable.
2. The independent variables should consist three or more independent groups. Mann-Whitney U Test is commonly used to test two groups within the population.

3. The observations in each group are independent and there is no relationship between the groups.

The stated H_0 of the test is that the data set comes from same distribution. Before getting into Kruskal-Wallis test, firstly we would explain the underlying theory of rank-sum test for two independent variables from Mann-Whitney U test (Mann and Whitney, 1947).

Mann-Whitney U test: rank sum test

Rank-sum test is a method ranking the raw data from the lowest (rank #1) to the highest value (rank #N), with tied ranks included. Tied ranks are assigned if there are two or more than two tied values in the raw data, thus the ranks are adjusted and equalized.

Define a population with 2 group of samples, group 1 and group 2, where n_1 is the size of observations in group 1 and n_2 is the size of observations in group 2. The sum rank of each group is defined with R_1 and R_2 respectively for group 1 and group 2. For any combination of n_1 and n_2, the maximum possible value of sum rank, R_{max}, in each group can be calculated as follow

$$R_{max_1} = n_1 n_2 + \frac{n_1(n_1+1)}{2} \tag{2.6a}$$

$$R_{max_2} = n_1 n_2 + \frac{n_2(n_2+1)}{2} \tag{2.6b}$$

Mann-Whitney proposed a parameter U which is equal to the difference between maximum possible value of rank sum, R_{max}, and the actual rank sum observed, R. The equation U for both groups follows

$$U_1 = R_{max_1} - R_1 \tag{2.7a}$$

$$U_2 = R_{max_2} - R_2 \tag{2.7b}$$

For any samples sizes, n_a and n_b, U parameter has identities:

$$U_1 + U_2 = n_1 n_2 \tag{2.8a}$$

$$U_1 = n_1 n_2 - U_2 \tag{2.8b}$$

$$U_2 = n_1 n_2 - U_1 \tag{2.8c}$$

$$U = \min(U_1, U_2) \tag{2.8d}$$

Depending on n_1, n_2, and level of significance, critical value of U, U_{crit}, can be calculated. If $U < U_{crit}$, then the test is significant and the null hypothesis is rejected.

In case where the number of samples, n_1 and n_2, are large (both equal to or greater than 5), U is calculated by using different approach. In this case, U approximates the normal distribution $N(\mu, \sigma)$, where

$$\mu = \frac{n_1 n_2}{s} \tag{2.9a}$$

$$\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{2.9b}$$

From these identities, standardized variable, z-values, can be calculated following

$$z = \frac{|U - \mu| - 0.5}{\sigma} \tag{2.10}$$

The value of - 0.5 is a correction for continuity to accommodate the sampling distributions of U which are discrete. Then, the probability value (p-value) of z is generated from normal distribution. If p-value < level of significance, then the null hypothesis is rejected.

Kruskal-Wallis test

The concept of Kruskal-Wallis test is quite similar with Mann-Whitney which also adapts the rank-sum test. Defined a population with k independent group of samples, where $n = (n_1, n_2, \ldots, n_k)$ represents the number of observation in each group, $R = (R_1, R_2, \ldots, R_k)$ represents the sum rank of the kth group, and $M = (M_1, M_2, \ldots, M_k)$ represents the mean of rank of the kth group. In addition, R_T is the sum of R of all k group and M_T is the sum of M of all k group, following:

$$R_T = \sum_{j=1}^{k} R_j \tag{2.11a}$$

$$M_T = \sum_{j=1}^{k} M_j \tag{2.11b}$$

Kruskal-Wallis measures a parameter $SS_{bg(R)}$ which is defined as the between-groups sum of squared deviates based on the rank value. The conceptual formula of $SS_{bg(RR)}$ is shown as

$$SS_{bg(R)} = \sum_{j=1}^{k} \left[n_j (M_j - M_T)^2 \right] \tag{2.12}$$

and the computational formula is shown as

$$SS_{bg(R)} = \sum_{j=1}^{k} \left[\frac{(R_j)^2}{n_j} - \frac{(R_T^2)}{N} \right] \tag{2.13}$$

where N is the size of data population.

Kruskal-Wallis hypothesis test is concluded by defining a test statistic H, which value is equal to the ratio between $SS_{bg(R)}$ and the mean of sampling distribution of $SS_{bg(R)}$,

$$H = \frac{SS_{bg(R)}}{N(N+1)/12} \tag{2.14}$$

An alternative way to write the formula H above is

$$H = \frac{12}{N(N+1)} \left(\sum_{j=1}^{k} \frac{(R_j)^2}{n_j} \right) - 3(N+1) \tag{2.15}$$

The p-value can be approximated by inputting the calculated H into chi-square distribution because H value has a close approximation to the chi-square distribution for $df = k - 1$, or following

$$H \sim \chi^2_{(k-1)} \tag{2.16}$$

where χ^2 is chi-squared distribution. If the returned p-value is less than level of significance (typically set at 5%), then we rejected the statement of the null hypothesis. Small values of p-value remove the doubt of the validity of H_0.

2.4 Kernel density estimation (KDE)

2.4.1 Motivation of univariate density estimation

The fundamental concept underlining the analysis of univariate data is the probability density function for non-parametric distribution (Simonoff, 1996). The density function of a random variable X which has probability density function $f(x)$ is shown as

$$P(a < X < b) = \int_a^b f(u)du \tag{2.17}$$

By using the definition of density function, an estimation of density function can be constructed. There are two types of probability density estimator: simple density estimator, which is often referred as histogram, and smooth density estimator.

Simple density estimator

Recall forward approximation of density function,

$$f(x) \equiv \frac{d}{dx}F(x) \equiv \lim_{h\to 0} \frac{F(x+h)-F(x)}{h}, \tag{2.18}$$

where $F(x)$ represents the cumulative distribution function of X.

Assume that density f consists random samples with size n which samples are independent and identically distributed, represented as $\{x_1,\ldots,x_n\}$. By dividing Equation 2.18 into a set of K bin numbers with width h and replace $F(x)$ with the empirical cumulative distribution function,

$$\hat{F}(x) = \frac{\#\{x_i \le x\}}{n}, \tag{2.19}$$

This equation leads histogram to be a density function estimator with each bin value equal to

$$\hat{f}(x) = \frac{(\#\{x_i \le b_{j+1}\} - \#\{x_i \le b_j\})/n}{h}, \qquad x \in (b_j, b_{j+1}], \tag{2.20}$$

where $x \in (b_j, b_{j+1}]$ is the boundaries of jth bin. In simpler way, density estimator of histogram can also be defined as

$$\hat{f}(x) = \frac{n_j}{nh}, \qquad x \in (b_j, b_{j+1}], \tag{2.21}$$

where n_j represents the number of observations in jth bin and bin width $h = b_{j+1} - b_j$.

Histogram is considered as the simplest method to estimate the distribution of univariate data. However, the shortcomings of using histogram are that the histogram is not giving a smooth estimation and not sensitive to f. Histogram may also distort depending on the bar width.

Smooth univariate estimator

Recall central approximation of density function,

$$f(x) \equiv \frac{d}{dx}F(x) \equiv \lim_{h\to 0} \frac{F(x+h)-F(x-h)}{2h}, \tag{2.22}$$

Different from histogram, the smooth estimator approaches the density function by estimating the derivative at each point x separately. By replacing $F(x)$ with empirical cumulative distribution,

$$\hat{f}(x) = \frac{\{\#x_i \in (x-h, x+h)\}}{2nh} \tag{2.23}$$

The equation above also can be written as

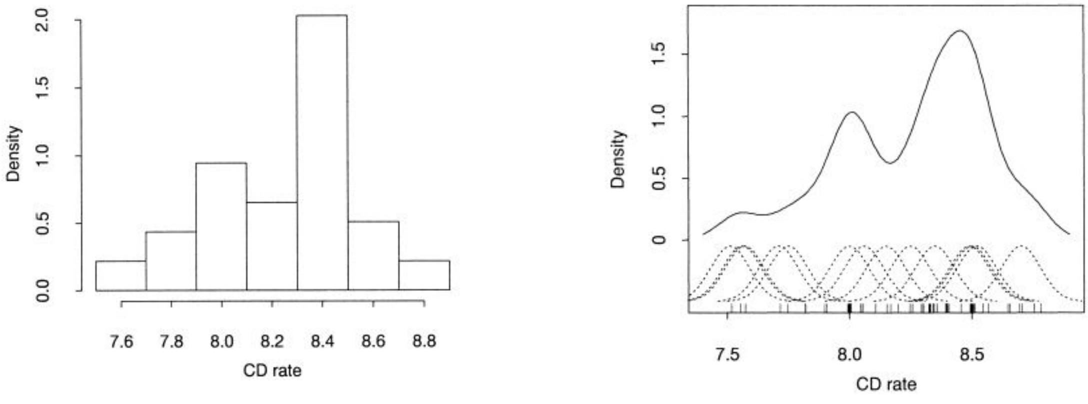
$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right), \tag{2.24}$$

where

$$K(u) = \begin{cases} \frac{1}{2}, & \text{if} -1 < u \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

The Equation 2.24 is the form of kernel density estimator, with uniform kernel function K. This estimator $\hat{f}(x)$ counts the percentage of the observations in each data point over the local neighbourhood which is close to the examined data point x. By merging all of the smooth kernel function at each data point, we will have a smooth density estimation for one population.

The comparison of simple and smooth density estimator in estimating density function can be seen in Fig. 2.8. The example given by Simonoff (1996) shows the comparison of probability density function from kernel estimator and histogram. Kernel estimator gives a estimation which is smoother compared to discreteness of the histogram.



(a)

(b)

Figure 2.8: (a) Simple density estimation, and (b) smooth density estimation (with underlying Gaussian kernel function) for certificated of deposit (CR) rates

2.4.2 Properties of kernel density estimator

Kernel density estimator is dependent on the kernel function, K, and the bandwidth, h. The bandwidth h is often referred as smoothing parameter which control the smoothness of the data function. A very small bandwidth will give undersmoothed estimation with more peaks and bumps, meanwhile a very large bandwidth will give over-smoothed graph. Bandwidth also have strong relation to bias and variance which creates a dilemma in selecting optimal bandwidth. A small bandwidth will reduce the bias of $\hat{f}(x)$, but it will trigger larger variance of $\hat{f}(x)$, and vice versa. The criterion on choosing optimal bandwidth mostly quantified through the measurement of mean squared error (MSE).

$$\begin{aligned} \text{MSE}\left[\hat{f}(x)\right] &= E_f\left[\hat{f}(x) - f(x)\right]^2 \\ &= \text{Bias}^2\left[\hat{f}(x)\right] + \text{Var}\left[\hat{f}(x)\right] \\ &- \frac{f(x)R(K)}{nh} + \frac{h^4\sigma_K^4\left[f''(x)\right]^2}{4} + O(n^{-1}) + O(h^6) \end{aligned} \tag{2.25}$$

By integrating MSE over the entire line, we will get MISE (integrated mean squared error). And the asymptotic MISE (AMISE) will follow

$$\text{AMISE}(h) = \frac{R(k)}{nh} + \frac{h^4\sigma_K^4 R(f'')}{4} \tag{2.26}$$

where $R(K) = \int K(u)^2\,\mathrm{d}u$, K satisfies condition $\int u^2 K(u)\mathrm{d}u = \sigma_K^2 > 0$, and f'' is the second derivative of density function f. The optimal bandwidth, h_0, is selected by minimizing AMISE through differential equation and resulting

$$h_0 = \left[\frac{R(K)}{\sigma_K^4 R(f'')}\right]^{1/5} n^{-1/5} \tag{2.27}$$

and the minimum AMISE follow

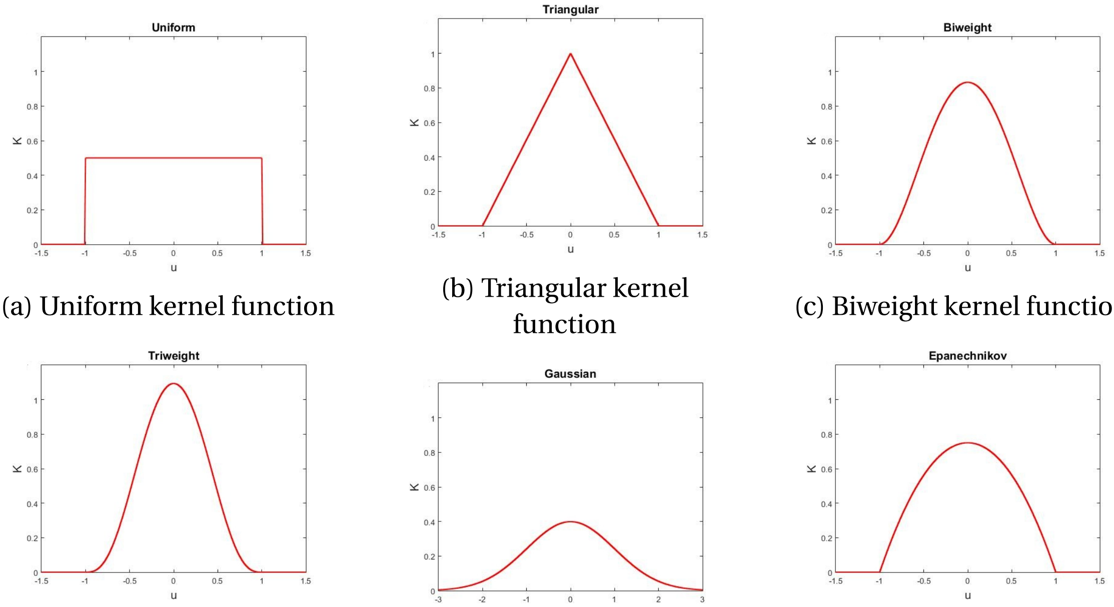
$$\text{AMISE}_0 = \frac{5}{4}\left[\sigma_K R(K)\right]^{4/5} R(f'')^{1/5} n^{-4/3} \tag{2.28}$$

The formula of h_0 consists an unknown density function f which value is out of the control of data analyst, thus this formula can not be applied directly. However, the term $[\sigma_K R(K)[^{4/5}$ can be minimized by choosing the optimal kernel function K, and this kernel function is often called as Epanechnikov kernel. Several other kernel functions which commonly user are shown in Table 2.2 and Fig. 2.9.

By comparing the inefficiency of each kernel function to Epanechnikov kernel, it is obvious that the selection of kernel functions is insensitive with MSE (Simonoff, 1996). Therefore, kernel function should be selected based on other consideration, such as properties of $\hat{f}$.

Table 2.2: Various kernel functions and forms

Kernel	**Form**
Uniform	$\frac{1}{2}$
Triangular	$k(u) = (1 - \lvert u \rvert)$
Biweight	$k(u) = \frac{15}{16}\left(1 - u^2\right)^2$
Triweight	$k(u) = \frac{35}{32}\left(1 - u^2\right)^3$
Gaussian	$k(u) = (2\pi)^{-1/2} e^{-u^2/2}$
Epanechnikov	$k(u) = \frac{3}{4}\left(1 - u^2\right)$



(a) Uniform kernel function

(b) Triangular kernel function

(c) Biweight kernel function

(d) Triweight kernel function

(e) Gaussian kernel function

(f) Epanechnikov kernel function

Figure 2.9: Distribution of $K(u)$ in various kernel functions

Choosing the optimal bandwidth

If the reference of density function f is based on the Gaussian function, then the Gaussian density can be substituted into equation 2.27, and resulting

$$h_0 = 1.059\sigma n^{-1/5} \tag{2.29}$$

This Gaussian reference density can also be used and converted into other types of kernel function. The optimal bandwidth of other kernel function, $h_{0,K*}$ satisfies the condition,

$$h_{0,K*} = c_{K*} h_{0,G}, \tag{2.30}$$

where

$$c_{K*} = \left[\frac{2\sqrt{\pi}R(K*)}{\sigma^4_{K*}}\right]^{1/5} \tag{2.31}$$

and $h_{0,G}$ is the optimal bandwidth of Gaussian kernel. This method is often referred as the Silverman Rule-of-Thumb of selecting optimal bandwidth. Depending on the true density, this method can give the optimal bandwidth if the true density is normal. However, if the true density is close to normal, the bandwidth will be close to optimal (Hansen, 2009).

2.5 Classification

In a population which consists several independent groups, very often we wish to look for the characteristics or features to separate the multivariate samples into the known groups. Based on the features, classification rule could be developed in order to identify and allocate an object from new observations into one of the groups.

Consider a population consists two sub-populations, denoted as π_1 and π_2. The probability density of each population is denoted as $f_1(x)$ and $f_2(x)$, with random variable of $X = \left(X_1, \ldots, X_p\right)$. Denote that Ω is the collection of all possible outcomes x, R_1 is the possible outcomes x which are classified as population π_1, and $R_2 = \Omega - R_1$ is the possible outcomes of x which are classified as population π_2.

The classification probabilities can be presented in the following table:

Table 2.3: Classification probabilities table

		Classified as:	
		π_1	π_2
True population:	π_1	$P(1\mid 1)$	$P(2\mid 1)$
	π_2	$P(1\mid 2)$	$P(2\mid 2)$

The probability misclassifying an object as a belonging to population π_2 when the object actually belong to population π_1 calculated as

$$P(2|1) = P(X \in R_2 | X \in \pi_1) = \int_{R_2} f_1(x)dx \tag{2.32}$$

and the probability misclassifying an object as a belonging to population π_1 when the actual belonging is population π_2 equals

$$P(1|2) = P(X \in R_1 | X \in \pi_2) = \int_{R_1} f_2(x)dx \tag{2.33}$$

In some cases, prior probability and costs of misclassification are taken account into classification rules. Prior probability is the probability of one population from prior observation and denoted as p_1 for prior probability of population π_1 and p_2 for prior probability of population π_2. The total of prior probability is equal to 1, $p_1 + p_2 = 1$. Costs of misclassification are defined as the prices to pay if an object is misclassified

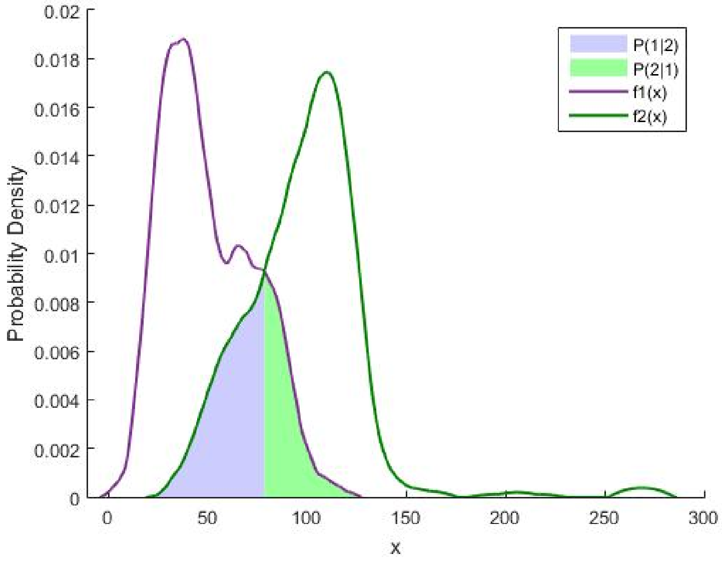


Figure 2.10: Description of misclassification regions $P(1|2)$ and $P(2|1)$. The purple shaded area indicates region of $P(1|2)$ and the light green shaded area indicates region of $P(2|1)$

and denoted as c_1 for the cost of classifying an object of π_1 as π_2 and c_2 for the cost of classifying an object of π_2 as π_1.

The classification rules are evaluated in terms of the *expected cost of misclassification* (ECM)

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \tag{2.34}$$

The optimal classification rule is calculated by minimizing the ECM, resulting

$$\begin{aligned} R_1 &= \left\{x \in \Omega; \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right\} \\ R_2 &= \left\{x \in \Omega; \frac{f_1(x)}{f_2(x)} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right\} \end{aligned} \tag{2.35}$$

Special classification rules prevail for conditions such as:

1. Equal (or unknown) prior probabilities: $p_1 = p_2$. The classification rule now depends on probability density ratio and cost ratio.

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}, \quad R_2 : \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)} \tag{2.36}$$

2. Equal (or undefined) misclassification cost: $c(1|2) = c(2|1)$. The classification rule now depends on prior probability and density ratio.

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p(2)}{p(1)}, \quad R_2 : \frac{f_1(x)}{f_2(x)} < \frac{p(2)}{p(1)} \tag{2.37}$$

3. Equal prior probabilities and equal misclassification cost: $p_1 = p_2, c(1|2) = c(2|1)$. The classification rule now only depends on probability density ratio.

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1, \quad R_2 : \frac{f_1(x)}{f_2(x)} < 1 \tag{2.38}$$

Chapter 3

Methodology and data

This chapter contains the overview of the methodology applied and the management of data which would be used in this study. Within data management section, the data source, the process of collecting data, and the data limitation are also discussed.

3.1 Methodology

This study was performed based on a set of systematic methods, which is referred as methodology. The methodology applied was summarized in a flowchart, described in Figure 3.1. The results from each process are presented and discussed further in this report.

3.2 Data management

3.2.1 Data source

The data used in this study was obtained from iQx software built by AGR company. The software provides well data management from approximately 6,000 wells in Norwegian Continental Shelf (NCS) which are grouped into ± 1693 geological blocks. The software records various types of well data, such as geological description, well schematic, surface logging, and also well logging (or geophysical) data. However, not all of the wells listed in this software have a complete set of data, which later became obstacles in our study. The detailed explanation of the obstacles is discussed in the next section after the data were extracted.

3.2.2 Filtering and collecting data

As explained in the previous section, there was a large number of wells recorded in iQx software. In order to keep the simplicity of this study, we focused working on a small number of wells which have similarity in the geological description. From the aspect of data quality, we selected wells which had a complete set geophysical data to aid any ambiguity in the result.

The filtering process to select the appropriate wells for the study was executed manually due to limitations in software functionality. Therefore, the existence and the

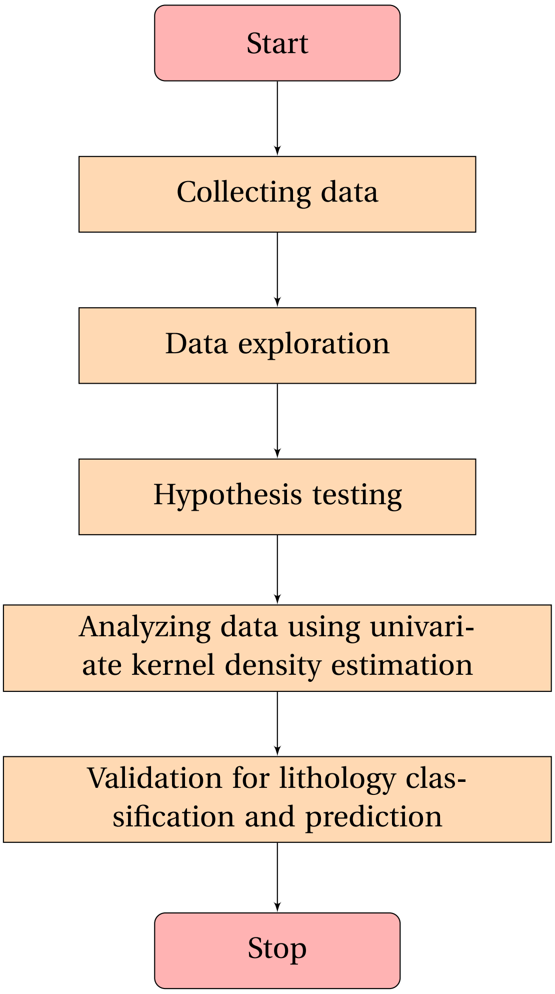

Start
Collecting data
Data exploration
Hypothesis testing
Analyzing data using univariate kernel density estimation
Validation for lithology classification and prediction
Stop

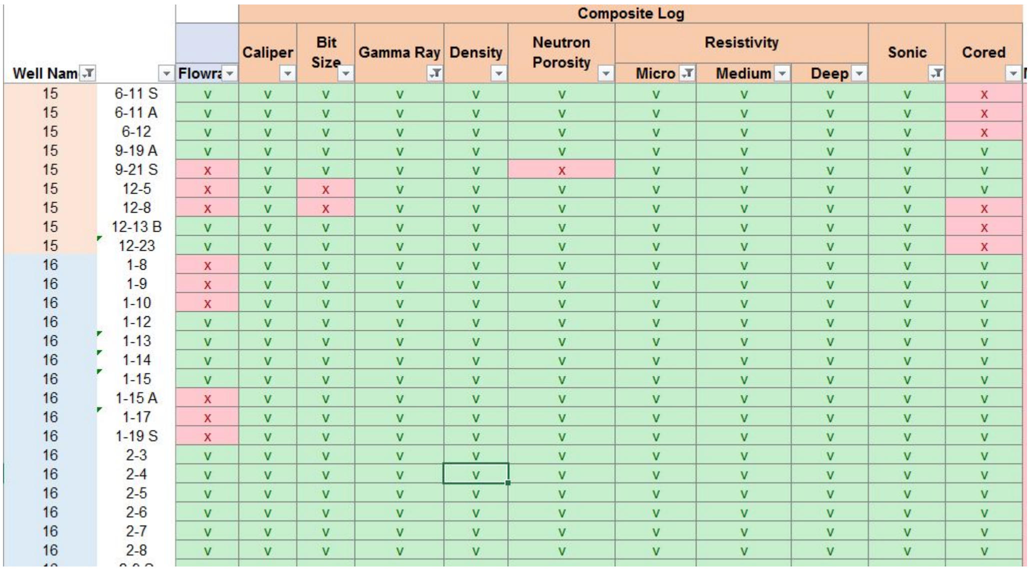
Figure 3.1: Flowchart of methodology in this study

quality of geophysical data from various wells were checked and recorded in a spreadsheet. Considering the large number of wells available and time limitation, the evaluation within this process was limited to 16 geological blocks.

From the evaluation, we chose 3 neighboring wells from Block 15, Well 15/5-7 A,Well 15/6-11 S, and Well 15/6-9 S. Apart from the availability and the good quality of logging data, these wells also had similar geological features. The evaluation also discovered that Block 15 has a large resource of wells which would be beneficial for validation process later in this study.

Once the filtering process was completed and the wells were chosen, we started extracting the data which related to the study, they were geophysical data, geological description and well schematic. The well schematic data provided information regarding casing size, hole size, and shoe depth. While the geological description provided information regarding formation distribution, characteristic, basal stereotype, depositional environment, and dominant lithology.

According to the geological description, Block 15 consisted ± 35 formations in total which were divided into 6 big formation groups. There were 4 major lithologies found in these formations, they were sandstone, shale, chalk, and carbonate lithology. However, the available geological description only provided one generalized description of formation for all of the selected wells and also lacked of detailed information (i.e. lithology mixture, minerals). The geological description of the formations is shown in



			Composite Log									
			Caliper	Bit Size	Gamma Ray	Density	Neutron Porosity	Resistivity			Sonic	Cored
Well Nam		Flowra						Micro	Medium	Deep		
15	6-11 S	v	v	v	v	v	v	v	v	v	v	x
15	6-11 A	v	v	v	v	v	v	v	v	v	v	x
15	6-12	v	v	v	v	v	v	v	v	v	v	x
15	9-19 A	v	v	v	v	v	v	v	v	v	v	v
15	9-21 S	x	v	v	v	v	x	v	v	v	v	v
15	12-5	x	v	x	v	v	v	v	v	v	v	v
15	12-8	x	v	x	v	v	v	v	v	v	v	x
15	12-13 B	v	v	v	v	v	v	v	v	v	v	x
15	12-23	v	v	v	v	v	v	v	v	v	v	x
16	1-8	x	v	v	v	v	v	v	v	v	v	v
16	1-9	x	v	v	v	v	v	v	v	v	v	v
16	1-10	x	v	v	v	v	v	v	v	v	v	v
16	1-12	v	v	v	v	v	v	v	v	v	v	v
16	1-13	v	v	v	v	v	v	v	v	v	v	v
16	1-14	v	v	v	v	v	v	v	v	v	v	v
16	1-15	v	v	v	v	v	v	v	v	v	v	v
16	1-15 A	x	v	v	v	v	v	v	v	v	v	v
16	1-17	x	v	v	v	v	v	v	v	v	v	v
16	1-19 S	x	v	v	v	v	v	v	v	v	v	v
16	2-3	v	v	v	v	v	v	v	v	v	v	v
16	2-4	v	v	v	v	v	v	v	v	v	v	v
16	2-5	v	v	v	v	v	v	v	v	v	v	v
16	2-6	v	v	v	v	v	v	v	v	v	v	v
16	2-7	v	v	v	v	v	v	v	v	v	v	v
16	2-8	v	v	v	v	v	v	v	v	v	v	v

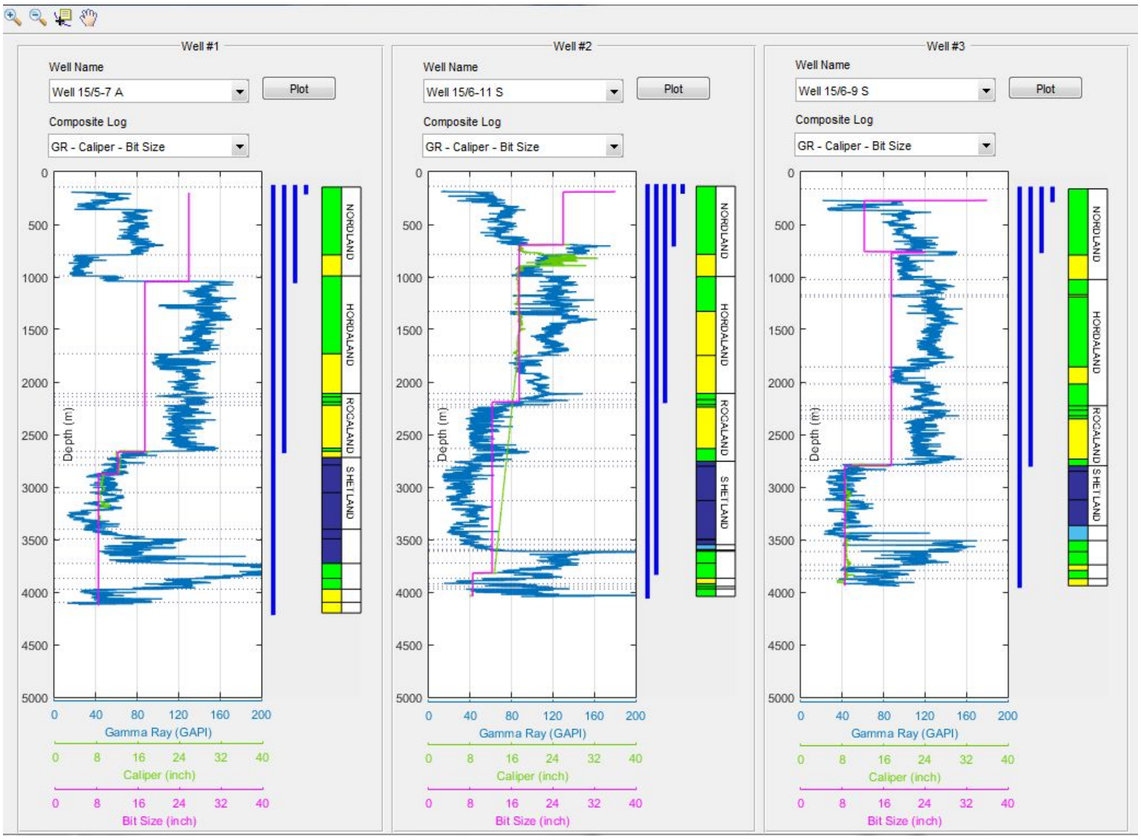
Figure 3.2: The spreadsheet for recording the availability of logging data in some of the wells

Appendix B.

After investigated the data attributes, we discovered that the geophysical data were only recorded based on the measured depth (MD), without the availability of true vertical depth (TVD)- and time-based geophysical data. The other data which were not available from the software were the information of the run wireline tools and well trajectory data. Such these data constraints would lead to limitations in the study analyses and results.

To begin with, we visualized all the extracted data into one figure of interface which constructed by using MATLAB, shown in Fig. 3.3. Eventually, we set some assumptions in this study:

1. The lithology information from the geological description represented the real lithology condition.
2. All the geophysical data were recorded from well-calibrated logging tools, thus the geophysical data were valid and represented real borehole condition.



Well #1

Well Name: Well 15/5-7 A

Composite Log: GR - Caliper - Bit Size

Well #2

Well Name: Well 15/6-11 S

Composite Log: GR - Caliper - Bit Size

Well #3

Well Name: Well 15/6-9 S

Composite Log: GR - Caliper - Bit Size

Figure 3.3: The interface for data visualization of Well 15/5-7 A, Well 15/6-11 S, and Well 15/6-9 S. In each well, the geophysical data were plotted in log traces, alongside casing data and geological description.

Chapter 4

Data exploration on GR data

This chapter contains the process of data exploration of GR data of Well 15/5-7 A, Well 15/6-11 S, and Well 15/6-9 Sand the results. This process was performed to check the quality of GR data and discover any variables grouping the GR data. From this process, we expected that the GR data would be valid to be used for further analysis. After the data exploration was completed, hypothesis testings were carried out in order to test and support the discovery from data exploration.

4.1 Data exploration results and discussions

Within the process of data exploration, we plotted the graphical description and calculated the numerical description of GR data which later would be used for observations. The graphical description included the data visualization by using histogram and boxplot while the numerical description included the measurement of data variability by calculating the mean, median, standard deviation, and IQR of the data. The bin width of the histogram was chosen from Scott rule (Chapter 2.2.1) while the boxplot visualization was adapted from Tukey method (Chapter 2.2.2).

4.1.1 GR data exploration under lithology grouping

In the theory of GR explained in Chapter 2.1.1, GR values are dependent on the types of lithology that present in the borehole and tends to have a similar value for one lithology group. According to this, the GR data description would be presented and observed based on lithology type. The graphical descriptions of GR data in Well 15/5-7 A,Well 15/6-11 S, and Well 15/6-9 S are presented in Fig. 4.1 - 4.3 with statistic description summarized in Table 4.1 - 4.3.

In Well 15/5-7 A, there were three different lithology types indicated, they were shale, sandstone, and chalk, while Well 15/6-11 S and Well 15/6-9 S had additional carbonate lithology. From the observation of histogram plots, GR of shale and sandstone from all of these wells had bimodal distribution which indicated by two major peaks presented in the histogram, while chalk and carbonate had uniform distribution. The observed bimodal distribution might indicate that the GR data of shale and sandstone lithology contained sub-groups.

Table 4.1: Statistic description of GR data in Well 15/5-7 A grouped according to the present lithology

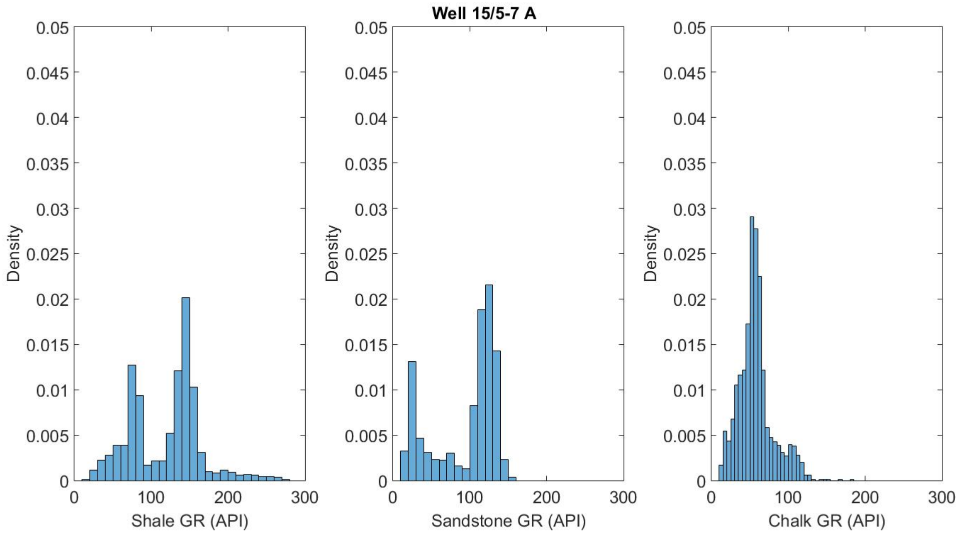
Lithology	**Mean**	**Median**	**Standard Deviation**	**IQR**
Chalk	57.39	55.18	22.67	21.20
Sandstone	94.71	115.66	41.63	73.45
Shale	118.89	132.88	45.78	68.69

Table 4.2: Statistic description of GR data in Well 15/6-11 S grouped according to present lithology

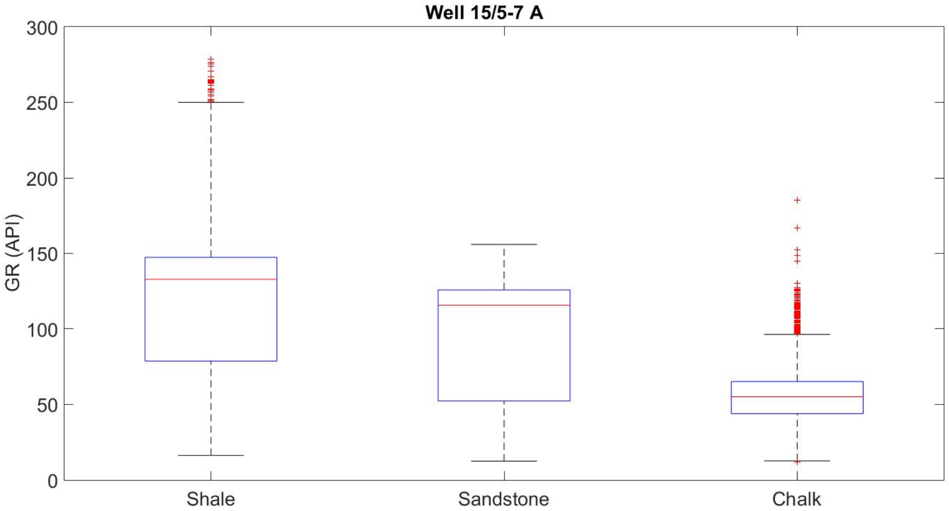
Lithology	**Mean**	**Median**	**Standard Deviation**	**IQR**
Carbonate	50.39	50.21	5.67	9.82
Chalk	35.66	35.62	10.42	14.01
Sandstone	91.66	100.05	31.33	55.28
Shale	97.47	92.31	33.85	52.89

Table 4.3: Statistic description of GR data in Well 15/6-9 S grouped according to present lithology

Lithology	**Mean**	**Median**	**Standard Deviation**	**IQR**
Carbonate	43.68	40.48	14.98	21.14
Chalk	45.06	45.66	10.05	13.25
Sandstone	102.58	104.32	18.39	19.12
Shale	116.99	122.42	21.58	22.88

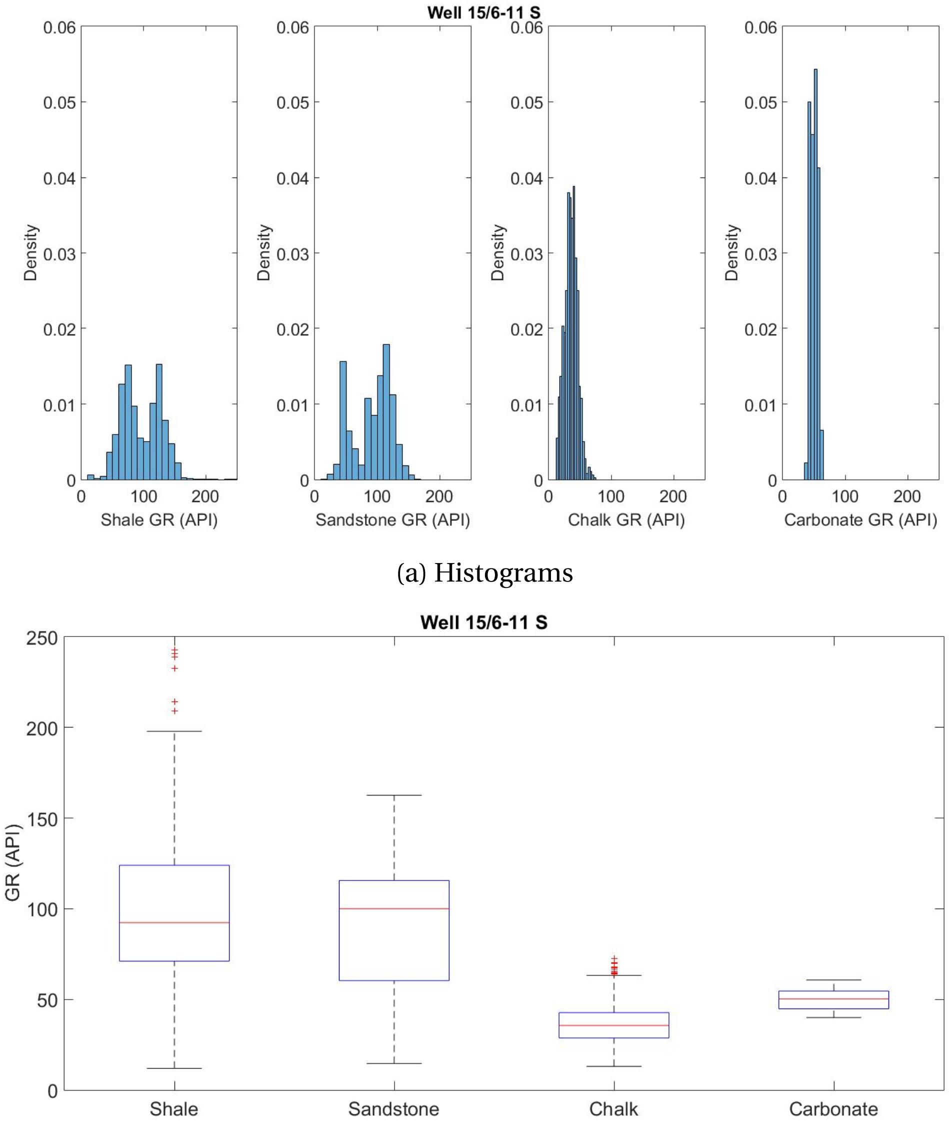


(a) Histograms



(b) Boxplots

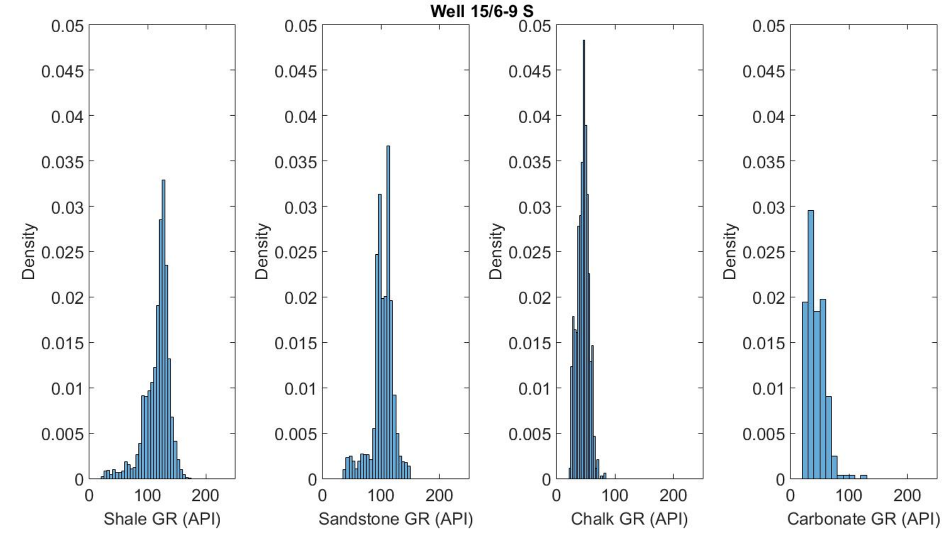
Figure 4.1: Graphical description of GR data in Well 15/5-7 A which grouped based on the lithology type



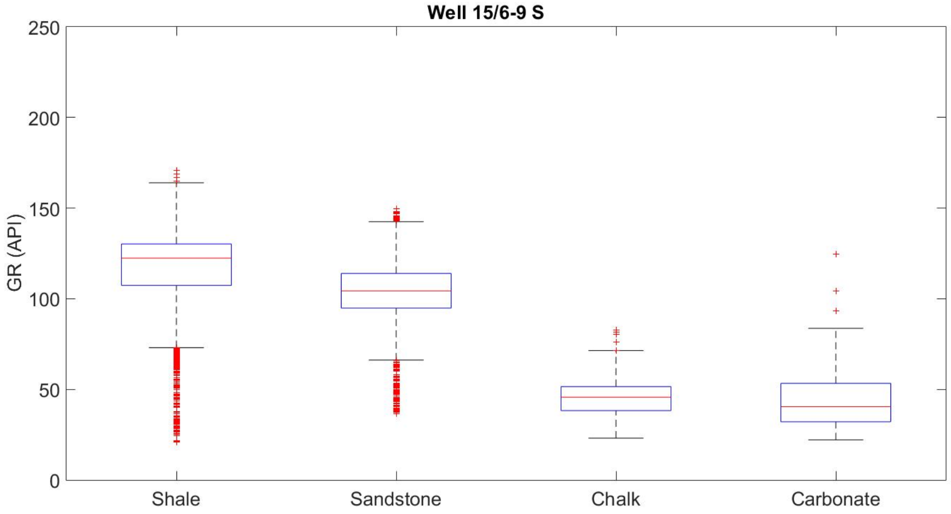
(a) Histograms

(b) Boxplots

Figure 4.2: Graphical description of GR data in Well 15/6-11 S which grouped based on the lithology type



(a) Histograms



(b) Boxplots

Figure 4.3: Graphical description of GR data in Well 15/6-9 S which grouped based on the lithology type

By looking into the GR distribution from histograms and boxplots, low GR values were observed in chalk and carbonate lithology while the GR values of shale and sandstone were varying. It was also discovered that lithology is a variable that divide the GR data into groups and each group appeared to be independent with each other. This presumption tested and proved later in the hypothesis test Chapter 4.2.1.

High variance of GR value in sandstone and shale lithology was detected from the standard deviation (σ) value, while carbonate and chalk had less variance. The widest span of GR values was detected in shale lithology with many outliers and long whiskers indicated from the boxplots. According to the discoveries, the GR data of shale and sandstone were not convincing due to indication of bimodal distribution and high variance of the data. We sensed that there were sub-groups might exist within each lithology group. Hence, another investigation was performed and the detailed explanation is provided in the next section.

4.1.2 GR data exploration under lithology grouping and hole size subgroup

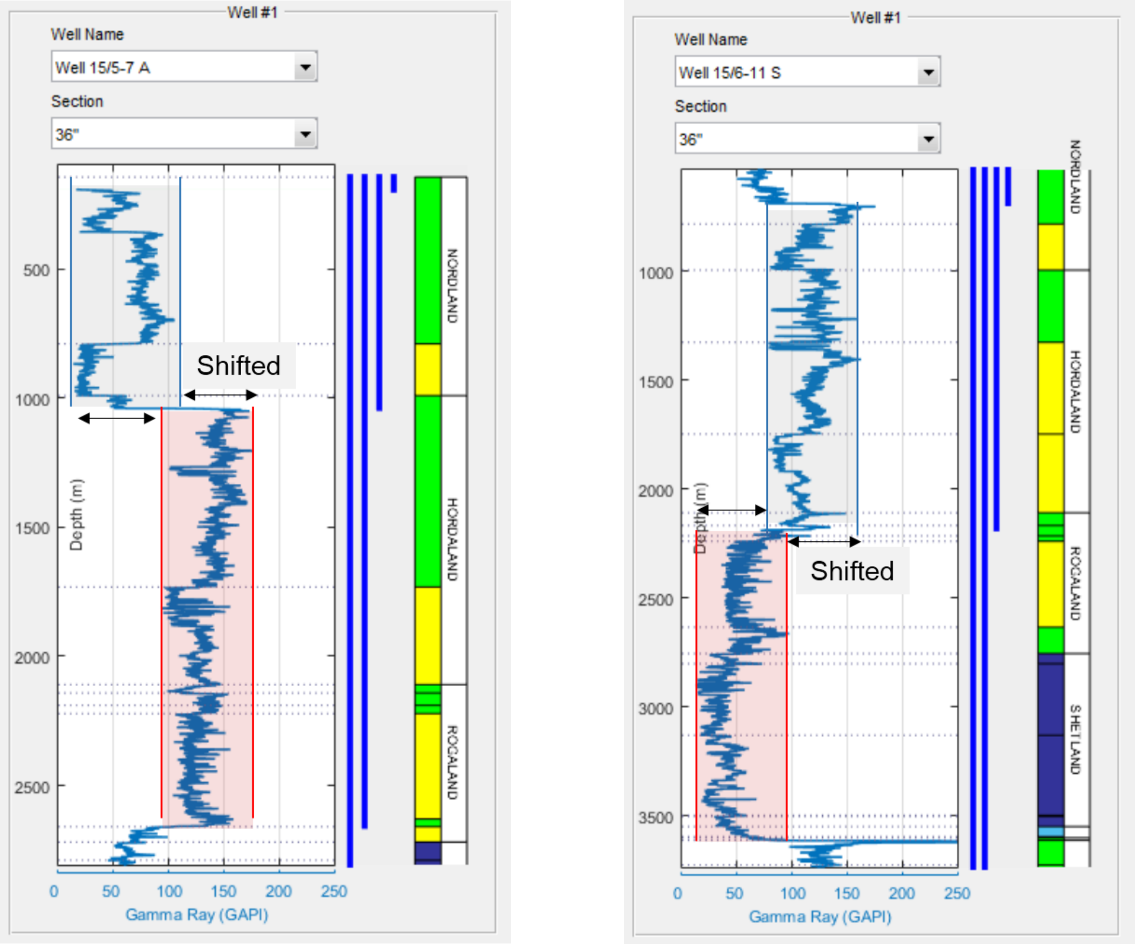
GR reading is affected by factors from borehole environment. Therefore, the quality of GR data depends on the correction was made or not. The error factors are: (i) tool eccentricity, (ii) hole size, (iii) mud weight, (iv) casing size, and (v) cement thickness. Referring to this theory, we presumed that the sub-groups within lithology group were emerged due to uncorrected GR data.

This presumption followed by an investigation to discover the variable of the subgroup which was performed by visualizing the GR data in log traces. By observing the log, we discovered that the values of GR in one hole size appeared to be shifted from GR values in the other hole size (see Fig. 4.4). The investigation was improved by constructing statistical graphs and calculating the statistic descriptions of GR data which grouped based on the lithology type and the hole size. The results are shown in Fig. 4.5 - 4.7 and Table 4.4 - 4.6.

By observing the minimum and maximum GR value of groups within shale and sandstone lithology, it was clearly seen that the span of GR value in each group was different from others and the variance of GR value was reduced. As an example, the GR value of sandstone lithology in Well 15/5-7 A without hole size grouping was ranged 12.47-155.91 API. But, by hole size grouping, now we discovered that the GR value of sandstone between 26" and 17 $\frac{1}{2}$" had a huge gap. Group 26" had fairly small GR value (13.64-50.57 API) compared to group 17$\frac{1}{2}$" (93.94-155.91 API)

Another approach proving that GR data distribution depends on the hole size was by observing the median of groups within one lithology type. Such an example, the median value of shale lithology in Well 15/6-11 S|: from group of 26" and 12 $\frac{1}{4}$" (70.89 API and 89.27 API, respectively) were rather small compared to median observed from group of 17$\frac{1}{2}$" and 8$\frac{1}{2}$" (125.39 API and 101.08 API, respectively).

Based on the histogram plots, shale and sandstone lithology showed better GR distribution with indication of unimodal distribution. In some cases, the hole size grouping resulted in symmetric distribution, such as groups of shale lithology with hole size 17$\frac{1}{2}$" from Well 15/5-7 A and from Well 15/6-9 S. Such this symmetric distribution had mean ($\bar{x}$) and median (M_d) values which were relatively closed ($\bar{x}$ = 142.69, M_d = 143.58 for shale 17$\frac{1}{2}$" Well 15/5-7 A, and $\bar{x}$ = 126.69, M_d = 126.68 for shale 17$\frac{1}{2}$" Well 15/6-9 S).



Well #1

Well Name

Well 15/5-7 A

Section

36"

Well #1

Well Name

Well 15/6-11 S

Section

36"

(a)

(b)

Figure 4.4: Shifted GR value from logging visualization: (a) 26" (blue area) and 17 $\frac{1}{2}$" (red area) in Well 15/5-7 A and b) 17 $\frac{1}{2}$" (blue area) and 12 $\frac{1}{4}$" (red area) in Well 15/6-11 S

Table 4.4: Statistic description of Well 15/5-7 A grouped according to the present lithology and subgroup of hole size

Well 15/5-7 A

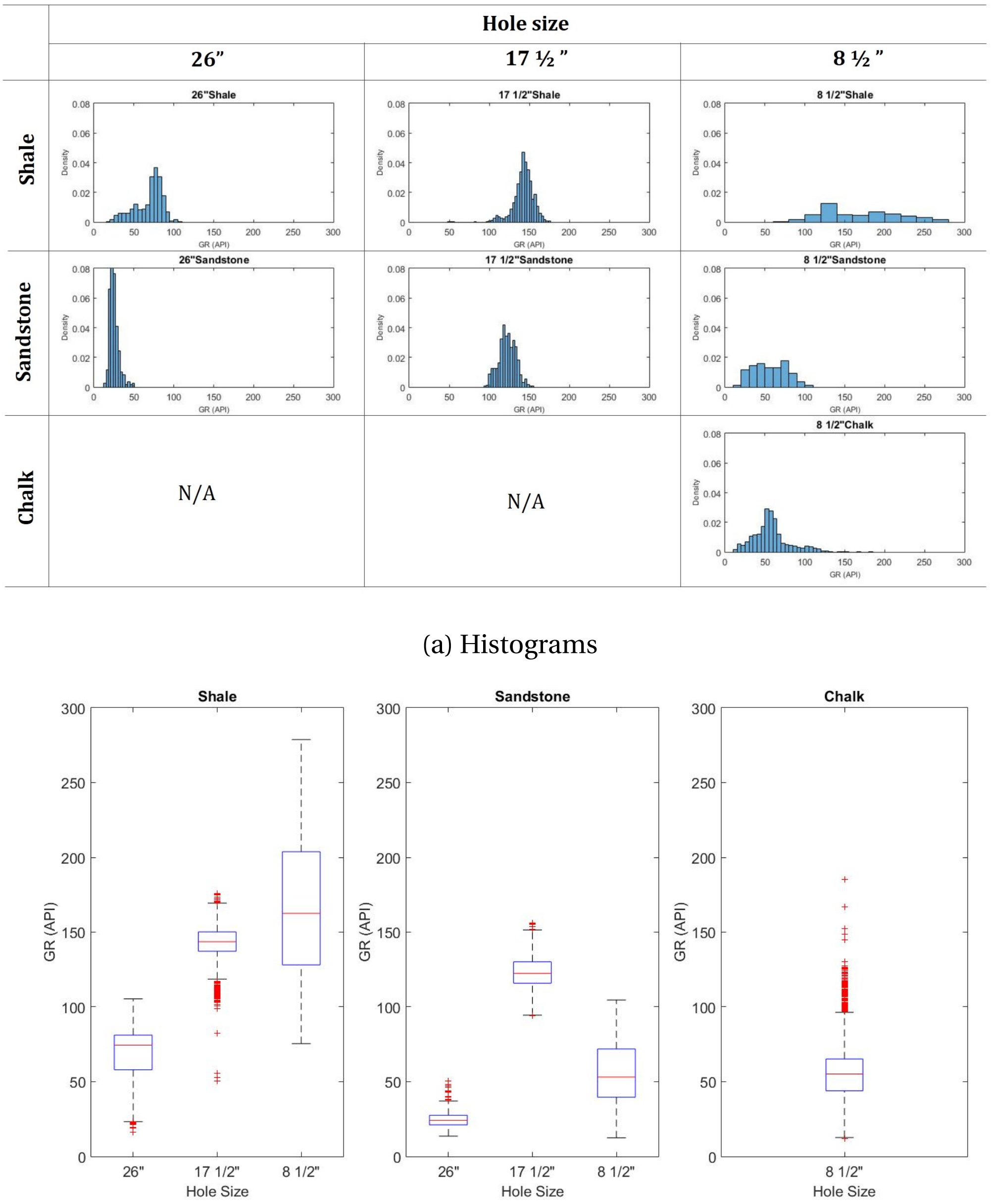
Lithology	**Hole Size**	**Mean**	**Median**	**St. Dev**	**IQR**	**Min**	**Max**
Shale	26"	69.02	74.42	17.48	23.13	16.28	105.38
	17 $\frac{1}{2}$"	142.69	143.58	12.74	12.95	50.53	175.64
	8 $\frac{1}{2}$"	169.02	162.51	48.54	75.60	75.47	278.76
Sandstone	26"	24.97	24.20	5.76	6.39	13.64	50.57
	17 $\frac{1}{2}$"	122.47	122.39	10.72	14.32	93.94	155.91
	8 $\frac{1}{2}$"	56.31	53.10	20.48	32.29	12.47	104.66
Chalk	26"	-	-	-	-	-	-
	17 $\frac{1}{2}$"	-	-	-	-	-	-
	8 $\frac{1}{2}$"	57.39	55.18	22.67	21.20	12.16	184.99

Table 4.5: Statistic description of Well 15/6-11 S grouped according to the present lithology and subgroup of hole size

Well 15/6-11 S							
Lithology	**Hole Size**	**Mean**	**Median**	**St. Dev**	**IQR**	**Min**	**Max**
Shale	26"	69.13	70.89	13.76	14.74	11.94	97.09
	17 $\frac{1}{2}$"	125.90	125.39	15.51	18.70	76.04	175.73
	12 $\frac{1}{4}$"	97.83	89.27	39.83	54.68	40.75	330.87
	8 $\frac{1}{2}$"	98.17	101.18	29.33	33.69	36.04	269.62
Sandstone	26"	-	-	-	-	-	-
	17 $\frac{1}{2}$"	110.16	110.82	16.02	22.78	77.37	162.63
	12 $\frac{1}{4}$"	50.50	46.54	9.67	13.79	35.75	84.67
	8 $\frac{1}{2}$"	50.52	42.33	21.77	39.23	14.63	109.02
Chalk	26"	-	-	-	-	-	-
	17 1/2"	-	-	-	-	-	-
	12 $\frac{1}{4}$"	35.66	35.62	10.42	14.01	12.99	72.67
	8 $\frac{1}{2}$"	-	-	-	-	-	-
Carbonate	26"	-	-	-	-	-	-
	17 $\frac{1}{2}$"	-	-	-	-	-	-
	12 $\frac{1}{4}$"	50.39	50.21	5.67	9.82	39.96	60.69
	8 $\frac{1}{2}$"	-	-	-	-	-	-

Table 4.6: Statistic description of Well 15/6-9 S grouped according to the present lithology and subgroup of hole size

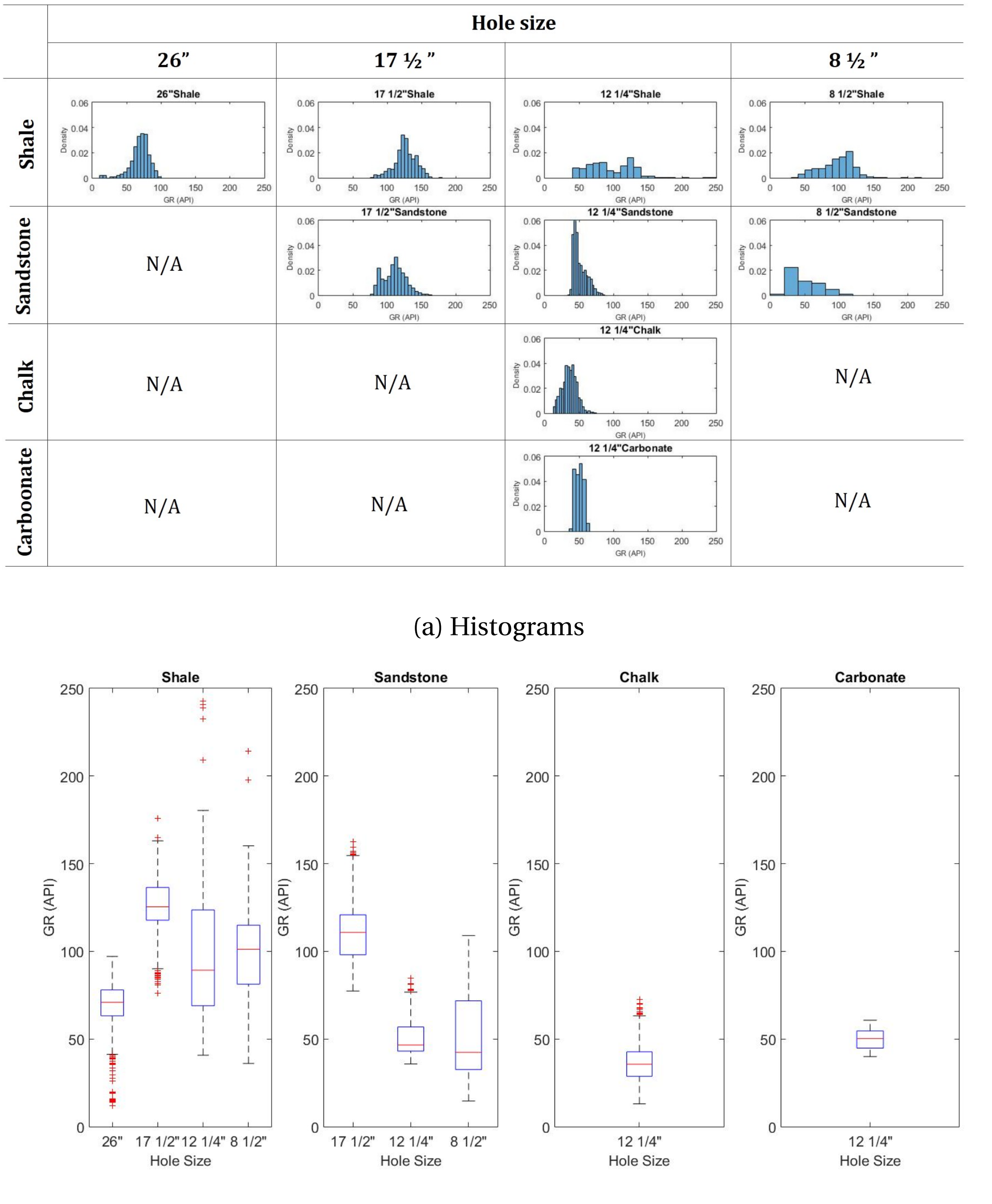
Well 15/6-9 S							
Lithology	**Hole Size**	**Mean**	**Median**	**St. Dev**	**IQR**	**Min**	**Max**
Shale	24"	94.55	98.57	21.39	14.66	20.84	133.09
	17 $\frac{1}{2}$"	126.69	126.68	8.78	10.79	85.16	156.34
	8 $\frac{1}{2}$"	113.29	116.03	29.06	44.30	40.05	170.61
Sandstone	24"	-	-	-	-	-	-
	17 $\frac{1}{2}$"	107.84	108.49	11.73	16.54	85.31	149.65
	8 $\frac{1}{2}$"	68.71	70.03	17.47	32.09	36.41	109.81
Chalk	24"	-	-	-	-	-	-
	17 $\frac{1}{2}$"	-	-	-	-	-	-
	8 $\frac{1}{2}$"	45.06	45.66	10.05	13.25	23.09	82.65
Carbonate	24"	-	-	-	-	-	-
	17 $\frac{1}{2}$"	-	-	-	-	-	-
	8 $\frac{1}{2}$"	43.68	40.48	14.98	21.14	22.16	124.79



(a) Histograms

(b) Boxplots

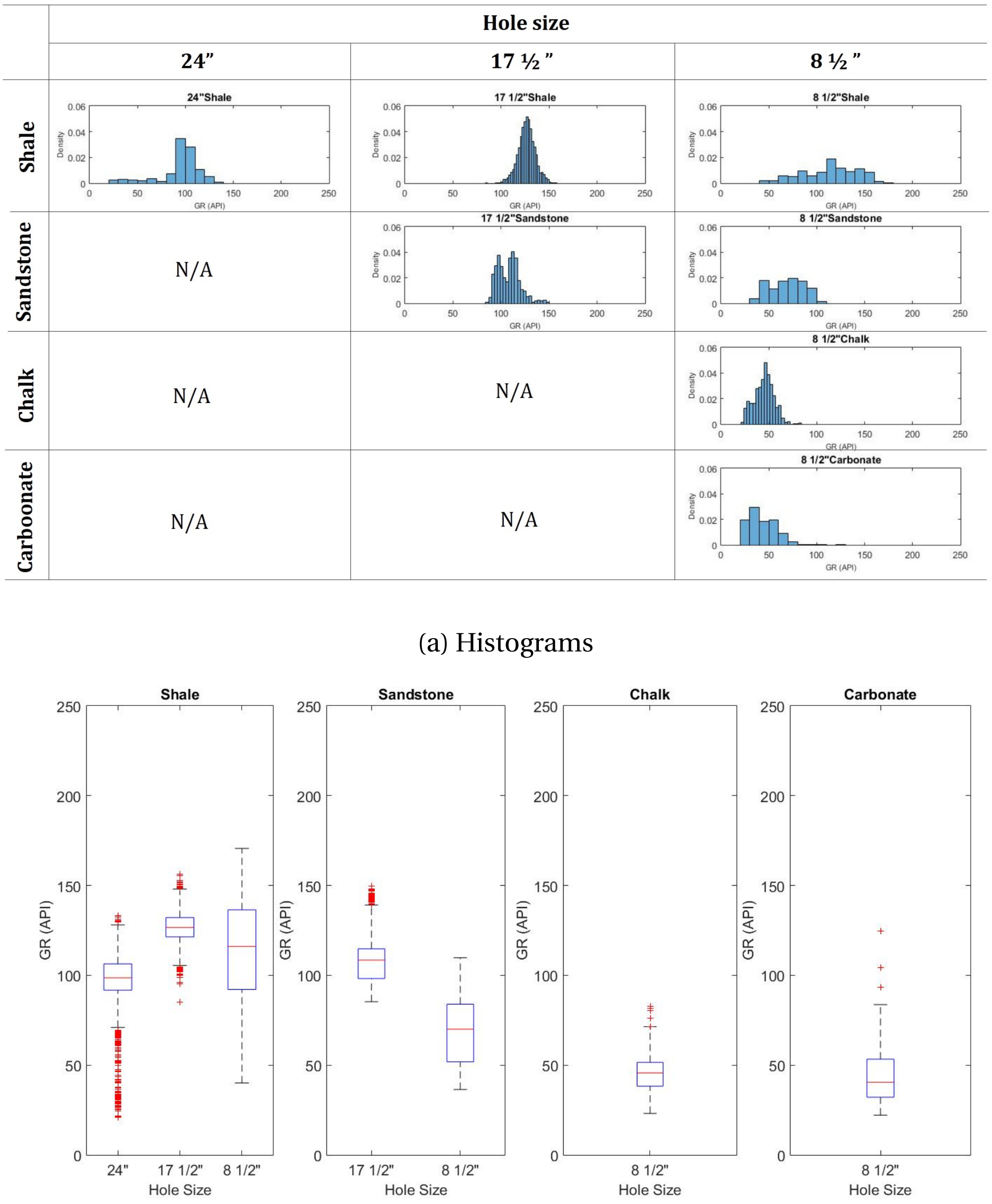
Figure 4.5: Graphical description of GR data in Well 15/5-7 A which grouped according to the lithology type and hole size



(a) Histograms

(b) Boxplots

Figure 4.6: Graphical description of GR data in Well 15/6-11 S which grouped according to the lithology type and hole size



(a) Histograms

(b) Boxplots

Figure 4.7: Graphical description of GR data in Well 15/6-9 S which grouped according to the lithology type and hole size

Summing up the investigation, hole size grouping in each lithology group had improved the GR data distribution significantly and it was proved that our presumption was correct. The results also indicated that each groups are independent with others. To prove these discoveries and the presumptions, we performed another hypothesis testing in Chapter 4.2.2.

Grouping the GR data based on the hole size can also remove other error factors, such as mud weight and casing size, because in the application usually the value of mud weight and casing size are constant during drilling one hole section. We contemplated on correcting the data based on the borehole environment, but the correction was not possible because we were constrained with the availability of GR tool description from the data source.

4.2 Hypothesis testing results and discussions

4.2.1 Hypothesis testing #1

The first hypothesis testing was directed to check and test our presumption that GR data contained an independent variable, which was lithology, from data exploration in Chapter 4.1.1. The stated hypothesis are following:

H_0 : the GR data had similar distribution and identical.

H_1 : lithology type affected GR data value and GR data behavior in each lithology was independent with the others.

The Kruskal-Wallis test was executed by using sample data from GR data and grouping variable of the lithology type. The level of significance for the test was set to 5%. The summary of p-value and mean rank is presented in Table 4.7 and the detailed result of ANOVA table is provided in Appendix C.

Table 4.7: P-value of the first hypothesis test for indicating any lithology group in GR data. The detailed lithology group investigated in each well is provided in Table 4.8

Well	**P-Value**
Well 15/5-7 A	$< 1.00 \times 10^{-323}$
Well 15/6-11 S	$< 1.00 \times 10^{-323}$
Well 15/6-9 S	$< 1.00 \times 10^{-323}$

The lithology effect on GR data has been evaluated by using Kruskal-Wallis H test. According to p-value given in each well, the null hypothesis was rejected due to p-value<0.05. As stated in the theory, the Kruskal-Wallis test is an omnibus test, thus we could not indicate which specific groups are statistically and significantly different with the others. However, observation of the mean rank value in each group can give a picture how each group differs from the others.

If a group has a mean rank value which relatively closed with the value from other groups, then these groups are considered identical. In Well 15/5-7 A, the mean rank

Table 4.8: Mean ranks of each lithology group in Well 15/5-7 A, Well 15/6-11 S, and Well 15/6-9 S from the first hypothesis test

Well	**Mean Group Rank**			
	Shale	Sandstone	Chalk	Carbonate
Well 15/5-7 A	5.129e+03	3.8037e+03	2.0172e+03	-
Well 15/6-11 S	4.8042e+03	4.4590e+03	1.0064e+03	2.1327e+03
Well 15/6-9 S	4.8096e+03	3.3632e+03	0.8340e+03	0.7671e+03

values of shale, sandstone, and chalk were significantly different. Meanwhile, sandstone and shale lithology in Well 15/6-11 S had quite similar mean rank (4804 and 4459, respectively). However at this state, we could not conclude that shale and sandstone lithology were identical because it was proved that there were subgroups of hole size within shale and sandstone lithology (see Chapter 4.1.2).

4.2.2 Hypothesis testing #2

The second hypothesis testing was conducted in order to test our presumption from Chapter 4.1.2 that the GR data not only contained independent variable of lithology but also hole size. The stated hypotheses are following:

H_0 : the GR data in lithology groups had similar and identical distribution.

H_1 : GR data in lithology groups were divided into another independent variable which was borehole size.

The level of significance was set to 5%. The results of p-value are summarized in Table 4.9 and the ANOVA table is provided in Appendix C.

Referring to the results of p-value in Table 4.9, all of the groups had p-value<0.05, thus the null hypothesis was rejected. this also proved that there were at least 2 subgroups existed in each lithology group. In order to acknowledge any similarity between the groups, we observed the mean rank value.

Observing groups in Well 15/5-7 A, we discovered that the shale lithology in $17\frac{1}{2}$" and 8 $\frac{1}{2}$" group had similar mean rank values, respectively 2313 and 2556. However, a contrast of the mean rank values between these groups was identified in sandstone lithology. This observation showed that shale from group $17\frac{1}{2}$" and 8 $\frac{1}{2}$" were identical but not for the sandstone lithology.

Another case from Well 15/6-11 S, both sandstone and shale had similar mean rank values which identified in hole $12\frac{1}{4}$" and $8\frac{1}{2}$" group. Thus, the GR data from these groups were identical for both sandstone and shale lithology and might be originated from the same population. However, we took a careful approach and did not merge the GR data of these groups in order to avoid errors.

In previous hypothesis testing, the observation of mean rank values of lithology groups were limited because indication of hole size sub-group. Now, after grouping GR data according to the hole size, we observed that lithologies in a group of hole size had a diverse mean rank value. A significant variation of the mean rank of each lithology was observed in the group of hole size 8 $\frac{1}{2}$" in Well 15/6-11 S with mean rank of shale = 1545, sandstone = 455, chalk = 795, carbonate = 75. This observation proved that each lithology group was independent and not identical with others.

Based on the observations of the results of hypotheses testing, the subsequent analysis on GR data would follow the discoveries from hypotheses test. Henceforth, the analysis would be performed and presented separately according to the group of lithology type and hole size.

Table 4.9: P-value result of the second hypothesis test. Empty (-) p-value results indicated that there are no hole size group found in the particular lithology group.

Well	**Lithology**	**P-Value**
Well 15/5-7 A	Shale	$< 1.00 \text{ x } 10^{-323}$
	Sandstone	$< 1.00 \text{ x } 10^{-323}$
	Chalk	-
Well 15/6-11 S	Shale	$< 1.00 \text{ x } 10^{-323}$
	Sandstone	$< 1.00 \text{ x } 10^{-323}$
	Chalk	-
	Carbonate	-
Well 15/6-11 S	Shale	$< 1.00 \text{ x } 10^{-323}$
	Sandstone	$1.061 \text{ x } 10^{-132}$
	Chalk	-
	Carbonate	-

Table 4.10: Mean rank for each category. Empty (-) mean rank results indicated that there are no mean rank for selected hole size and lithology group.

Well	**Hole Size**	**Lithology**			
		Shale	Sandstone	Chalk	Carbonate
Well 15/5-7 A	26"	652	224	-	-
	17 $\frac{1}{2}$"	2313	1598	-	-
	8 $\frac{1}{2}$"	2556	584	1010	-
Well 15/6-11 S	26"	784	-	-	-
	17 $\frac{1}{2}$"	2435	1886	-	-
	12 $\frac{1}{4}$"	1545	455	795	47
	8 $\frac{1}{2}$"	1570	417	-	-
Well 15/6-11 S	26"	757	-	-	-
	17 $\frac{1}{2}$"	2572	1043	-	-
	8 $\frac{1}{2}$"	1883	152	570	145

4.3 Concluding remarks

- Data exploration on GR data has been performed. The observation of the results showed that there were at least two group types within the GR data distribution, they were lithology type and borehole size.

- Hypothesis tests were conducted by using Kruskal-Wallis H test to support the discoveries from data exploration.

- Two hypothesis tests were run and the p-value from the results had shown that GR data was dependent to the lithology type and the borehole size.

Chapter 5

Univariate KDE analysis on GR data

This chapter introduces the application of KDE on GR data to get the estimation of probability density. Later on, a validation process was performed to assess the effectiveness of GR variable on classifying lithology by using the results acquired from KDE application. The results and discussions from validation process are provided in a separated section in this chapter.

5.1 KDE analysis on GR data

The KDE application was performed in order to get a smooth probability density estimation of non-parametric distribution of GR data. The kernel estimation was applied for wells: Well 15/5-7 A, Well 15/6-11 S, and Well 15/6-9 S to get the probability density based on Equation 2.24.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right), \qquad \text{(2.24 revisited)}$$

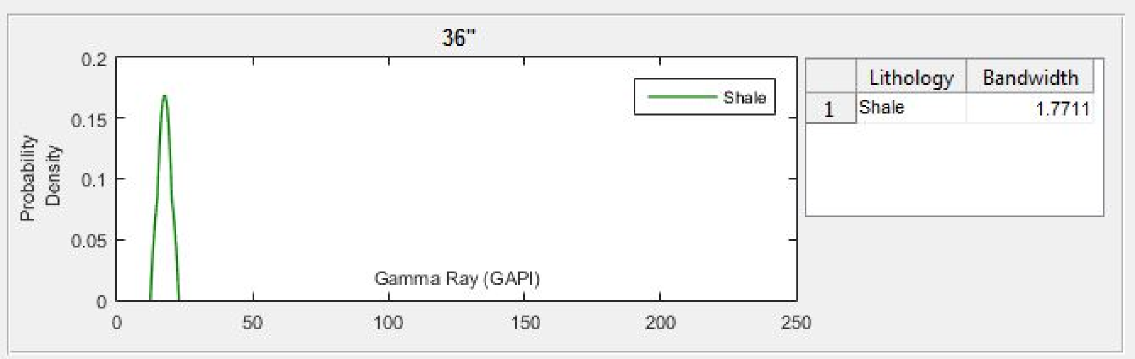
The kernel function applied for this analysis was Epanechnikov function by considering that this function minimizes the AMISE value (see Chapter 2.4.2). However, it was stated that the choice of kernel function does not affect the result significantly (Silverman, 1981).

The estimation was carried out by using MATLAB R2015A built-in function, `ksdensity`, which returns the estimation of the probability density and the sample vector (Deveaux, 1999). The optimal bandwidth of kernel smoothing, h, was also calculated and selected by using this function automatically. The default of the optimal bandwidth from this function is based on a distribution of normal densities. The values of the optimal bandwidth for each data group are provided in the result section together with the results of probability density plots.

5.1.1 Results

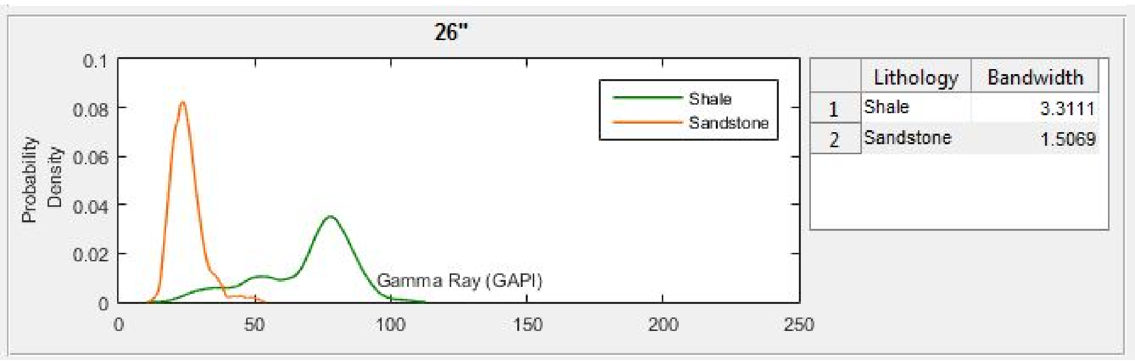
The results of probability density plot for each well are provided based on hole size. The lithology groups which presented in each hole size group were plotted together in one axis with different line colors (indicated in the legend).

Well 15/5-7 A



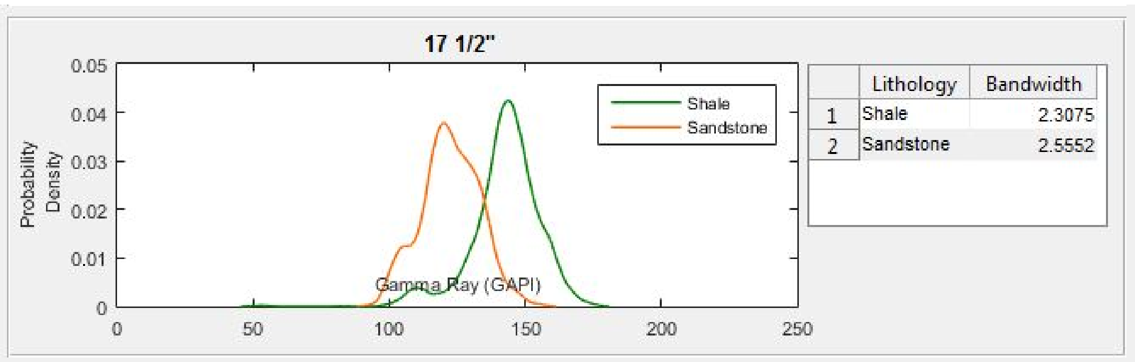
	Lithology	Bandwidth
1	Shale	1.7711

(a)



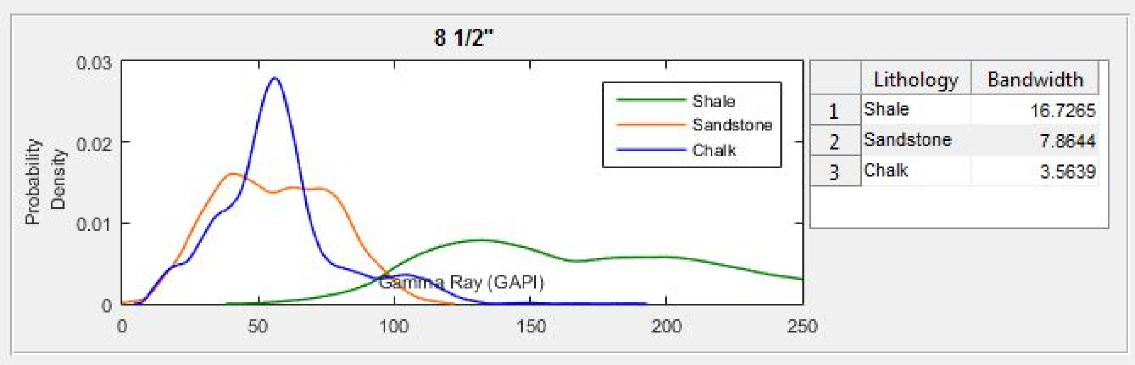
	Lithology	Bandwidth
1	Shale	3.3111
2	Sandstone	1.5069

(b)



	Lithology	Bandwidth
1	Shale	2.3075
2	Sandstone	2.5552

(c)

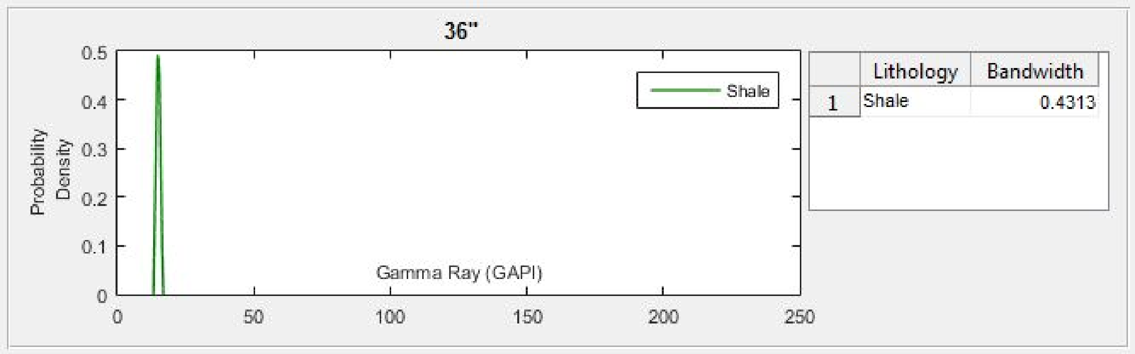


	Lithology	Bandwidth
1	Shale	16.7265
2	Sandstone	7.8644
3	Chalk	3.5639

(d)

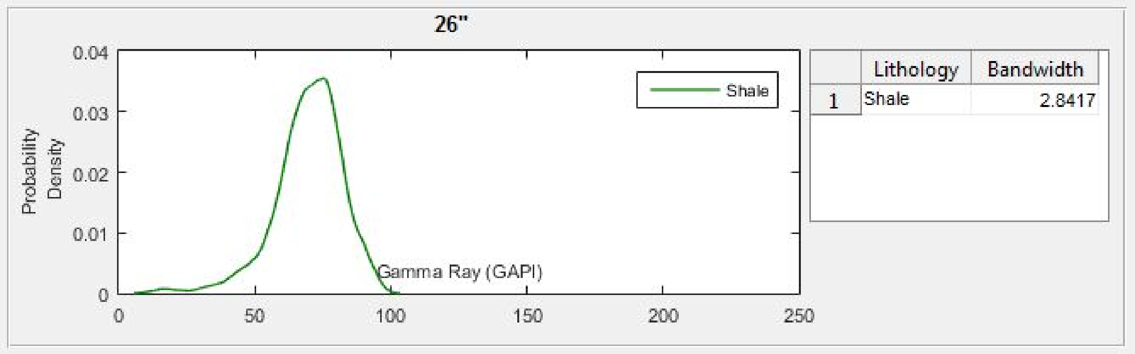
Figure 5.1: Density probability plots with KDE and kernel bandwidth results of Well 15/5-7 A for present lithology in borehole size: (a)36", (b)26", (c)17 $\frac{1}{2}$", and (d)8$\frac{1}{2}$"

Well 15/6-11 S



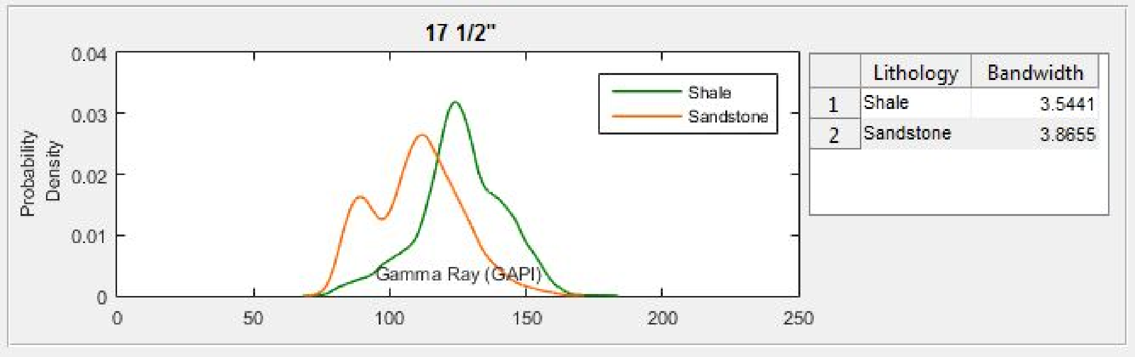
	Lithology	Bandwidth
1	Shale	0.4313

(a)



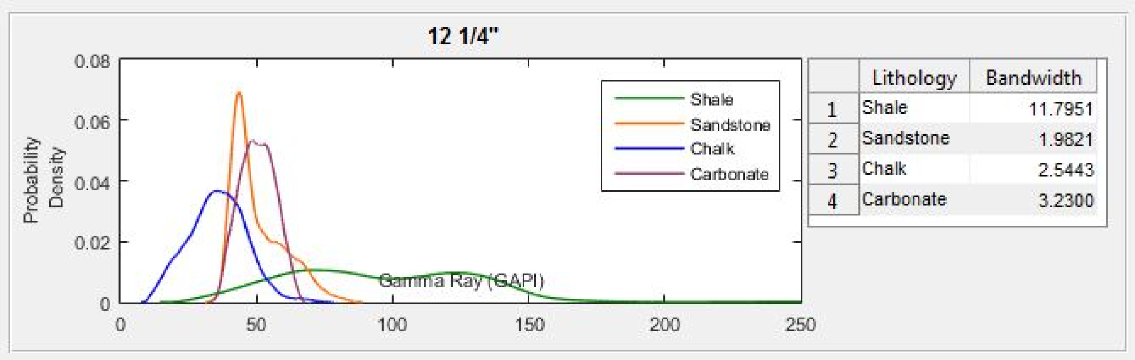
	Lithology	Bandwidth
1	Shale	2.8417

(b)



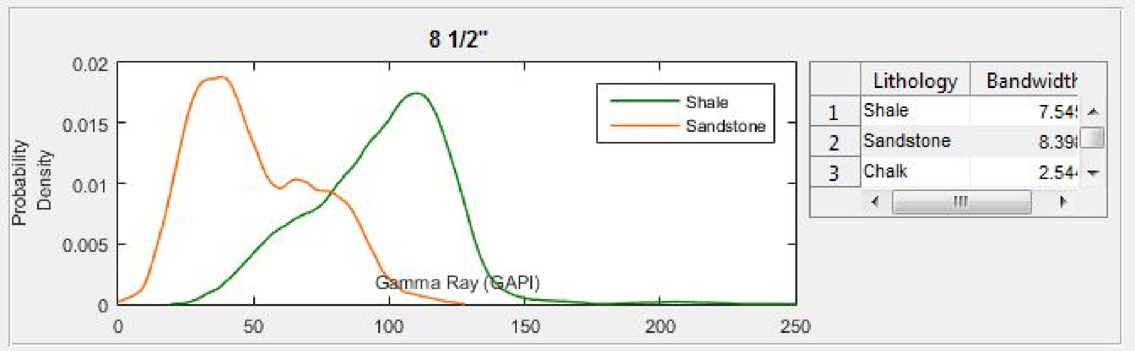
	Lithology	Bandwidth
1	Shale	3.5441
2	Sandstone	3.8655

(c)



	Lithology	Bandwidth
1	Shale	11.7951
2	Sandstone	1.9821
3	Chalk	2.5443
4	Carbonate	3.2300

(d)

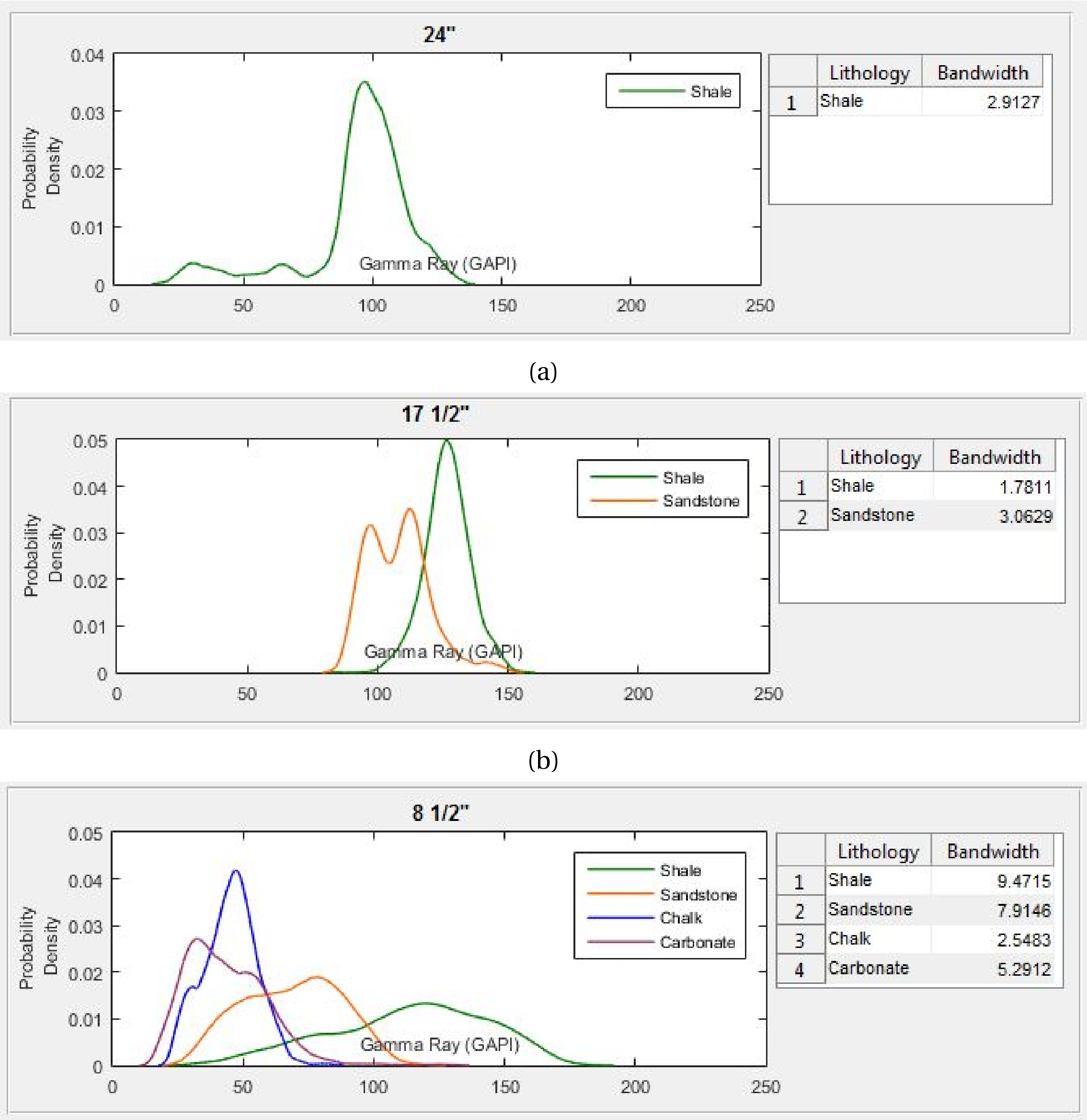


	Lithology	Bandwidtl
1	Shale	7.54!
2	Sandstone	8.39!
3	Chalk	2.54

(e)

Figure 5.2: Density probability plots with KDE and kernel bandwidth results of Well 15/6-11 S for present lithology in borehole size: (a)36", (b)26", (c)17 $\frac{1}{2}$", (d)12 $\frac{1}{4}$", and (e)8$\frac{1}{2}$"

Well 15/6-9 S



	Lithology	Bandwidth
1	Shale	2.9127

(a)

	Lithology	Bandwidth
1	Shale	1.7811
2	Sandstone	3.0629

(b)

	Lithology	Bandwidth
1	Shale	9.4715
2	Sandstone	7.9146
3	Chalk	2.5483
4	Carbonate	5.2912

(c)

Figure 5.3: Density probability plots with KDE and kernel bandwidth results of Well 15/6-9 S for present lithology in borehole size: (a)24", (b)17 $\frac{1}{2}$", and (c)8$\frac{1}{2}$"

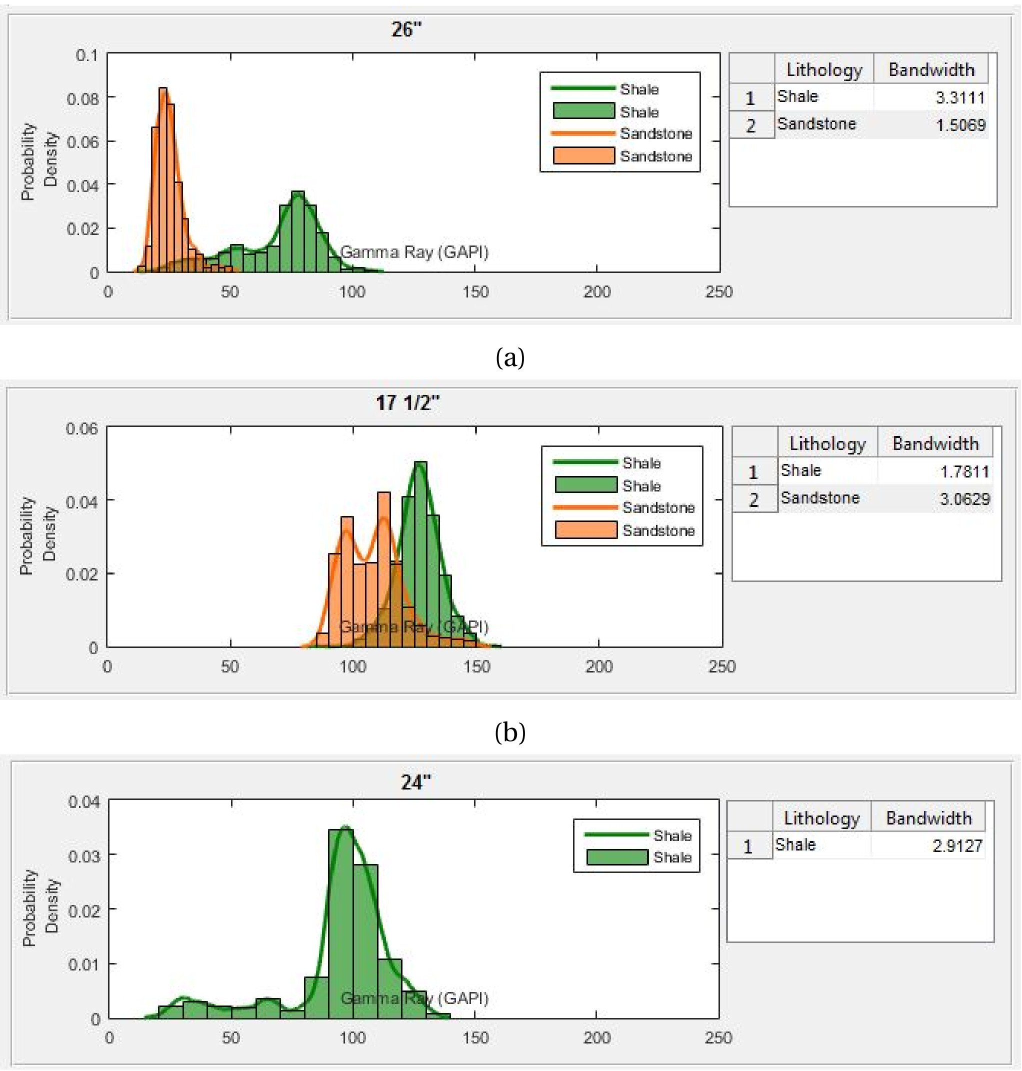
5.1.2 Discussions

There are two discussion points presented within this discussion section. The first discussion point discusses the results from histogram and kernel estimator. While, the second discussion point discussed about the GR data distribution for each lithology.

Simple and smooth probability density estimator

The plots of probability density estimated by kernel estimator in the result section showed smoothed probability density lines along GR values. In addition, we also discovered that the kernel estimator was able to match the non-parametric distribution of GR data. To assess the effectiveness of kernel estimator over histogram in estimating probability density, we plotted data of several groups using kernel estimator and histogram in one figure, shown in Fig. 5.4. In this comparison, the histogram was plotted by using the same bin width method as in Chapter 4.1, which was Sturges method.

In figure 5.4a, both histogram and kernel estimator returned similar estimation of the probability density and were able to approach the non-parametric distribution of GR data. In another investigation of figure 5.4b, we observed a bimodal distribution of GR in sandstone lithology with two adjacent modes. The histogram returned poor estimation of probability density around these modes which was indicated by the abrupt change of probability density values on GR values between these modes. Meanwhile, kernel estimator returned a better estimation which indicated by the gradual changes between these modes. The better estimation from kernel estimator was due to the ability of this estimator to calculate the superimpose effect of GR values in the local neighborhood around these modes. Fig. 5.4c showed that the kernel estimator was also able to approach skewed distribution.



	Lithology	Bandwidth
1	Shale	3.3111
2	Sandstone	1.5069

(a)

	Lithology	Bandwidth
1	Shale	1.7811
2	Sandstone	3.0629

(b)

	Lithology	Bandwidth
1	Shale	2.9127

(c)

Figure 5.4: Comparison of histogram and KDE in estimating probability density in (a)Well 15/5-7 A , section 26", (b)Well 15/6-9 S , section 17 $\frac{1}{2}$", and (c) section 24"

As stated in theory, kernel estimator depends on the smoothing parameter, h, just as histogram depends on the bin width, thus, the probability density distribution from histogram and kernel depends on how these parameters are chosen. However, by observing the results in general, it was clearly seen that the visualization of probability density from kernel estimator was much better compared to the discreteness of the histogram.

Despite of that, histogram has another weakness comparing distribution of several group samples. The histogram of each sample could not be plotted and merged together in one figure unless the samples have a uniform bandwidth. Because histogram plot contains discrete bins, the varying bandwidths of samples in one figure can cause a subjective comparison between one sample to others. This weakness of histogram can be avoided by using the kernel estimator because the probability density returned by kernel estimator is continuous.

Data distribution of GR from probability distribution

Observing the probability density of groups which contain two or more than two lithologies, there were overlapped distributions of lithologies which forming region(s). This region indicated that there were multiple lithologies exist for any GR value located within it. Moreover, the size of the overlapped regions affects the misclassification rate. The bigger the size of this region, more misclassification will likely occur.

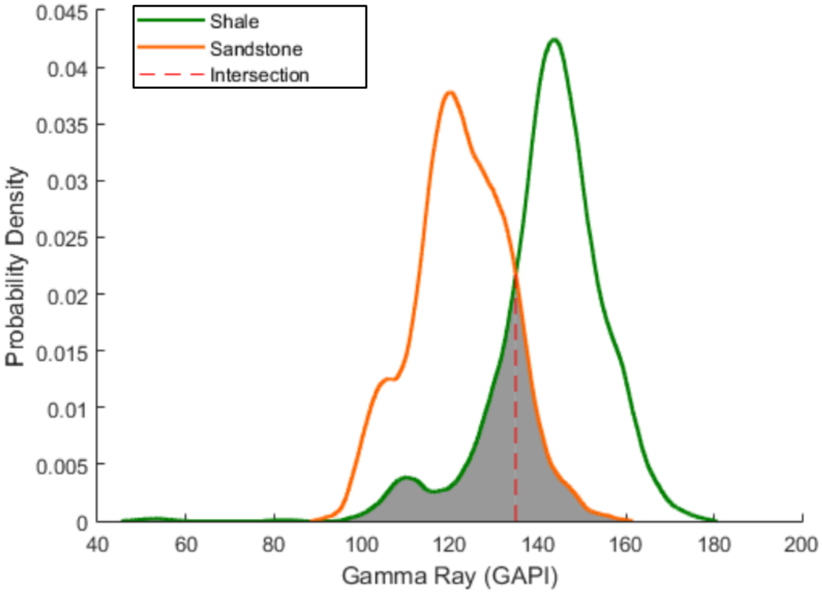
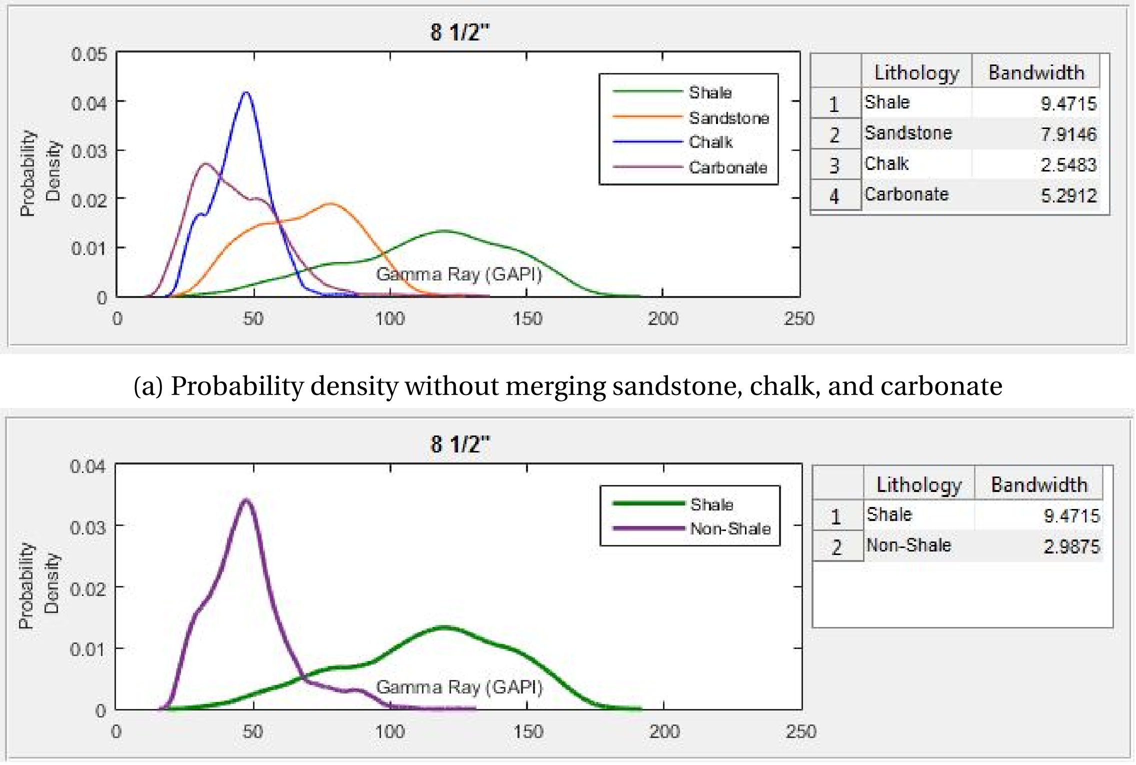


Figure 5.5: Probability density of Well 15/5-7 A for hole size $17\frac{1}{2}$". The gray area is the overlapping distribution between shale and sandstone lithology.

Group of $17\frac{1}{2}$" from Well 15/5-7 A was dominated with high GR value of shale and low GR value of sandstone. The separation between these lithologies was clear if we observed the peak of probability density of each lithology. However, there was an overlapped region formed because of the left-skewed distribution from shale lithology. Similar cases also found in other groups which contain shale and sandstone lithology.

In another case, several overlapped regions were formed within groups which contain additional chalk and/or carbonate lithology. This typical case was identified in group of 8 $\frac{1}{2}$" from Well 15/5-7 A and Well 15/6-9 S and 12 $\frac{1}{4}$" from Well 15/6-11 S. These multiple overlapped regions were formed due to low GR value of chalk, carbonate, and sandstone lithology and the separation of these lithologies was difficult to be identified. This discovery was consistent with the discussion from GR data exploration in Chapter 4.

Based on this, chalk, carbonate, and sandstone lithology were grouped into one group, which was non-shale lithology group, because these lithologies had identical value of GR. This decision was also made to reduce the complexity of classification of these lithologies. Henceforth, the lithology classification was simplified into shale and non-shale lithology group. This type of classification was in accordance with the usage of GR data which is shale identification. The results of probability density plots categorized into shale and non-shale lithology are shown in Appendix D.



	Lithology	Bandwidth
1	Shale	9.4715
2	Sandstone	7.9146
3	Chalk	2.5483
4	Carbonate	5.2912

(a) Probability density without merging sandstone, chalk, and carbonate

	Lithology	Bandwidth
1	Shale	9.4715
2	Non-Shale	2.9875

(b) Probability density with merging sandstone, chalk, and carbonate into non-shale lithology group

Figure 5.6: Comparison of probability density from group of 8$\frac{1}{2}$" in Well 15/6-9 S which are (a) not-merged and (b) merged

A significant change of probability density distribution was observed in groups which contain more than two lithologies (group of 8 $\frac{1}{2}$" from Well 15/5-7 A and Well 15/6-9 Sand 12 $\frac{1}{4}$"). Meanwhile, there was no significant change indicated in the groups which only contain shale and sandstone lithologies. Fig. 5.6 showed the alteration by merging chalk, carbonate, and sandstone lithologies into a non-shale group. The overlapped regions in Fig. 5.6a were reduced and simplified into one overlapped region in figure Fig. 5.6b due to the skewed distribution of non-shale lithology group.

So, what is the explanation of the overlapped lithology distribution in term of geological aspect? Based on the theory of GR, the GR tool works by measuring the emitted radioactivity of rock in the borehole. Even though shale appertains as a lithology with high radioactivity, there are some cases when high radioactivity is failed to be indicated as shale, such as in the shaly sandstone formation, uranium bearing formation, high radioactivity sandstone formation. These failed cases often cause misinterpretation which leads to lithology misclassification. This condition explained why the overlapped regions formed between shale and non-shale lithologies. There are methods that have been proposed to improve the lithology classification in cases mentioned above, such as the usage of P_e measurement or quantitative analysis to indicate minerals in shaly sandstone (Poupon and Gaymard, 1970), uranium measurement from spectral GR logging to indicate uranium-rich formation, and potassium reading to indicate sandstone with mica formation (Ellis and Singer, 2010). However, the investigation on these cases was beyond the scope of this study.

5.2 Validation of probability density on GR data by using KDE

From the results and discussions in the previous section, kernel estimator was proved to have the capability to estimate probability density of GR data which has non-parametric distribution. Based on this discovery, we performed a validation into the models which were constructed from GR data by using kernel estimator. The motivation of validation was to evaluate the capability of GR variable to discriminate lithologies by testing unknown dataset into the model and investigating how the model will generalize to the testing dataset. The goal from this process was to estimate the accuracy of the GR model in practice (lithology classification and prediction for this case).

Prior to the validation process, the model must be constructed from a known dataset (or referred as *training dataset*). In addition to the model, the dataset to be tested into the model (or referred as *testing dataset*) which was taken from unknown data also should be prepared. Once the models and testing data were prepared, the validation was started by classifying elements in the testing data toward the model based on the classification rule.

Within the classifying process, the number of observations of elements in testing data which correctly classified and misclassified was noted in the confusion matrix. The confusion matrix is a table that reports the number of false positives, false negatives, true positives, and true positives (see Table 5.1). From the final result of confusion matrix, the misclassification rate can be calculated following

$$\text{Misclassification rate} = \frac{\text{FP} + \text{FN}}{\text{Total number of observation}}, \tag{5.1}$$

where total number of observation = TN + FP + FN + TP.

From a brief explanation above, there are three main properties to be set within the validation process: the models (testing dataset), the testing dataset, and the classification rule.

The classification rule

Referring the theory of classification in Chapter 2.5, prior probability and cost of misclassification can be taken account in classification rule. In this study, we neglected the misclassification cost, thus $c(1|2) = c(2|1)$. However, we would investigate the

Table 5.1: Confusion matrix table of 2 sub-population, π_1 and π_2

		Predicted	
		π_1	π_2
Actual	π_1	True Negative (TN) : Number of observations correctly classified as π_1 that belong to π_1	False Positive (FP): Number of observations incorrectly classified as π_2 that belong to π_1
	π_2	False Negative (FN): Number of observations incorrectly classified as π_1 that belong to π_2	True Positive (TP): Number of observations correctly classified as π_2 that belong to π_2

effect of prior probability and compared the result with another result which neglected the effect of prior probability. Therefore, there were two classification rules applied in this study, they were rules from Eq. 2.38 and Eq. 2.37.

The models (training dataset)

The models selected in this study were originated from GR data of Well 15/5-7 A, Well 15/6-11 S, and Well 15/6-9 S. The models in each well were grouped based on the hole size to get in accordance to results from data exploration. Moreover, we removed groups which only contain one type of lithology from the models to avoid any error since the purpose of this process was lithology classification.

According to the selected classification rules from the explanation above, the models were constructed into two different types:

1. **Model 1, model without prior probability.** This model was constructed by assuming that prior probabilities between two lithology groups were equal (p(1) = p(2) = 0.5). This model would agree with the classification rule from Eq. 2.38.

2. **Model 2, model with prior probability.** This model was constructed by taking account the effect of prior probabilities. The prior probability values were determined by calculating the number of observations of shale and non-shale lithology from the geological description of testing dataset. The number of observations then standardized into 1 in order to fulfill the condition $p(1)+p(2) = 1$. This model complied the classification rule from Eq. 2.37.

The testing dataset

In this study, the testing data were not randomly selected, instead the data were selected on basis of depth span. This adjustment was based on the fact that rock in formation is formed as a bed or layer, thus term of rock lithology related to group of rock in particular depth range.

Based on the source of testing data, there were 3 different types of experiment which performed in this study:

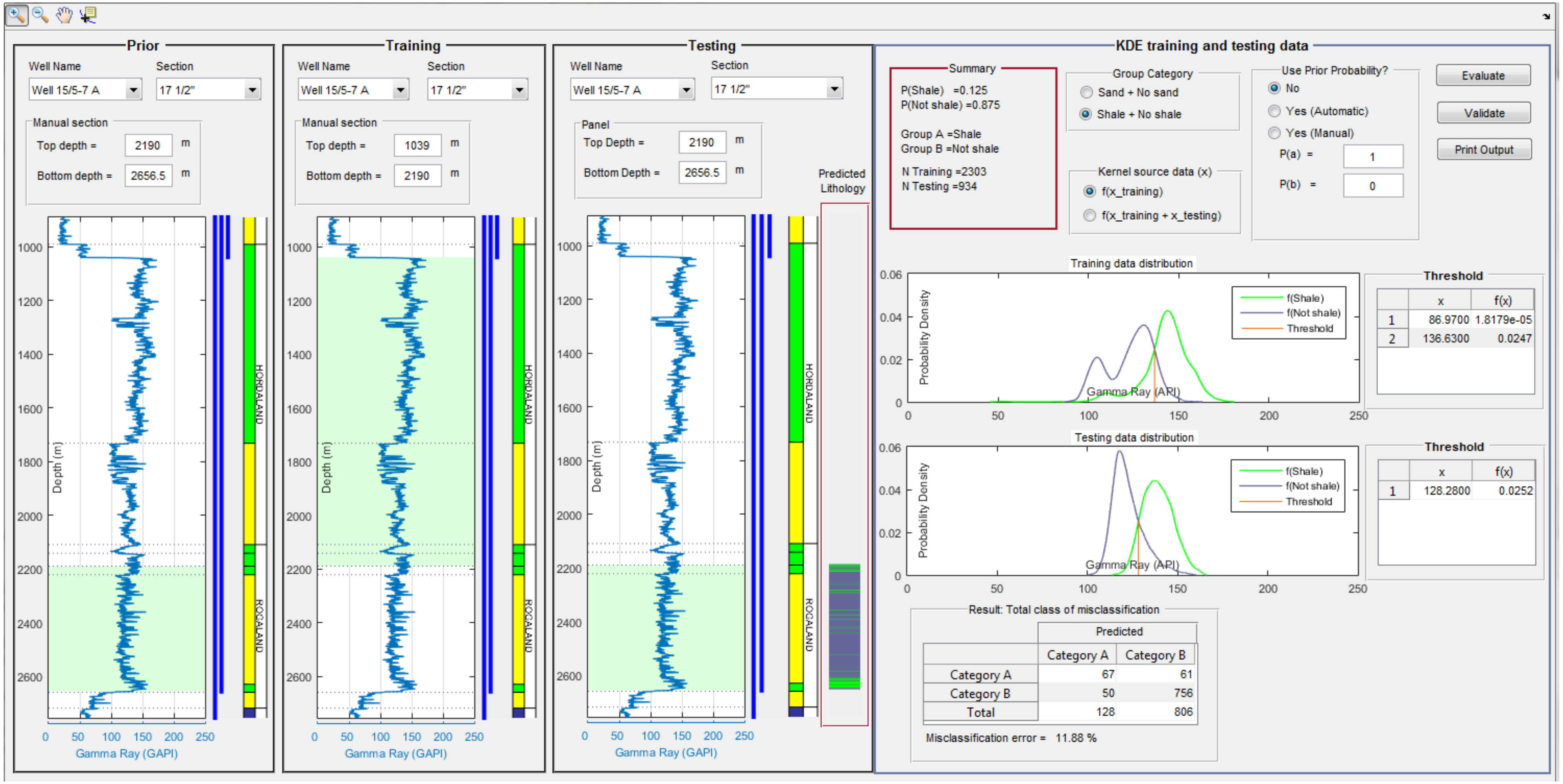
1. **Experiment 1, a validation within groups in the same well.** The source of testing data for this experiment was generated from the same source of training data.

The ratio of training and testing dataset was set to ± 70 % /30 %, respectively. The section which selected for training and testing data was not randomly selected, but the selection was based on the consideration if we apply this experiment in the field. In the practice, we would have a training dataset from formations that have been drilled and with known lithology. Meanwhile, the testing dataset was taken from the formations that recently been drilled with unknown lithology. According to this, the testing dataset would always be taken from a section in a greater depth than the depth of the section for testing data. Hence, the GR data from the upper depth region was set as training data and the lower depth region was set as testing data.

2. **Experiment 2, a validation with neighboring wells within Block 15**. The testing data would be picked from neighboring wells in Block 15 and tested according to the hole size. One additional well from Block 15, Well 15/6-12, was used in the validation process for group of 12 $\frac{1}{4}$" in Well 15/6-11 S.

3. **Experiment 3, a validation with wells from different block**, which is Block 16. The wells used in this experiment were Well 16/1-14, Well 16/2-7, and Well 16/2-13 A. Similar to experiment 2, the testing data would be tested according to the hole size.

The validation process was performed by using a program which constructed by using MATLAB. In addition to the program, we also built an interface using MATLAB GUI toolbox to improve the speed of analysis and provide visualizations of the results. This interface provides several functionalities, such as preview of probability plots for training and testing data, misclassification error calculation, lithology prediction plot, intersection points table, and confusion table. Moreover, the interface allows the user to select preferred depth of training and testing data, and prior probability data source.



Prior

Well Name: Well 15/5-7 A — Section: 17 1/2"

Manual section
Top depth = 2190 m
Bottom depth = 2656.5 m

Training

Well Name: Well 15/5-7 A — Section: 17 1/2"

Manual section
Top depth = 1039 m
Bottom depth = 2190 m

Testing

Well Name: Well 15/5-7 A — Section: 17 1/2"

Panel
Top Depth = 2190 m
Bottom Depth = 2656.5 m

KDE training and testing data

Summary
P(Shale) =0.125
P(Not shale) =0.875

Group A =Shale
Group B =Not shale

N Training =2303
N Testing =934

Group Category
○ Sand + No sand
● Shale + No shale

Kernel source data (x)
● f(x_training)
○ f(x_training + x_testing)

Use Prior Probability?
● No
○ Yes (Automatic)
○ Yes (Manual)
P(a) = 1
P(b) = 0

Threshold (training)

	x	f(x)
1	86.9700	1.8179e-05
2	136.6300	0.0247

Threshold (testing)

	x	f(x)
1	128.2800	0.0252

Result: Total class of misclassification

	Predicted	
	Category A	Category B
Category A	67	61
Category B	50	756
Total	128	806

Misclassification error = 11.88 %

Figure 5.7: Preview of the interface for validation process built by using MATLAB

5.2.1 Results

Experiment 1

1. Well 15/5-7 A

Table 5.2: Summary of data and results for validation within each hole section in Well 15/5-7 A

Hole size	**Training data**		**Testing data**		**Prior probability**		**Misclassification Rate (%)**	
	Depth (m)	**N**	**Depth (m)**	**N**	**Shale**	**Not-shale**	**Model 1**	**Model 2**
26"	194 - 850	1313	850 - 1038.5	378	0.254	0.746	2.12	11.38
17 $\frac{1}{2}$"	1039 - 2180	2283	2180 - 2656.5	954	0.144	0.856	13	12.58
8 $\frac{1}{2}$"	2657 - 3800	2287	3800 - 4119	639	0.541	0.459	11.27	9.86

2. Well 15/6-11 S

Table 5.3: Summary of data and results for validation within each hole section in Well 15/6-11 S

Hole size	**Training data**		**Testing data**		**Prior probability**		**Misclassification Rate (%)**	
	Depth (m)	**N**	**Depth (m)**	**N**	**Shale**	**Not-shale**	**Model 1**	**Model 2**
17 $\frac{1}{2}$"	690 - 1730	2081	1730 - 2181	903	0.158	0.842	12.62	15.84
12 $\frac{1}{4}$"	2181.5 - 3320	2278	3320 - 3816.5	994	0.447	0.553	13.88	12.07

3. Well 15/6-9 S

Table 5.4: Summary of data and results for validation within each hole section in Well 15/6-9 S

Hole size	**Training data**		**Testing data**		**Prior probability**		**Misclassification Rate (%)**	
	Depth (m)	**N**	**Depth (m)**	**N**	**Shale**	**Not-shale**	**Model 1**	**Model 2**
17 $\frac{1}{2}$"	753 - 2180	2855	2180 - 2785.5	1212	0.368	0.632	43.48	38.28
8 $\frac{1}{2}$"	2786 - 3590	1609	3590 - 3942	705	0.684	0.352	23.69	25.39

Experiment 2

1. Well 15/5-7 A as training data

Table 5.5: Summary of data and results for validation of Well 15/6-11 S and Well 15/6-9 S , by using Well 15/5-7 A as training data

Hole size	**Training data: Well 15/5-7 A**		**Testing Data**			**Prior probability**		**Misclassification Rate (%)**	
	Depth (m)	**N**	**Well**	**Depth (m)**	**N**	**Shale**	**Not-shale**	**Model 1**	**Model 2**
26"	194 - 1038.5	1690	Well 15/6-11 S	187 - 689.5	1006	1	0	2.19	0
17 $\frac{1}{2}$"	1039 - 2656.5	3236	Well 15/6-11 S	690 - 2181	2983	0.334	0.666	34.43	34.56
			Well 15/6-9 S	753 - 2785.5	4066	0.607	0.393	52.26	48.55
8 $\frac{1}{2}$"	2657 - 4119	2925	Well 15/6-11 S	3817 - 4043	453	0.768	0.232	33.77	22.3
			Well 15/6-9 S	2786 - 3942	2313	0.276	0.724	8.13	11.59

2. Well 15/6-11 S as training data

Table 5.6: Summary of data and results for validation of Well 15/5-7 A and Well 15/6-9 S , by using Well 15/6-11 S as training data

Hole size	**Training data: Well 15/6-11 S**		**Testing Data**			**Prior probability**		**Misclassification Rate (%)**	
	Depth (m)	**N**	**Well**	**Depth (m)**	**N**	**Shale**	**Not-shale**	**Model 1**	**Model 2**
26"	187 - 689.5	1006	Well 15/5-7 A	194 - 1038.5	1690	0.763	0.237	23.67	23.67
17 $\frac{1}{2}$"	690 - 2181	2983	Well 15/5-7 A	1039 - 2656.5	3236	0.514	0.486	35.97	36.53
			Well 15/6-9 S	753 - 2785.5	4066	0.607	0.393	15.1	15.15
12 $\frac{1}{4}$"	2181.5 - 3816.5	3271	Well 15/6-12	2754 - 3628.5	1750	0.01	0.99	2.17	1.14
8 $\frac{1}{2}$"	3817 - 4043	453	Well 15/5-7 A	2657 - 4119	2925	0.17	0.83	12.27	6.39
			Well 15/6-9 S	2786 - 3942	2313	0.276	0.724	7.57	8

3. Well 15/6-9 S as training data

Table 5.7: Summary of data and results for validation of Well 15/5-7 A and Well 15/6-11 S , by using Well 15/6-9 S as training data

Hole size	Training data: Well 15/6-9 S		Testing Data			Prior probability		Misclassification Rate (%)	
	Depth (m)	N	Well	Depth (m)	N	Shale	Not-shale	Model 1	Model 2
17 $\frac{1}{2}$"	753 - 2785.5	4066	Well 15/5-7 A	1039 - 2656.5	3236	0.514	0.486	35.51	35.97
			Well 15/6-11 S	690 - 2181	2983	0.334	0.666	29	28.03
8 $\frac{1}{2}$"	2786 - 3942	2313	Well 15/5-7 A	2657 - 4119	2925	0.17	0.83	26.39	13.64
			Well 15/6-11 S	3817 - 4043	453	0.768	0.232	18.98	14.79

Experiment 3

1. Well 15/5-7 A as training data

Table 5.8: Summary of data and results for validation of wells in Block 16 by using Well 15/5-7 A as training data

Hole size	**Training data: Well 15/6-9 S**		**Testing Data**			**Prior probability**		**Misclassification Rate (%)**	
	Depth (m)	**N**	**Well**	**Depth (m)**	**N**	**Shale**	**Not-shale**	**Model 1**	**Model 2**
$17\frac{1}{2}$"	1039 - 2656.5	3236	Well 16/1-14	371 - 1477.5	2214	0.642	0.358	41.00	38.44
			Well 16/2-7	700 - 1771.5	2144	0.628	0.372	62.78	62.64
$8\frac{1}{2}$"	2657 - 4119	2925	Well 16/1-14	2228 - 2548	641	0.716	0.284	36.97	26.21
			Well 16/2-7	2098 - 2498	801	0.648	0.352	11.74	12.36
			Well 16/2-13 A	2487 - 2772	571	0.595	0.405	30.65	30.30

2. Well 15/6-11 S as training data

Table 5.9: Summary of data and results for validation of wells in Block 16 by using Well 15/6-11 S as training data

Hole size	**Training data: Well 15/6-9 S**		**Testing Data**			**Prior probability**		**Misclassification Rate (%)**	
	Depth (m)	**N**	**Well**	**Depth (m)**	**N**	**Shale**	**Not-shale**	**Model 1**	**Model 2**
$17\frac{1}{2}$"	690 - 2181	2983	Well 16/1-14	371 - 1477.5	2214	0.642	0.358	29.49	28.27
			Well 16/2-7	700 - 1771.5	2144	0.628	0.372	50.98	45.24
$12\frac{1}{4}$"	2181.5 - 3816.5	3271	Well 16/1-14	1478 - 2227.5	1500	0.703	0.297	12.87	14.00
			Well 16/2-7	1772 - 2097.5	652	0.189	0.811	55.83	47.55
			Well 16/2-13 A	717 - 2486.5	3540	0.674	0.326	25.82	25.51
$8\frac{1}{2}$"	3817 - 4043	453	Well 16/1-14	2228 - 2548	641	0.716	0.284	19.34	22.93
			Well 16/2-7	2098 - 2498	801	0.648	0.352	19.35	24.84
			Well 16/2-13 A	2487 - 2772	571	0.595	0.405	22.42	29.95

3. Well 15/6-9 S as training data

Table 5.10: Summary of data and results for validation of wells in Block 16 by using Well 15/6-9 S as training data

Hole size	**Training data: Well 15/6-9 S**		**Testing Data**			**Prior probability**		**Misclassification Rate (%)**	
	Depth (m)	**N**	**Well**	**Depth (m)**	**N**	**Shale**	**Not-shale**	**Model 1**	**Model 2**
$17\frac{1}{2}$"	753 - 2785.5	4066	Well 16/1-14	371 - 1477.5	2214	0.642	0.358	49.10	48.24
			Well 16/2-7	700 - 1771.5	2144	0.628	0.372	80.88	77.66
$8\frac{1}{2}$"	2786 - 3942	2313	Well 16/1-14	2228 - 2548	641	0.716	0.284	20.12	21.53
			Well 16/2-7	2098 - 2498	801	0.648	0.352	23.22	23.10
			Well 16/2-13 A	2487 - 2772	571	0.595	0.405	26.80	28.72

5.2.2 Discussions

There are 3 main points discussed in this section, they are the results of misclassification rate, the effect of prior probability, and the explanation of how the validation method for lithology prediction can be applied in the field. Each point is discussed in the separated section.

Misclassification error from the experiments

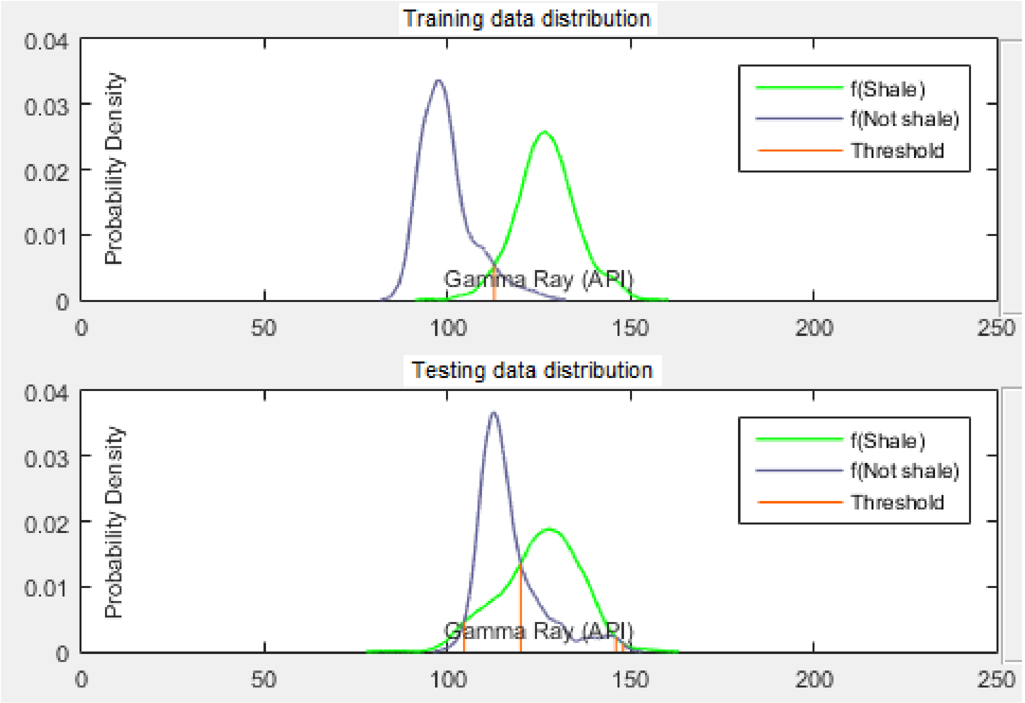
The misclassification rate in each experiment is summarized in Table 5.11. In general, the table showed that the misclassification rate increased significantly if the well sources between the models and the testing data were different. Testing the models by using datasets from the neighboring wells in experiment 2 increased the averaged misclassification rate up to 7 % for model 1 and 3 % for model 2 compared to the rate in experiment 1. Meanwhile, testing the models by using datasets from wells in different block increased the averaged misclassification rate up to 17% for model 1 and 15% for model 2 compared to the results in experiment 1.

Table 5.11: Table of minimum, maximum, and averaged value of misclassification error in each model and experiment

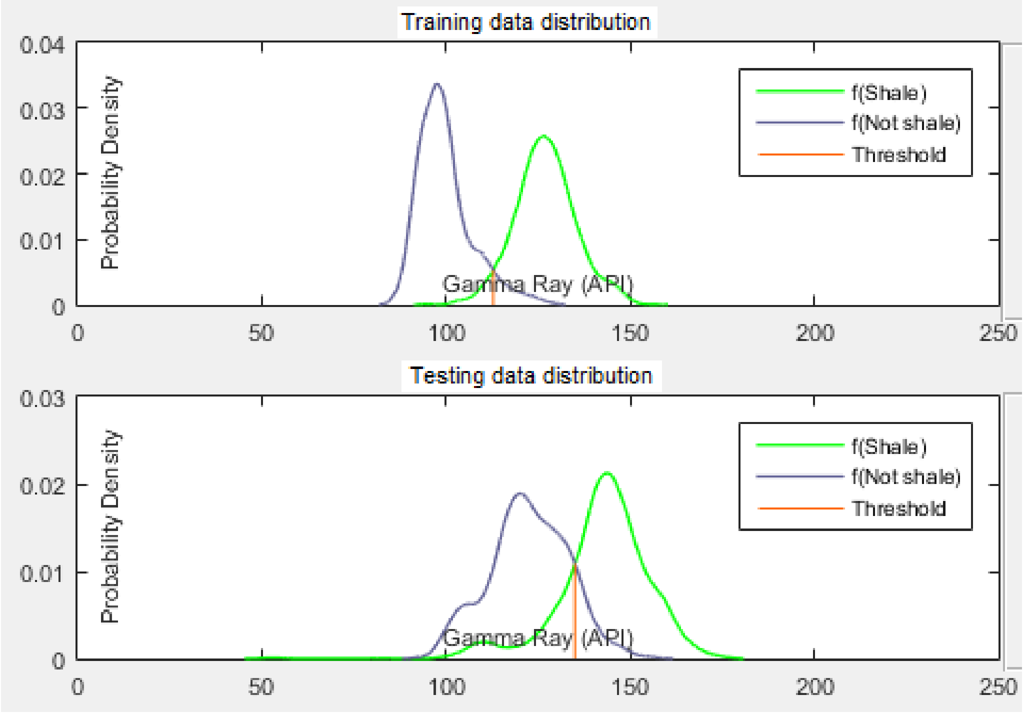
Experiment	Minimum error(%)		Maximum error(%)		Averaged error(%)	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model2
Experiment 1	2.12	9.86	43.38	38.28	18.54	19.02
Experiment 2	7.57	6.39	52.26	48.56	24.74	21.93
Experiment 3	11.74	12.36	80.88	77.66	35.39	34.54

The increasing rates of misclassification in experiment 2 and 3 were due to the difference of GR data distribution between the model and the testing data. Several examples of experiment 1, 2, and 3 were taken from model of $17\frac{1}{2}$" in Well 15/6-9 S and the distribution between the model and testing dataset are presented in Fig. 5.8. A contrast between the distribution of each dataset was improved from experiment 1 to experiment 3. If we observed the peak of the probability density of each lithology, it could be seen that the GR value from testing data in experiment 2 were shifted away from the GR value of the model and created a gap of GR value. A greater gap of testing data distribution was indicated in experiment 3.

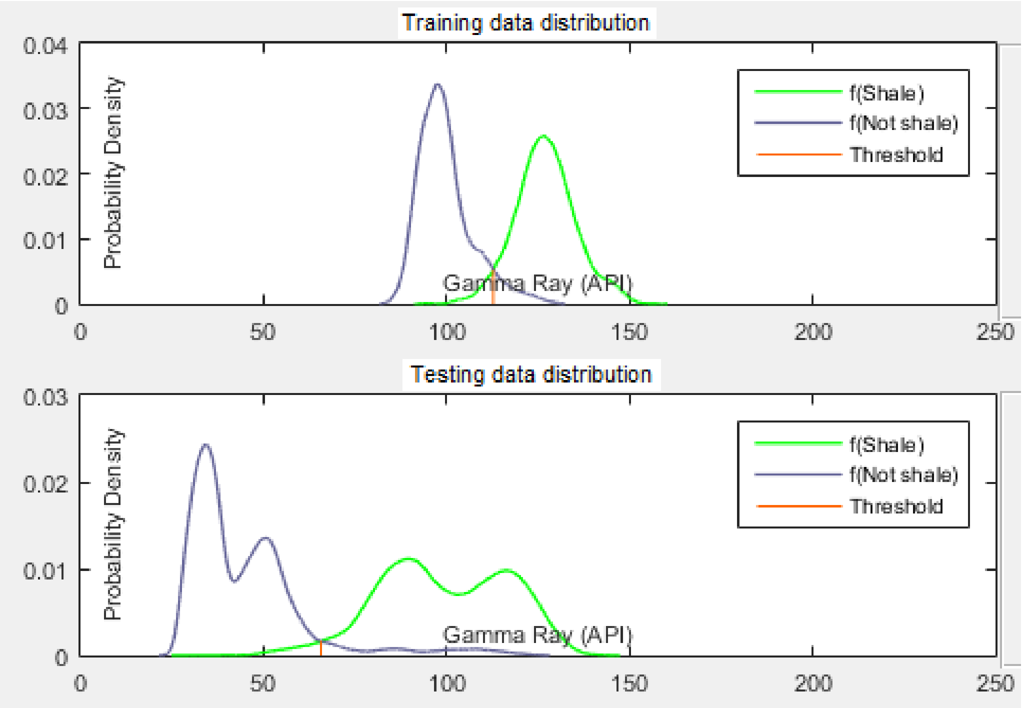
Summing up the observations above, we believed that the increased of misclassification rate was related to the uncertainty factors on GR data. And we were certain that the further away the location of wells used as the testing data source, the uncertainty was increased. We identified some of the possible source of uncertainty, they were geological factors, borehole environment, and GR tool. The uncertainty from geological factors includes the mineral content, lithology mixture, and depositional environments of the formations. However, this geological factor was believed to have a small contribution to the classification in this study because the wells tested in experiment 2 and 3 located in the same sub-basin, Viking Graben, and GR values of one formation are less likely to change in vertical direction.



(a) Experiment 1, with testing data from Well 15/6-9 S(2180 - 2785.5 m)



(b) Experiment 2, with testing data from Well 15/5-7 A



(c) Experiment 3, with testing data from Well 16/1-14in Block 16

Figure 5.8: Training and testing data distributions from different experiments for group of $17\frac{1}{2}$". The testing data was from Well 15/6-9 S

Borehole environment and GR tools were considered as the most contributing factors to this misclassification. We dismissed the effect of hole size because the models were tested according to the hole size. However, factors such as drilling fluid, the tool position, hole cavity, cement type, and the tool size could contribute as uncertainty factors, notably with the lack of correction from the source and data limitation.

All of these uncertainties rise through the imperfect knowledge. However, the effect from these uncertainties could be minimized if we integrate data from other source. One of the example is the usage of data from geological description in addition to GR data for lithology classification. This example was applied in this study and the application was presented by adding prior probability into the classification rule. The detailed discussion of the results from this application is shown in the next section.

Effect of prior probability on misclassification error

Within this discussion point, we would discuss the result of validation from model 1 and 2, the way prior probability working in reducing misclassification rate, and conditions that lead model 2 fails to reduce the misclassification rate. Referring to Table 5.11, model 2 from experiment 2 and 3 produced less misclassification rate, with difference around 3% and 1% respectively, compared to model 1. Meanwhile, model 2 from experiment 1 produced error 0.5% higher compared to model 1.

Based on the number of cases, 4 out of 7 total cases (4/7) in experiment 1 showed that model 2 yielded less misclassification rate compared to model 1. A similar event also occurs in several cases from experiment 2 and 3, with the number of cases 7/13 and 11/18 respectively. This observation discovered that more than half of the cases in each experiment proved that model 2 produced less misclassification rate.

The unique values of prior probability in model 2 indicate the true distribution of lithologies of the training data. This was unmistakable because the prior probability values were generated based on the number of observation of each lithology from the geological description of training data. This was different from model 1 which assumed that the lithologies have an equal distribution.

So, in what way that the true distribution affects the classification? By recalling and modifying the equation from Eq. 2.37, we will get following equation

$$\begin{aligned} R_1 &: \frac{f_1(x)}{f_2(x)} \frac{p(1)}{p(2)} \quad \geq 1 \\ R_2 &: \frac{f_1(x)}{f_2(x)} \frac{p(1)}{p(2)} \quad < 1 \end{aligned} \tag{5.2}$$

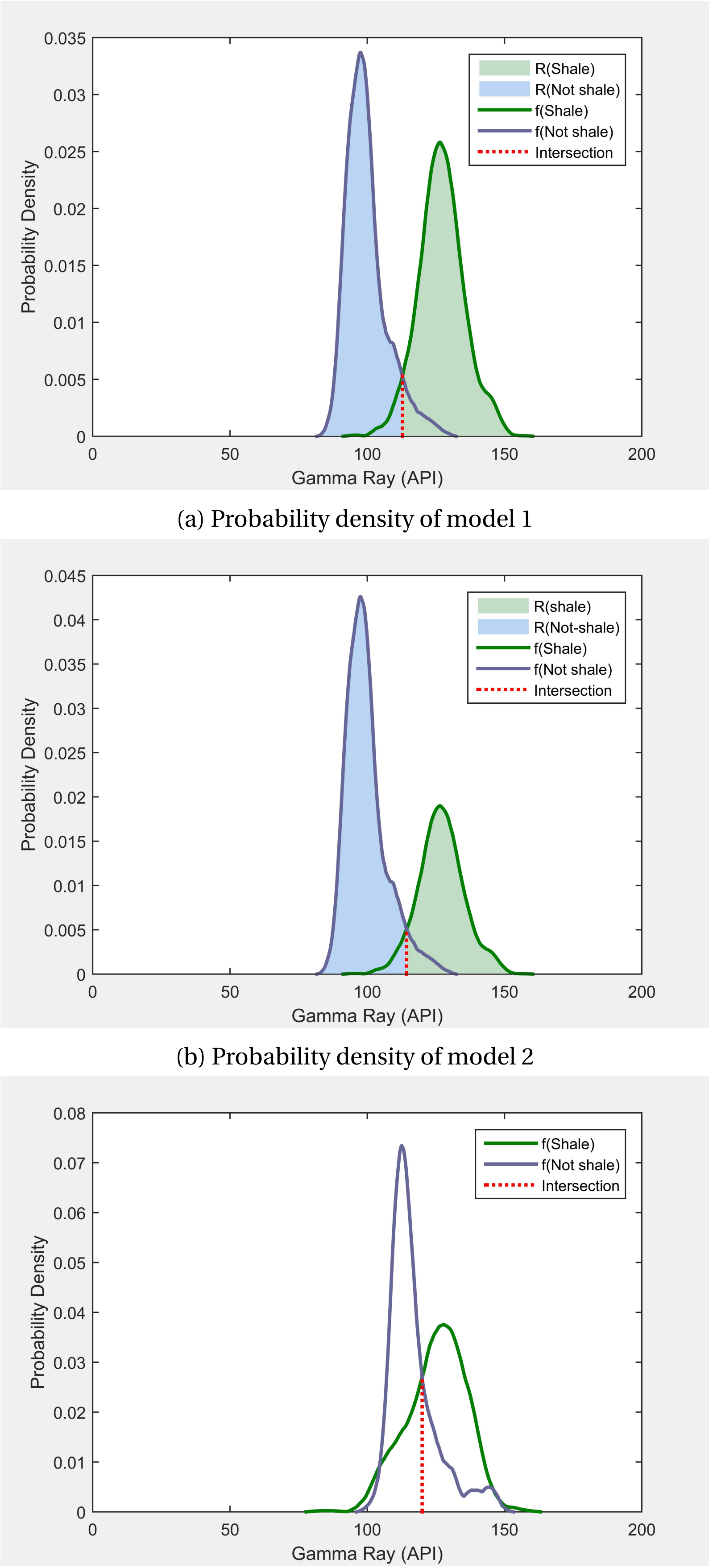
According to the equation above, the true distribution from the prior probability will affect the value of probability density of the models for each lithology group. And as the further effect, the data distribution of the model for each lithology would also change. In order to get a better understanding of this explanation, we took an example of experiment 1 from group of $17\frac{1}{2}$" in Well 15/6-9 S.

In model 1, the values of prior probability of shale and not shale lithology were equal, p(shale) = p(not-shale) = 0.5. By multiplying these values to the probability density, the model would produce probability density plot shown in Fig. 5.9a. Referring to the plot and the classification rule, all GR values which fulfill the condition of $\frac{f_1(x)}{f_2(x)} \frac{p(1)}{p(2)} \geq 1$ fall within the green area which indicates shale region, R(shale). Meanwhile, all GR values which fulfill the condition of $\frac{f_1(x)}{f_2(x)} \frac{p(1)}{p(2)} < 1$ fall within the blue area which indicates not-shale region, R(not-shale).

In model 2, the values of prior probabilities were calculated and the results were p(shale)=0.368 and p(not-shale)=0.632. By multiplying these values into the probability density of each lithology, we generated model 2 which shown in Fig. 5.9b. Because the value of p(shale) is smaller compared to the value of p(not-shale), the area of R(shale) is also smaller than the area of R(not-shale). In addition, the distribution of probability density of model 2, f(shale) and f(not-shale), resembled the distribution of training data (Fig. 5.9c) much closer compared to model 1.

By observing the plots from both model 1 and 2, the red dashed line which is formed at the intersection of f(shale) and f(not-shale) separated R(shale) and R(not-shale) very clearly. According to this, we discovered that the x-value of this intersection point acted as a discriminator between these two lithology groups. Meanwhile, the y-value of this intersection point influenced the size of overlapped region.

In order to understand in what extent these values can reduce the misclassification rate, we summarized the x- and y- values of the intersection points from the example in Table 5.12. The table shows that model 2 approached the x-values of testing data much closer than model 1. Meanwhile, the less y-value was given from model 2. From this observation, we believed that the more identical the x-values between the model and the training data, less misclassification rate would be yielded, especially by knowing that the x-value behave as discriminator between the lithology groups. Meanwhile, the smaller the y-value, the smaller the overlapped region of the model thus the classification from the model would be more precise. Even though the difference of these values between model 1 and 2 was rather small, the misclassification error reduced by 5%.



(a) Probability density of model 1

(b) Probability density of model 2

(c) Probability density of testing data

Figure 5.9: Plot of probability density for case in experiment 1: $17\frac{1}{2}$" in Well 15/6-9 S

Table 5.12: The values of the intersection points of the models and the testing data from experiment 1, group of $17\frac{1}{2}$" in Well 15/6-9 S

Data	**Intersection points**	
	x-values (API)	**y-values**
Model 1	112.8	0.0053
Model 2	114.3	0.0051
Testing data	120.01	0.0135

The classification by using model 2 did not always produce less misclassification rate but instead, model 1 produced a better result. We sensed that the main reason of this event was the difference of the shape of data distribution between the model and the testing. This reason indeed is related to the reason causing high misclassification rate from the previous discussion point. However, the impact of this reason for this case was somewhat different. We observed that in some cases adding prior probability increased the contrast of x-value between the model and the testing data which led to misclassification rate increment. Two examples showing this event are shown in Fig. 5.10 and Fig. 5.11.

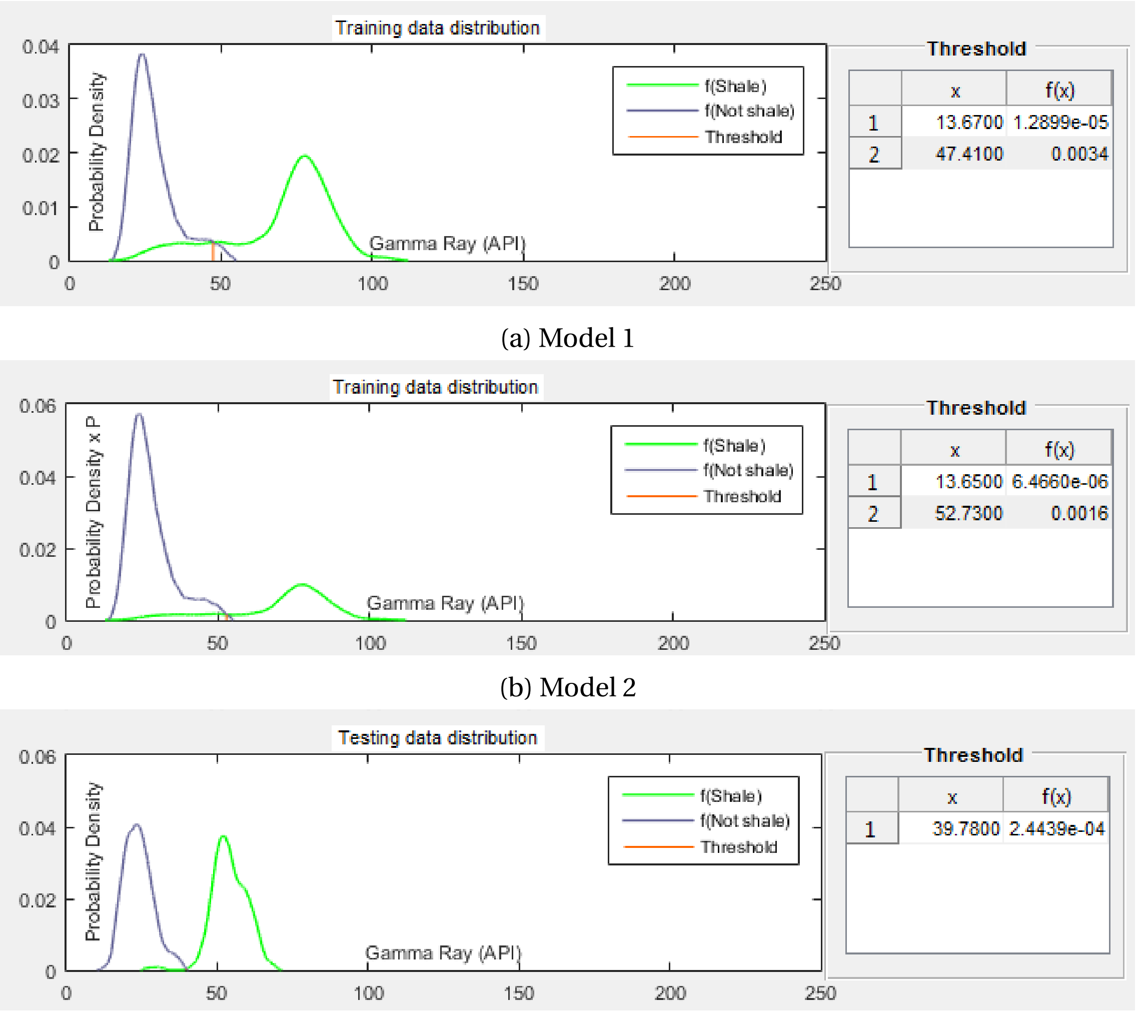
Application of validation process in lithology prediction

The discussions of model validation above give insights of how GR data can be processed for lithology classification. Within this discussion point, we also presented the best way to apply the method in the practice.

The experiments results discovered that the classification by using training and testing data which originated from the same source was the most optimum in minimizing the misclassification rate. According to this, we suggested that the training data for this application will be generated from the drilled section which lithology type of this section is acknowledged and confirmed by other interpretation methods.

However, this may come with limitation during drilling formation at the beginning of a section because no data from drilled formation is available. Therefore, we recommended the usage of data from the neighboring well as the training data during drilling the top of a section. And in order to reduce the errors from uncertainty factors, it is required to select the neighboring well which fairly identic to the current drilled well. Once the information from the drilled formation is considered adequate, the training data can be switch and use data from the drilled formation.

Meanwhile, the testing data will be taken from the section that just been drilled. The usage of prior probability was also recommended to improve the results and the value of prior probability can be calculated from the prediction of geological description of the current well. Once the data is classified toward the training data, the classification results can be verified with the results from other interpretation method to check any misclassification from the proposed method. In our vision, we sensed that the application of this classification method in the field could improve the lithology interpretation and optimize the drilling operation.



Threshold

	x	f(x)
1	13.6700	1.2899e-05
2	47.4100	0.0034

(a) Model 1

Threshold

	x	f(x)
1	13.6500	6.4660e-06
2	52.7300	0.0016

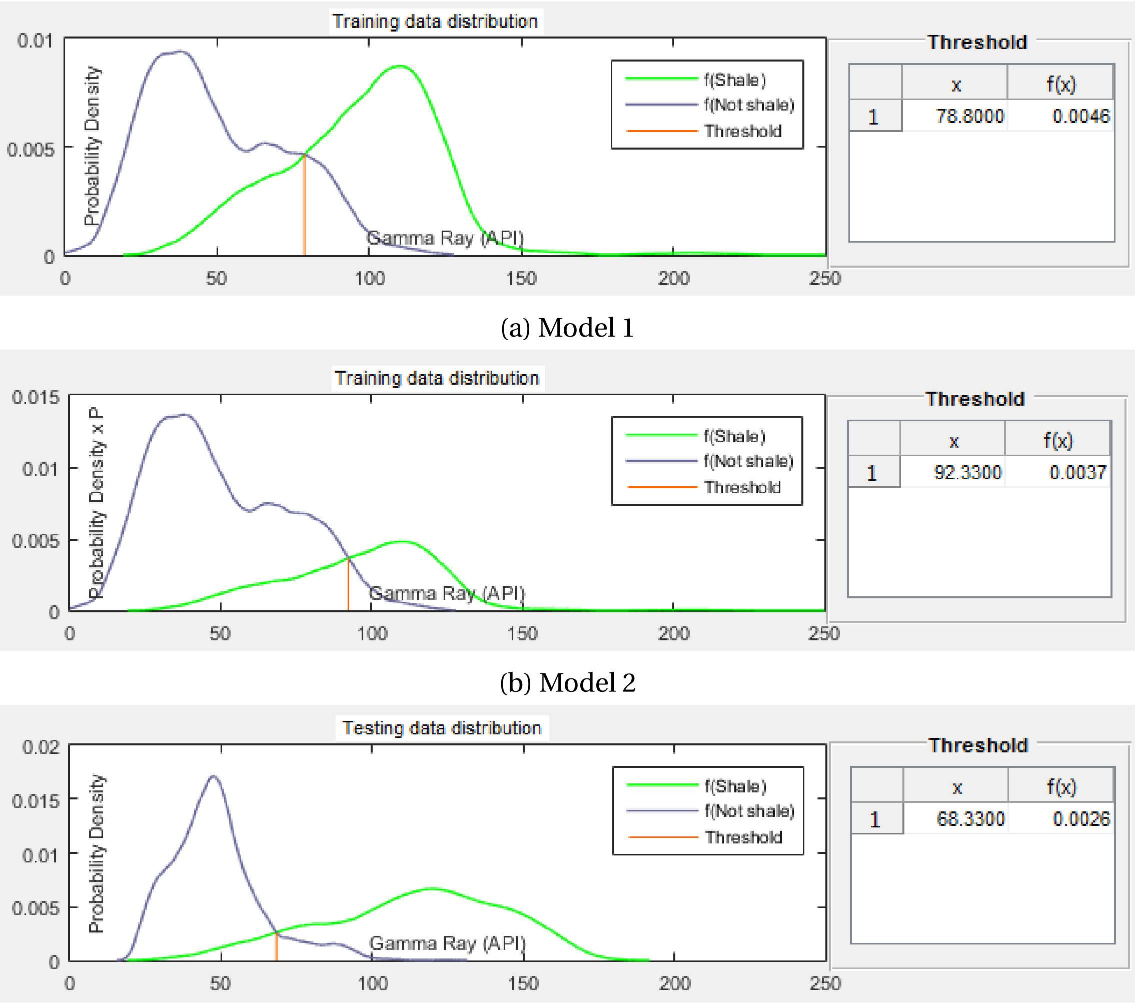
(b) Model 2

Threshold

	x	f(x)
1	39.7800	2.4439e-04

(c) Testing data

Figure 5.10: The f(shale) of the models is more skewed to the left compared to the f(shale) of the testing data. Multiplying the probability density with prior probabilities in model 2 shifted the x-value of intersection point to right side (52.73 GR API), further away from the x-value in testing data (29.78 GR API)



Threshold

	x	f(x)
1	78.8000	0.0046

(a) Model 1

Threshold

	x	f(x)
1	92.3300	0.0037

(b) Model 2

Threshold

	x	f(x)
1	68.3300	0.0026

(c) Testing data

Figure 5.11: The f(not-shale) in training data has bimodal distribution and f(shale) has right-skewed distribution. Meanwhile, f(not-shale) of testing data has right-skewed distribution and f(shale) is less skewed than f(shale) in training data. Adding prioir probability shifted the intersection point to 92.3 GR API, further away from the intersection in the testing data (68.3 GR API)

5.3 Preliminary study of bivariate analysis

Bivariate analysis is a statistical analysis which involves two different variables. It is usually performed to improve the univariate analysis and determine empirical relationship between the two variables. In this study, we carried out this bivariate analysis as a preliminary exploration which still requires improvement. The motivation of this analysis was to improve the lithology classification by adding another variable of well logging, which was neutron log. In the previous chapter, the experiments results showed that univariate analysis of GR was insufficient to classify lithology as indicated with the overlapping distribution of the lithology groups. We selected neutron log as the second variable due to the capability of neutron log in distinguishing lithology, as explained in Chapter 2.1.2.

The analysis was performed by plotting GR and neutron data together in a scatter plot, with GR as the x-axis and neutron as the y-axis. Adjacent to the scatter plot, there are 2 plots of probability density which estimated by kernel estimator. The GR probability density plot is located at the bottom side of scatter plot, while the neutron probability density plot is located at the left side of the scatter plot. The data points in scatter plot are marked according to its lithology type. The data for analysis was generated from wells in Block 16, Well 16/1-14, Well 16/2-7, and Well 16/2-13 A. The results of the plots for these wells are shown in Appendix E.

Within this section, we would discuss the results of plots from two different groups: $8\frac{1}{2}$" from Well 16/1-14 and $17\frac{1}{2}$" from Well 16/2-7 (Fig. 5.12 and Fig. 5.13). Both figures show that the overlapping distribution of lithologies indicated from one variable could be distinguished by another variable. In other words, the GR and neutron variables complemented each other. This is because GR and neutron tool have different principal of measurement, thus each tool has different sensitivity in the particular rock types. The results also showed better separation of lithologies and we expected that the misclassification rate could be reduced.

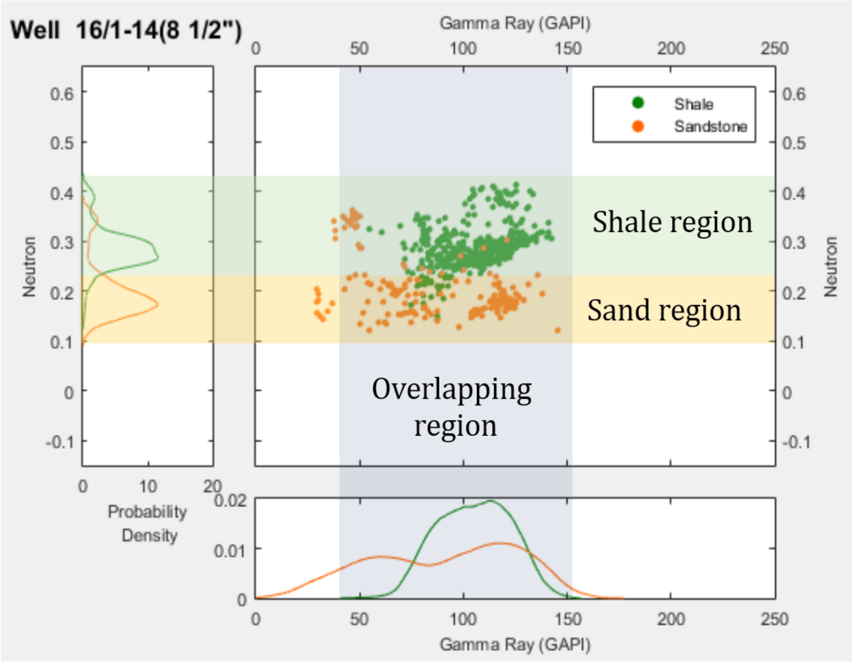


Figure 5.12: The overlapped distribution of shale and sandstone lithology from GR could be separated by neutron data. Shale is dominated with high neutron value, while sandstone has small neutron value.

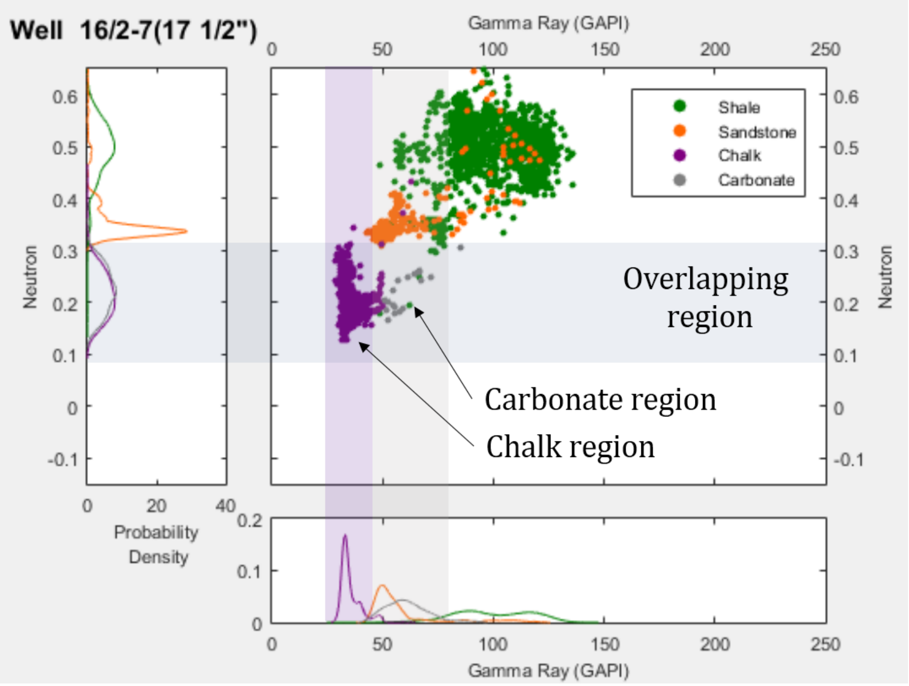


Figure 5.13: The overlapped region of chalk and carbonate lithologies from neutron could be separated by using GR. Chalk has smaller GR value than to carbonate.

5.4 Concluding remark

- The kernel density estimator with Epanechnikov kernel function has been applied to approximate the probability density of GR data. It was shown that the kernel estimator was able to return the non-parametric distribution of GR data. Moreover, the probability density from kernel estimator were considered better over the probability density from histogram.

- The validation was applied to assess the GR data in lithology classification. From the investigation, the quality of classification depends on the distribution of training and testing data.

- The prior probability values which extracted from geological description were effective in reducing the misclassification rate.

- The classification method from this study can be applied for lithology interpretation in the practice. The application was considered useful to reduce time of lithology interpretation during drilling operation.

- The preliminary study of bivariate analysis proved that by using two variables of GR and neutron, the lithology classification was improved. The most significant result was indicated from the separation of overlapped distribution of one variable.

Chapter 6

Conclusions

In this study, a quantitative analysis of GR data for lithology classification and prediction has been performed. Prior to the main analysis of this study, data exploration and kernel application on GR data were carried out. The results from these processes were exploited for the main analysis which was validation as a part of an assessment of variable GR for lithology classification.

The data exploration was executed by visualization of statistical graphs: histogram and boxplot, and analysis of statistical descriptive. At the beginning of the process, GR data were grouped based on the lithology type. As the results were investigated, the discovery showed that GR data contain sub-groups which originated from variable of hole size. This was indicated from the high variance of GR within the lithology groups and the contrast of GR values between different hole sizes during visualization GR in log trace. We sensed that the sub-groups were emerged due to uncorrected GR data.

In order to support the discoveries, we tested the hypotheses from data exploration by using Kruskal-Wallis H test. The results of p-value and mean rank value from these tests agreed with the discoveries from data exploration. The results also proved that GR data depended on lithology and hole size variables. The correction on GR data was not possible due to inadequate informations from the data source, thus the GR data for further analysis remained to be grouped based on the hole size.

The kernel density estimator was applied in this study to approximate the probability density of GR data which further would be employed for lithology classification. The kernel estimator was chosen because of the capability of kernel estimator to return a continuous probability density and approach the non-parametric distribution of GR data. The type of kernel function applied in this study is Epanechnikov function. The results of probability density from kernel application were preferred over histogram due to the discreteness of histogram. The results of the probability plots also showed that the lithology group within GR data was more convenient to be grouped into shale and not-shale lithology. This was due to sandstone, chalk, and carbonate lithologies had similar GR value range which was smaller than the shale lithology. In addition to that, error from misclassifying of these 3 lithologies could be prevented.

The lithology classification and validation were performed by using 2 different models in 3 different experiments. The results showed that the misclassification rate was reduced significantly in experiment 1 which model and testing data were taken from the same data source. Meanwhile, model 2 which added prior probability value within the classification rule had less misclassification rate compared to model 1 which neglected the prior probability value.

Summing up the observations of the main analysis, the main cause of lithology misclassification was due to the difference of data distribution between the model and the testing data. The most significant difference in data distribution was indicated in experiments which testing data was taken from different wells and/or different block. We sensed that the difference in data distribution was caused by uncertainty factors. However, the effect of uncertainty factors could be reduced or even dismissed if the data quality could be improved.

Considering that data quality was beyond the scope of this study, we suggested to perform the experiment by using a corrected data logging. We also suggested to carry out a similar study which lithology was classified by using the relative value of GR log, thus the necessity of correction could be neglected and GR can be analyzed without grouping the data based on the hole size.

The application of the method proposed in this study has been explained. The application would give opportunities to predict and classify the lithology from GR reading. Since the prediction was still limited and only processed the reading from the tool location inside the borehole, a further work to improve the prediction beyond the tool was recommended. In addition, the prediction could be improved by using the posterior probability calculation from Bayes formula. The prediction will be more precise because this formula calculates the conditional probability which are useful for decision making (Ott and Longnecker, 2010).

The results from bivariate analysis proved that the lithology classification could be improved by using 2 variables: GR and neutron logging. We recommended to continue the study of bivariate analysis through implementation of bivariate kernel estimation and validation. It was also possible to combine data from mud logging and well logging within the bivariate analysis and study the the most optimum variables combination. In addition to that, discriminate analysis was also suggested so that the classification could be enhanced.

Appendix A

Acronyms

MWD Measurement while Drilling

LWD Logging while Drilling

WOB Weight on Bit

GR Gamma Ray

ROP Rate of Penetration

GNT Gamma Ray/Neutron Tool

SNP Sidewall Neutron Porosity Tool

CNL Compesanted Neutron Log

EDA Exploratory Data Analysis

KDE Kernel Density Estimation

IQR Interquartile Range

MSE Mean Squared Error

MISE Integrated Mean Squared Error

AMISE Assymptotic MISE

ECM Expected Cost of Misclassification

NCS Norwegian Continental Shelf

TVD True Vertical Depth

Appendix B

Geological Description

The geological description is summarized in a table and consists information of formation characteristic, dominated lithology, and depositional environment. The geological description was provided in the iQx software.

Group	Formation	Lithology Description
Nordland	Utsira	The upper boundary of this formation contains claystones, while the formation itself mostly contains sandstone. The formation probably represents shallow marine shelf sandstones.
Hordaland	Skade	The formation is dominated with sandstone and the lower boundary is characterized with decreasing gamma-ray into claystones overlying Hordaland Group. The formation was deposited in open marine environment.
	Grid	The majority lithology comprised in this formation is sandstone. The upper boundary usually contains claystones from Hordaland Group.
Rogaland	Balder	Laminated shales is indicated at the upper part of formation and decreasing to lower boundary separating Sele formation, often with glauconitic overlying sediments. The formation was deposited in a deep marine setting.
	Sele	The upper boundary separating Balder formation has an improved reading in GR and the lower boundary has more sandy composition. The formation was deposited in a deep marine setting.
	Lista	In areas where Lista formation is overlain by Sele formation, there is no distinct changes indicated.

Rogaland	Heimdal	The formation has majority of sandstone lithology. the upper boundary is usually defined by mixture of shale from Lista formation. The formation was deposited as submarine fans derived from sand accumulations on the shallow shelf. The turbidity currents formed shale layers partly in Heimdal formation.
	Våle	This formation is dominated with shale lithology and was deposited in a marine environment.
	Ty	Ty formation is dominated with sandstone. The lower boundary is indicated with decreasing gamma ray into Shetland Group. The depositional environment of this formation is deep marine fan system.
Shetland	Ekofisk	This formation is dominated with chalk lithology. The depositional environment is open marine with deposition of calcareous debris flow and turbidities.
	Tor	The formation has layer thin in the Norwegian sector and consists of chalk lithology. The depositional environment is similar with Ekofisk formation.
	Hod	The boundary between Hod and Blodøks formation is often indicated with change in gamma ray readings which increase toward Blodøks formation. The depositional environment is open marine with deposition of cyclic pelagic carbonates and distal turbidities.
	Blodøks	The upper boundary is characterized with decreasing gamma ray due to more chalk indicated in the Hod formation. The formation was deposited in anoxic conditions.
	Hidra	The upper boundary between Hidra and Blodøks formation is characterized with lithology changing from clhalk lithology to mudstone. The gamma ray reading is decreasing towards Hidra formation. The depositional environment of this formation is an open marine with perioditic origin.

Cromer Knoll	Rødby	The upper boundary of Rødby formation and Shetland Group has characteristic of decreasing gamma ray reading toward Rødby formation. The dominated lithology in this formation is chalk. The Rødby formation has depositional environment of an open marine with reddish sediments and oxygenated environment.
	Sola	The lower boundary is indicated with increasing gamma-ray reading from sandy sediments into shaly Sola formation. The depositional environment of this formation is a marine with alternating anoxic and oxic conditions.
	Åsgard	This lower boundary of this formation is indicated with increasing gamma-ray towars underlying sediments which consists of rich claystones and shales. The formation was deposited in an open marine with low-energy shelf environment.
Viking	Draupne	The formation has a clear break in upper boundary with high response gamma ray. The formation was deposited in a marine environment with restricted circulation in the bottom and anaerobic condition. Any sandstone indicated in the formation was originated from turbiditic.
	Heather	The formation is dominated with silty claystone and was deposited in an open marine environment. Anomalously high gamma ray reading is indicated in upper boundary.
Vestland	Sleipner	Sleipner formation is dominated with shaly rock. The upper part of this formation marks with transition into shales of the Viking Group or the sandstones of Hugin formation. The formation was deposited in a continental fluviodeltaic sequence.
	Hugin	This formation is dominated with sandstone lithology and has clear log break in the upper boundary with decreasing gamam ray reading. The depositional environment of this formation is shallow marine with occasional influence of continental fluviodeltaic conditions.

No group defined	Skagerrak	The formation was probably deposited in a prograding system of alluvial fans and dominated with sandstone lithology.
Dunlin	Cook	The upper and lower boundary of this formation are indicated with decreasing gamma ray reading. The sandstone in this formation represents redeposited sands from the edge of the shelf.
Statfjord	Statfjord	The top of formation is on the contact with Dunlin Group and often consists of calcareous sandstones and dark shales and siltstone. The formation was deposited at lower alluvial plain and braided stream deposits.

Appendix C

ANOVA Table Results of Hypotheses Test

The ANOVA table showed the summary of hypothesis testing. The calculated variables are described below.

Source indicates the source of the variation in data. There are 3 sources provided, Groups, Error and Total. Groups indicates the data variation due to the factor of interest, or variation in the populations being compared. Error means the variation within each groups being compared, while Total means the total variation in population data.

df means the degrees of freedom in the source. The formulas to calculate the df shown as

$$df_{Groups} = k - 1 \tag{C.1a}$$

$$df_{Error} = N - k \tag{C.1b}$$

$$df_{Total} = N - 1 \tag{C.1c}$$

, where k is the number of groups in the source and N is the number of measurements in the source.

SS means the sum of squares in the source. SS_{Groups} calculates the sum of squares between treatment groups, which formula is shown in Equation 2.12. SS_{Error} indicates the sum of squares within groups which following

$$SS_{Error} = \sum_{i=1}^{N} \sum_{j=1}^{k} \left(y_{ij} - M_j\right)^2 \tag{C.2}$$

, where y is the data point or measurement in data source, $y = \{y_1, y_2, \ldots, y_N\}$. SS_{Total} is equal to the sum of SS_{Groups} and SS_{Error}.

MS is the mean square of each source and calculated following

$$MS_{Groups} = \frac{SS_{Groups}}{df_{Groups}} \tag{C.3a}$$

$$MS_{Error} = \frac{SS_{Error}}{df_{Error}} \tag{C.3b}$$

Chi-sq is the distribution which approximated by H test value as shown in Equation 2.14 - Equation 2.16.

Table C.1: ANOVA table of hypothesis testing #1 in Well 15/5-7 A

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.2382e+10	2	6.1912e+09	2.4091e+03	0
Error	2.7975e+10	7850	3.5637e+06		
Total	4.0358e+10	7852			

Table C.2: ANOVA table of hypothesis testing #1 in Well 15/6-11 S

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.7061e+10	3	5.6869e+09	3.4302e+03	0
Error	2.1356e+10	7721	2.7659e+06		
Total	3.8416e+10	7724			

Table C.3: ANOVA table of hypothesis testing #1 in Well 15/6-9 S

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.7038e+10	3	5.6793e+09	3.7934e+03	0
Error	1.5930e+10	7337	2.1711e+06		
Total	3.2967e+10	7340			

Table C.4: ANOVA table of hypothesis testing #2 for shale lithology in Well 15/5-7 A

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	2.406e+09	2	1.203e+09	2.423e+03	0
Error	1.019e+09	3448	2.957e+05		
Total	3.425e+09	3450			

Table C.5: ANOVA table of hypothesis testing #2 for sandstone lithology in Well 15/5-7 A

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	7.857e+08	2	3.929e+08	1.658e+03	0
Error	3.434e+08	2381	1.442e+05		
Total	1.129e+09	2383			

Table C.6: ANOVA table of hypothesis testing #2 for shale lithology in Well 15/6-11 S

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.373e+09	3	4.577e+08	1.644e+03	0
Error	1.269e+09	3161	4.014e+05		
Total	2.642e+09	3164			

Table C.7: ANOVA table of hypothesis testing #2 for sandstone lithology in Well 15/6-11 S

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.269e+09	2	6.346e+08	1.836e+03	0
Error	7.214e+08	2877	2.508e+05		
Total		2879			

Table C.8: ANOVA table of hypothesis testing #2 for shale lithology in Well 15/6-9 S

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	2.298e+09	2	1.149e+09	1.665e+03	0
Error	3.316e+09	4066	8.155e+05		
Total	5.614e+09	4068			

Table C.9: ANOVA table of hypothesis testing #2 for sandstone lithology in Well 15/6-9 S

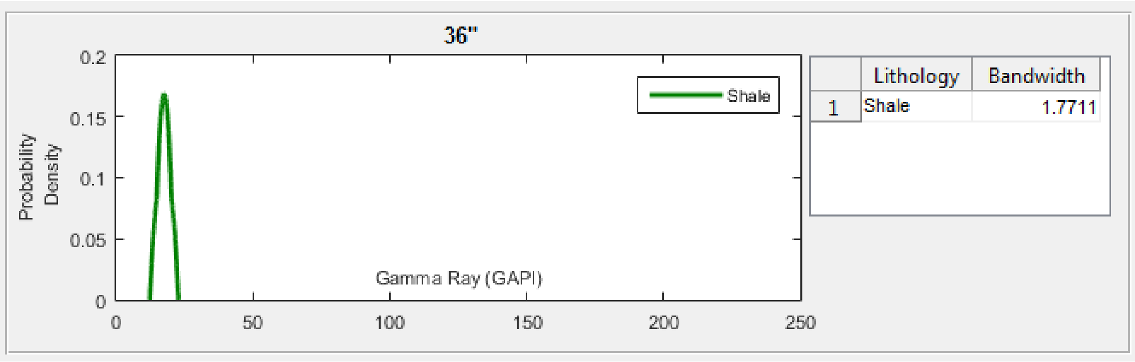
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.707e+08	1	1.707e+08	6.009e+02	1.061e-132
Error	3.535e+08	1844	1.917e+05		
Total	5.242e+08	1845			

Appendix D

Results of GR probability density of two categories: shale and non-shale

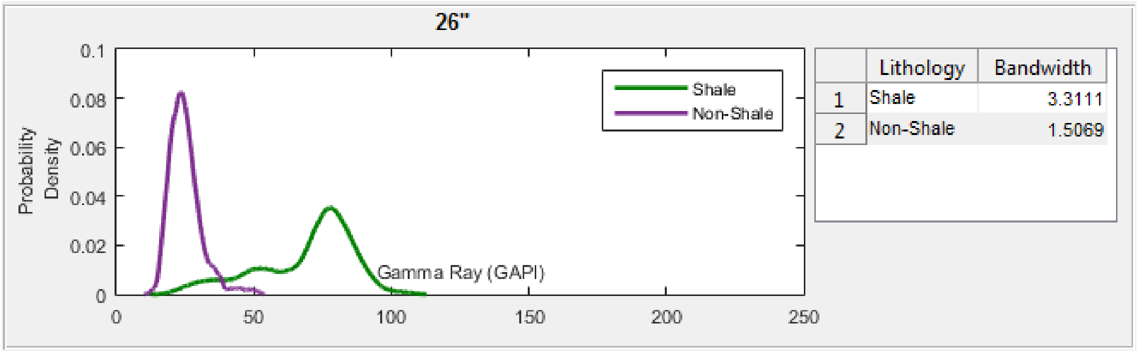
The results of probability density of GR estimated by kernel estimator for two categories shale and non-shale lithology are provided in this section. The results were generated from GR reading in each hole section of Well 15/5-7 A, Well 15/6-11 S, and Well 15/6-9 S.

D.1 Well 15/5-7 A



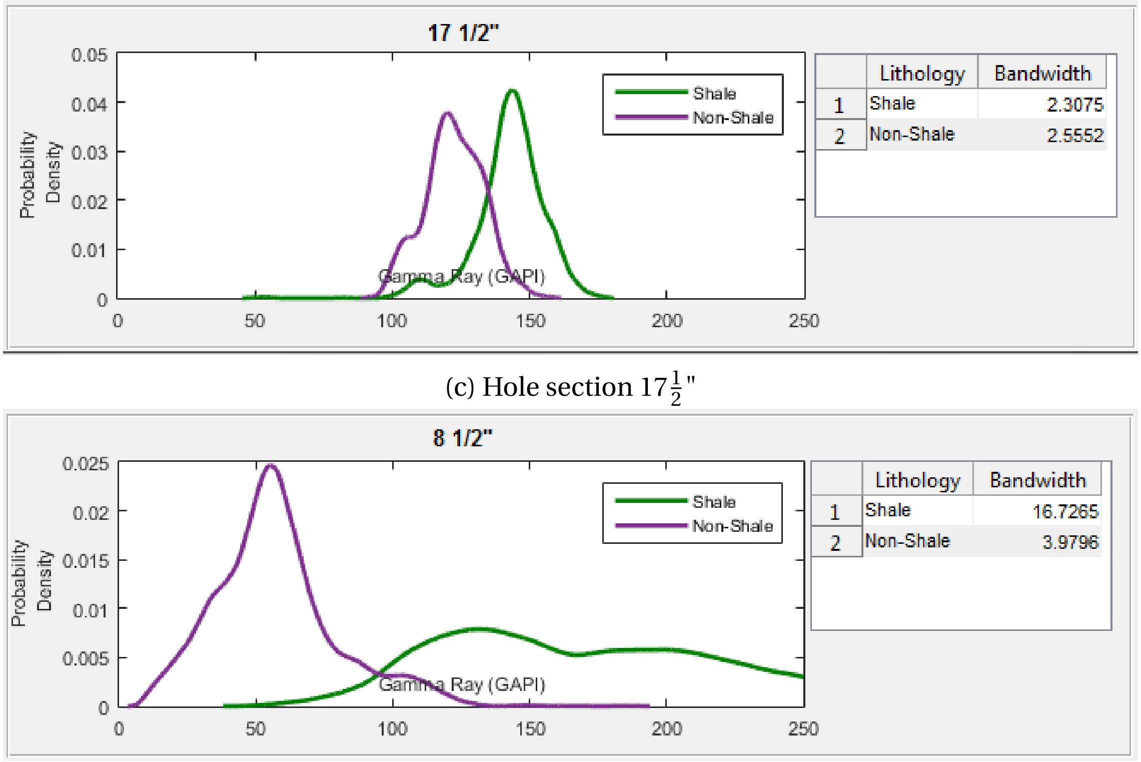
	Lithology	Bandwidth
1	Shale	1.7711

(a) Hole section 36"



	Lithology	Bandwidth
1	Shale	3.3111
2	Non-Shale	1.5069

(b) Hole section 26"



	Lithology	Bandwidth
1	Shale	2.3075
2	Non-Shale	2.5552

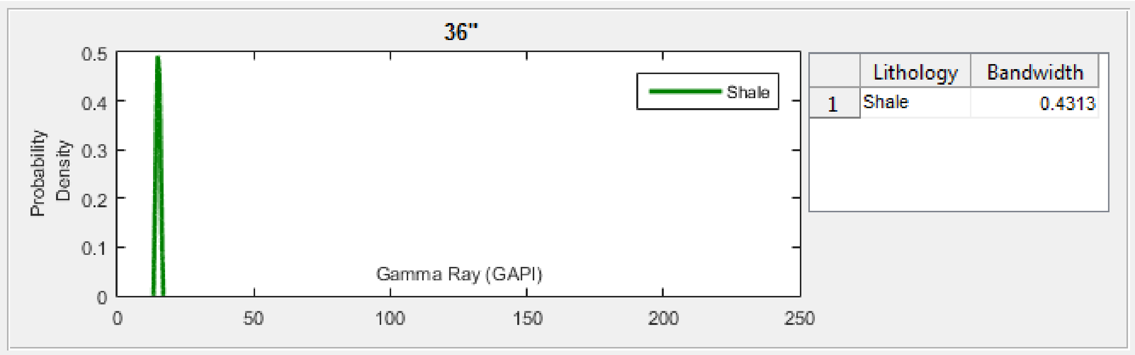
(c) Hole section $17\frac{1}{2}$"

	Lithology	Bandwidth
1	Shale	16.7265
2	Non-Shale	3.9796

(d) Hole section $8\frac{1}{2}$"

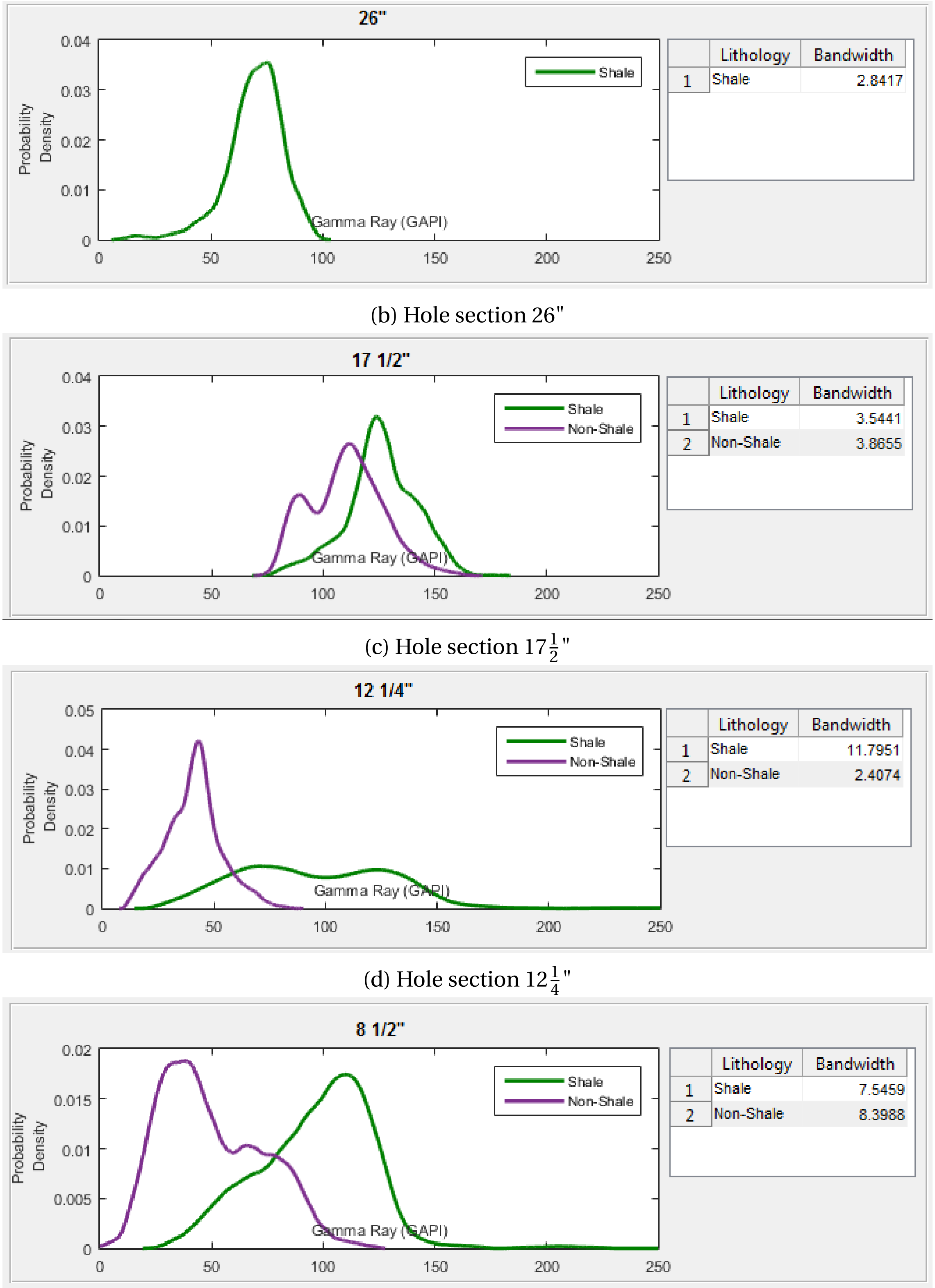
Figure D.1: Probability density results of each hole section in Well 15/5-7 A using kernel estimator grouped into shale and non-shale lithology

D.2 Well 15/6-11 S



	Lithology	Bandwidth
1	Shale	0.4313

(a) Hole section 36"



	Lithology	Bandwidth
1	Shale	2.8417

(b) Hole section 26"

	Lithology	Bandwidth
1	Shale	3.5441
2	Non-Shale	3.8655

(c) Hole section $17\frac{1}{2}$"

	Lithology	Bandwidth
1	Shale	11.7951
2	Non-Shale	2.4074

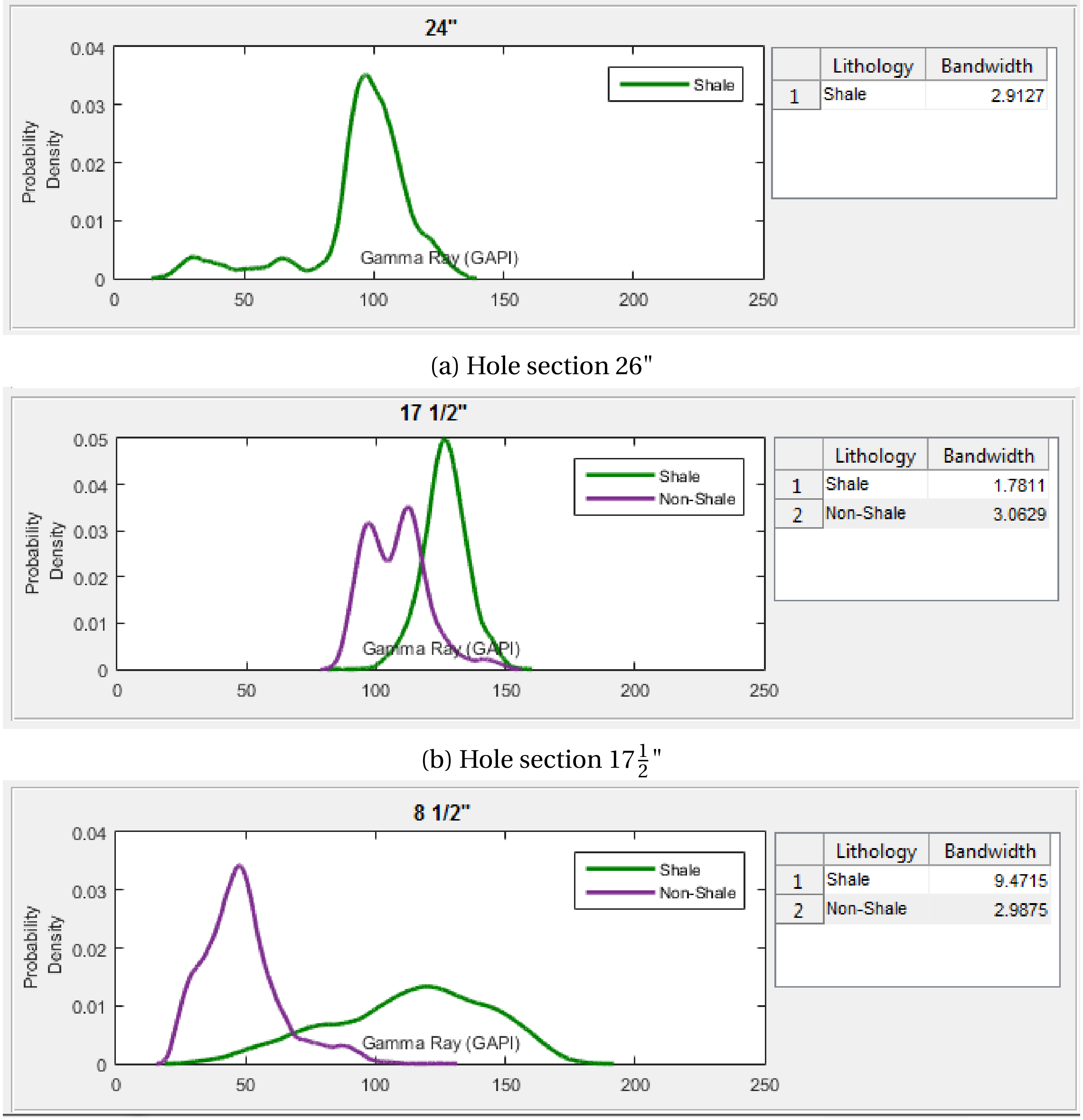
(d) Hole section $12\frac{1}{4}$"

	Lithology	Bandwidth
1	Shale	7.5459
2	Non-Shale	8.3988

(e) Hole section $8\frac{1}{2}$"

Figure D.2: Probability density results of each hole section in Well 15/6-11 S using kernel estimator grouped into shale and non-shale lithology

D.3 Well 15/6-9 S



	Lithology	Bandwidth
1	Shale	2.9127

(a) Hole section 26"

	Lithology	Bandwidth
1	Shale	1.7811
2	Non-Shale	3.0629

(b) Hole section $17\frac{1}{2}$"

	Lithology	Bandwidth
1	Shale	9.4715
2	Non-Shale	2.9875

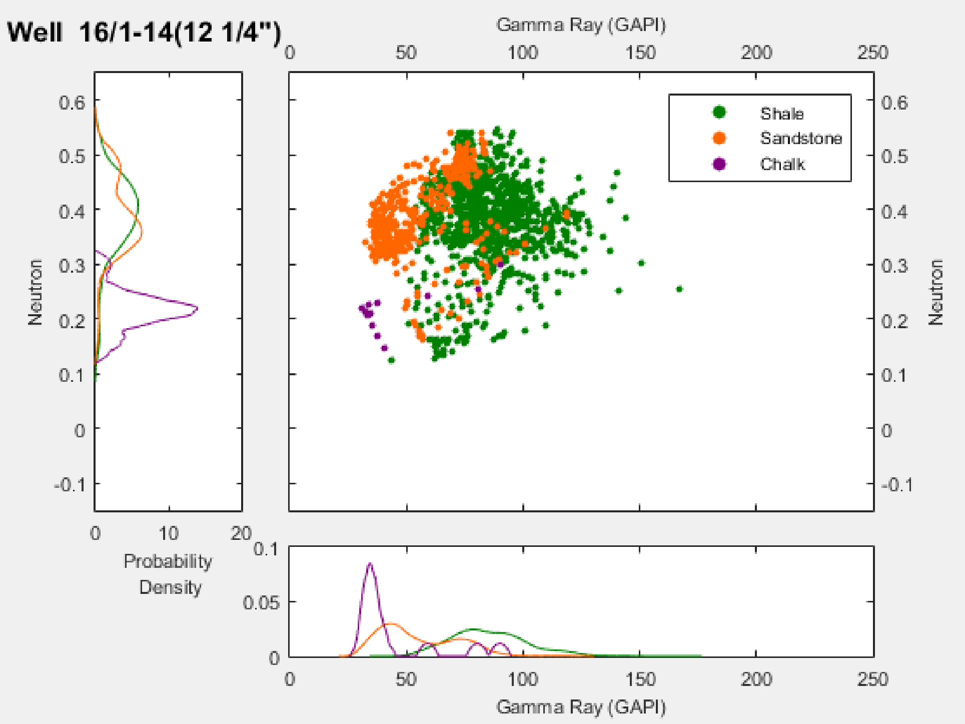
(c) Hole section $12\frac{1}{4}$"

Figure D.3: Probability density results of each hole section in Well 15/6-9 S using kernel estimator grouped into shale and non-shale lithology

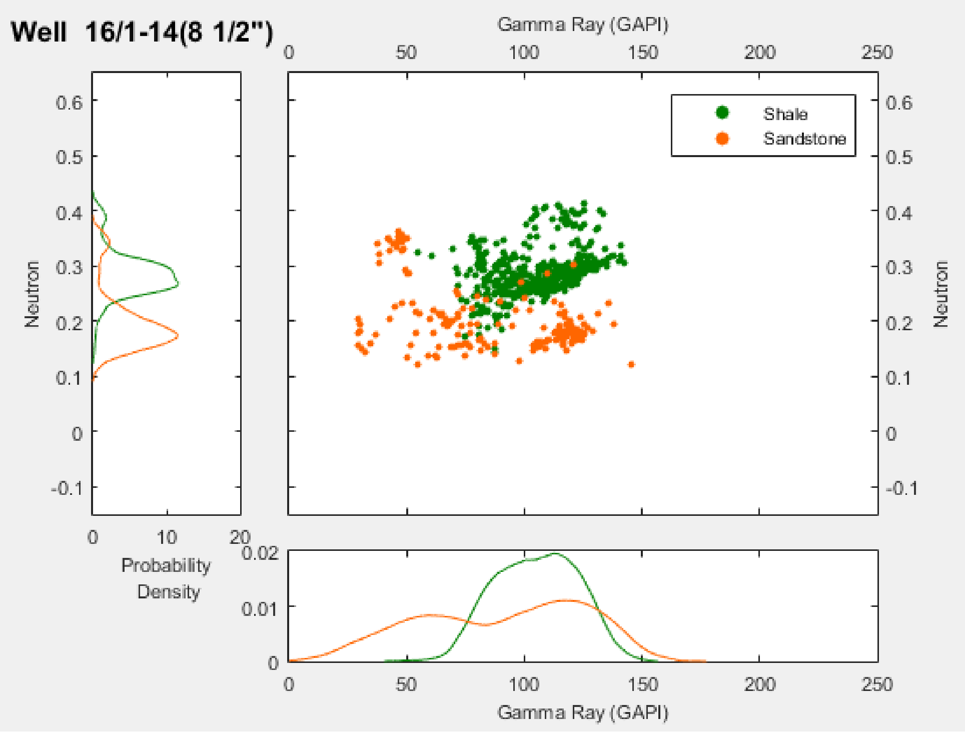
Appendix E

Results of bivariate analysis of GR and neutron

E.1 Well 16/1-14



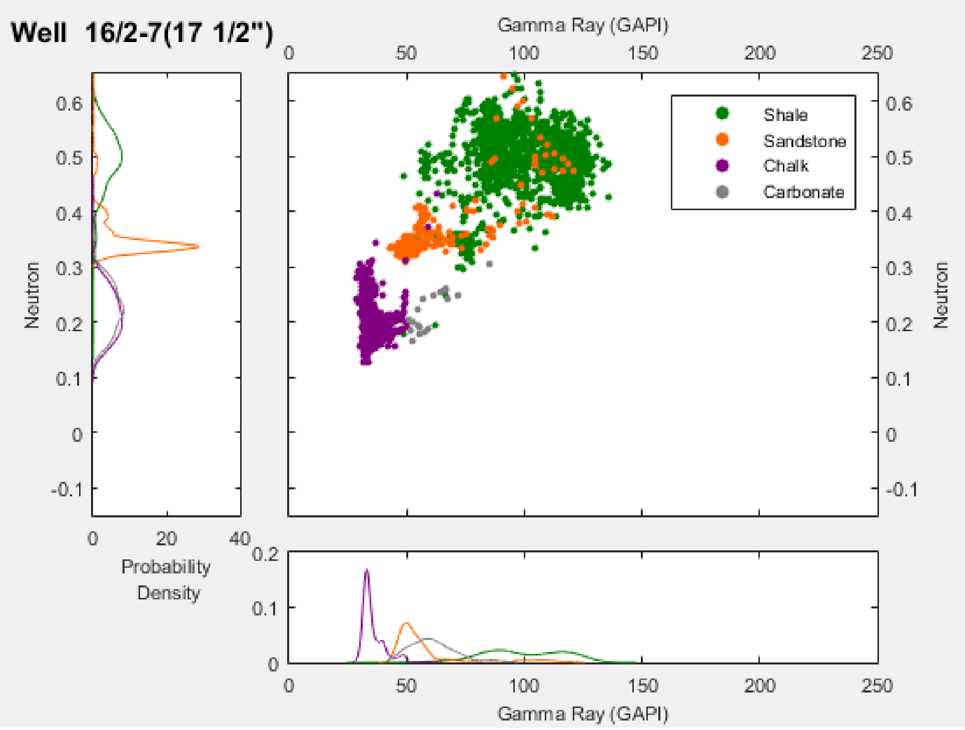
(a)



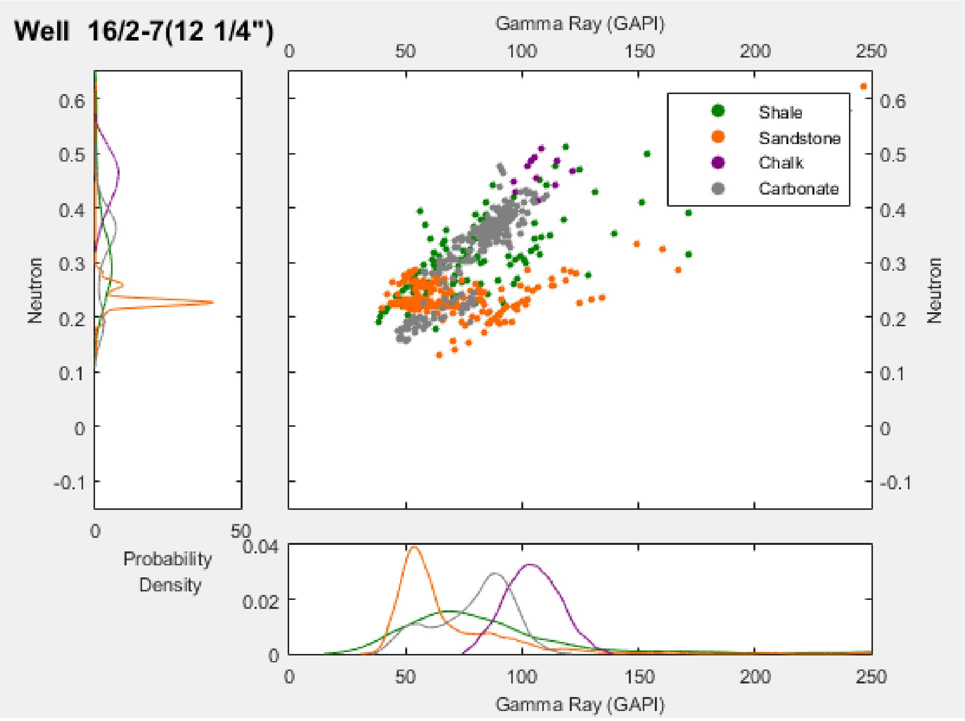
(b)

Figure E.1: Scatter plot and probability density plots of GR and neutron for Well 16/1-14 in hole section: (a) 12 $\frac{1}{4}$" and (b) 8$\frac{1}{2}$"

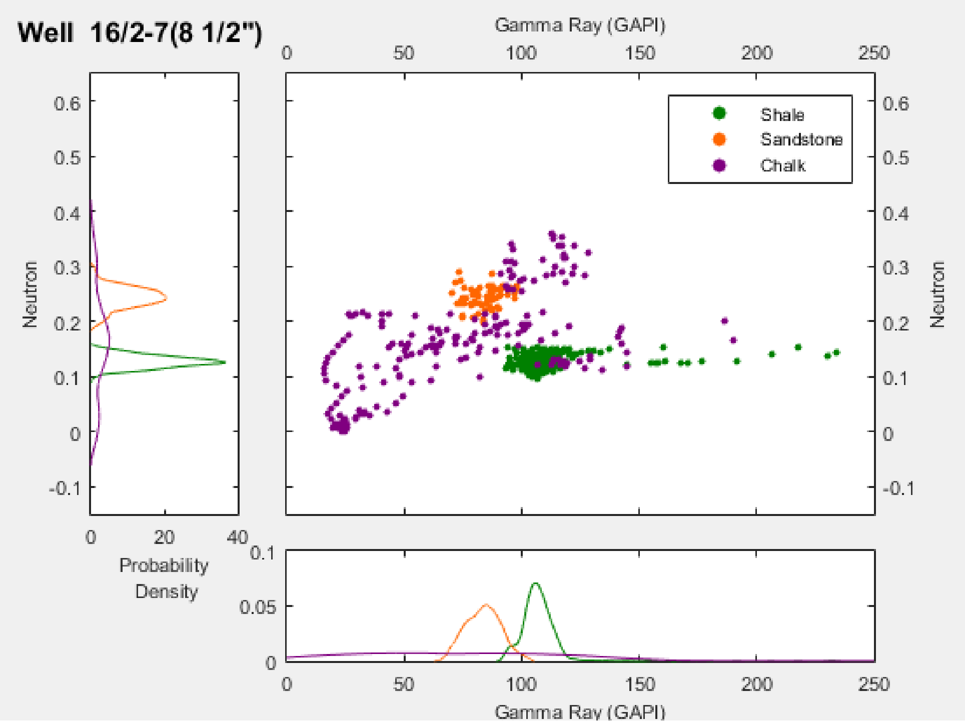
E.2 Well 16/2-7



(a)



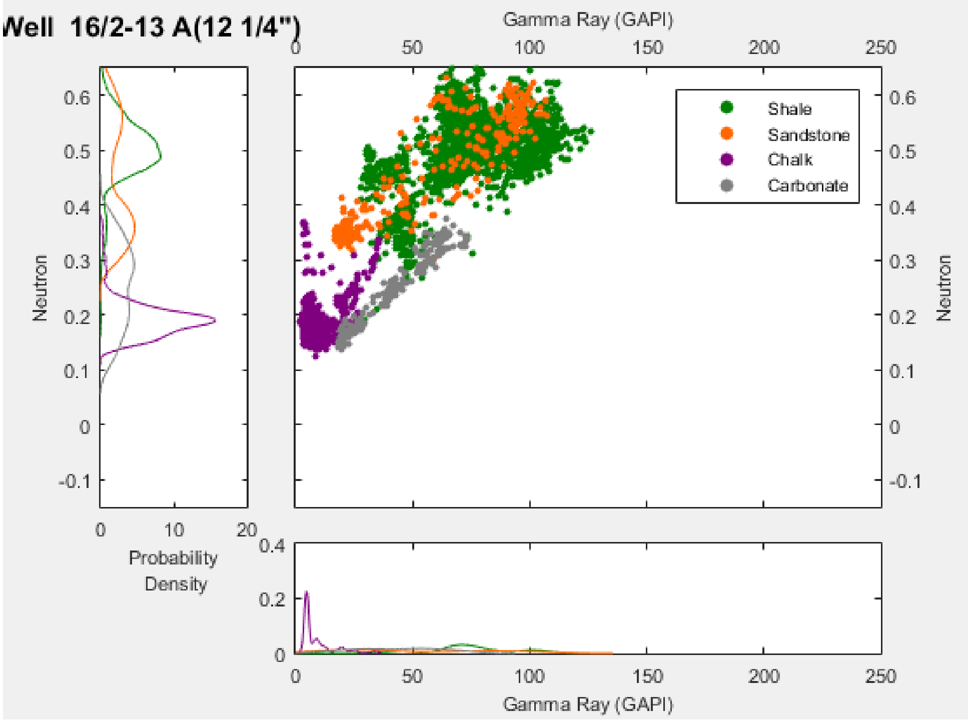
(b)



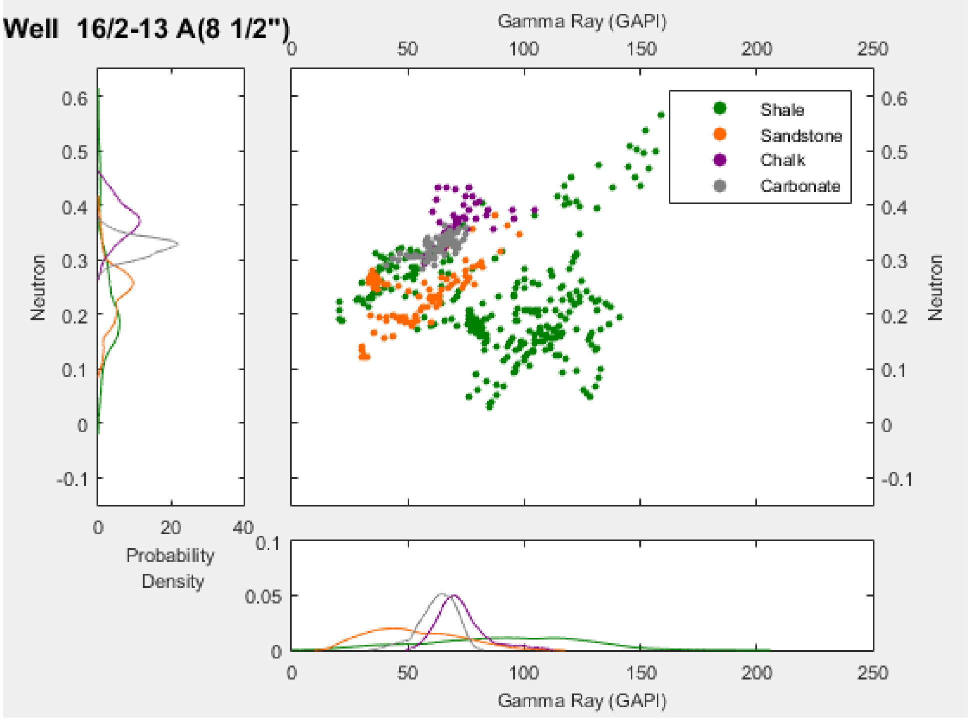
(c)

Figure E.2: Scatter plot and probability density plots of GR and neutron for Well 16/2-7 in hole section: (a) $17\frac{1}{2}$", (b) 12 $\frac{1}{4}$", and (c) $8\frac{1}{2}$"

E.3 Well 16/2-13 A



(a)



(b)

Figure E.3: Scatter plot and probability density plots of GR and neutron for Well 16/2-13 A in hole section: (a) 12 $\frac{1}{4}$" and (b) 8$\frac{1}{2}$"

Bibliography

Bateman, R. M. (2012). *Openhole log analysis and formation evaluation.* Society of Petroleum Engineers SPE, Richardson, Tex, 2nd ed. edition.

Bonner, S. et al. (1992). Logging While Drilling: A Three-Year Perspective. *Oilfield Review*, 4(3).

Bourgoyne, A. T., Millheim, K. K., and Chenevert, M. E. (1985). *Applied Drilling Engineering.* Society of Petroleum Engineers, Richardson.

Burke, J., Campbell, Jr., R., and Schmidt, A. (1969). The Litho Porosity Cross Plot: A New Concept For Determining Porosity And Lithology From Logging Methods. In *SPWLA-1969-Y*, SPWLA. Society of Petrophysicists and Well-Log Analysts.

Busch, J., Fortney, W., and Berry, L. (1987). Determination of Lithology From Well Logs by Statistical Analysis. *SPE-14301-PA*.

Clavier, C., Hoyle, W., and Meunier, D. (1971). Quantitative Interpretation of Thermal Neutron Decay Time Logs: Part I. Fundamentals and Techniques. *SPE-2658-A-PA*.

Clavier, C. and Rust, D. (1976). Mid Plot:a New Lithology Technique. *SPWLA-1976-vXVIIn6a2*.

Delfiner, P., Peyret, O., and Serra, O. (1987). Automatic Determination of Lithology From Well Logs. *SPE-13290-PA*.

Deveaux, R. D. (1999). Applied smoothing techniques for data analysis. *Technometrics*, 41(3):263–263.

Dewan, J. T. (1986). Open-Hole Nuclear Logging - State Of The Art. In *SPWLA-1986-MM*, SPWLA. Society of Petrophysicists and Well-Log Analysts.

Ellis, D. V. and Singer, J. M. (2010). *Well logging for earth scientists.* Springer, Dordrecht.

Frigge, M., Hoaglin, D. C., and Iglewicz, B. (1989). Some Implementations of the Boxplot. *The American Statistician*, 43(1):50.

Gardner, J. S. and Dumanoir, J. (1980). Litho-Density Log Interpretation. In *SPWLA-1980-N*, SPWLA. Society of Petrophysicists and Well-Log Analysts.

Glover, D. (2001). *Petrophysics MSc Course Notes.* University of Aberdeen.

Hansen, B. E. (2009). Lecture Notes on Nonparametrics.

Kruskal, W. H. and Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Laosripaiboon, L., Saiwan, C., and Prurapark, R. (2015). Reservoir Characteristics Interpretation by Using Down-Hole Specific Energy With Down-Hole Torque and Drag. In *OTC-25890-MS*, OTC. Offshore Technology Conference.

Larionov, V. (1969). Borehole Radiometry.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Mwenifumbo, C. (1993). Kernel Density Estimation In The Analysis And Of Borehole Geophysical Data. *SPWLA-1993-v34n5a3*.

Ott, L. and Longnecker, M. (2010). *An introduction to statistical methods and data analysis*. Brooks/Cole Cengage Learning, Australia ; United States, 6th ed edition.

Poupon, A. and Gaymard, R. (1970). The Evaluation Of Clay Content From Logs. In *SPWLA-1970-G*, SPWLA. Society of Petrophysicists and Well-Log Analysts.

Privitera, G. J. (2015). *Statistics for the behavioral sciences*. SAGE, Los Angeles, second edition edition.

Provost, Jr., C. (1987). A Real-Time Normalized Rate of Penetration Aids in Lithology and Pore Pressure Prediction. In *SPE-16165-MS*, SPE. Society of Petroleum Engineers.

Schlumberger Educational Services (1989). *Schlumberger: Cased Hole Log Interpretation Principles/Applications*. Houston.

Schlumberger Wireline & Testing (1998). *Log interpretation charts*. Schlumberger Wireline & Testing, Sugar Land, Texas.

Scott, D. W., editor (1992). *Multivariate Density Estimation*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Scott, D. W. (2004). Multivariate density estimation and visualization. Papers / Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE) 2004,16.

Serra, O., Delfiner, P., and Levert, J. C. (1985). Lithology Determination From Well-Logs: Case Studies. In *SPWLA-1985-WW*, SPWLA. Society of Petrophysicists and Well-Log Analysts.

Silverman, B. (1981). Density estimation for univariate and bivariate data. In *Interpreting multivariate data*, pages 37–53. Wiley: Chichester, v edition.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. Springer series in statistics. Springer, New York.

Stieber, S. (1970). Pulsed Neutron Capture Log Evaluation - Louisiana Gulf Coast. In *SPE-2961-MS*, SPE. Society of Petroleum Engineers.

Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co, Reading, Mass.

Wilson, R. (1955). Mud Analysis Logging And Its Use In Formation Evaluation. In *SPE-587-G*, SPE. Society of Petroleum Engineers.

Ziaja, M. and Roegiers, J.-C. (1998). Lithology Diagnosis Based on the Measurements of Drilling Forces and Moments at the Bit. In *SPE-47799-MS*, SPE. Society of Petroleum Engineers.