Automatic Lithology Prediction from Well Logging Using Kernel Density Estimation

A.N. Corina[a,*], S. Hovda[a]

[a] *Norwegian University of Science and Technology (NTNU), Trondheim, Norway*

Abstract

Technologies of real-time data measurement during drilling operation have kept the attention of petroleum industries in the past years, especially with the benefit of real-time formation evaluation through logging-while-drilling technology. It is expected that most of the logging data will be recorded in real-time operation. Hence, application of automated lithology prediction tool will be essential.

An automatic method to predict lithology from borehole geophysical data was developed. It was solved as a multivariate classification problem with multidimensional explanatory variables. The learning algorithm combines kernel density estimates and a classification rule that is based on these estimates. The goal of this work is to test the method on a univariate variable and validate the prediction accuracy by calculating the misclassification rates. In addition, the results will be established as a baseline for application in practice and future developments for multivariate variables analysis.

Gamma-ray from wireline logging is selected as the variable to describe two lithology groups of shale and not-shale. Data from six wells in the Norwegian Continental Shelf were extracted and examined with aids of explorative data analysis and hypothesis testing, and then divided into a training- and test data set. The selected algorithm processed the training data into models, and later each element of test data was assigned to the models to get the prediction. The results were validated with cutting data, and it was proved that the models predicted the lithology effectively with misclassification rates less than 15 % at its lowest and average of ±31%. Moreover, the results confirmed that the method has a promising prospect as lithology prediction tool, especially in real-time operation, because the non-parametric approach allows real-time modeling with fewer data assumptions required.

Keywords: Real-time drilling data, gamma ray, statistical classification, kernel density estimation, non-parametric data, lithology prediction

* Corresponding author

Email addresses: anisa.corina@ntnu.no (A.N. Corina), sigve.hovda@ntnu.no (S. Hovda)

1. Introduction

The process of lithology identification is traditionally executed using data from cutting visualization, core inspection, or wireline logging. And today, many new technologies are advancing and replacing the manual process into a more automated process, such as high-speed telemetry. This development means that more types of borehole geophysical data are measured in the real-time operation, and hence lithology identification methods are expected to be more straightforward and precise than the traditional methods. This motivates the development of an automated method of lithology prediction.

The early technique of lithology interpretation was accomplished using qualitative approach through identification of log separations or unique trends between several well log curves visually without the requirement of calculations. In practice, this technique provides quick evaluations, especially over a depth of interval which is consistent. However, the application becomes demanding for complex lithologies identification that requires large dataset analysis and depends on the geological history of the area (Ellis and Singer, 2007).

The advanced progress of modern computers has stimulated the development of quantitative methods of lithology identification with improved speed and accuracy. There are wide variations of mathematical techniques adapted as lithology identification tool, such as clustering (Wolf and Pelissier-Combescure, 1982; Ye and Rabiller, 2000), fuzzy logic (Cuddy et al., 1997; Saggaf and Nebrija, 2003), and neural networks (Benaouda et al., 1999; Maiti et al., 2007). One of the early studies that implemets statistical probability method with combination of clustering and classification technique for lithofacies determination was accomplished by Delfiner et al. (1987). Since then, many other studies were carried out in similar manners, including studies by Busch et al. (1987) and Coudert et al. (1994). Those studies came in conclusion that the classification technique based on probability density was promising for lithology prediction and the statistical methods were suitable for handling large databases. However, the assumption of normal (Gaussian) distribution for the density probability function was believed to be strict for modeling non-parametric data.

Modeling the non-parametric data that are infinite-dimensional is best approached using non-parametric statistic technique. The application is convenient for dataset that grows in size – i.e. a dataset whose final structure of data distribution is yet unknown–, such as model from real-time dataset. In statistic probability, the estimation of probability density function of non-parametric data is usually accomplished using kernel density estimator. It is also an excellent tool for estimating univariate, bivariate, or trivariate data, even when the number of data points is relatively low (Silverman, 1986). Kernel density estimator has also been applied to solve geophysical and geologicals problem in the past (Mwenifumbo, 1993; Mwenifumbo et al., 2004). Mwenifumbo (1993) specifically applied the estimator on well logging data and proved that the results of probability density function were precise in showing the major features of each lithofacies.

Until recently, the automated lithology predictions that based on statistical probability density did not take account of non-parametric modeling, meaning that the assumptions were not practicable on real-time dataset. Therefore, in this study we attempted to develop a lithology prediction method using a classification technique based on probability density function of explanatory variables, which was estimated using kernel density estimator. The selected classification technique implemented a classification rule, or classifier, to generate the final classification models. Two types of classifiers were presented in this study, one of which implemented prior probability value.

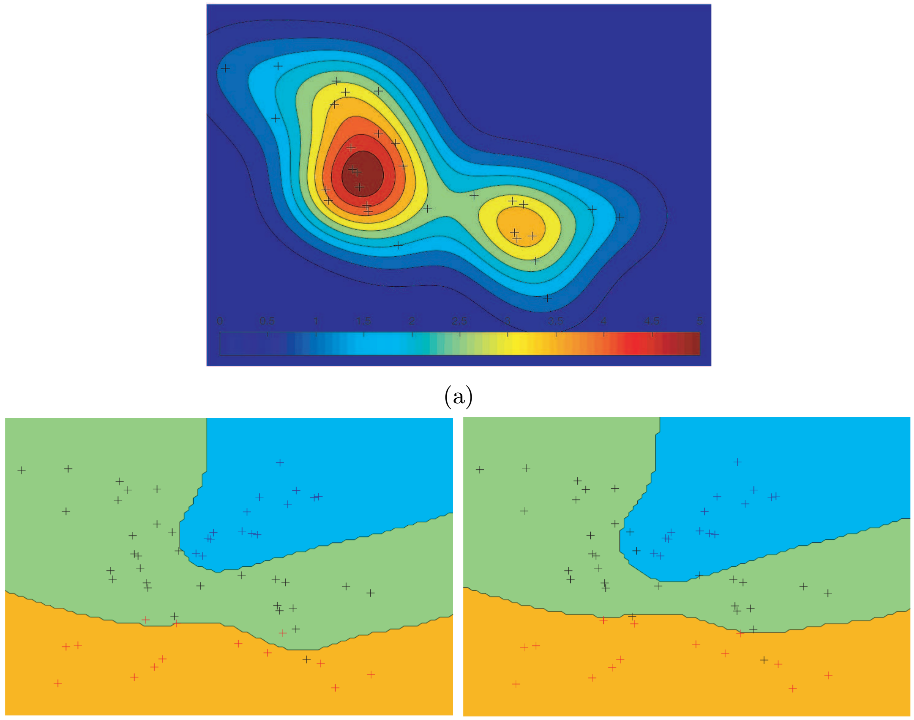
To give a brief overview of the proposed method, we presented a set of two-dimensional data with 30 points (black, red, and blue points) as a contour plot of the probability density functions, estimated by kernel density estimator in Fig. 1a. Fig. 1b describes a trinary classification rule, which neglected the prior probability, based on two-dimensional data, dividing the data into three different classes marked with the green, blue, and yellow region. If the classification rule was modified, by taking prior probability into account, some regions expanded or shrunk depends on the probability value of the particular region (see Fig. 1c). Notice that there are some black points now classified into the blue region after the classification rule was modified.

One of the principal aims in this study is to test the proposed learning algorithm by using a univariate data, which is gamma ray log, and acquire the accuracy given by the models from classifying new observations to lithology groups of shale and not-shale. Our methods to select the data and how to employ them into the learning algorithm are described in detail prior the test. Another of our aims is to present the application of the proposed method in practice as a baseline for petroleum engineers to implement, especially in real-time operation.

2. Dataset description

The data used in this study was from six wells located in Norwegian Continental Shelf. The wells are situated at the eastern part of the South Viking Graben with three wells from Block 15, situated at Gina Krog field within Ve sub-basin, and three wells from Block 16, situated at Ivar Aasen field within the Gudrun Terrace (Fig. 2). The configuration of the South Viking Graben is mainly due to the Callovian-Ryazinian rift event. The South Viking Graben has a steep bounding with a small terrace to the east (The Gudrun Terrace). The Gudrun Terrace is dominated with shallow marine deposition on the basin flanks, with terrace topography. The fault bounding the graben to the west was active during the regressive phase of Lower Oxfordian, while sediment gravity flowed to the grabenal area. The Ve sub-basin is located at the grabenal area with a thick section of Cretaceous (Steel et al., 1995).

The available data included gamma-ray logs, well schematic, geological descriptions, and mud logging. In this study, we chose gamma ray log as the explanatory variable to distinguish shale and not-shale lithology because it is a reliable shale detector and the tool is commonly run in combination with high pulse telemetry. Gamma ray tool measures the composition of the natural-occurring isotopes contained in the rocks, such as potassium, uranium, and



(a) (b) (c)

Fig. 1: The 2-dimensional multivariate analysis: (a) probability density function from kernel density estimation, (b) group region based on classification rule without prior probability, and (c) group region based on classification rule with prior probability.

thorium (Ellis and Singer, 2007). Due to high content of radioactive mineral in shale, the tool is effective to identify shale (Schlumberger Educational Services, 1989). However, the tool is sensitive to several borehole environment factors, such as hole diameter, borehole quality (e.g. caving or washout), mud weight, casing properties, and cement thickness. In addition to borehole environment factors, false gamma ray reading can be caused of the tool offset from the hole center during the tool running.

Both geological description and mud logging data contained information of lithology description, but each was given by different sources. The lithology information from geological descriptions is a rough estimation given by geologists prior drilling operation. Meanwhile, lithology information from mud logging is obtained based on cutting visualization during drilling operation. Hence, the mud logging data has better accuracy than geological descriptions. Both lithology information showed that the wells were composed of four major lithologies: sandstone, shale, carbonate, and chalk. Within the study, sandstone, carbonate, chalk, and other minor lithologies were grouped into non-shale lithology.

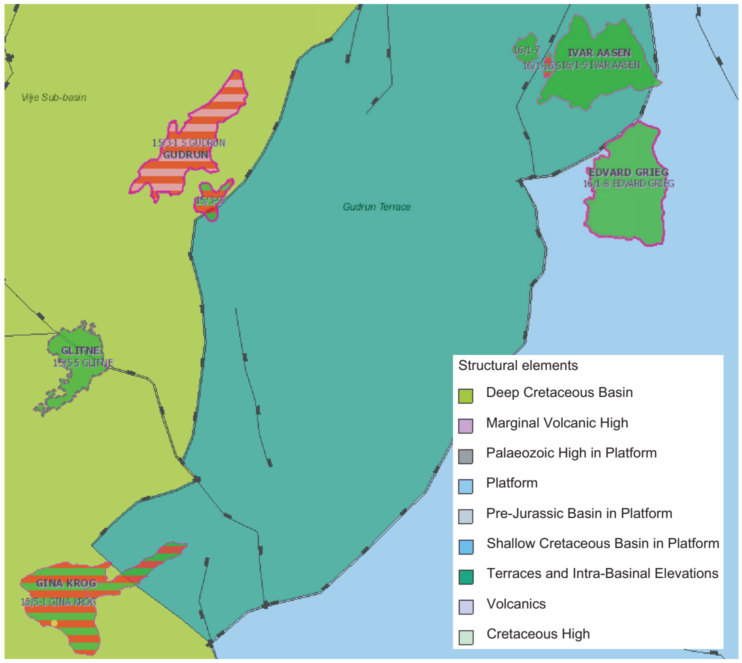


Fig. 2: Location of the selected wells at the South Viking Graben: Ve Sub-basin and Gudrun Terrace (Norwegian Petroleum Directorate, 2017)

3. Data exploration

Data exploration of the gamma-ray dataset was carried out using explanatory data analysis and hypotheses testing. This approach allowed us to identify the characteristic of the gamma ray dataset in describing lithology, and hence it was relevant for the modeling task. Moreover, with the lack of information on gamma ray tool properties, this approach would also be a countermeasure for any neglected calibration offset of the tool or the missing corrections of gamma ray reading.

3.1. Exploratory data analysis

The exploratory data analysis was comprised of the numerical descriptions of mean, median, and standard deviation, and graphical descriptions of boxplots and histograms. The boxplot visualization was adapted from Tukey method that illustrates three quartiles value indicated by three lines forming a box and extreme values or outliers indicated by whiskers perpendicular to the quartile lines (Frigge et al., 1989). In addition, the histogram bin width was calculated following Scott rule (Scott, 1992).

Table 1: Statistic description of gamma ray data of each lithology in Well 15/5-7 A by: (a) ungrouping and (b) grouping according hole size

Lithology	**Mean**	**Median**	**St. Dev**
Shale	112.92	127.61	43.65
Non-shale	82.79	75.54	36.16

(a)

Lithology	**Hole size**	**Mean**	**Median**	**St. Dev**
Shale	17 1/2”	138.42	139.37	13.07
	8 1/2”	104.06	88.38	47.46
Non-shale	17 1/2”	118.52	118.01	11.73
	8 1/2”	58.63	56.77	16.32

(b)

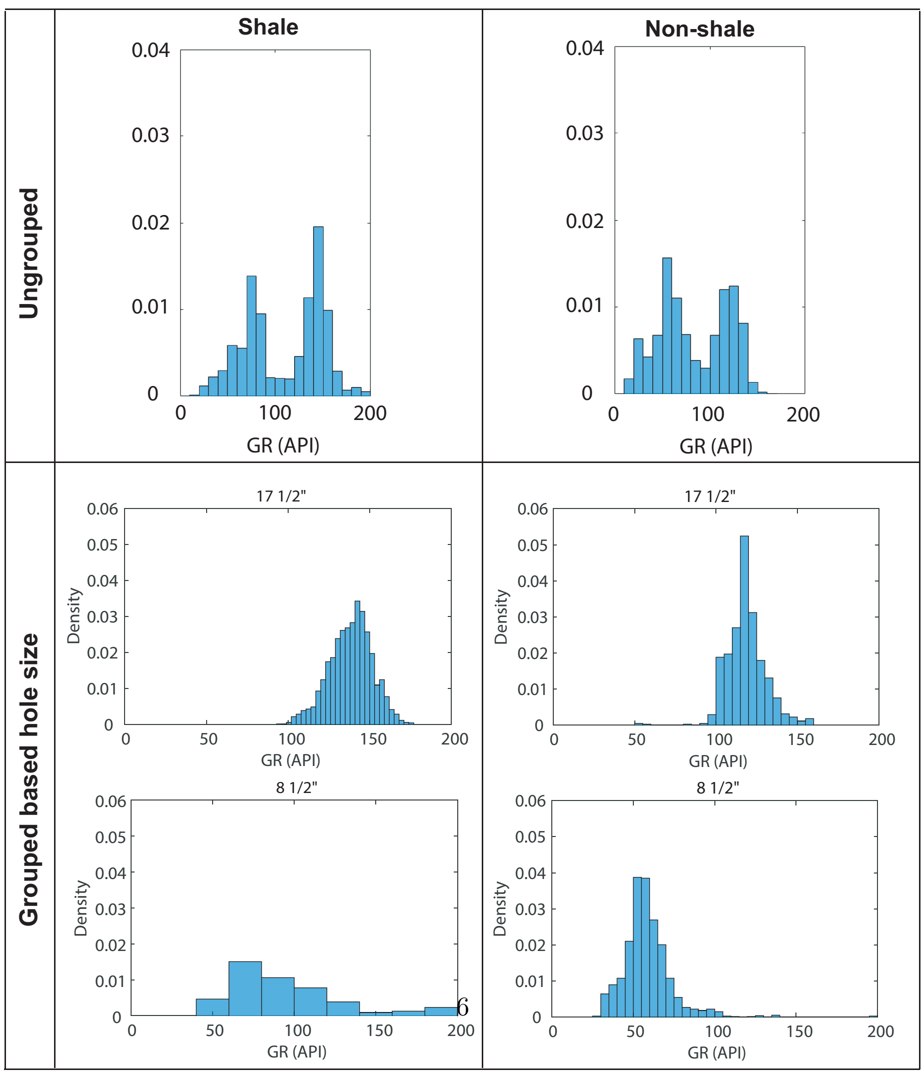


Fig. 3: Comparison of gamma-ray data of Well 15/5-7 A when ungrouped and grouped based on hole size, visualized in: (a) histogram and b) boxplot

Based on the result of one example from Well 15/5-7 A , high variance and bimodal distributions of gamma ray value were detected in both shale and non-shale lithology (see the ungrouped plots in Fig. 3 and Table 1a). After plotting the data in log traces, it appeared that the gamma ray logs shifted from one hole section to others (Fig. 4). Because of data limitation, the source of error factors could not be recognized, hence clustering the gamma ray based on the hole size was considered as the most relevant attempt to reduce data variation. Improvements of data distribution were observed by hole size grouping as each group had reduced standard deviations (see the grouped plots in Fig. 3 and Table 1b). In addition, it was observed from the histogram and boxplot that lithology data distributions in each hole size group were not symmetrical and the shapes did not follow the normal distribution.

3.2. Hypothesis testing

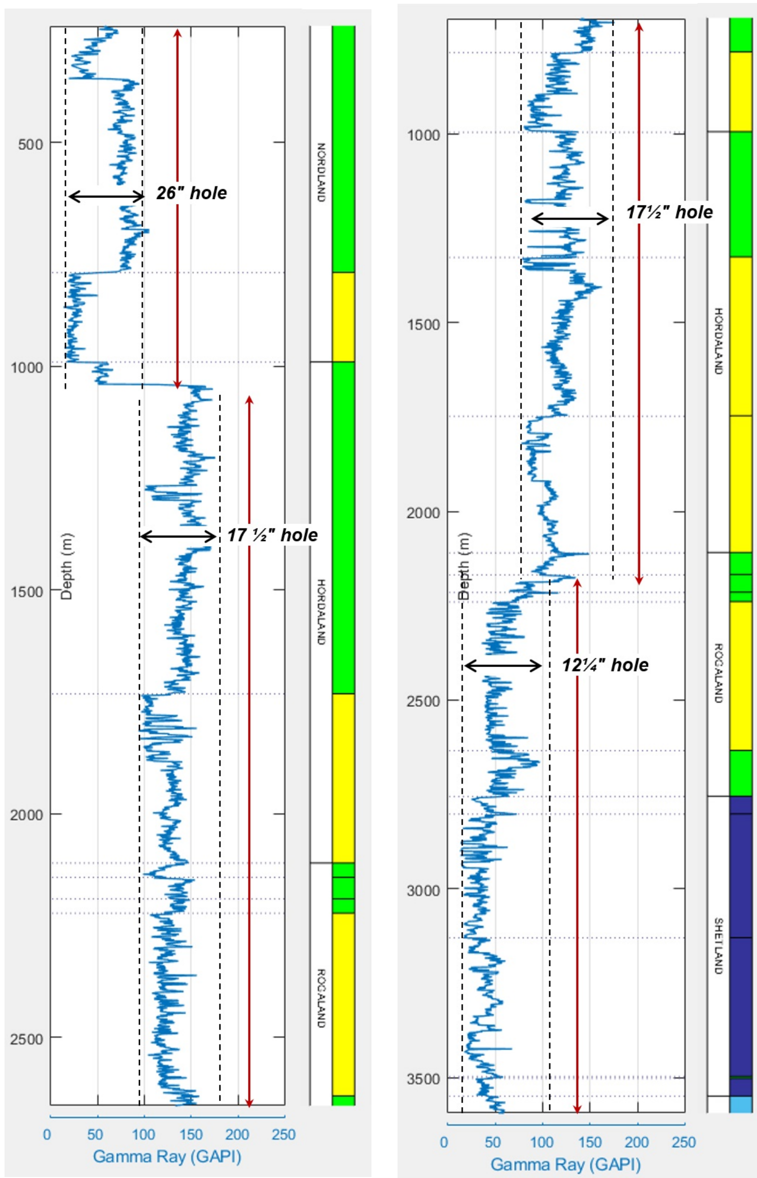
The result of exploratory data analysis above indicated that the gamma-ray data of one lithology type in a hole section could not be used interchangeably with the same lithology type in other hole sections for the same well. However, the process of exploratory data analysis tended to be visually qualitative and mostly concentrated on the comparison of the statistical properties and the data distribution. Thus, drawing a conclusion from explanatory data analysis by itself was considered inadequate, advancing us to perform hypothesis testing.

Hypothesis testing is a method for testing a hypothesis of a group within a population (Privitera, 2015). Hypothesis testing tests the null hypothesis (H_0) – a statement of a population parameter that is assumed to be true – whether it is likely to be true or not. The statement that opposes the null hypothesis is called the alternative hypothesis (H_1). This study adapted the Mann-Whitney test, a rank-based test which evaluates if there are any independent variables contained between two sets of non-parametric data. If the probability value (p-value) given from the test is less than the level of significance, then the null hypothesis will be rejected (Mann and Whitney, 1947).

In this test, the null hypothesis was the distribution of gamma ray of two groups of hole section is equal. Each lithology group in one hole section was tested toward other hole section with the level of significance at 5%. The test was repeated for a different combination of groups because more than two hole sections appeared in one well. The results, summarized in Table 2, showed that the returned probabilities from the combinations of all of the wells were lower than the level of significance, and hence the null hypothesis was rejected. In other words, gamma ray data between two groups of hole section were independent of each other. Based on data exploration, we concluded that the modelling task was better performed for each hole size of the well.

4. Approach of the machine learning algorithm for lithology prediction

Classification is an instance of learning the model f that projects the observed variables, x, to one of the predefined group, y. The process employs



(a) (b)

Fig. 4: Shifted gamma ray value from logging visualization: (a) 26" and 17 1/2" in Well 15/5-7 A and b) 17 1/2" and 12 1/4" in Well 15/6-11 S

a learning algorithm that implements classification, also known as classification rule, to identify the best fit model that provides a relationship between the attribute set and the class labels from the input data. Before classifying

Table 2: Example of hypothesis testing result for wells in Ivar Aasen field

Well	Lithology	Section #1	Section #2	P Value
15/5-7 A	Shale	17 1/2”	8 1/2”	2.90×10^{-257}
	Non-shale	17 1/2”	8 1/2”	$< 2.251 \times 10^{-308}$ *
15/6-11 S	Shale	17 1/2”	12 1/4”	7.71×10^{-292}
		17 1/2”	8 1/2”	3.81×10^{-170}
		12 1/4”	8 1/2”	2.81×10^{-160}
	Non-shale	17 1/2”	12 1/4”	$< 2.251 \times 10^{-308}$ *
		17 1/2”	8 1/2”	5.32×10^{-67}
		12 1/4”	8 1/2”	1.19×10^{-67}
15/6-9 S	Shale	17 1/2”	8 1/2”	$< 2.251 \times 10^{-308}$ *
	Non-shale	17 1/2”	8 1/2”	$< 2.251 \times 10^{-308}$ *

* The smallest positive normalized floating point number in IEEE ® double precision.

new observation, the *training dataset*, which consists of the observation whose groups are known, is trained to develop the models. Afterwards, the models are employed to predict the group of new observations whose groups are unknown, also called as *test data*. Then, the prediction of test data will be validated with the expected output for model evaluation.

The type of classification rule proposed in this paper was based on probability density function, and hence the probability density estimation from training data was required. Based on the data exploration above, the gamma-ray dataset had a non-parametric distribution, and hence kernel density estimation was suitable to generate the probability density function. Descriptions of the kernel density estimation and the classification rule are explained in this section.

4.1. Probability density function from kernel density estimation

The fundamental concept underlying the analysis of univariate data is the probability density function for non-parametric distribution. Different from the parametric approach which implements strong assumptions, the non-parametric approach uses relatively weak assumptions. Thus, the non-parametric approach can get the true pattern of the data and identify any subgroups within the data (Simonoff, 1996).

Kernel density estimation is an expansion of histogram method, the simplest method to estimate probability density. Because histogram method returns a discrete result and does not sensitive to probability density function f, the smoothing method, such as kernel density estimation, is more favorable to return a continuous probability density. Study also showed that this method was suitable to estimate borehole geophysical data, especially on data with fat-tailed distribution and analysis of multivariate data (Mwenifumbo, 1993).

The density function of a random variable X which has probability density function $f(x)$ is shown as below

$$P(a < X < b) = \int_a^b f(u)du \tag{1}$$

for any constants a and b.

Let $\{x_1, ..., x_n\}$ represent a random sample of size n from the density f. For univariate density estimation, the empirical cumulative distribution function gives:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2}$$

The degree to which the data are smoothed is dependent on the smoothing parameter, or bandwidth, h. The optimal bandwidth value is obtained by minimizing the mean square error. Even though there is no objective method to determine it, several approaches have been studied (Simonoff, 1996). The kernel function, K, is a non-negative function, and the area underneath the function integrates to 1. Different forms of kernel function are available, and the choice of kernel function is beyond the topic of this study (Silverman, 1986; Simonoff, 1996).

In this study, the process of estimation was performed using a MATLAB R2015A function, `ksdensity`, which returns the estimation of the probability density evaluated at equally spaced points x_i that cover the range of the input data of x (Bowman and Azzalini, 1997). The kernel function applied was Epanechnikov function and the optimal bandwidth was given from `ksdensity` function automatically, of which value is calculated based on the distribution of normal densities.

4.2. Classification scheme based on probability density

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 $X = (X_1, \dots, X_p)$. Denote that Ω is the collection of all possible outcomes x. As $f_1(x)$ and $f_2(x)$ usually overlap, some points of Ω can belong to π_1 and π_2, with different probability values. In order to divide Ω into two non-overlapping regions R_1 and R_2 ($R_1 \cup R_2 = \Omega$ and $R_1 \cap R_2 = \emptyset$), the probability of misclassification must be minimum.

For a new observation x_0, a rule is exist to allocate x_0 to π_1 if the probability value from π_1 is greater that probability value of x_0 from π_2, or to allocate x_0 to π_2 if the opposite holds. Based on this criterion, then R_1 is the set of possible outcomes of x such that $f_1(x) > f_2(x)$ and R_2 is the set of possible outcomes of x such that $f_1(x) < f_2(x)$. The classification rule is, therefore:

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1, \qquad R_2 : \frac{f_1(x)}{f_2(x)} < 1 \tag{3}$$

If equality holds, x_0 is allocated to one of the group randomly. This type of classification rule is also known as *likelihood ratio rule* (Cios et al., 2007).

In case the prior probability information is available, the classification rule from probability density can be combined with prior probabilities. The prior probabilities represent initial knowledge about how likely each class may emerge without any help of any further information about the object, or without information from explanatory variable x. Denote by $p(1)$ the prior probability that x_0 belongs to π_1 and $p(2)$ the prior probability that x_0 belongs to π_2, the classification rule will become,

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p(2)}{p(1)}, \qquad R_2 : \frac{f_1(x)}{f_2(x)} \quad < \frac{p(2)}{p(1)} \tag{4}$$

The results from classification are validated toward the expected results, which then summarized in a confusion matrix, a table that reports the number of false positive (FP), false negative (FN), true positive (TP), and true negative (TN), see Table 3. From the observed numbers, the misclassification rate can be calculated following.

$$\text{Misclassification rate} = \frac{\text{FP} + \text{FN}}{\text{TN + FP + FN + TP}}, \tag{5}$$

Table 3: Confusion matrix table of two 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

5. Simulations of lithology prediction and discussions

Once the proposed method was coded together using MATLAB R2015A, simulations of lithology prediction were carried out by model testing.Two types of model testing were run to understand the extent of the models in predicting accurate lithology using different test dataset. In the first test (Test 1), each model that was trained from a portion of the dataset from one particular well was tested using the rest of dataset from the same well. Meanwhile, each model in the second test (Test 2) was trained from a complete dataset from one particular well. Then, the models were tested using dataset from the neighboring wells.

In both tests, we used two approaches of classifications: (1) classification adopting likelihood ratio rule (Equation 3) and (2) classification adopting the rule that regards prior probability values (Equation 4), respectively named as rule #1 and rule #2 for ease of reference. The prior probability for rule #2 was calculated based on the number of observations of shale and non-shale lithology from the geological description of the test set, which then normalized to 1 to fulfill the condition $p(1) + p(2) = 1$. Afterwards, the result from the prediction were verified with lithology data taken from cuttings, and then summarized in the confusion matrix. Within the context of the present paper, the accuracy of the prediction was reported in term of percentage of misclassification rates. This approach was consistent with the large size of test set ($>$ 450 samples). And to correspond the result from data exploration, the test had to be performed on the models from training data that had equivalent hole size.

5.1. Test 1

Table 4: Misclassification rates of Test 1 for rule #1 and #2 applied

Well	**Hole size (")**	**Training data**		**Testing data**		**Misclassification Error (%)**	
		Depth (m)	**N**	**Depth (m)**	**N**	**Rule #1**	**Rule #2**
15/5-7 A	17 1/2	1039-2180	2283	2180-2657	954	35.74	32.18
	8 1/2	2657-3800	2287	3800 -4119	639	10.33	9.86
15/6-11 S	17 1/2	690 - 1730	2081	1730-2181	903	78.74	86.38
	12 1/4	2182-3320	2278	3320-3817	994	23.74	25.50
15/6-9 S	17 1/2	753-2180	2855	2180-2785	1212	44.88	30.78
	8 1/2	2786-3590	1609	3590-3942	705	30.78	44.26
16/1-4	17 1/2	371-1145	1531	114-1477	666	64.86	64.26
	12 1/4	1478-2002	1049	2002-2227	452	21.24	20.35
16/2-7	17 1/2	700-1450	1481	1450-1772	644	31.99	31.37
16/2-13 A	12 1/4	717-1955	2441	1955-2487	1064	26.97	12.03
Average						**36.93**	**35.70**

Error $<$ 15% , Error 15 $-$ 35%, and Error $>$ 35%

The model testings in Test 1 were carried out using dataset from wells in Gina Krog and Ivar Aasen field. In each well, the dataset of each hole section were split into 70% of training data and 30% of test data. The training data was taken from the top depth of a hole section down to 70% of the total depth of a hole section, while the rest 30% was set as testing data, see illustration in Fig. 5. The scheme of dataset allocation was adjusted to be in-line with the purpose of this current study. Even though the gamma-ray value is independent of depth, this scheme was made to correspond the process of real-time prediction in practice, with details explained in Chapter 6. The model testing result from Test 1, with total of 10 cases, is shown in Table 4. The misclassification rates for this test were fairly low, reaches down to ±10%, and the most often returned misclassification rate is ± 31% for both applied rules. Meanwhile, there are only two cases had high misclassification rates over 60%.

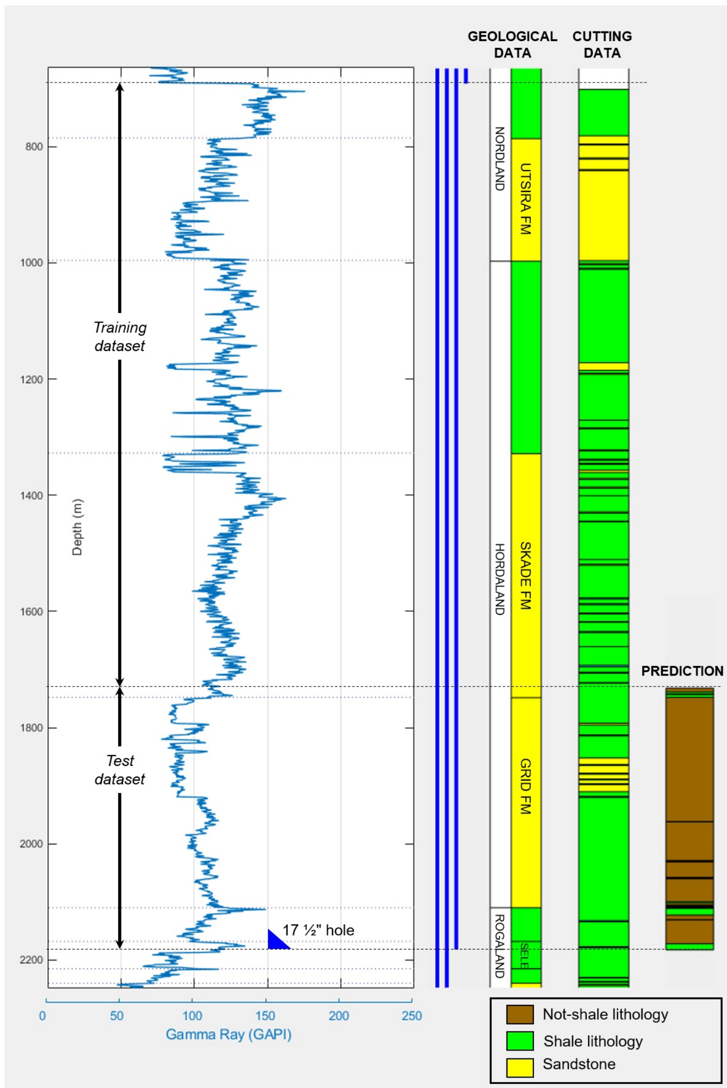


Fig. 5: Data division of training and test dataset of well 15/6-11S 17 1/2". The training dataset was taken from Nordland group and upper part of Hordaland group, while test dataset was from lower part of Hordaland group. Most shale layers in the Grid formation were poorly predicted as not-shale using rule due low gamma-ray reading.

The results showed that the high misclassification rates were mainly occured on tests that contain shale-sandstone layers, found in test 15/6-11 S (17 1/2") – as shown in Fig. 5 – and 16-1/4 (17 1/2"), specifically on the Grid formation which is the member of Hordaland Group. The geological information confirmed that Grid formation has a soft sediment deformation that produces sand bodies with poor connectivity. This finding also suggested that the sand beds mixed with the shale beds, which are the main lithology of Hordaland Group. Hence, the shale in Grid formation had lower gamma-ray compared to other shale beds from other formations within the Hordaland Group.

In general, the application of rule #2 decreased the average misclassification rates compared to the application of rule #1. However, the accuracy improvement was not significant. In addition, the application of this rule did not meet our expectancy to improve prediction on complex shale-sandstone bed. When the rule was applied to test well 16-1/4, the misclassification rate only decreased by 0.6%, and when applied to test well 15/6-11 S, the misclassification rate only increased by 7%. In the latter case, the increasing misclassification was due to false lithology data from geological interpretation, as seen in the geological data of Grid formation in Fig. 5.

5.2. Test 2

The models for Test 2 were trained using the complete dataset of each hole section of three wells from Gina Krog field. Then, the models were tested using dataset from: (a) the neighboring wells located in the same field as the models, Gina Krog field, and (b) wells located in another field, Ivar Aasen field.

Table 5: Misclassification rates of the first test in Test 2, with test set from Gina Krog field

Model	**Hole size**	**Rule #1**			**Rule #2**		
		15/5-7A	15/6-11S	15/6-9 S	15/5-7A	15/6-11S	15/6-9 S
15/5-7 A	17 1/2"	N/A	58.45	40.56	N/A	65.04	34.55
	8 1/2"	N/A	26.93	30.26	N/A	26.93	29.05
15/6-11 S	17 1/2"	25.37	N/A	30.28	25.56	N/A	32.19
15/6-9 S	17 1/2"	20.60	44.41	N/A	21.00	51.33	N/A
	8 1/2"	21.06	29.14	N/A	22.94	42.60	N/A
Average			**32.706**			**43.278**	

Error $<$ 15% , Error 15 $-$ 35%, and Error $>$ 35%

From testing the models with the dataset from Gina Krog field (Table 5), more than half of the cases returned misclassification rate below 30.5% for both applied classification rules. Misclassification rates above 35% were mostly found when testing dataset from Well 15/6-11 S, especially on hole size 17 1/2". A consistent misclassification was found for Skade and Grid formation with shale misclassified as sandstone. Even though all models of 17 1/2" hole section were

Table 6: Misclassification rates of the second test in Test 2, with test set from Ivar Aasen field

Model	**Hole size**	**Rule #1**			**Rule #2**		
		16/1-14	16/2-7	16/2-13A	16/1-14	16/2-7	16/2-13A
15/5-7 A	17 1/2”	32.29	62.38	N/A	30.87	60.17	N/A
	8 1/2”	50.55	75.53	53.42	52.42	76.03	57.27
15/6-11 S	17 1/2”	25.46	36.86	N/A	24.86	32.77	N/A
	12 1/4”	11.47	36.50	21.35	16.07	35.52	23.57
15/6-9 S	17 1/2”	32.56	54.24	N/A	30.92	49.29	N/A
	8 1/2”	40.56	70.29	45.71	56.16	80.52	70.75
Average		**35.119**			**46.479**		

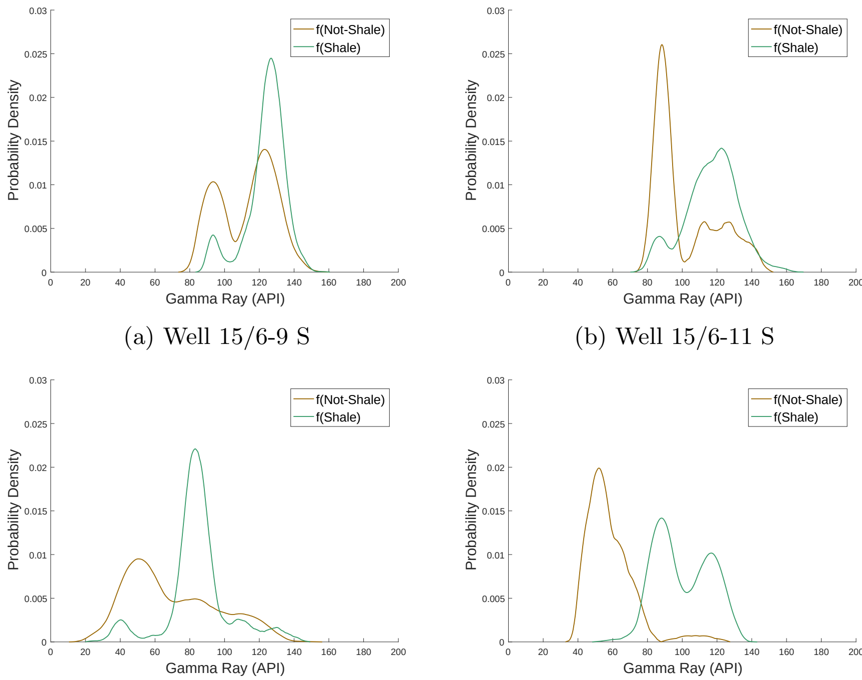
Error $< 15\%$, Error $15 - 35\%$, and Error $> 35\%$

also trained using dataset from Grid formation, the prediction on this shaly sandstone section was still challenging. Meanwhile, the accuracy of prediction from the application of rule #2 in most cases did not improve significantly and the averaged misclassification rate even increased compared to the results with rule #1 applied.

Less accuracy was observed when the models were tested using the dataset from Ivar Aasen field, with more than half of cases returned misclassification over 35 % (Table 6). In the most cases, the false prediction was due to shale misclassified as the not-shale lithology. Unlike the misclassification due to shaly-sandstone layers in the previous case, the misclassification in the current case was mainly due to the difference of gamma-ray data distribution between the models and test dataset. Comparing the gamma-ray probability density function of Hordaland formation group from wells at Gina Krog and Ivar Aasen field, we found that the shale reading from wells in Ivar Aasen was generally lower than wells in Gina Krog, see Fig. 6. In addition, the peaks of probability densities for both lithologies lie down on the different gamma-ray values, and the data range for each lithology was different. The discrepancy was presumed due to the sensitivity of the tool factors to the borehole environments. Indeed, it is common that wells in one field are exclusively drilled and logged in similar manners, but it is rarely done for wells in different fields. Hence, factors such as tool diameter and offset, mud weight, and cement thickness, caused inconsistency of gamma-ray reading from field to field.

5.3. Summary of results

Several lessons from tests above were learned regarding the automated lithology prediction method with gamma ray log. First, the method was successfully applied on univariate variable of gamma-ray and produced models that predicted lithology in two different tests with fair accuracy. In addition, we observed that the models in both test had high sensitivity to capture the change of thin lithology layers, as shown in Fig. 7. Second, the current models were limited by the



(a) Well 15/6-9 S

(b) Well 15/6-11 S

(c) Well 16/1-14

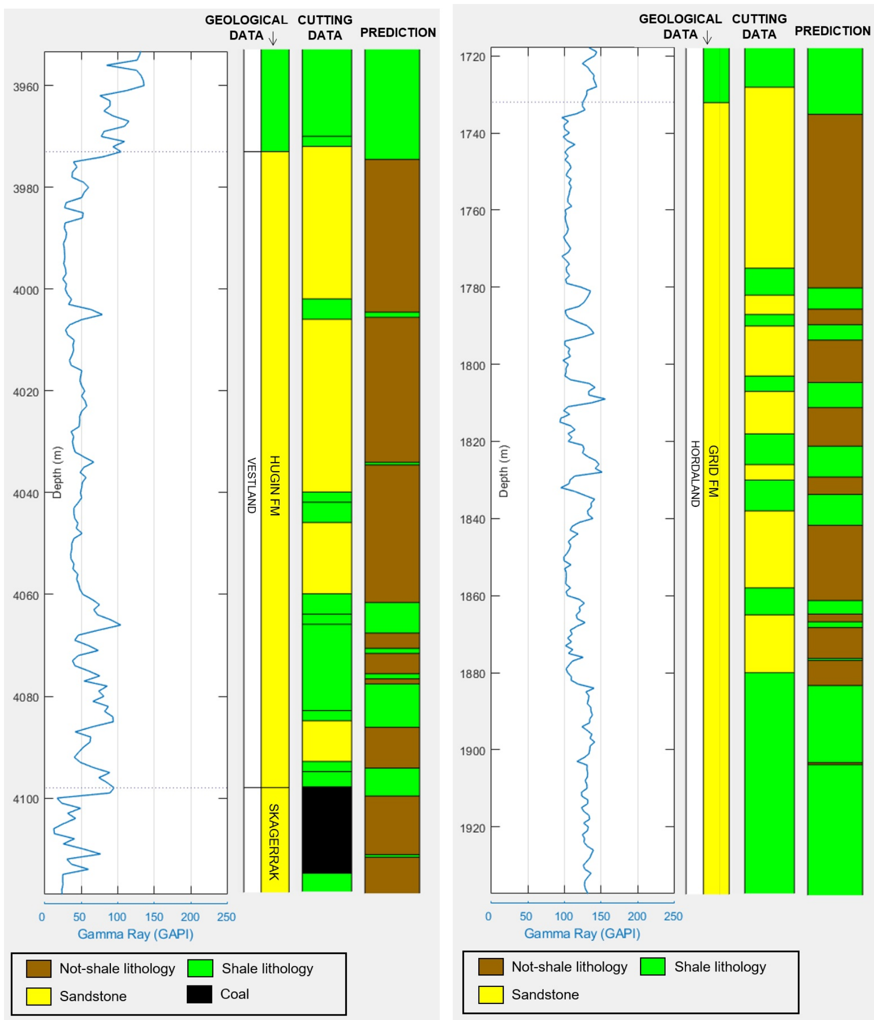
(d) Well 16/2-7

Fig. 6: Gamma-ray distribution as probability density function of Hordaland formation group in wells from Ivar Aasen- (a & b) and Gina Krog (c & d) field, estimated using kernel density estimation

tool sensitivity from borehole environmental factors. Without including those error factors in the models, the prediction would only be valid for wells with the same hole size or wells from the same field as the models. Another source of weakness in the models was the prediction limitation on the complex lithology, such as shale-sandstone mixture, that relatively had low gamma-ray reading. Lastly, the contribution of geological interpretation to increase prediction accuracy was not significant, especially in Test 2. It was still unclear whether the biggest cause was due to the poor lithology estimation from geological interpretation, or the large testing dataset size that reduced the sensitivity of prior probability, or the combination of both.

6. Application of lithology prediction method in practice

The tests above proved that models developed using the proposed algorithm could give accurate prediction, and hence the method is valid to be implemented in practice. The implementation can be done in multiple ways depending on the problems to be solved. In this paper, however, we highlighted the application in the most excellent way the proposed method can offer.



(a) Test 1, Well 15/5-7 A S, 8½"

(b) Test 2, Well 15/5-7 A S from model Well 15/6-9 S, 17½"

Fig. 7: Lithology predicition on thin layers: a) layers of shale (4,000-4,500 m) and sandstone-shale-coal (4,085-4,115 m) and b) layers of shale-sandstone (1,735-1,885 m), are predicted correctly.

The application of non-parametric technique within the method means that the modeling can be processed continuously to update the classification models everytime new elements of training data are observed. This type of modeling is very suitable for any operation in the field that implements mud-pulse telemetry system to obtain real-time data from borehole. Such as in drilling operation, the training data can be taken from the real-time log reading of the drilled section that the lithology has been verified with valid information, such as cutting visualizations. As the drilling ongoing on a particular well and the models are updating, the prediction can be made for the undrilled section of the same well. The process of prediction following the suggested approach was reasonably represented by the data employment in Test 1, where the prediction was made using training dataset taken from the same well. In Test 1, the training data from the 70 % of the uppermost depth can be presumed as the drilled formation, while the test data from the 30 % of remaining depth can be presumed as the undrilled formation.

Furthermore, the modeling can also be achieved without using real-time training data, for example by using history data from the neighboring wells. This way of application was closely represented by the process in Test 2 that used the training data from the neighboring wells for prediction. This approach of modeling can be applied to aid the prediction from the real-time data modeling, specifically at the beginning of real-time operation when the size of training data is insufficient to be modeled.

7. Next steps

A number of possible future studies using the proposed algorithm are apparent. In the next step, it would be necessary to improve and develop the method by modeling more explanatory variables using more sophisticated techniques of kernel density estimation (Hovda, 2014). Adding and combining more variables would enhance the features of each lithology, especially for complex mixture, such as shaly sandstone. For example, spectral gamma-ray log is relevant for describing the feature of mineral contents, while resistivity log is relevant for describing the feature of fluid contents. Therefore, the dimension of lithology groups that are inspected can be increased.

A further investigation is suggested to examine the sensitivity of different logging tools toward error factors – such as drillstring mechanical effect, borehole quality, drilling fluid type. By acquiring the error factors, corrections can be included together in the algorithm, and automatically assigned during the modeling. Therefore, the prediction made by models will not be subjective for specific conditions, such as hole sizes or well location. Lastly, a greater focus on applying the method in practice, as suggested in the previous chapter, could provide definite evidence of the method's effectivity.

8. Conclusion

An automated lithology prediction method was outlined in this paper. A univariate version that uses the gamma-ray log was evaluated in terms of its misclassification rates. Among the run tests, the most accurate predictions were found for gamma-ray models to predict: (a) dataset from the same well as the training data, as indicated in Test 1, and (b) dataset from the wells in the same field as the training data. More than half of the cases in the predictions mentioned above returned misclassification rate less than 31%. These results are viewed as meeting the initial goal of providing accurate lithology prediction using the developed method. Despite the good accuracy, the non-parametric technique applied in the method is suitable for data modeling without the need to set initial assumptions of training data distribution, allowing the models to expand. The method is believed to be an effective tool applied in the field, especially for real-time operation.

9. Acknowledgments

We thank Simen Steiro Strømslid for the help in digitizing the cutting data.

References

Benaouda, D., Wadge, G., Whitmarsh, R. B., Rothwell, R. G., MacLeod, C., 1999. Inferring the lithology of borehole rocks by applying neural network classifiers to downhole logs: an example from the ocean drilling program. Geophysical Journal International 136 (2), 477–491. URL `+http://dx.doi.org/10.1046/j.1365-246X.1999.00746.x`

Bowman, A. W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis. Oxford Statistical Science Series. Oxford University Press.

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

Cios, K. J., Pedrycz, W., Swiniarski, R. W., Kurgan, L., 2007. Data Mining: A Knowledge Discovery Approach. Vol. 26. Springer US.

Coudert, L., Frappa, M., Arias, R., 1994. A statistical method for litho-facies identification. Journal of Applied Geophysics 32 (2), 257 – 267. URL `http://www.sciencedirect.com/science/article/pii/0926985194900264`

Cuddy, S., et al., 1997. The application of the mathematics of fuzzy logic to petrophysics. In: SPWLA 38th Annual Logging Symposium. Society of Petrophysicists and Well-Log Analysts.

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

Ellis, D. V., Singer, J. M., 2007. Well logging for earth scientists. Vol. 692. Springer.

Frigge, M., Hoaglin, D. C., Iglewicz, B., Feb. 1989. Some Implementations of the Boxplot. The American Statistician 43 (1), 50. URL `http://www.jstor.org/stable/2685173?origin=crossref`

Hovda, S., 2014. Using pseudometrics in kernel density estimation. Journal of Nonparametric Statistics 26 (4), 669–696. URL `https://doi.org/10.1080/10485252.2014.944524`

Maiti, S., Krishna Tiwari, R., Kmpel, H.-J., 2007. Neural network modelling and classification of lithofacies using well log data: A case study from ktb borehole site. Geophysical Journal International 169 (2), 733–746. URL `+http://dx.doi.org/10.1111/j.1365-246X.2007.03342.x`

Mann, H. B., Whitney, D. R., Mar. 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. URL `http://projecteuclid.org/euclid.aoms/1177730491`

Mwenifumbo, C., Sep. 1993. Kernel Density Estimation in The Analysis and of Borehole Geophysical Data. SPWLA-1993-v34n5a3.

Mwenifumbo, C., Elliott, B., Jefferson, C., Bernius, G., Pflug, K., 2004. Physical rock properties from the athabasca group: designing geophysical exploration models for unconformity uranium deposits. Journal of Applied Geophysics 55 (1), 117 – 135, non-Petroleum Applications of Borehole Geophysics.
URL http://www.sciencedirect.com/science/article/pii/S0926985103000740

Norwegian Petroleum Directorate, 2017. Norwegian petroleum directorate factmaps. http://gis.npd.no/factmaps/html_21/, accessed: 2017-01-05.

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

Saggaf, M. M., Nebrija, E. L., 2003. A fuzzy logic approach for the estimation of facies from wire-line logs. AAPG Bulletin 87 (7), 1223.
URL +http://dx.doi.org/10.1306/02260301019

Schlumberger Educational Services, 1989. Schlumberger: Cased Hole Log Interpretation Principles/Applications. Houston.

Scott, D. W. (Ed.), Aug. 1992. Multivariate Density Estimation. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.
URL http://doi.wiley.com/10.1002/9780470316849

Silverman, B. W., 1986. Density estimation for statistics and data analysis. Vol. 26. CRC press.

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

Steel, R., Felt, V., Johannesson, E., Mathieu, C., 1995. Sequence Stratigraphy on the Northwest European Margin. Norwegian Petroleum Society Special Publications. Elsevier Science.

Wolf, M., Pelissier-Combescure, J., Jan. 1982. Faciolog - Automatic Electrofacies Determination. In: SPWLA-1982-FF. Society of Petrophysicists and Well-Log Analysts, SPWLA.

Ye, S.-J., Rabiller, P., Jan. 2000. A New Tool For Electro-Facies Analysis: Multi-Resolution Graph-Based Clustering. SPWLA.