UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

VINICIUS MEDEIROS GRACIOLLI

# A Novel Classification Method Applied to Well Log Data Calibrated by Ontology-based Core Descriptions

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Profa. Dra. Mara Abel
Advisor

Porto Alegre, February 2018

**CIP – CATALOGING-IN-PUBLICATION**

Graciolli, Vinicius Medeiros

A Novel Classification Method Applied to Well Log Data Calibrated by Ontology-based Core Descriptions / Vinicius Medeiros Graciolli. – Porto Alegre: PPGC da UFRGS, 2018.

64 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2018. Advisor: Mara Abel.

I. Abel, Mara. II. Título.

*“Welcome to the show,*
*the great finale is finally here.”*
— Blackie Lawless

# ACKNOWLEDGMENTS

I would like to thank first and foremost my advisor Mara Abel for all her support prior and during this work. I would also like to thank my colleagues Eduardo, Lucas, Júlia, Renata and Elias who helped me process the data cordially ceded by ANP and Petrobras. This work would also not be possible without the help of Luiz Fernando De Ros who provided support with the geological part of this work.

I would also like to thank Jean-François Rainaud and his wife Marie-Laure for being such gracious hosts and providing me with the opportunity to spend a wonderful time in France.

# CONTENTS

[1] STRATALEDGE is a trademark of ENDEEPER Co.

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| KNN | K-Nearest Neighbor |
| NN | Neural Network |
| BNN | Bayesian Neural Network |
| HMM | Hidden Markov Model |
| SVM | Support Vector Machine |
| LDA | Linear Discriminant Analysis |
| CRF | Conditional Random Field |
| NB | Naïve Bayes |
| ABD | Automatic Bedding Discriminator |
| DT | Sonic interval transit time |
| GR | Gamma Ray |
| NPHI | Neutron porosity |
| RHOB | Bulk density |
| DRHO | Density correction |
| ILD | Induction resistivity |
| LAS | Log ASCII Standard |

# LIST OF FIGURES

# LIST OF TABLES

# ABSTRACT

A method for the automatic detection of lithological types and layer contacts was developed through the combined statistical analysis of a suite of conventional wireline logs, calibrated by the systematic description of cores.

The intent of this project is to allow the integration of rock data into reservoir models. The cores are described with support of an ontology-based nomenclature system that extensively formalizes a large set of attributes of the rocks, including lithology, texture, primary and diagenetic composition and depositional, diagenetic and deformational structures. The descriptions are stored in a relational database along with the records of conventional wireline logs (gamma ray, resistivity, density, neutrons, sonic) of each analyzed well. This structure allows defining prototypes of combined log values for each lithology recognized, by calculating the mean and the variance-covariance values measured by each log tool for each of the lithologies described in the cores. The statistical algorithm is able to learn with each addition of described and logged core interval, in order to progressively refine the automatic lithological identification.

The detection of lithological contacts is performed through the smoothing of each of the logs by the application of two moving means with different window sizes. The results of each pair of smoothed logs are compared, and the places where the lines cross define the locations where there are abrupt shifts in the values of each log, therefore potentially indicating a change of lithology. The results from applying this method to each log are then unified in a single assessment of lithological boundaries.

The mean and variance-covariance data derived from the core samples is then used to build an n-dimensional gaussian distribution for each of the lithologies recognized. At this point, Bayesian priors are also calculated for each lithology. These distributions are checked against each of the previously detected lithological intervals by means of a probability density function, evaluating how close the interval is to each lithology prototype and allowing the assignment of a lithological type to each interval.

The developed method was tested in a set of wells in the Sergipe-Alagoas basin and the prediction accuracy achieved during testing is superior to classic pattern recognition methods such as neural networks and KNN classifiers. The method was then combined with neural networks and KNN classifiers into a multi-agent system. The results show significant potential for effective operational application to the construction of geological models for the exploration and development of areas with large volume of conventional wireline log data and representative cored intervals.

**Keywords:** Core-log integration, geophysical log, core description, lithology interpretation.

# RESUMO

Um método para a detecção automática de tipos litológicos e contato entre camadas foi desenvolvido através de uma combinação de análise estatística de um conjunto de perfis geofísicos de poços convencionais, calibrado por descrições sistemáticas de testemunhos.

O objetivo deste projeto é permitir a integração de dados de rocha em modelos de reservatório. Os testemunhos são descritos com o suporte de um sistema de nomenclatura baseado em ontologias que formaliza extensamente uma grande gama de atributos de rocha. As descrições são armazenadas em um banco de dados relacional junto com dados de perfis de poço convencionais de cada poço analisado. Esta estrutura permite definir protótipos de valores de perfil combinados para cada litologia reconhecida através do cálculo de média e dos valores de variância e covariância dos valores medidos por cada ferramenta de perfilagem para cada litologia descrita nos testemunhos. O algoritmo estatístico é capaz de aprender com cada novo testemunho e valor de log adicionado ao banco de dados, refinando progressivamente a identificação litológica.

A detecção de contatos litológicos é realizada através da suavização de cada um dos perfis através da aplicação de duas médias móveis de diferentes tamanhos em cada um dos perfis. Os resultados de cada par de perfis suavizados são comparados, e as posições onde as linhas se cruzam definem profundidades onde ocorrem mudanças bruscas no valor do perfil, indicando uma potencial mudança de litologia. Os resultados da aplicação desse método em cada um dos perfis são então unificados em uma única avaliação de limites litológicos.

Os valores de média e variância-covariância derivados da correlação entre testemunhos e perfis são então utilizados na construção de uma distribuição gaussiana n-dimensional para cada uma das litologias reconhecidas. Neste ponto, probabilidades a priori também são calculadas para cada litologia. Estas distribuições são comparadas contra cada um dos intervalos litológicos previamente detectados por meio de uma função densidade de probabilidade, avaliando o quão perto o intervalo está de cada litologia e permitindo a atribuição de um tipo litológico para cada intervalo.

O método desenvolvido foi testado em um grupo de poços da bacia de Sergipe-Alagoas, e a precisão da predição atingida durante os testes mostra-se superior a algoritmos clássicos de reconhecimento de padrões como redes neurais e classificadores KNN. O método desenvolvido foi então combinado com estes métodos clássicos em um sistema multi-agentes. Os resultados mostram um potencial significante para aplicação operacional efetiva na construção de modelos geológicos para a exploração e desenvolvimento de áreas com grande volume de dados de perfil e intervalos testemunhados.

**Palavras-chave:** integração perfil-testemunho, perfil geofísico, descrição de testemunho, interpretação litológica.

# 1 INTRODUCTION

This work details the development of a novel technique for data classification applied to geological data. This classification consists of associating numerical data samples with classes representing specific rock types. The framework developed attempts to emulate the behavior of the geologist commonly assigned to this task by segmenting data prior to the use of a classification algorithm; as opposed to classifying the samples independently, as it is commonly done in applications that attempt to solve this task automatically.

In the process of petroleum exploration and development, massive amounts of data are generated with the goal of locating possible oil reservoirs and determining their viability. This data can span scales ranging from the continental to the microscopic. When a rock formation containing oil reserves is found, this data is used to build 3D reservoir models that can be used to simulate the production potential of the reservoir. The goal of this work is to ultimately enrich these models by providing a tool that can be used to infer additional information from the available data, which can then be added to the model in locations where this information is missing.

The main datasets used to create these models come from seismic imaging and well data. Seismic imaging consists of taking images of the subsurface of the selected location that are produced using seismic echography; by laying acoustic sensors on the ground at regular intervals and detonating explosive charges these subsurface images can be constructed by measuring the time it takes for the acoustic waves to penetrate the rock formation and be reflected back to the sensors. These sensors can be spread across as several square kilometers, and with these images, which reflect the density and shape of the underlying rock formations.

Well data includes geophysical information and rock samples. This geophysical information is presented as wireline logs, which are numerical measurements taken from tools lowered down the well during or after perforation. These tools can gauge many different properties of the rock formation such as resistivity, conductivity, sonic transit time or gamma ray emissions, producing reports such as the one depicted in figure 1.1 that shows a single well with several geophysical measures aligned by depth. With these logs, properties such as porosity, grain size or the presence of fluids can be inferred based on the response of the log. When there are several available wells in the same sedimentary basin, they can be integrated to support the generation of a 3D model of the reservoir.

The recovery of rock samples allows direct access to rocks located many hundreds of meters underground. There are many types of rock samples that can be recovered and analyzed; detritus that leaves the borehole during drilling for instance, can inform the type of rock present in the drilled area within a reasonably precise depth. More detailed samples are also recovered from the well during and after drilling, such as sidewall cores, which are samples taken from wall of the borehole via mechanical or percussion drilling

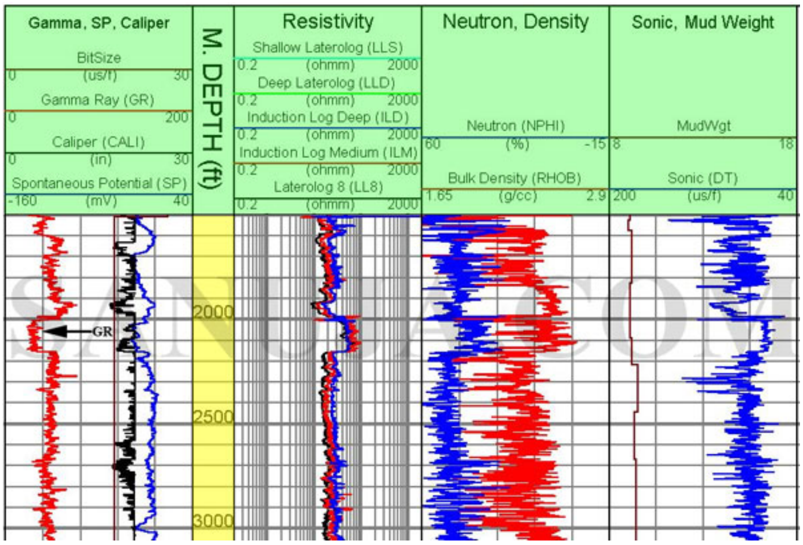


Figure 1.1: Example of a well log report. Each line represents a different geophysical measurement, as outlined in the image header. The vertical axis measures the depths at which the measurements were taken. Source (SENANYAKE, 2016. Available at http://sanuja.com/blog/what-is-a-well-log. Accessed Mar 22 2018)

at specific depths. Objective measurements can also be taken by submitting a rock sample to laboratorial analysis, to calculate attributes such as density, microporosity and permeability.

The most valuable sample that can be extracted from the well however, is the core sample, which is a whole cylindrical section of the borehole which is extracted during drilling without being destroyed in the process. Figure 1.2 shows a box containing core samples collected from a well for further description. The recovery of a core sample is an expensive process, since it requires the drilling to stop, and the regular drilling head to be replaced by a special drill with a hollow center to allow for the recovery of the core. For that reason, a borehole spanning several kilometers may only have a couple hundred meters of cores recovered. As the core is retrieved relatively intact, it can offer important information about the structure of the rock formation, as well as allow the analysis of the contacts between each layer of rock. These samples can be analyzed by a geologist which generates a qualitative description of the rock formation based on the perceived characteristics of the rock sample. The qualitative description is a hand-made task that produces reports similar to those presented in Fig XX

This work focuses on the correlating lithology data, which is a qualitative assessment of the rock characteristics recorded in the core descriptions, with readings taken from wireline logs. The goal is to be able to extrapolate these rock characteristics based on the much more common wireline log data, which can span most of the depth of the well. This is a pattern recognition problem, where a series of numerical measurements taken at a specific depth in the borehole are labeled as a lithology, based on the readings taken at depths with previously known lithologies. This work uses a novel approach where prior to being labeled, the samples are grouped based on their context within the borehole; by detecting sudden changes in the log values, likely depths at which a lithological change may occur set the boundaries between these groups. This grouping process is shown to increase the prediction accuracy when used in tandem with the method developed to label the samples analyzed.

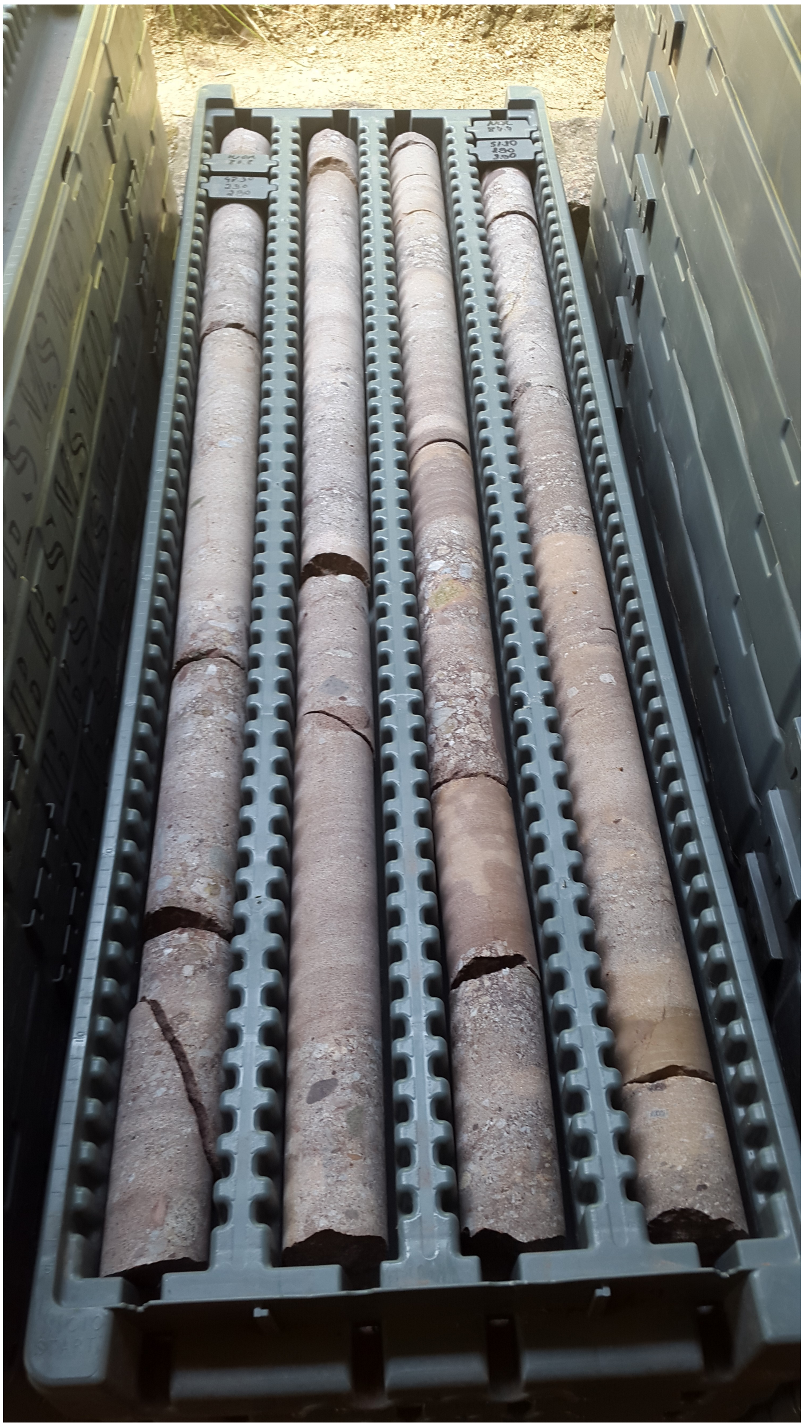
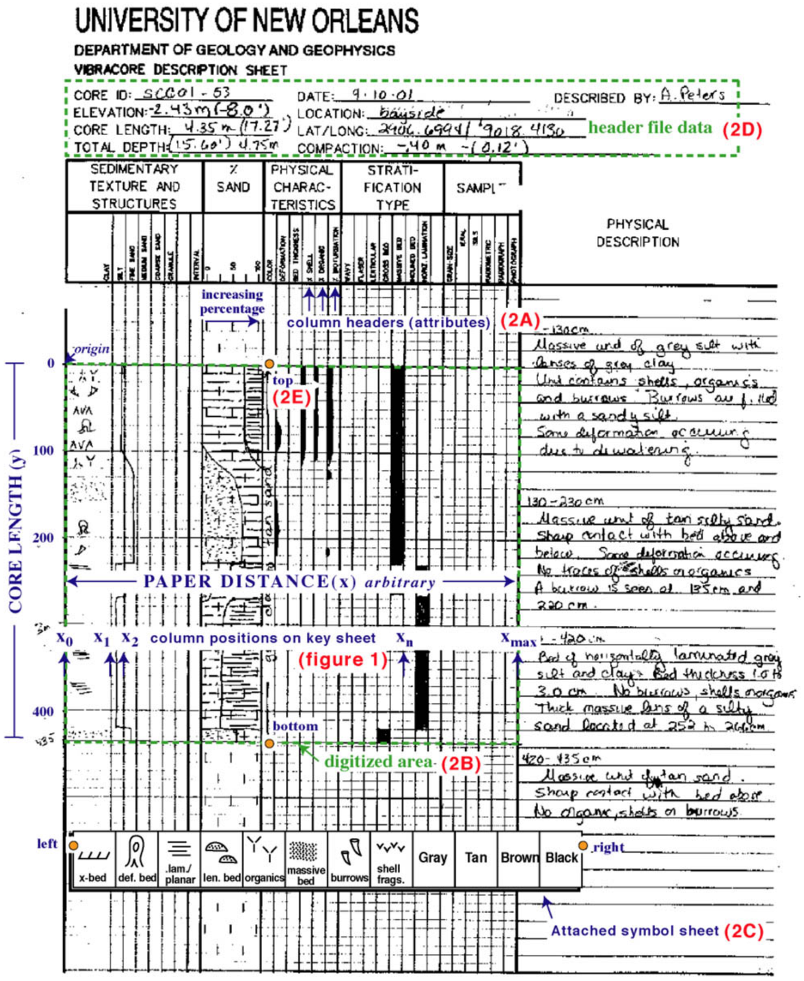


Figure 1.2: Example of a box containing a sequence of core samples labeled by depth.



UNIVERSITY OF NEW ORLEANS
DEPARTMENT OF GEOLOGY AND GEOPHYSICS
VIBRACORE DESCRIPTION SHEET

CORE ID: SCC01 - 53
ELEVATION: -2.43m (-8.0')
CORE LENGTH: 4.35 m (17.27')
TOTAL DEPTH: (15.60') 4.75m
DATE: 9.10.01
LOCATION: bayside
LAT/LONG: 2906.6994 / 9018.4130
COMPACTION: -.40 m -(0.12')
DESCRIBED BY: A. Peters

PHYSICAL DESCRIPTION

- 130cm
Massive unit of grey silt with lenses of grey clay
Unit contains shells, organics and burrows. Burrows are filled with a sandy silt.
Some deformation occuring due to dewatering.

130 - 230 cm
Massive unit of tan silty sand. Sharp contact with bed above and below. Some deformation occuring.
No traces of shells or organics
A burrow is seen at 135cm and 220 cm.

- 420 cm
Bed of horizontally laminated grey silt and clays. Bed thickness 1.0 to 3.0 cm. No burrows, shells or organics
Thick massive lens of a silty sand located at 252 to 266cm

420 - 435cm
Massive unit of tan sand.
Sharp contact with bed above
No organics, shells or burrows.

Figure 1.3: Example of a non-standardized core description made with freehand drawings and text. The labels show column attributes (2A), area being digitized (2B - outlined green box), attached symbol menu (2C), header information (such as latitude, longitude, and elevation) (2D - dashed green box), and locations of first four digitized points: top and bottom of core, and left and right side of symbol menu (2E). Columns are recognized by their distance from origin (x-value), and vertical dimension (y) is either core length or depth of core penetration relative to a datum. Source: (USGS, 2010. Available at https://pubs.usgs.gov/ds/542/. Accessed Mar 22 2018.)

This work relies on three hypotheses. First, that we can increase the classification accuracy on lithology prediction by separating the data in segments corresponding to homogeneous rock sections. Second, that the boundaries between these rock sections should be detectable by a sudden change in log value. And third, that each lithology has a distinct log signature.

The method used for classification is a gaussian classifier based on supervised learning, where the inference process is calibrated by a training set composed of a set of wireline logs and core descriptions which provide the lithology labels for sampled depths in the wireline logs. The framework developed was trained and tested with real well data from the Sergipe-Alagoas basin, and the classification accuracy was tested against other classic machine learning methods for pattern recognition such as fuzzy classifiers, neural networks and k-nearest neighbors classifiers. The results show that the method developed achieves a classification accuracy comparable to or greater than these other algorithms.

By inferring this lithology data in well scale, this information can later be extrapolated to a larger scale by integrating this information into reservoir models built from seismic imaging in uncored sections, where there is no lithology information.

This work is organized in the following structure. First, the current methods for lithology identification methods are analyzed and their shortcomings are identified. Then we present a summary my previous work and how it relates to the method developed. Later, the datatypes which are used in this work are introduced and explained. Then, the data used for testing is presented and analyzed. Finally, the framework for lithology interpretation developed is presented, tested and compared to other classification methods, from which we draw our final conclusions.

# 2 STATE OF THE ART

Predicting lithology from wireline logs is not a novel idea; much work has been done in this area attempting to predict lithotypes from log values. Where most of these works fall short is when it comes to the testing data. Synthetic datasets are often used, and they can show a vastly different reality from the average borehole. Even when done with real data, samples may be classified not by actual inspection of the rock formation, but by grouping the samples through clustering algorithms.

Neural networks are a popular solution for pattern recognition problems such as this, which boils down to recognizing a pattern of log signatures and assigning a lithology value based on these measurements. Neural networks work as a series of nodes passing along numerical values through weighted connections. By training the network with a set of data containing inputs and expected outputs, these weights are adjusted by means of a backpropagation algorithm until the output of the neural network is in line with the expected output.

Reid (REID; LINSEY; FROSTICK, 1989) describes an approach that he has called the *Automatic Bedding Discriminator*, a method to detect boundaries between lithologies based on gamma ray logs. This is done by using moving means to detect sudden changes in the log data, which characterize a change in lithology. The method can then discriminate between rocks with larger or smaller grain sizes based on the deflection of the log at these change points, this is possible due to the gamma ray log used being highly sensitive to grain size. This method forms the basis of the first part of this work, which allows the separation of the logged interval into discrete sections.

Coudert (COUDERT; FRAPPA; ARIAS, 1994) uses gaussian distributions to build prototypes in a similar way to the method developed in this work. However, Coudert's gaussian distributions are all one-dimensional, as opposed to the multivariate distributions used in this work. Once the prototypes have been calculated by correlating log values and rock samples, they are compared to the prototypes through a probability density function. To increase prediction accuracy, Coudert also uses Bayesian priors and rules based on geological principles to determine lithology on uncertain situations. Due to the use of these rules, Coudert's method is not as flexible as this work when it comes to adaptation to different rock characteristics or other problems.

Brereton (BRERETON; GALLOIS; WHITTAKER, 2001) uses a clustering method that distributes sampled points in a color space based on the readings of the wireline logs, and then assigns a lithological significance to these points based on the area of the color space that was assigned to them. While it discerns the most contrasting changes in lithology with reasonable accuracy, it also detects a large number of lithological changes not described in the validation data. While the article claims that these are subtle variations not easily detected by the geologist doing the core description, it also means the results can only be truly validated against the rock sample themselves, not against descriptions that can not be revisited.

Li (LI; ANDERSON-SPRECHER, 2006) compared a naive Bayes classifier with linear discriminant analysis (LDA) and found both methods to perform adequately on a set of data from three well consisting of gamma-ray (GR), neutron porosity (NPHI), formation density (RHOB), and deep resistivity (LLD) logs and core descriptions from which five distinct facies, which were not entirely based on lithology were identified. The top prediction accuracy of Li's work reached 81.2% with the linear discriminant analysis approach.

Al-Anazi (AL-ANAZI; GATES, 2010) shows an approach based on a support vector machine (SVM). Support vector machines can generate mapping functions through supervised learning which allows the samples to be separated by a hyperplane in n-dimensional space. Al-Anazi's work focuses on predicting permeability, and suffers from the lack of hard rock data, as validation is made through comparisons with known electrofacies.

Gifford (GIFFORD; AGAH, 2010) uses neural networks, along with other learning algorithms such as k-nearest neighbors (KNN) classifiers in a multi-agent system. In this approach, the problem is solved by multiple independent modules, each using a different method. These results are then integrated into a final output, resulting in a higher accuracy than any individual method used. The complete system presented in the article achieves a top accuracy of 84.3%.

The problem with neural network based approaches is that the training is expensive in terms of processing power, requiring a large dataset to avoid overfitting. While it can produce good results, the series of calculations learned and performed by a neural network are often seen and treated as a black box, for peering inside it reveals a series of low level processes that are difficult to be understood by a human reader.

Bosch in (BOSCH; LEDO; QUERALT, 2013b) describes a fuzzy logic based method that was implemented as a MATLAB routine for the task of facies classification. In this method, membership functions are calculated from the training data in order to classify a validation set by measuring the degree of membership of sample to each lithology in a manner similar to the method developed in this work. In Bosch's work, his method was tested using synthetic data. Since the implementation of Bosch's work was made public, it has been tested with the data used in this work for the purposes of comparison with the method developed.

Ojha (OJHA; MAITI, 2013) presents a Bayesian Neural Network (BNN) based approach that optimizes the starting weights of his neural network, decreasing the training time needed until the neural network starts to produce reasonable results. The network is then trained using a data set derived from clustering and statistical analysis of wireline log data. In the presence of 10% red noise, the method presented in the article achieves an average accuracy of 67.38%.

Jeong (JEONG et al., 2014) uses a Hidden Markov Model (HMM) and a Conditional Random Field (CRF) based approach to tackle lithology prediction. A HMM can be seen as a sequential version of a naive Bayes (NB) classifier, which learns how to classify data through a joint distribution of the training data. On the other hand, a CRF can be seen as a sequential version of a logistic reversion classifier, which learns from conditional distribution. While Jeong's work managed up to 82% prediction accuracy, it was tested using synthetic data.

Some limitations are show to be very prevalent in the methods presented in the literature, namely the usage of low quality or synthetic data, and the fact that all these methods treat each data sample individually, ignoring the context in which they are inserted. The method described in this work introduces a new approach where the data is segmented into facies prior to classification. This segmentation is done using a moving-means based algorithm derived from Reid's work, which was adapted to work with multiple logs instead of relying solely on the gamma ray readings. These segments are then labeled by comparing the samples contained within against multivariate gaussian distributions derived from correlation between well logs and core descriptions in a training set. This approach has the benefit of increasing classification accuracy by analyzing the data samples within the context of a contiguous body of rock, instead of isolated datapoints.

# 3 PREVIOUS WORK IN THIS PROJECT

The framework described in this work bases its first step in the research previously presented in (GRACIOLLI, 2014), where wireline logs are used to determine lithological boundaries which can then be correlated to descriptions of core samples in order to correct any possible offsets between the core and log depths resulting from faulty data acquisition.

The method developed uses the boundary information obtained from application of that previous work as a way to enhance the accuracy of pattern recognition methods when applied to the task of lithological identification in a novel way which is not explored by the methods currently used in this task.

This segmentation method is based on the work done by Reid (REID; LINSEY; FROSTICK, 1989) which is briefly mentioned in the previous chapter, but expanded in order to accommodate working with multiple wireline logs, instead of being restricted to the gamma ray log. Reid's work, as well as the enhancements developed in (GRACIOLLI, 2014) are described in the next section.

## 3.1 Automatic Bedding Discriminator

Reid's boundary detection algorithm is dubbed the *Automatic Bedding Discriminator*. His method starts from the assumption that a lithological change can be characterized in the log reading by a sudden change in log values. Reid's work uses exclusively gamma ray logs to build a boundary assessment; which is one of the most common logs taken from a borehole. The Gamma Ray log responds to the organic matter content of the rock and has a high correlation to grain size, which makes it useful to detect intercalated sandstone-shale layers.

The first step in detecting these sudden changes is to deal with signal noise. Log readings can be affected by a wide number of variables such as borehole size or the composition of the drilling mud used during perforation. Even small scale changes in lithology that are not detected by the geologist on a core sample or do not characterize a clear change in lithology can be picked up by the logging tools and appear as slight changes in log value.

This noise is dealt with by applying a centered moving mean to the log data. The value of a centered moving mean at the data point $d_p$ with a window of size $n$ is defined as the average value of the $n$ closest data points to point $d_p$ (including $d_p$). This results in a loss of data, so care must be taken to not use a window that is too large, which will discard meaningful log features; nor a window that is too small, which won't effectively filter out variations induced by noise. Reid's work suggests a window of around 1m for this step, since information is rapidly lost with windows larger than 2m.

Next, a moving mean with a much larger window is applied to the original log, with the aim of deriving a curve that shows the general trend of the log; for this, Reid suggests a window of approximately 10m. By comparing both filtered logs, the points where there are sudden changes in log values can be determined by checking at which positions both logs intersect; in essence, where the log increases or decreases more than it's general trend. This is done simply by checking at each data point if the status quo of which filtered log is higher than the other is either maintained or inverted, if the previously lower-valued log turns becomes the higher-valued log, that means the logs have intersected. A visual representation of the result of this process can be seen in figure 3.1.

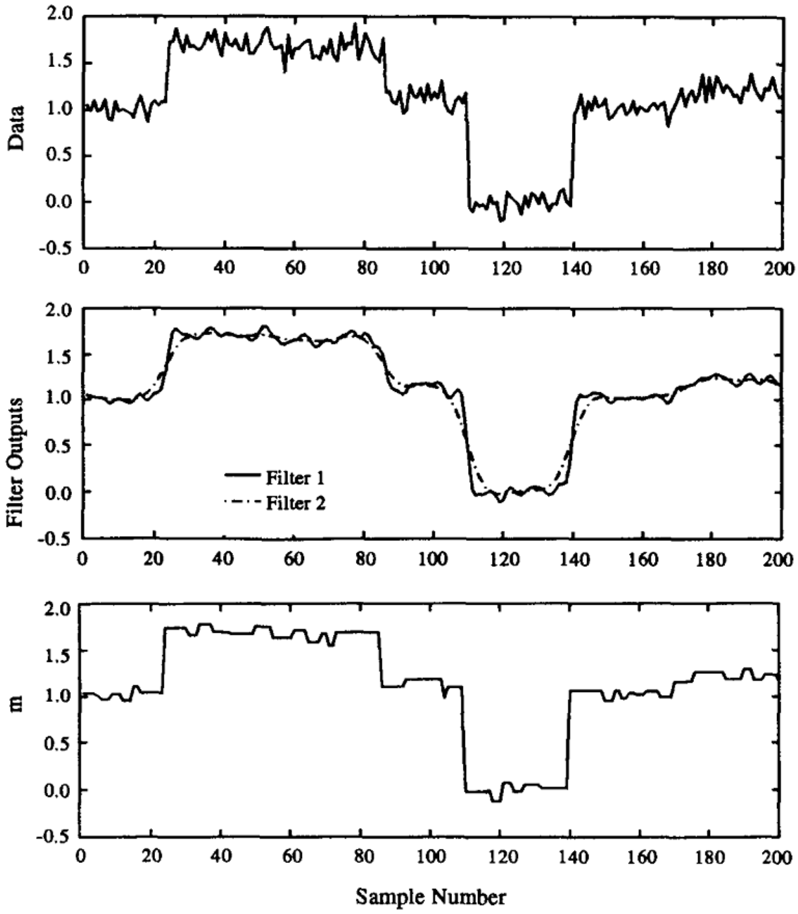


Figure 3.1: Result of moving means being used to detect change points in an arbitrary dataset. Filter 1 is obtained by applying a small moving window to the original data, Filter 2 by applying a larger one. The final result displayed is obtained by averaging all values between each intersection. Source (MACDOUGALL; NANDI, 1997)

Another consideration taken in *Reid's Automatic Bedding Discriminator* is with sections of the log where there is little change across a long section of measured depths; in these cases the values of both smoothed logs may be very similar, and very small alterations in the log value may register as an intersection between the smoother logs. In order to deal with this situation and cull the resulting false positives from the assessment, Reid establishes a threshold of 4 API[1] and disregards detected changes in lithology resulting from a change in log value lower than this threshold.

[1] API is a measurement originated from the petroleum industry, and is the standard unit of measurement for gamma ray logs

Reid's method then attempts to classify the facies found in the previous step of his algorithm by analyzing the deflection on the log curve at the points where a bedding contact was detected. By assessing the sign and magnitude of the change in log value, it is possible to estimate the increase or decrease in grain size which can differentiate between shale and sandstone.

## 3.2 Enhancements

Reid's *Automatic Bedding Discriminator* results in an assessment of break points for a single gamma ray log, but the method can be expanded in order to take into account other logs that may also have important lithological significance, such as porosity, density and resistivity logs (OJHA; MAITI, 2013); by applying it to multiple logs and then integrating the results in an unified assessment. Before this can be done, however, some issues must be taken into consideration: first, wireline logs have varying degrees of representativeness for lithological assessment (KRYGOWSKI, 2003); second, the same bedding contact is not likely to be detected at the exact same depth across multiple logs.

The first problem can be solved by assigning a weight $w$ to each log, which is checked against a user defined threshold $t$ when declaring break points: if the sum of the weights of all logs accusing a break at depth $d$ is equal or greater than the threshold $t$, we say there is a break at depth $d$. The ideal weights can be affected by the type of sedimentary terrain. Considering that the field available for validation in this work is mostly composed by siliciclastic rocks with minor presence of carbonates, the following guidelines for assigning weights were proposed by the inquired geologists:

- Gamma Ray logs are the most representative for lithological changes, and thus should have the highest weight.
- Density and porosity related logs are also highly representative and thus should have weights close or equal to highest weight.
- Resistivity logs should have weights around half of the highest value.
- The remaining logs are not representative enough for lithological assessment, and thus, should have weights equal to zero.

The second problem can be addressed by defining a window of size $ws$ around the breaks detected by each log, and then checking not which logs are accusing a break at a given depth $d$, but which logs have a window overlapping depth $d$. If enough logs have breaks sufficiently near a given depth, as defined by their overlapping windows, and the sum of weights $w$ of these logs are equal or greater than $t$, a break point is declared on the depth defined by the average of the depths of the logs involved, weighted by their respective $w$ weights. An example of this procedure can be seen in figure 3.2

Once the individual results are compared, and depths that pass the threshold test are declared as bedding contacts, we have an automated unified assessment of heterogeneities in the rock formation. This assessment is more accurate than one derived exclusively from the gamma ray log as it takes into account log characteristics other than organic matter content and grain size which are expressed in the gamma ray log.

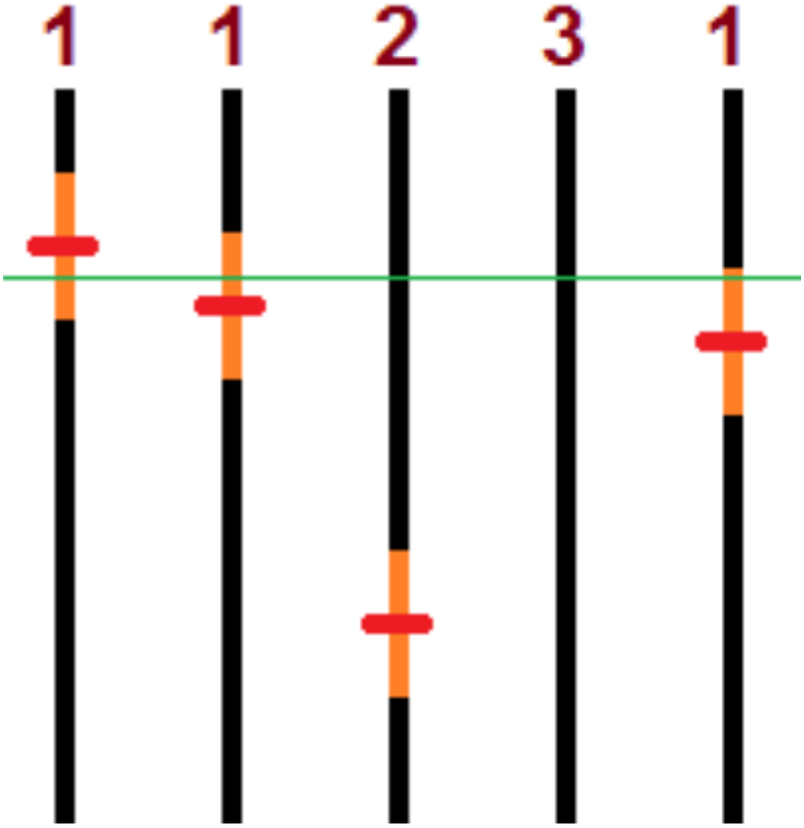


Figure 3.2: Example showing the integration of break point results. Each vertical black line represents a log from the same well, where break points were detected at the red markers. The numbers on top of each log corresponds to the weight of that log. The orange sections on the logs represent the depths encompassed by the agreement window $ws$. If we consider a weight of 3 to declare breaks, a break will be recorded in the final assessment on the depth corresponding to the mean depth of the three breaks crossed by the green line, as the sum of the three logs with breaks within that depth's tolerance window is greater or equal than 3.

# 4 DOMAIN ONTOLOGY OF SEDIMENTARY FACIES

The branch of philosophy known as ontology, which is sometimes equated to metaphysics, is the field of study which deals with the nature and structure of reality. Aristotle defined ontology as the study of attributes intrinsic to things (GUARINO; OBERLE; STAAB, 2009). As such, an ontological study is not concerned with modeling reality under a perspective constrained by data and experiments, but with providing a description of the things present in the domain of interest. This means it is completely valid to study the ontology of dragons for example, even though dragons are fictitious beasts, they can be described in terms of concepts and relations.

In the context of computer science and software engineering, an ontology can be seen as a data artifact which specifies the concepts and relations that exist within the universe available for a given information system. In an ontology describing dragons for example, the concepts needed to describe our universe would include *dragon*, *wings*, *scales*, *ability to breathe fire*, *hoard* and *hero*, along with the required relations to link these concepts, such as *possesses*, *guards* and *fights*. These concepts and relations are organized into a hierarchical taxonomy, with *drake* for example being a subclass or specialization of *dragon*. These general concepts can then be instantiated to refer to specific actors.

A computational ontology has been defined as "a formal, explicit specification of a shared conceptualization" (STUDER; BENJAMINS; FENSEL, 1998). To fulfill these requirements this means the ontology should be written in a language that is machine readable, so it is formal. This allows the ontology to be queried and parsed by an application, which allows the system to answer questions such as *"does an instance of dragon possesses a hoard?"* by analyzing the actors involved and concepts that link them. Many languages exist today to encode these ontologies, such as OWL, KIF and OntoUML.

This definition also requires that our concepts and relations should interpreted correctly and consistently so our specification can be explicit. The effective way to ensure this is to constrain the interpretation of the language used by the means logical axioms that allow the possible states of the universe in our specification to be modeled while also minimizing the possibility of modeling unintended, illegal states. For example, the relations *fights* and *possesses* can be differentiated by specifying the relation *fights* as irreflexive, intransitive and symmetrical and the relation *possesses* as irreflexive, intransitive and asymmetrical. This is reflected in the languages described earlier, which tend to be based on predicates and first-order logic.

The last point made in the definition presented is that the ontology should be shared. This means the concepts and relations specified in the model should express a consensus instead of an individual view; since as a collection of structured knowledge, an ontology is only useful if information it models is agreed upon by all the users. Many top-level ontologies and domain ontologies have been created with this purpose of knowledge sharing. Top-level ontologies such as DOLCE and UFO deal with describing the most basic concepts needed to represent and categorize various entities; they define for example, the difference between countable and uncountable subjects, concrete things versus abstract things.

A domain ontology can be defined as a collection of concepts and relations pertaining to a specific domain. Domain ontologies are usually extended from a known top-level ontology and deal with describing the concepts relevant to the domain of interest. We have for example the National Cancer Institute Thesaurus (NCIT) ontology, which defines over one hundred thousand terms related to the medical sciences. The core description data used in this work is backed by another domain ontology focused on the field of geology, this allows for standardized and unambiguous descriptions of the rock characteristics apparent in the samples.

When making core descriptions, geologists rely on drawings for expressing what they observe in the rock, since the available vocabulary for describing outcrops or rock samples is in many cases incomplete or ambiguous. In a previous project, it was studied how to deal with this visual knowledge in order to provide the best support for capturing sedimentary facies descriptions for stratigraphic interpretation, keeping in mind that computers require propositional information for processing.

## 4.1 The Strataledge®[1] Ontology

Lorenzatti in (ONTOLOGICAL PRIMITIVES FOR VISUAL KNOWLEDGE, 2010) proposed a hybrid representation approach for ontologies, which was demonstrated in a domain ontology for macroscopic description of sedimentary facies as a pair composed by an icon that visually resembles the visual aspect of sedimentary feature and a propositional descriptor, as can be seen in figure 4.1.

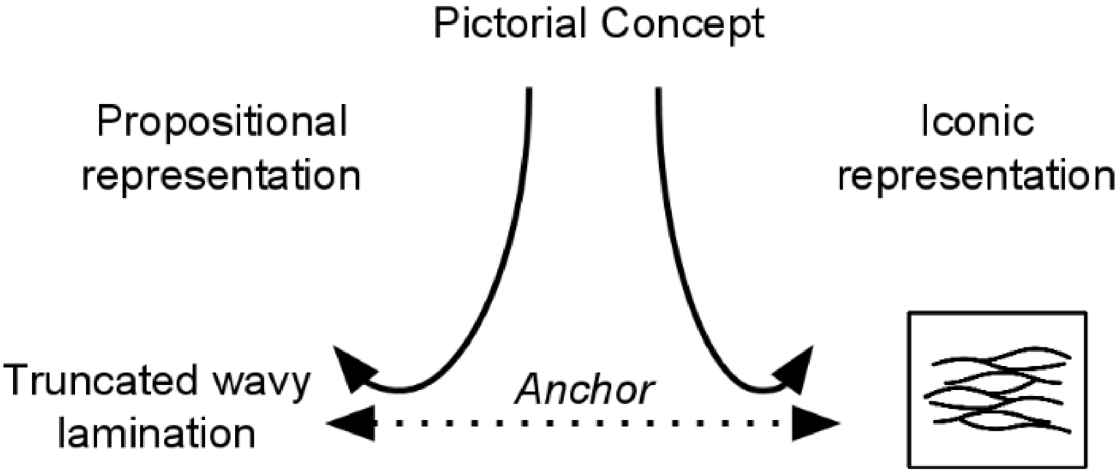


Figure 4.1: Representation model of sedimentary facies built by a propositional term and a pictorial icon. The icon resembles the visual aspect of the facies. Source: (ONTOLOGICAL PRIMITIVES FOR VISUAL KNOWLEDGE, 2010)

Later on, Endeeper (STRATALEDGE: CORE DESCRIPTION SYSTEM, 2012) took advantage of this proposal and formalized an extensive domain ontology for facies description of all types of rocks covering more than 750 geological features and 300 icons.

[1] STRATALEDGE is a trademark of ENDEEPER Co.

The ontology was developed following the principle of foundational ontologies (GUIZZARDI, 2005) and it covers all the textural, structural, palentological and lithological aspects of igneous, metamorphic and sedimentary rocks, including the metassomatic, cataclastic and chemical less common types. The descriptive capability of the formal vocabulary provides the needed semantic content for the geologist to capture the aspects of the rock for stratigraphic interpretation. We can see in figure 4.2 a small example of the knowledge model of sedimentary features.

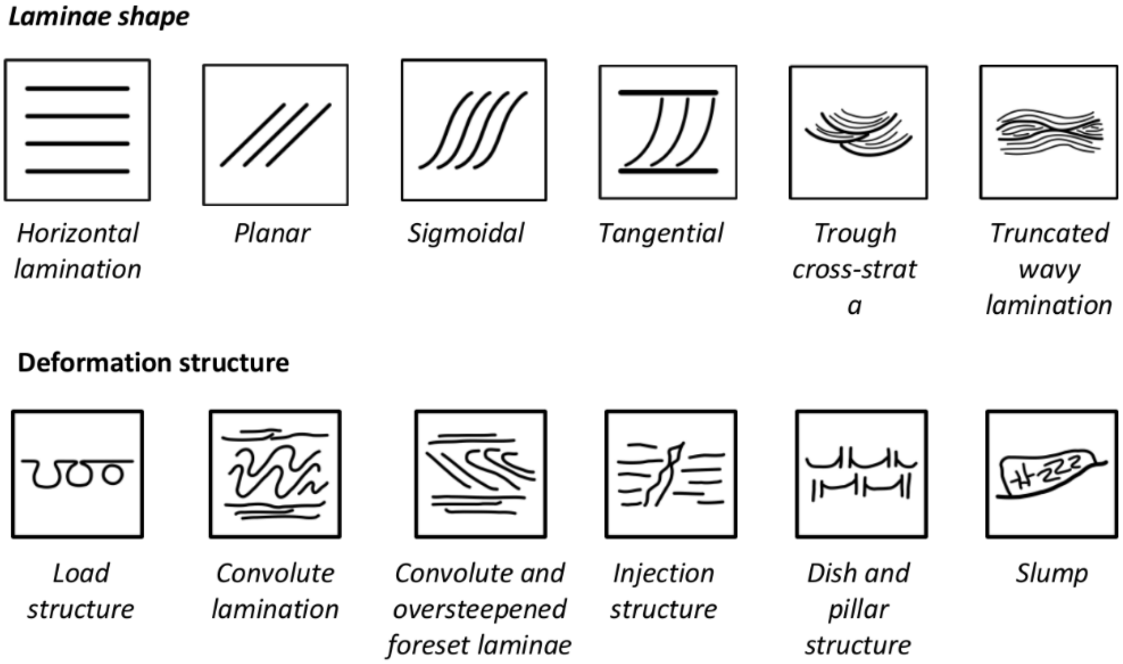


Figure 4.2: Examples of the rock facies domain ontology, showing the dual representation of geological features. Source: (ONTOLOGICAL PRIMITIVES FOR VISUAL KNOWLEDGE, 2010)

The ontology originally proposed by Lorenzatti was further extended and refined by Carbonera in (CARBONERA, 2012) to support automatic interpretation of depositional processes. The author describes the features that define a sedimentary facies which are further used to discriminate the sedimentary units. These features were used by Strataledge® in this work to segment and describe the core samples. A brief overview of the sedimentary facies concept and its attributes as defined by the ontology is presented in Annex 1.

This extensive controlled vocabulary is embedded in an application for description of cores and columnar outcrops. The Strataledge® system produces standardize descriptions that are stored as records in a database, eliminating ambiguities and reducing subjectivity of the description process. This capability allows computer algorithms to process the information extracting automatic geological interpretation like those described in this work.

The Strataledge descriptions can then be exported in a XML format that allows for easy data processing, or as an SVG profile image file that can be shared for human inspection. The core sample descriptions used in this work are expressed in the Strataledge format, which allows for easy correlation between the wireline logs and the core samples.

# 5 WIRELINE LOGS

Wireline logs are measurements taken from boreholes by using tools that may be lowered in one at a time or as a series of sequentially connected tools. These measurements can be either taken during drilling by using logging while drilling (LWD) methods where the logging tools are integrated into the drilling head, or more commonly; after the drilling is done and the tools are lowered into the well one at a time or as a series of connected tools.

When done after drilling, the logging is commonly done while the well is still uncased, which means it has not yet been cemented and the pipe has not yet been inserted. This gives the tool access to the bare rock, which results in less obstructions and more precise readings. While not as common, logs made after the borehole is cased are still possible and are sometimes performed.

This chapter describes the most commonly used well logs (KRYGOWSKI, 2003) and the structure of the file used to record the data resulting from the logging process.

## 5.1 The LAS File

The Log ASCII Standard (LAS) file, developed by the Canadian Well Logging Society(CRANGLE, 2007) is currently the industry standard for storage of wireline log data and is organized as such:

1. A header informing the version of the LAS file.

2. A section containing metadata, such as the identity of well that is logged, its geographical coordinates, elevation, logged depth, among others.

3. A section listing which logs are present in the file, each entry is composed of a mnemonic and an optional description.

4. A section containing the data itself, in the form of a list of space separated numerical values pertaining to each of the logs listed in the previous section at each of the logged depths.

## 5.2 Correlation Logs

The logs described in this section are usually used for correlation with other logs, as well as to differentiate between reservoir and non-reservoir formations.

### 5.2.1 Spontaneous Potential

Also known as SP, the spontaneous potential log measures the voltage from electrical currents resulting from the difference in salinities between water in the rock formation and the drilling mud in the well. As such, its values are can vary highly from well to well based on the drilling mud used. It can only be run in uncased wells and in the presence of water or water-based drilling mud.

This log can expresses the presence of a reservoir as a sharp change in value (either positive or negative) from an arbitrary yet stable baseline value. This log can also show the presence (but not the magnitude) of permeability in the rock formation. Depositional environment can also be inferred from the shape of the log curve. The presence of hydrocarbons or shale content in the rock formation will also cause a small deflection in the log value.

The mnemonic most often used for this log is SP, and the measurements are usually taken in millivolts (mV).

### 5.2.2 Gamma Ray

One of the most common logs, this tool measures the emission of gamma rays from naturally occurring thorium, potassium and uranium present in the rock formation. The tool may either take a single measurement from all these three elements, or discrete measurements from each one of them (in which case, it is referred to as a spectral gamma ray log). These tools have no restriction on cased or uncased boreholes, or the type of fluid present in the well.

Gamma ray measurements correlate to the amount of organic matter present in the rock formation. High values can therefore indicate a source rock, which is rich in organic matter, or a fracture where soluble uranium compounds have been deposited. Gamma ray readings are also highly correlated with shale content, and therefore, can be a good indicative of grain size.

The mnemonic used for this log is usually GR or some variation thereof. Spectral gamma ray logs may be divided in 3 logs, identified as THOR, URAN, POTA, or TH, U, K, based on the element being tracked. Measurements are taken in API or ppm.

### 5.2.3 Caliper

The caliper measures the diameter of the borehole, most commonly through the use of arms that extend from the tool. As with the gamma ray log, it is a very common log that has no operational constraints.

This log in particular has no direct correlation to rock type, and it is used mostly as input for environmental corrections on other logs.

The most common mnemonics used for the caliper log are CAL or CALI, measurements are taken in centimeters or inches.

## 5.3 Porosity Logs

The logs used in this section are mostly used to estimate the porosity of a given rock formation. It is important to note that none of these logs measure porosity directly, the estimation of porosity is usually obtained through the interpretation of the combination of two or three of these logs.

### 5.3.1 Sonic

This tool consists of a transmitter that emits sonic pulses that are then received by two or more receivers located on the same tool. The time differential between each receiver detecting the sonic pulse is called the transit time, or $\Delta T$. This tool can only be run in uncased boreholes containing a non-gaseous medium.

The sonic log is used in conjunction with neutron and density data as an estimator of lithology, it can also be a good indicator of the mechanical properties of the borehole, such as formation strength, permeability and porosity.

The usual mnemonic used for sonic logs is DT, and the measurements are taken in $\mu$sec/ft or $\mu$sec/m.

### 5.3.2 Density

The tools in this category emit gamma rays from a chemical source towards the rock formation, two detectors in the tool count the number of returning gamma rays, which are related to the density of electrons in the rock formation.

Through combination with the neutron log, these density logs can be used to estimate lithology, gas presence, clay content and formation mechanical properties.

Density logs can refer to a variety of different log curves, such as bulk density (RHOB, DEN or ZDEN, measured in g/cm$^3$ or kg/m$^3$), density porosity (DPHI, PHID or DPOR, measured in % or v/v decimal), density correction (DRHO, measured in g/cm$^3$ or kg/m$^3$) or photoelectric effect (PE, Pe, PEF, measured in b/e).

### 5.3.3 Neutron

The tool emits high energy neutrons from a chemical source which are slowed down by the nuclei in the rock formation. Two detectors in the tool count either the number of returning neutrons or gamma rays, which are inversely proportional to the amount of hydrogen residing in the rock formation. Since this hydrogen resides inside the pores of the formation, this measurement is related to the porosity of the rock.

This log can estimate porosity taking a specific lithology such as limestone as a baseline, corrections should be made to estimate porosity for other lithologies through the use of charts or other algorithms. By combining this log with density and sonic logs, it is possible to estimate lithology, presence of gas and clay content.

The common mnemonics for the neutron porosity log are NPHI, PHIN and NPOR, and the measurements are expressed as % or v/v decimal.

## 5.4 Resistivity

These logs measure the electrical resistivity of the rock formation, and can give an indicator of the fluid saturation in the rock formation.

### 5.4.1 Induction

The tool contains transmitter coils which induce an alternating current in the rock formation, the response is then sensed in both magnitude and phase by receiver coils built into the tool. This response is the conductivity of the rock formation, which is the inverse of the resistivity. This tool can only be run in an uncased borehole.

Induction logs can be used to calculate formation restivity, fluid saturation, diameter of invasion and geopressure.

Mnemonics associated to the curves generated by induction logging tools include ILD, RILD, ILM, RILM, LLR, SGRD and SFL, which are all measured in ohm.meter.

### 5.4.2 Laterolog

This tool creates a horizontal disk-shaped current around the borehole by focusing a low frequency current through the use of an electrode array. Resistivity can be measured by monitoring the current passing through the tool. This tool can only be run in an uncased borehole filled with water or water-based mud.

As with the induction logs, laterolog measurements can be used to calculate formation resistivity, fluid saturation, diameter of invasion and geopressure.

Mnemonics associated with laterologs are DLL, LLD, RLLD, SLL, LLS, RLLS and Rxo, which are all measured in ohm.meter.

### 5.4.3 Microresistivity

Also know ans Rxo, this tool forces an electrical current into the rock formation using electrodes mounted on pads which are pressed against the borehole wall. Some microresitivity tools focus the current using electrodes similar to the ones used in laterolog tools. This tool can only by run in an uncased borehole filled with water or water-based mud.

Rapid curve movement in this log can be an indicator of fractures. The relationship between microresistivity and other resistivity logs can be in indication of permeability. Microresistivity measurements can also be used to calculate flushed zone formation resistivity and water saturation. This log is also useful to identify very thin beds.

Microresistivity curve mnemonics include MNOR, MINV, MSFL and MLL, which are all measured in ohm.meter.

# 6 DESCRIPTION OF THE TESTING DATA

The greatest hurdle to overcome during this work was the obtention of quality testing data. Well data is highly confidential and therefore, going through the proper channels and obtaining the required clearance to work with the data takes time. Additionally, once the data is made available it is often useless for the process described in this work due to the datasets containing few logs, short or inexisting cored sections or simply due to the low quality of the data itself. Frequently the datasets acquired presented one extremely dominant lithology, usually sandstone, being over 80% of the cored section. Such biased datasets make it difficult to create reliable lithology prototypes, as the less prevalent lithologies can present a number of samples that are not representative enough for reliable statistical analysis.

Another issue arises from certain lithotypes being highly variable when it comes to log responses due to their intrinsic properties. Heterolites, which are rock types characterized by the successive intercalation of thin layers of high and low grain size deposits are a good example of this. A section of rock may be characterized as a sand/clay heterolite and the log readings on this rock might be more similar to clay or sand depending on the ratio of clay to sand in this rock, which might lead to miscategorization. Evaporites such as anhydrite are also very troublesome due to interaction with water changing the chemical composition of the rock and drastically altering the resulting log readings. Occurrences of anhydrite can therefore have highly unique log signatures depending on the environment on which they are situated. All these variations are common in real data and bring additional difficulties for the automatic recognition of rock types.

The data which is provided for scientific research also tends to be older, and while Strataledge provides a rich and well structured vocabulary with which to make the rock descriptions, the data captured in the field with direct access to the rock samples available for these experiments was not originally described with the semantic richness offered by Strataledge. Therefore, the Strataledge descriptions available for testing are translations of descriptions made using other less precise tools, which means the Strataledge toolkit is not being used to its full effect. Needless to say this confidentiality makes it extremely difficult to replicate the results of similar works since there is generally no access to their data.

Despite these issues, a suitable test case was found in a sedimentary environment where the cored section of the wells involved presented enough variability in lithotypes with enough samples to derive reasonable prototypes from once these lithologies were grouped.

The studied sedimentary succession was deposited in the Sergipe-Alagoas basin in northeastern Brazil. The wells used in this work belong to the Carmopolis field, where the depositional environment is interpreted as an alternation between fan deltaic systems

Table 6.1: This table shows the number of samples for each lithology in the three wells chosen for the test case.

| | | Well 1 | Well 2 | Well 3 |
|---|---|---|---|---|
| Group 1 | Clay/silt heterolite | 2 | 8 | 19 |
| | Shale | 27 | 36 | 41 |
| | Siltstone | 33 | 24 | 44 |
| Group 2 | Sand/silt/clay heterolite | 0 | 41 | 61 |
| | Sand/clay heterolite | 0 | 1 | 0 |
| | Sand/silt heterolite | 16 | 0 | 0 |
| Group 3 | Sandstone | 164 | 158 | 300 |
| Group 4 | Conglomerate | 71 | 55 | 151 |

and associated alluvial fans and braid deltas prograding into lakes under arid/semiarid climate and increasing marine influence (AZAMBUJA FILHO et al., 1980) (CANDIDO; WARDLAW, 1985). In this environment, conglomerates, sandstones and mudrocks were deposited in fining-upward cycles. The reservoirs in this sedimentary unit are constituted of conglomerates and sandstones that occur at the base of the depositional cycles. Fining-upward amalgamated sequences of these cycles are interbedded with shales, marls, and calcilutites containing anhydrite nodules and stromatolitic laminites replaced by dolomite and anhydrite.

## 6.1 Testing Data Analysis

The method proposed was tested using real data from three exploration wells in the Sergipe-Alagoas basin. Each well had geophysical logs available measuring sonic transit time (DT), gamma ray (GR), resistivity (ILD), neutron porosity (NPHI), bulk density (RHOB) and density correction (DRHO); as well as core descriptions converted from hand-made descriptions into the Strataledge format. We can see this data plotted as scatter plots in figures 6.3 6.4 6.5. In these graphs, each box shows the samples present in each well plotted in a two-dimensional plane where each axis corresponds to one of the logs. In these graphs it is possible to see the high degree of correlation between certain logs, as well as how samples of the same lithology tend to be grouped together, this is especially noticeable in the conglomerates and sandstones in wells 2 and 3. A breakdown of the samples present in this data can be seen in table 6.1, it shows a clear dominance of sandstones and conglomerates among every well.

The difference in log signatures between certain lithotypes was determined to not be significant enough for reliable discrimination, so these lithotypes were grouped in four different groups by a geologist based on grain size and lithological similarity. The method then classifies each sampled depth as one of these groups, not as a specific lithology. The lithotypes present in the testing data were grouped as follows:

1. Group 1 - Claystone-Siltstone: Clay/silt heterolite, Shale, Siltstone

2. Group 2 - Siltstone-Sandstone: Sand/clay heterolite, Sand/silt heterolite, Sand/silt/clay heterolite

3. Group 3: Sandstone

4. Group 4: Conglomerate

The grouped data can be seen plotted in figures 6.6 6.7 6.8. Box plots showing the differences of the log readings for each lithology can be seen in figure 6.1, box plots for the lithology groups can be seen in figure 6.2. These box plots show how the sampled values for each log is distributed in each lithology. The line in the middle of each box indicates the median value of that log for that lithology; the box represents the range in which 50% of samples are located; the "whiskers" extend to the value of the highest and lowest data points that are not outliers; outliers are marked by + sign and correspond to readings that are at least three scaled median absolute deviations away from the median. In these figures we can clearly see how the conglomerates have a very distinctive distribution compared to the other lithologies.

Additionally, a synthetic dataset was gracefully granted by the authors of (JEONG et al., 2014). This synthetic dataset presents three cases with 1600 samples each belonging to one of three artificial lithologies with two associated artificial log values. This data can be seen plotted in figures 6.9, 6.10 and 6.11. These cases come from scenario 3 described in (JEONG et al., 2014), which is the most complete scenario presented, where noise has been added to the data to better simulate the environmental conditions from a real well. Even in their most true-to-reality form, these graphs show readings with a much lower correlation, and much more well-defined groups than the real data presented.

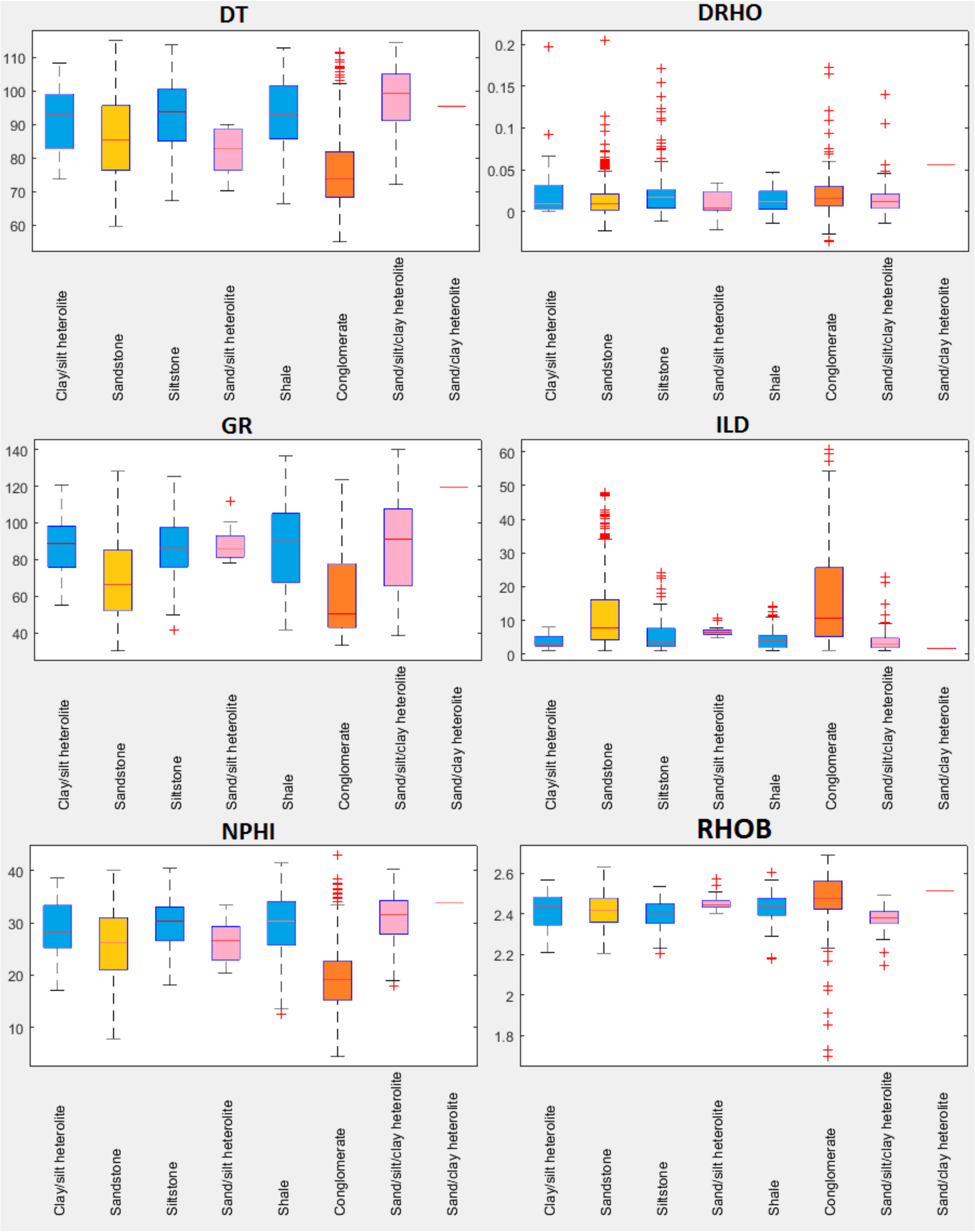


Figure 6.1: Box plot of the log values for each of the lithologies present in the testing data. The logs present in this data are Gamma Ray (GR), Sonic (DT), Density (RHOB, DRHO) and Neutron (NPHI). The lithologies are color coded based on the group to which they were assigned. The sand/clay heterolite belongs to group 2 (pink), and contains only one sample. While the other two lithologies in group 2 seem to have contrasting signatures, this is due to the low number of samples on the Sand/silt heterolite not being representative enough to create an accurate model

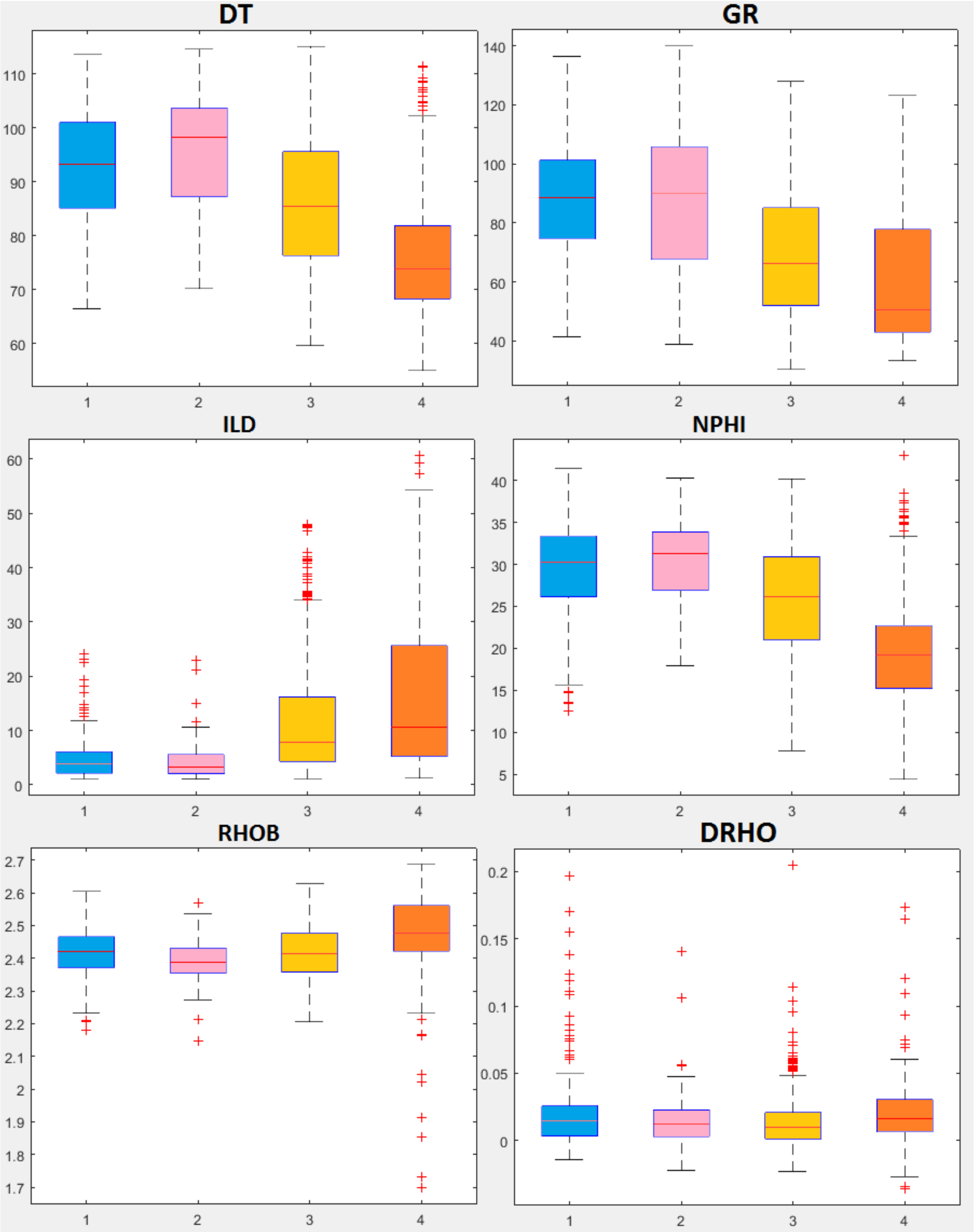


Figure 6.2: Box plot of the log values for each of the lithologies present in the testing data. Notice the contrast in distributions between groups 3 and 4 (sandstone and conglomerates) and the other lithologies.

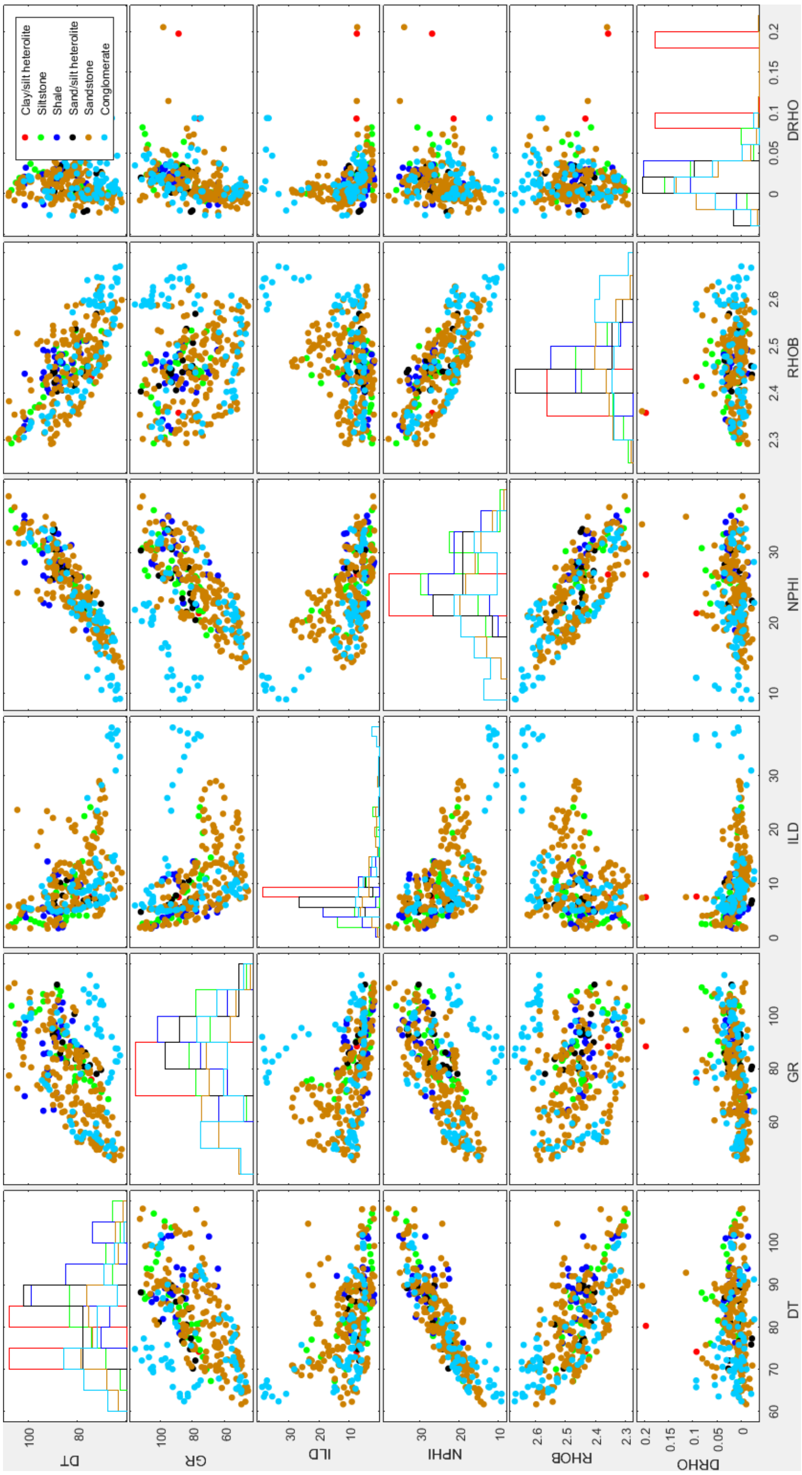


Figure 6.3: Plot of the samples contained in Well 1. Each square presents a scatterplot of two logs, the diagonal shows histograms of the distribution of each of the logs.

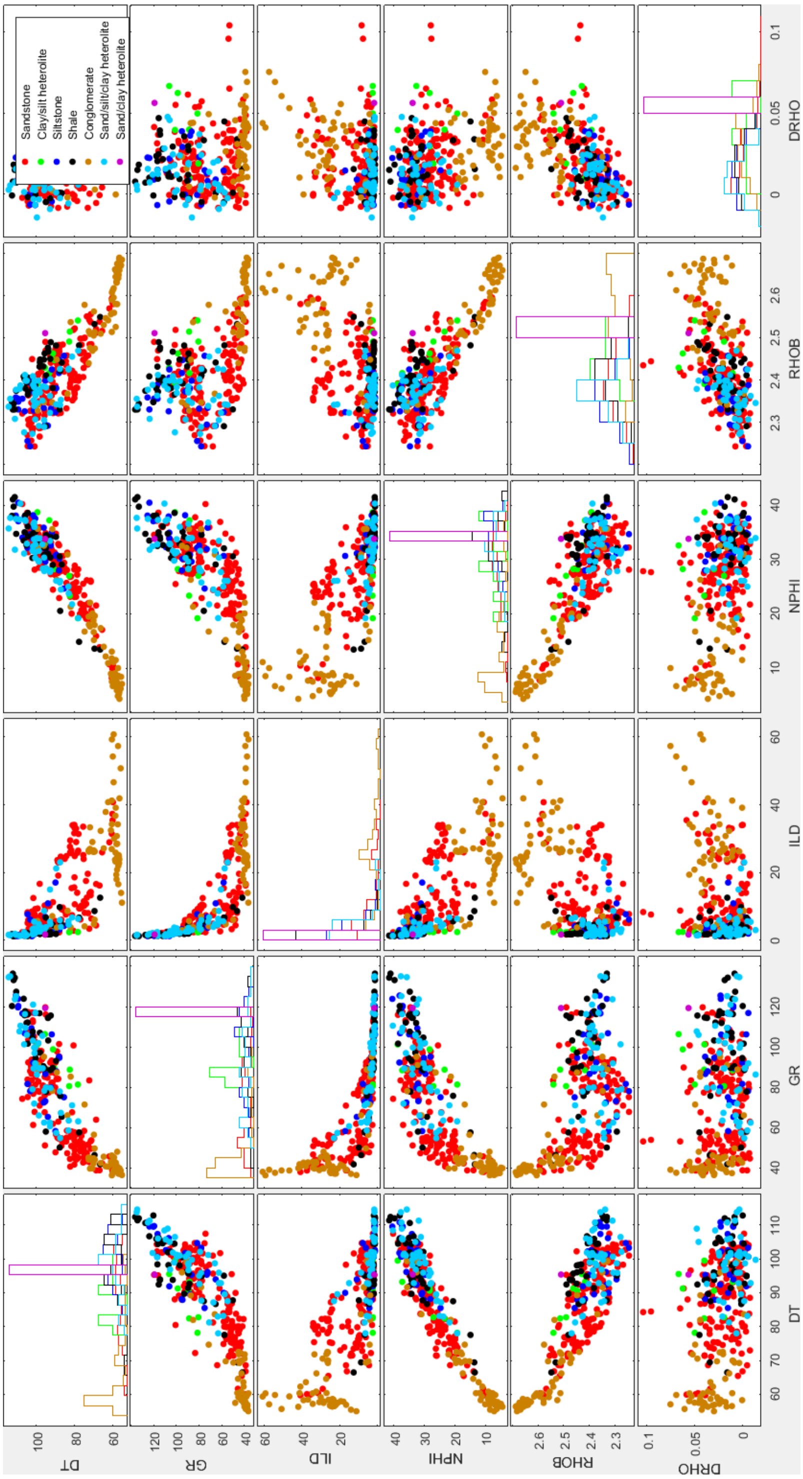


Figure 6.4: Plot of the samples contained in Well 2. Each square presents a scatterplot of two logs, the diagonal shows histograms of the distribution of each of the logs.

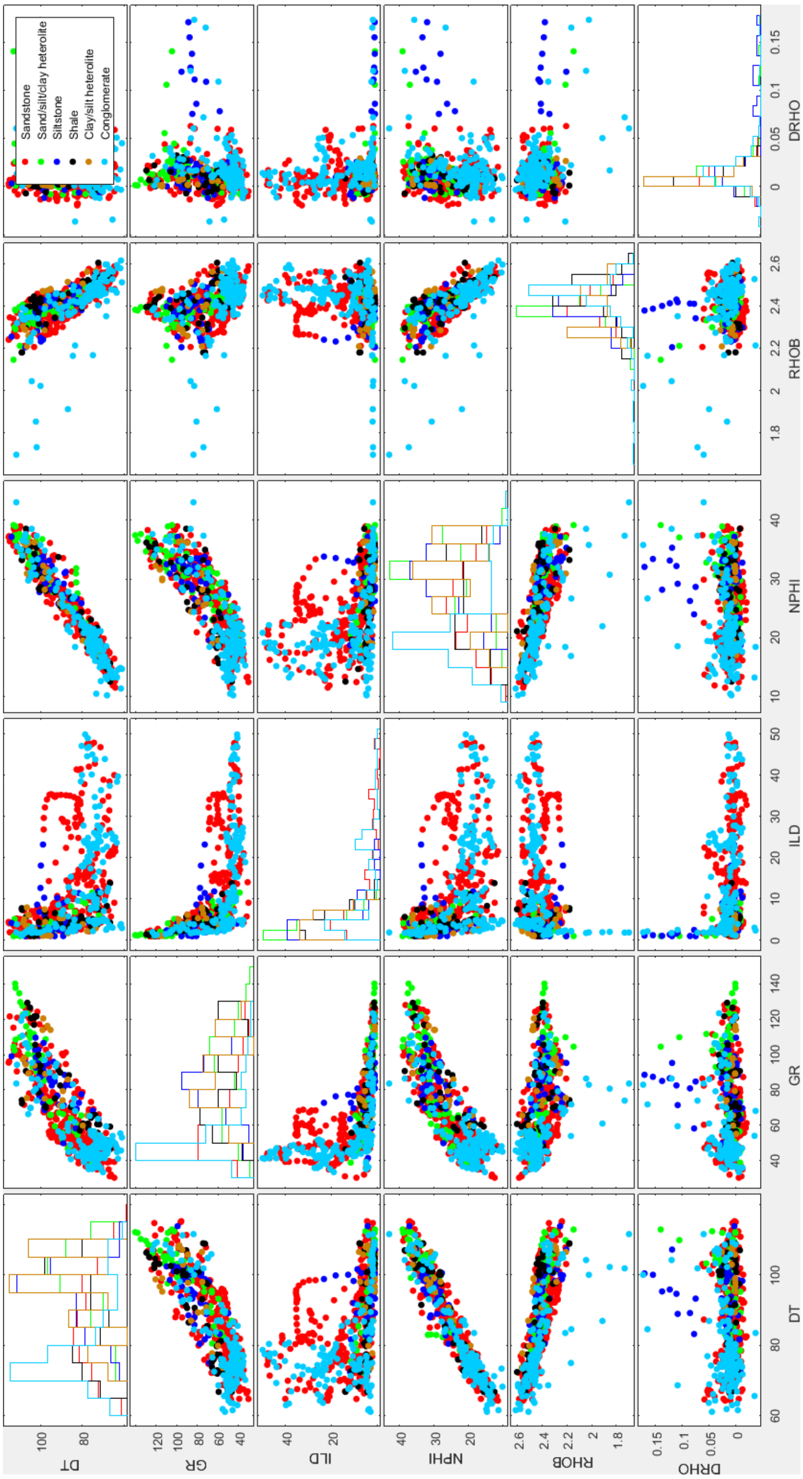


Figure 6.5: Plot of the samples contained in Well 3. Each square presents a scatterplot of two logs, the diagonal shows histograms of the distribution of each of the logs.

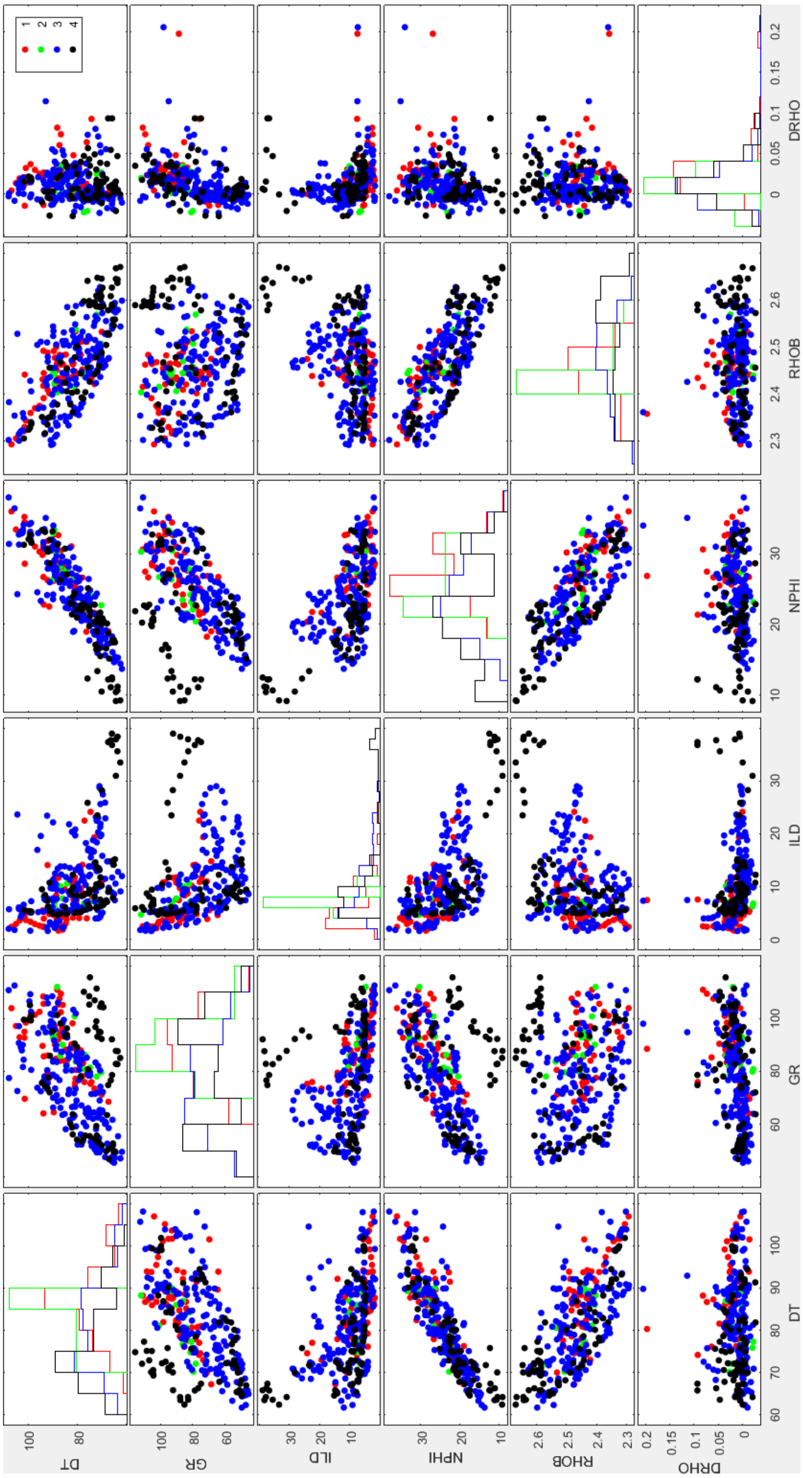


Figure 6.6: Plot of the samples contained in Well 1 grouped in the way described in this chapter. Each square presents a scatterplot of two logs, the diagonal shows histograms of the distribution of each of the logs.

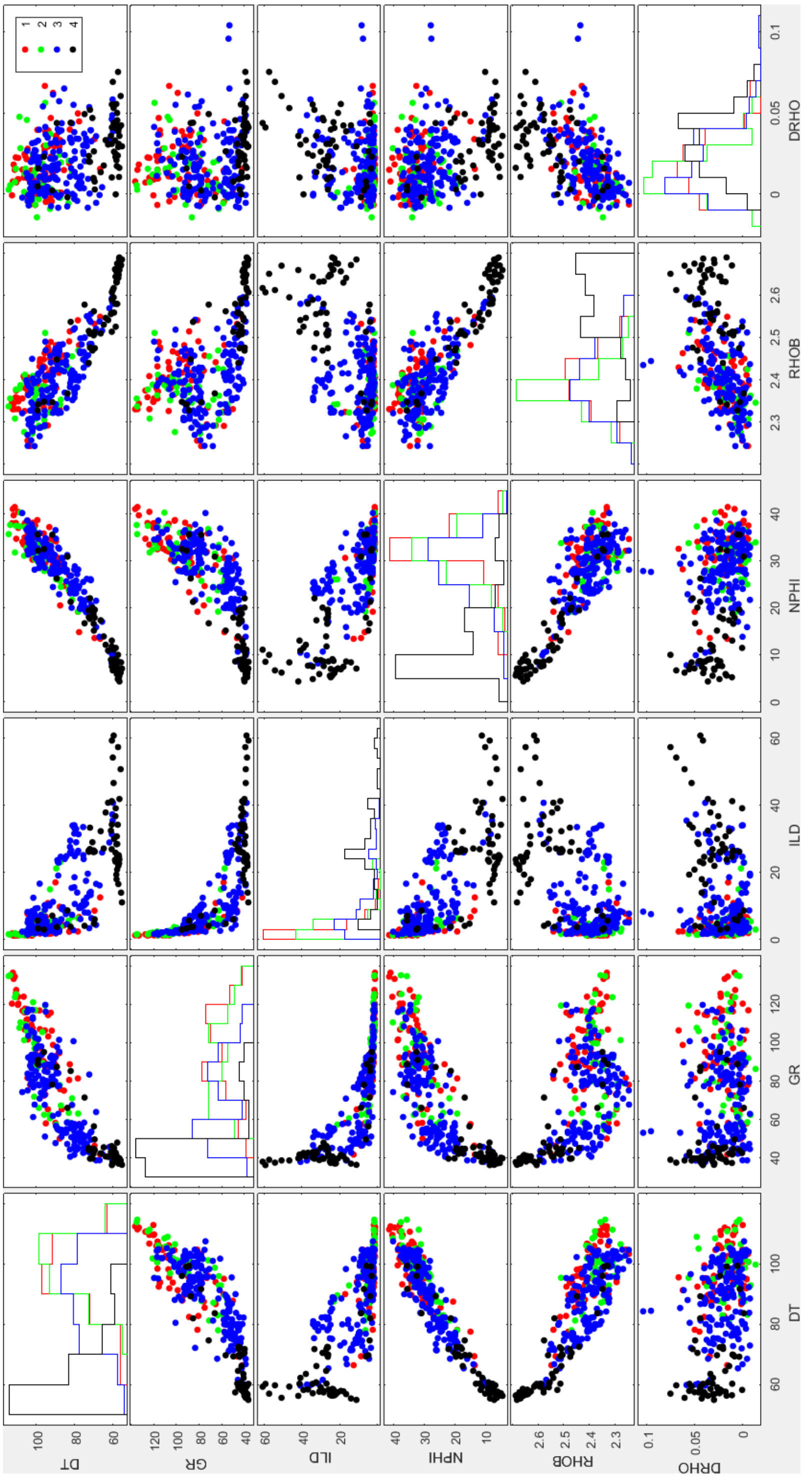


Figure 6.7: Plot of the samples contained in Well 2 grouped in the way described in this chapter. Each square presents a scatterplot of two logs, the diagonal shows histograms of the distribution of each of the logs.

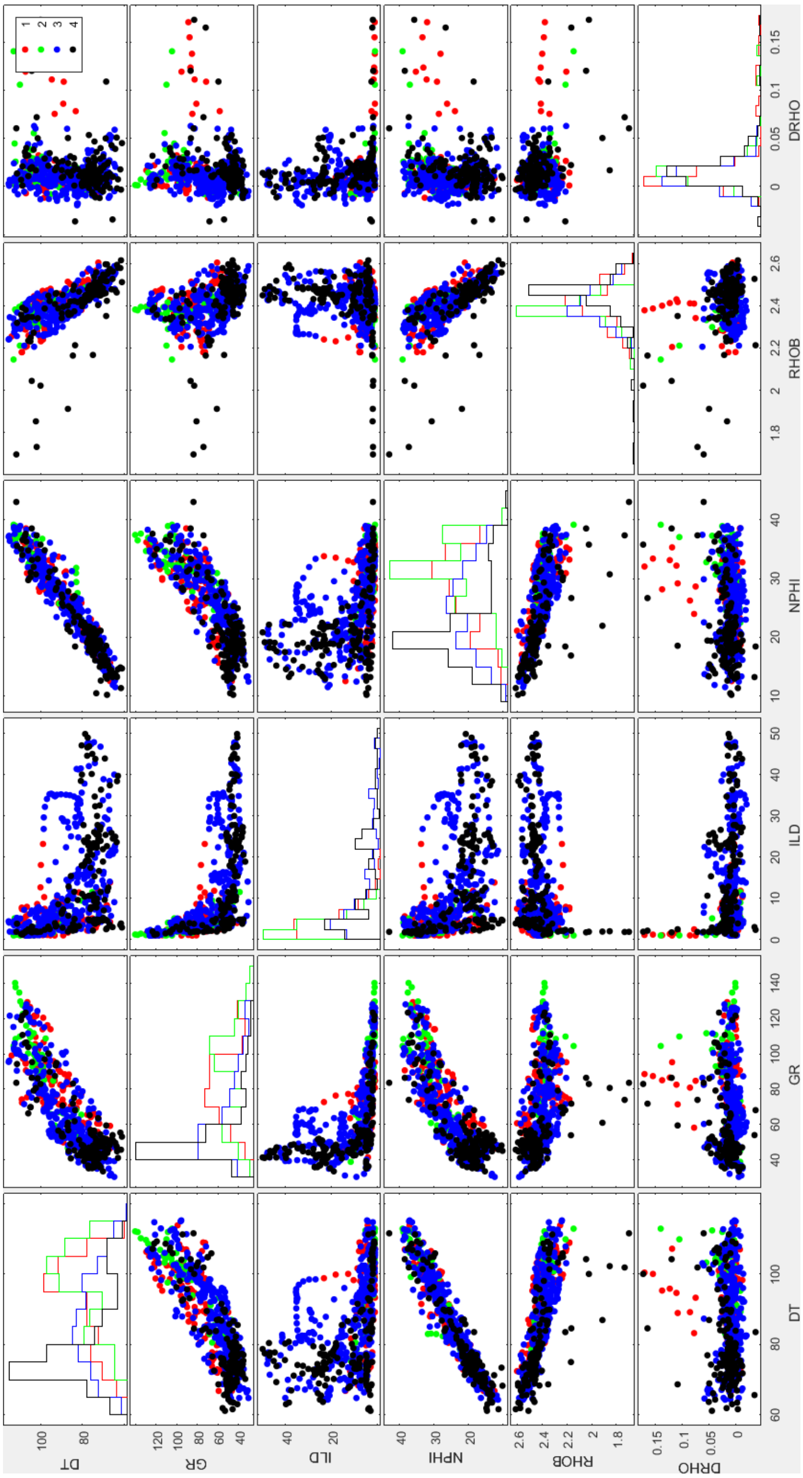


Figure 6.8: Plot of the samples contained in Well 3 grouped in the way described in this chapter. Each square presents a scatterplot of two logs, the diagonal shows histograms of the distribution of each of the logs.

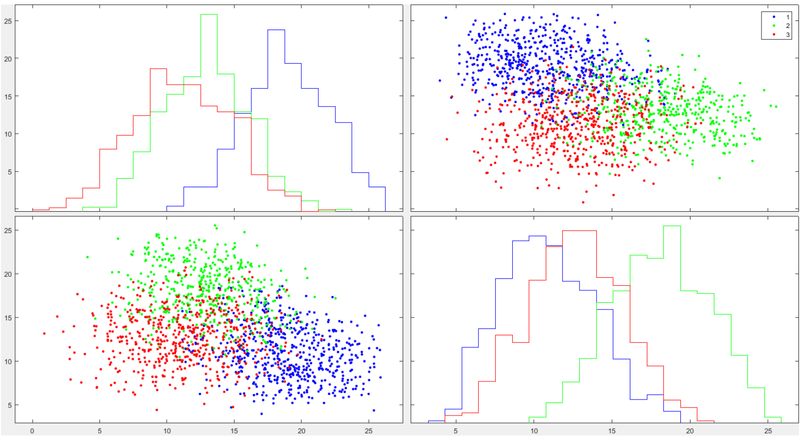


Figure 6.9: Plot of the samples contained in the first synthetic dataset. Each square presents a scatterplot of two artificial logs, the diagonal shows histograms of the distribution of each of the logs.

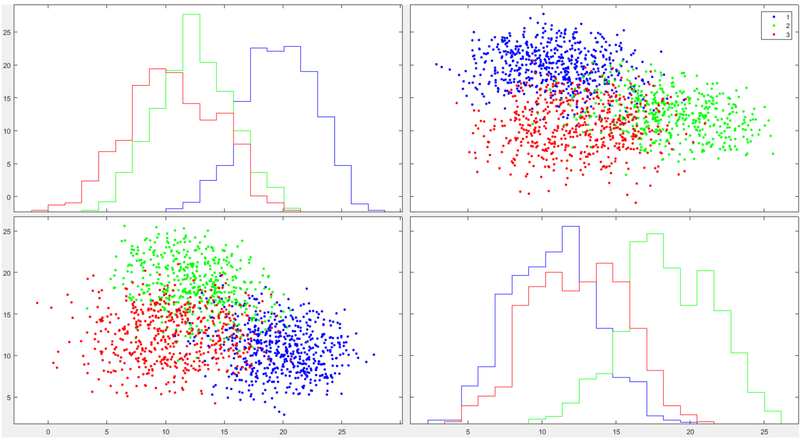


Figure 6.10: Plot of the samples contained in the second synthetic dataset. Each square presents a scatterplot of two artificial logs, the diagonal shows histograms of the distribution of each of the logs.

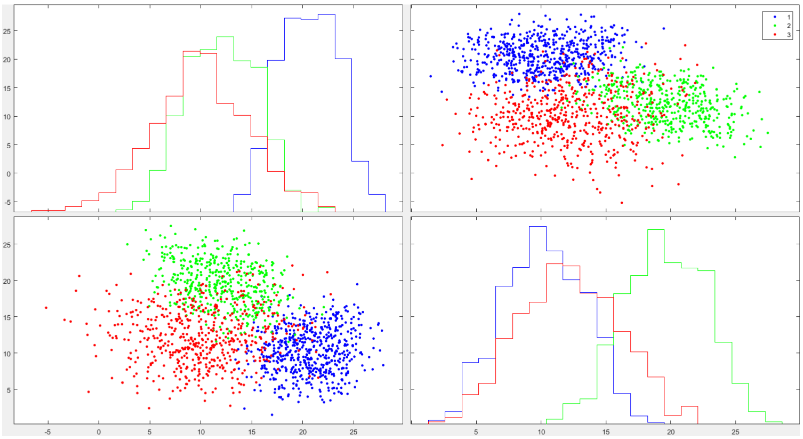


Figure 6.11: Plot of the samples contained in the third synthetic dataset. Each square presents a scatterplot of two artificial logs, the diagonal shows histograms of the distribution of each of the logs.

# 7 METHODOLOGY

Having the standardized core description and the corresponding geophysical logs, our method is divided into three steps: segmenting the log data into lithology intervals, building lithological prototypes by matching core and log data, and finally assigning lithologies to the segmented intervals by comparing them to the lithology prototypes. The method attempts to extrapolate rock data taken from cores ()which is rare and expensive), from log data (which is relatively cheap and abundant). A workflow of this process can be seen in figure 7.1, the steps taken in this workflow are detailed later.

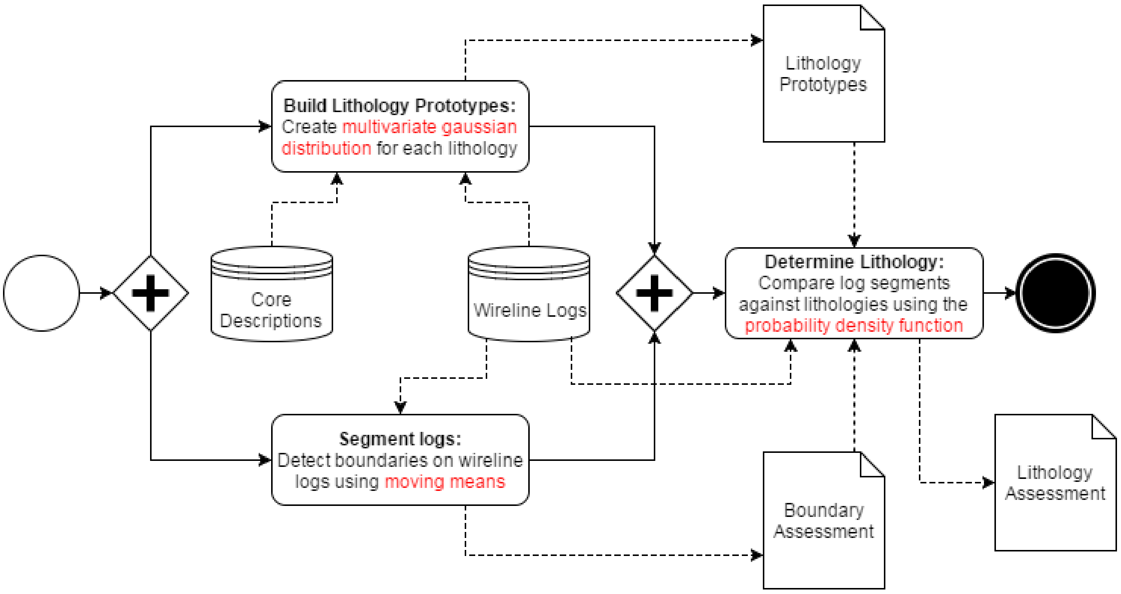

Lithology Prototypes
Build Lithology Prototypes:
Create multivariate gaussian distribution for each lithology
Core Descriptions
Wireline Logs
Determine Lithology:
Compare log segments against lithologies using the probability density function
Segment logs:
Detect boundaries on wireline logs using moving means
Boundary Assessment
Lithology Assessment

Figure 7.1: This framework shows an overall view of our method for lithology detection, the processes involved and what data artifacts are used and generated.

## 7.1 Detecting Intervals

In order to extract lithological significance from wireline logs, the most intuitive way of doing this is to first segment the log into regions where the log readings stay within roughly the same value. This is the way geologists work while trying to identify rock formations from wireline logs, and our work attempts to emulate this process with automatic detection of these change points.

Starting from the assumption that a change in lithology can be characterized in the log record by a sudden change in the measured value, we detect these changes in lithology using the work previously described in (GRACIOLLI, 2014) and presented in chapter 3.

The assessment generated from this procedure can then be edited or not by the user using the program interface and the intervals defined by these break points are passed to the second part of the method, which will estimate the likely lithology of each interval based on the log values encompassed by each of them.

## 7.2 Determining Lithology

Once the intervals of interest have been determined, the log values measured at those depths can be analyzed and a lithological significance to the interval in question can be assigned. But first, a representation of each lithotype must be created and expressed in terms of numerical values, which can then be compared with the log readings taken at each interval.

With the standardized Strataledge core descriptions, the rock characteristics at any depth of a cored interval can be checked easily, this allows the parsing large amounts of data automatically. By checking the depth column in the log files, the rest of the core readings can be associated with a corresponding lithology if that depth at that well correlates to a cored interval, a prototypical log signature for each lithology can then be defined based on the readings taken at depths corresponding to that lithology. The set of logs and core descriptions used to create these prototypes is referred in this work as the **training set**.

The prototypes are created by calculating the average and covariance values between all available logs for each lithology. With these values, a log signature corresponding to a certain lithology can be expressed as a multivariate Gaussian distribution defined by the average vector $\mu$ and the covariance matrix $\Sigma$. Considering a matrix $D_{m \times n}$ where each of the $m$ rows corresponds to a sampled depth in the wireline log data, with $n$ measured log values, the average vector $\mu$ can be defined as expressed in the equation 7.1, and each of the values of the covariance matrix $\Sigma_{n \times n}$ can be calculated as defined in equation 7.2.

During this step, the Bayesian prior for each lithology is also calculated. The prior of a lithology $L$ is expressed in 7.4, where $n$ is the number of samples of lithology $L$, and $m$ is the total number of samples in the training data. This prior represents the probability for a data point belonging to a given lithology before being analyzed. This prior is useful to give more weight to lithologies that are common in the testing data.

$$\mu = [\sum_{i=1}^{m} \frac{D_{i,1}}{m}, \sum_{i=1}^{m} \frac{D_{i,2}}{m}, \sum_{i=1}^{m} \frac{D_{i,3}}{m}, ..., \sum_{i=1}^{m} \frac{D_{i,n}}{m}] \tag{7.1}$$

$$\Sigma_{j,k} = \sum_{i=1}^{m} \frac{(D_{i,j} - \mu_j)(D_{i,k} - \mu_k)}{m} \tag{7.2}$$

$$pdf(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)) \tag{7.3}$$

$$P(L) = \frac{n}{m} \tag{7.4}$$

$$updf(x, \mu, \Sigma) = \sum_{i=1}^{m} \frac{pdf(x_i, \mu_i, \Sigma_{(i,i)})}{m} \tag{7.5}$$

Each interval defined in the first step can then be checked to determine which lithology is the best fit for each interval. This is done by checking every point inside the interval against each lithology by means of a probability density function, defined in equation 7.3. The equation receives as input the point $x$ that is being checked, represented as a vector of all log values at that depth, and the average vector $\mu$ and covariance matrix $\Sigma$ of the lithology that is being checked. The result of this function measures how close the point is to that lithology, this result is then multiplied by the prior of the corresponding lithology to obtain the final score indicating the likelihood of that point belonging to that lithology. Once every point inside the interval is compared to every lithology encountered in the training data, the interval's lithology is declared to be the one that scored higher across every point in the interval.

A univariate variation of this method was also implemented, where the covariance between the logs is disregarded during the calculation, this can be done simply by calculating the average of the probability density function for each individual value of $x$, $\mu$ and the diagonal of $\Sigma$. This univariate variation of the $pdf$ is defined in 7.5.

The original method developed was tested against other pattern matching algorithms on the same datasets. These algorithms are:

1. A KNN classifier, an implementation of which was used as one of the agents described in (GIFFORD; AGAH, 2010). A $K$-Nearest Neighbours classifier works by assigning a label to a data point based on the distance of that point to the $K$ nearest points in the training set. By verifying which labels are the most prevalent in the surrounding points, a label to the point analyzed can be assigned based on this assessment. There are many ways to calculate this distance and weigh labels of the neighbors in this estimate. The version implemented in this work uses euclidean distance and equal weights to all neighbors.

2. An artificial neural network composed of 500 nodes on the hidden layer and trained using scaled conjugate gradient. Different neural network implementations have been widely used in this task, such as in (GIFFORD; AGAH, 2010) (OJHA; MAITI, 2013). The most common neural network implementations consist of a series of nodes (also called neurons) organized in three layers, where each node feeds its output to every node in the next layer. The first layer is the input layer, each node in the input layer corresponds to one value of the input data, in our case, each log reading at the depth to be analyzed. The second layer is the hidden layer, where the number of neurons is chosen based on the task. A hidden layer size of 500 was chosen as a compromise between prediction accuracy and training time. The last layer is the output layer, with one node for each possible prediction label, in our case, one node for each lithology. The neural network is trained by feeding it a training set consisting of pre-labeled data points adjusting the weights of the connections and the bias of each node until the output of the network classifies the data points with the expected label at a reasonable degree of accuracy. Neural networks have been successfully used in many classification problems similar to the one presented in this work. As the starting values for the weights of the network are randomized at the start of each training session, multiple training sessions may result in varying degrees of categorization accuracy.

3. A fuzzy logic based approach outlined in (BOSCH; LEDO; QUERALT, 2013a). In this method, membership functions are calculated from the training data in order to classify a validation set by measuring the degree of membership of sample to each lithology in a manner similar to the method developed in this work.

The method developed was then combined with the KNN classifier and the neural network in a multi-agent system. In a multi-agent system, the problem is solved multiple times by independent methods, the results are then compared to reach a final, unified assessment. Each sample is labeled with the lithology that labeled by the most agents. For example, if 3 agents label sample $S$ as lithology $A$ and 2 agents label sample $S$ as lithology $B$, we label sample $S$ as lithology $A$. In the case of a tie, the sample is labeled as the lithology detected by the first agent to be queried between all the agents involved in the tie. These agents are always queried in the same order.

This methodology aims to verify the classification accuracy of the method developed by comparing how it performs against other algorithms, subjecting them to tests using both real and synthetic data. The goal is to validate the hypothesis that the segmentation of the log data leads to an increase in classification accuracy.

# 8 RESULTS

The methodology previously described has been applied to the both the synthetic data and the real data previously mentioned in this work.

As the method relies on separating the log data in intervals, three different methods of making this assessment were defined and dubbed as follows:

1. Complete split: each sampled depth is considered an independent facies, this is the same as skipping the segmentation step of the method, and it is useful to assess the impact of this step in the final result.

2. Perfect split: this is a facies assessment created by checking the facies limits in the testing data, and represents a best case scenario where the facies segmentation is done with $100\%$ accuracy.

3. Auto split: this is the moving means method derived from (REID; LINSEY; FROSTICK, 1989) and developed in (GRACIOLLI, 2014) as described in this work.

First, the method was tested using the synthetic dataset provided by the authors of (JEONG et al., 2014). Three test cases were constructed from the three datasets in the scenario analyzed, where each test case would use one of the datasets as the validation case, and the remaining two datasets as the training set. These cases were tested against both the univariate and multivariate versions of the algorithm, assisted by the three types of segmentations previously described, as well as the method described in (BOSCH; LEDO; QUERALT, 2013a). The results can be seen in table 8.1

These results show that the premise that the segmentation of the dataset leads to increased accuracy holds, as the perfect split outperformed the complete split in every test case and achieved an accuracy in line with the results obtained in (JEONG et al., 2014) where this dataset was originally used. However, the segmentation method proved to

Table 8.1: Classification accuracy achieved on the synthetic dataset with different variations of the method developed, compared to the method described in (BOSCH; LEDO; QUERALT, 2013a).

| | Fuzzy Logic | Univariate | | | Multivariate | | |
|---|---|---|---|---|---|---|---|
| | Based Approach | Complete | Auto | Perfect | Complete | Auto | Perfect |
| Dataset 1 | 0.5244 | 0.7644 | 0.5669 | 0.8313 | 0.7769 | 0.5356 | 0.8275 |
| Dataset 2 | 0.5463 | 0.8006 | 0.5738 | 0.8525 | 0.8325 | 0.5444 | 0.8625 |
| Dataset 3 | 0.6200 | 0.7850 | 0.5676 | 0.8306 | 0.8750 | 0.5381 | 0.8900 |

Table 8.2: Classification accuracy achieved on the real dataset with different variations of the method developed, compared to the method described in (BOSCH; LEDO; QUERALT, 2013a).

| | Fuzzy Logic | Univariate | | | Multivariate | | |
|---|---|---|---|---|---|---|---|
| | Based Approach | Complete | Auto | Perfect | Complete | Auto | Perfect |
| Well 1 | 0.4984 | 0.5304 | 0.5240 | 0.5240 | 0.4185 | 0.7029 | 0.7157 |
| Well 2 | 0.3901 | 0.5077 | 0.4906 | 0.4892 | 0.5015 | 0.6074 | 0.6409 |
| Well 3 | 0.4042 | 0.4954 | 0.4876 | 0.4892 | 0.6378 | 0.6532 | 0.6749 |

Table 8.3: Classification accuracy in each well using the method developed on a complete split test with and without the use of the Bayesian prior.

| | Without Bias | With Bias |
|---|---|---|
| Well 1 | 0.2971 | 0.4185 |
| Well 2 | 0.4211 | 0.5015 |
| Well 3 | 0.4367 | 0.6378 |

be ill-suited for this particular synthetic dataset. Since the data is generated from randomly seeded samples, it does not behave like readings from a real borehole; there are no smooth, small variances within readings from the same facies contrasting with sharp and sudden increases or decreases when the rock type changes. For this reason, the best segmentations achieved through exhaustive testing still produced results with an accuracy well under the other types of splits. In this particular case, the segmenting step of the algorithm works best when using windows which are close in value, which result in a larger number of segments in the limited number of logs available in this dataset.

Next, the method was tested against the data extracted from real boreholes. With three wells in the testing data, three different test cases were built, each one using one of the wells as validation data, and the other two as training data. Due to the close geological distance between these wells, they present similar characteristics and can be used to create prototypes which are valid between them. These were tested against the method developed and the method outlined in (BOSCH; LEDO; QUERALT, 2013a). These tests were using the complete and perfect splits, and both the regular multivariate version of the method and the univariate variation. The results can be seen in 8.2. The impact of the Bayesian prior was also analyzed by performing complete split tests with and without the bias on each test case, the results of this test can be seen in 8.3

The results show that the multivariate version of the algorithm produces a more accurate prediction than both the univariate version and Bosch's approach when supported by a good segmentation of the log data. It also shows that the univariate version is actually negatively impacted by the segmentation. Since it disregards the correlation between the log values, the univariate version has a lower variance on the degree of certainty it assigns labels. This is further exacerbated when grouping multiple samples, which causes the higher degree of misclassification since it is harder for the correct lithology to emerge as a clear winner once the probabilities are summed up.

Well 1 as the validation set presented the best results when supported by a quality facies segmentation; this is due to the higher number of outliers and distribution overlap present in well 1 creating less reliable prototypes when inserted into the training data; when these samples are used only as validation this is not an issue: it is easier to classify

Table 8.4: Classification accuracy in each well using a KNN classifier, the number after KNN specifies how many neighbors the classifier takes into account. Using more than 40 neighbors makes little sense considering the number of samples in the training data. The NN line stands for the classification accuracy of the neural network.

| | KNN5 | KNN10 | KNN20 | KNN40 | NN |
|---|---|---|---|---|---|
| Well 1 | 0.3801 | 0.4281 | 0.4600 | 0.4792 | 0.5320 |
| Well 2 | 0.4427 | 0.4303 | 0.4861 | 0.5108 | 0.5273 |
| Well 3 | 0.4513 | 0.4643 | 0.4513 | 0.4562 | 0.4894 |

outliers using reliable prototypes than classify quality data using unreliable prototypes. The wells were then tested using auto split, the parameters used in the segmentation greatly affect the final result. After extensive experimentation involving multiple wells, it was ascertained that the most accurate segmentations are obtained by using large windows for both moving means filters, and a break threshold lower than any of the weights ascribed to any of the logs. In essence, trusting breaks detected in any logs completely, but requiring large changes in the logs to determine a change in lithology. This goes against the guidelines suggested by Reid (REID; LINSEY; FROSTICK, 1989), however, since multiple logs are being used instead of just the gamma ray, the use larger windows is feasible since the loss of detail incurred by a more extensive smoothing in a single log will be offset by the detections made in other logs. By using these guidelines and experimenting with the segmentation parameters, an accuracy of $70.29\%$ on well 1 could be achieved with moving mean windows of sizes $100$ and $300$, a result which is very close to the one obtained using the perfect split. Windows used on well 2 were $150/300$ and on well 3 $130/280$.

The other two methods were then tested in each well, without the use of segmentation, and the results can be seen in 8.4. Neural network results vary with the random initial value of the nodes during training, but the average accuracy after 10 independent training sessions can be seen in the table. The neural network outperformed the KNN classifier in all test cases, but both were still inferior to the method developed in the presence of a reliable assessment of break points.

Next, a multi-agent system was built using the methods previously tested. In this implementation, each agent classifies the samples in the test data as one of the four lithological groups based on the same training data. The label for each sample is then determined to be the one that was selected the most between all the agents.

The multi-agent system that presented the most consistent results was composed of the following agents:

- Two agents implementing the method developed in this work.
- One agent implementing the neural network previously tested.
- One agent implementing a cascade forward version of the neural network previously tested. In this version of the neural network, the input nodes are also linked to nodes in the output layer in addition to the nodes in the hidden layer.
- One agent implementing a KNN40 classifier as previously tested.

Since the method developed obtained the best results, it is implemented twice to act as a tie-breaker in case of indecision between an even number of agents. The results

Table 8.5: Results from the application of the multi-agent system assembled on the three real-data test cases

| | Univariate | | Multivariate | |
|---|---|---|---|---|
| | Complete Split | Perfect Split | Complete Split | Perfect Split |
| Well 1 | 0.5335 | 0.5240 | 0.5240 | 0.6645 |
| Well 2 | 0.5356 | 0.5170 | 0.5759 | 0.6254 |
| Well 3 | 0.4659 | 0.4594 | 0.4627 | 0.4753 |

from this multi-agent system can be seen in table 8.5. These results are in line with the ones previously obtained on this data set. Wells 1 and 2 presented increases in prediction accuracy on the complete split cases, but a decrease on the perfect split cases due to the other methods disregarding the log segmentation.

The results show that the method developed benefits greatly from the segmentation of well data. Despite the difficulty in finding effective parameters for the automatic segmentation method, the tests done with the perfect split show a top increase in accuracy of up to 71% in one of the test cases.

# 9 CONCLUSION

The objective of this work is to is to present a method for lithological interpretation through the calibration of wireline logs via correlation with core sample descriptions; with the goal of enriching reservoir models by integrating this data in a larger scale.

To this end, a novel method was developed that proposes the segmentation of the log data prior to the use of a classification algorithm. By segmenting the log data into regions of interest using a moving-means based algorithm, these segments can then be labeled by a gaussian classifier that takes into account a whole section of the rock formation, instead of a single sample. This segmentation was shown to improve the final classification accuracy of the algorithm significantly.

The results show that the method developed can achieve an accuracy comparable or better than current methods used for lithology classification. This work shows that the log segmentation performed previous to the classification proved to increase the accuracy of the method significantly, and while it is dependent on a large number of parameters, it can be assisted by a geologist determining the break points manually or semi-manually, improving the final results of the classification. By utilizing signal processing algorithms prior to the classic pattern recognition methods we can increase the prediction accuracy in situations such as the one described in this work, where the context of the data matters; the samples do not exist in a vacuum, as they are measurements from a contiguous section of rock, and therefore should be treated as such as most of the current methods for lithology prediction propose.

Another strength of the method proposed is its modularity. The framework presented can easily accommodate changes to each individual process without changing the overall behavior of the method. This leaves the method open to improvements in the way that new methods can be developed to segment the logs, build prototypes or classify samples which can then be easily plugged in the framework. This was already demonstrated to an extent in this work in the way the multi-agent system was implemented by adding parallel methods that are later integrated in the categorization section of the framework.

It is also worth noting that the method itself is completely divorced from any notion of lithology, the only thing determining that the program classifies lithologies are the numerical models fed into it. We can just as easily create statistics based on other features whose description is supported by domain ontology, such as hydrocarbon levels or create models that reflect other ways to categorize rock formations, such as petrofacies. The techniques developed in this work are also not tied to the field of geology, and can be used in any other similar problem that attempts to classify sections of continuous data.

Limitations arise from the high number of parameters required by the segmentation algorithm. Log readings can also vary wildly between different basins, reservoirs and wells; which makes hinder the construction of more universal lithology prototypes. The method must use a training set with data recovered from locations near the testing well, this can be a problem if the sampling in the region is low-quality or low-volume. Biased sampling when recovering the core samples can also lead to misclassification on depths where the rock data was not reflected in the core samples.

A different segmentation proposal or a method to infer the best possible parameters for segmentation more efficiently can make the method more reliable and result in a product with easier interaction with the end user. The method can also be explored to identify different rock characteristics or to be applied to a different field altogether in the task of classifying sequential numerical data. Testing different mathematical models for the prototypes could also prove to be beneficial.

Finally, this work should also be revisited at a later date when more quality data becomes available. More reliable core descriptions captured with the support of Strataledge instead of merely translated into Strataledge could be used to create more accurate prototypes and eliminate issues with more ambiguous lithologies where there is a weaker consensus between geologists describing them.

# REFERENCES

AL-ANAZI, A.; GATES, I. Support vector regression for porosity prediction in a heterogeneous reservoir: a comparative study. **Computers & Geosciences**, [S.l.], v.36, n.12, p.1494–1503, 2010.

AZAMBUJA FILHO, N.; ABREU, C.; HORSCHUTZ, P.; CANDIDO, A.; RAMOS, E. Estudo sedimentológico, faciológico e diagenético dos conglomerados do campo petrolífero de Carmópolis: xxxi congresso brasileiro de geologia. **Anais**, [S.l.], v.1, p.240–253, 1980.

BOSCH, D.; LEDO, J.; QUERALT, P. Fuzzy logic determination of lithologies from well log data: application to the ktb project data set (germany). **Surveys in Geophysics**, [S.l.], v.34, n.4, p.413–439, 2013.

BOSCH, D.; LEDO, J.; QUERALT, P. Fuzzy logic determination of lithologies from well log data: application to the ktb project data set (germany). **Surveys in Geophysics**, [S.l.], v.34, n.4, p.413–439, 2013.

BRERETON, N.; GALLOIS, R.; WHITTAKER, A. Enhanced lithological description of a Jurassic mudrock sequence using geophysical wireline logs. **Petroleum Geoscience**, [S.l.], v.7, n.3, p.315–320, 2001.

CANDIDO, A.; WARDLAW, N. Reservoir geology of the Carmopolis oil field, Brazil. **Bulletin of Canadian Petroleum Geology**, [S.l.], v.33, n.4, p.379–395, 1985.

CARBONERA, J. L. Raciocínio sobre conhecimento visual: um estudo em estratigrafia sedimentar. , [S.l.], 2012.

COUDERT, L.; FRAPPA, M.; ARIAS, R. A statistical method for litho-facies identification. **Journal of applied geophysics**, [S.l.], v.32, n.2-3, p.257–267, 1994.

CRANGLE, R. D. Log ASCII Standard(LAS) Files for Geophysical Wireline Well Logs and Their Application to Geologic Cross Sections through the Central Appalachian Basin. **Open-file Report. U. S. Geological Survey**, [S.l.], p.14, 2007.

GIFFORD, C. M.; AGAH, A. Collaborative multi-agent rock facies classification from wireline well log data. **Engineering Applications of Artificial Intelligence**, [S.l.], v.23, n.7, p.1158–1172, 2010.

GRACIOLLI, V. M. Bedding contact detection: a moving mean-based approach. , [S.l.], 2014.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: **Handbook on ontologies**. [S.l.]: Springer, 2009. p.1–17.

GUIZZARDI, G. **Ontological foundations for structural conceptual models**. [S.l.]: CTIT, Centre for Telematics and Information Technology, 2005.

JEONG, J.; PARK, E.; HAN, W. S.; KIM, K.-Y. A novel data assimilation methodology for predicting lithology based on sequence labeling algorithms. **Journal of Geophysical Research: Solid Earth**, [S.l.], v.119, n.10, p.7503–7520, 2014.

KRYGOWSKI, D. A. Guide to petrophysical Interpretation. **Austin Texas USA**, [S.l.], 2003.

LI, Y.; ANDERSON-SPRECHER, R. Facies identification from well logs: a comparison of discriminant analysis and naïve bayes classifier. **Journal of Petroleum Science and Engineering**, [S.l.], v.53, n.3-4, p.149–157, 2006.

MACDOUGALL, S.; NANDI, A. K. Hybrid Bayesian procedures for automatic detection of change-points. **Journal of the Franklin Institute**, [S.l.], v.334, n.4, p.575–597, 1997.

OJHA, M.; MAITI, S. Sediment classification using neural networks: an example from the site-u1344a of {IODP} expedition 323 in the bering sea. **Deep Sea Research Part II: Topical Studies in Oceanography**, [S.l.], n.0, p.–, 2013.

ROCHA COSTA, A. C. da; VICARI, R. M.; TONIDANDEL, F. (Ed.). **Ontological Primitives for Visual Knowledge**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p.1–10.

REID, I.; LINSEY, T.; FROSTICK, L. E. Automatic bedding discriminator for use with digital wireline logs. **Marine and petroleum geology**, [S.l.], v.6, n.4, p.364–369, 1989.

SENANYAKE, S. **What is a Well Log?** 2016. Available at http://sanuja.com/blog/what-is-a-well-log. Accessed Mar 22 2018.

STRATALEDGE: core description system. Porto Alegre, Brazil: Endeeper, 2012.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. **Data & knowledge engineering**, [S.l.], v.25, n.1-2, p.161–197, 1998.

USGS. **U.S. Geological Survey Data Series 542**. 2010. Available at https://pubs.usgs.gov/ds/542/. Accessed Mar 22 2018.

# ANNEX 1

The following text is an excerpt translated from Carbonera's master dissertation as mentioned in chapter 4, which details the concepts and structure of a section of the domain ontology used by Strataledge to create the core descriptions used in this work.

## 8.1 Domain ontology presentation

The fundamental concept of the domain is that of Sedimentary Facies. This concept offers the identity criteria to all its instances, once geologists are capable of distinguishing one instance from the other through observation of its visual characteristics. This concept also offers the unity criteria to its instances. This criteria involves the observation of discontinuities on the visual characteristics of the facies and allows for the geologist to discriminate between the different units of the Sedimentary Facies. Furthermore, this concept is rigid, once Sedimentary Facies instances can't stop being Sedimentary Facies instances, unless they cease to exist. In this ontology, a Sedimentary Facies is considered a *Kind*, according to UFO.

A Sedimentary Facies is characterized through sixteen distinct qualities, which are considered *Quality Universals*, according to UFO. Each of these Quality Universals is associated with a quality structure which represents the space of possible values. Each of these qualities, along with the respective associated value space, is detailed as follows:

- **Mode:** Defies the most frequent grain size present in the rock. This is a textural property of extreme importance, since it characterizes distinct sedimentary environments and provides information regarding depositional processes. Its quality structure contains: Silt, Clay, Very fine sand, Fine sand, Medium sand, Coarse sand, Very coarse sand, Gravel, Granule, Pebble, Block and Boulder.

- **Larger grain size:** Measurement of the largest grain size observed in a sedimentary facies. Its quality structure contains: Silt, Clay, Very fine sand, Fine sand, Medium sand, Coarse sand, Very coarse sand, Gravel, Granule, Pebble, Block and Boulder.

- **Smaller grain size:** Measurement of the smallest grain size observed in a sedimentary facies. Its quality structure contains: Silt, Clay, Very fine sand, Fine sand, Medium sand, Coarse sand, Very coarse sand, Gravel, Granule, Pebble, Block and Boulder.

- **Base grain size:** Mode of the size measurement of the particles observed at the base of a sedimentary facies. Its quality structure contains: Silt, Clay, Very fine sand, Fine sand, Medium sand, Coarse sand, Very coarse sand, Gravel, Granule, Pebble, Block and Boulder.

- **Top grain size:** Mode of the size measurement of the particles observed at the top of a sedimentary facies. Its quality structure contains: Silt, Clay, Very fine sand, Fine sand, Medium sand, Coarse sand, Very coarse sand, Gravel, Granule, Pebble, Block and Boulder.

- **Sorting:** Refers to the degree of uniformity between the size of the particles composing the facies. Sediments whose grains possess a predominantly uniform size are classified as well sorted. Sediments which contain grains with different diameters are considered poorly sorted. Its quality structure contains: Very well sorted, Well sorted, Moderately sorted, Poorly sorted and Very poorly sorted.

- **Sphericity:** Indicates the degree of approximation between a sedimentary particle and a sphere. Low sphericity particles are elongated and are indicative of the source area. Its quality structure contains: High, Medium, Low.

- **Roundness:** Corresponds to an attribute independent from sphericity, which indicates the degree of angularity of the corners of the sedimentary particles. Well rounded particles are commonly indicative of prolonged mechanical abrasion. Its quality structure contains: Very angular, Angular, Subangular, Subrounded, Rounded and Well-rounded.

- **Color:** Mainly indicates the composition of the minerals in a rock and consequently in a facies. Isolated however, this attribute is not a sufficient property to discriminate between facies. Its quality structure contains 119 colors defined nominally according to the standard rock table NBS/ISCC RC.

- **Lithology:** Corresponds to classification of the rock that composes the facies, as a result of combination of chemical, mineralogic and textural aspects of the facies. Its quality structure contains 141 different lithotypes.

- **Bed gradation:** Indicates the gradation mode of the grain sizes along a facies, from base to top. Normal gradation indicates a reduction on granulometry from the base to the top of the rock layer. Inverse gradation indicates a gradation that is the opposite of the normal, with higher diameter grains on the top, gradating to lower diameter grains on the base. The facies can also present inexistent gradation, when it shows homogeneity between the grain sizes from the base to the top. Its quality structure contains: Inexistent, Normal, Inverse.

- **Laminae gradation:** Indicates an aspect analogous to bed gradation, considering however, the laminae that form the structures. In essence, it indicates the grain size gradation between two laminae of the structure that compose the facies. The measurement of this gradation is taken considering the direction of the force responsible for the creation of the laminae, which can be identified through observation of the smallest angle between the laminae and the base of the structure. Its quality structure contains: Inexistent, Normal, Inverse.

- **Specific grains:** Indicates the presence of certain types of specific grains, important for future predictions. Specific grains can be intraclasts, when they are reworked fragments of poorly consolidated sediments inside the same depositional basin, through the action of currents, landslides, etc; or extraclasts, when they are sedimentary fragments whose origins are external to the sedimentary environment. Its quality structure contains: Extraclasts, Intraclasts.

- **Fabric orientation:** Indicates the orientation of the grain that form the facies. Its quality structure contains: Chaotic, Imbricated, Heterogeneous, Homogeneous, Parallel, Sub-parallel, Without orientation.

- **Fabric support:** indicates the support of the sediment in the facies. Its quality structure contains: Crystalline fabric, Bioconstruction undifferentiated fabric, Bioconstruction baffle fabric, Bioconstruction encrusting fabric, Bioconstruction rigid framework fabric, Grain-supported, Matrix-supported, Cement-supported, Grain to cement-supported, Matrix to cement-supported, Grain to matrix-supported.,

- **Geometry:** Refers to the geometrical forms external to the layer which contains the facies, independent of its internal organization, along a determined lateral extension, as a result of sediment deposition. Often geometry is hard to reconstruct, due to the limitation of exposure. In rock outcrops, the description of geometry is based mainly in the observation of the thickness of the facies and their lateral continuity, in core samples however, the reconstruction of geometry is generally not possible. Its quality structure contains: Tabular, Lenticular concave-convex, Lenticular convex top plain base, Lenticular plain top convex base, Sigmoidal, Wedge and Irregular.

In figure 9.1 we present a fragment of the domain ontology representing the concept of sedimentary facies and its qualities.

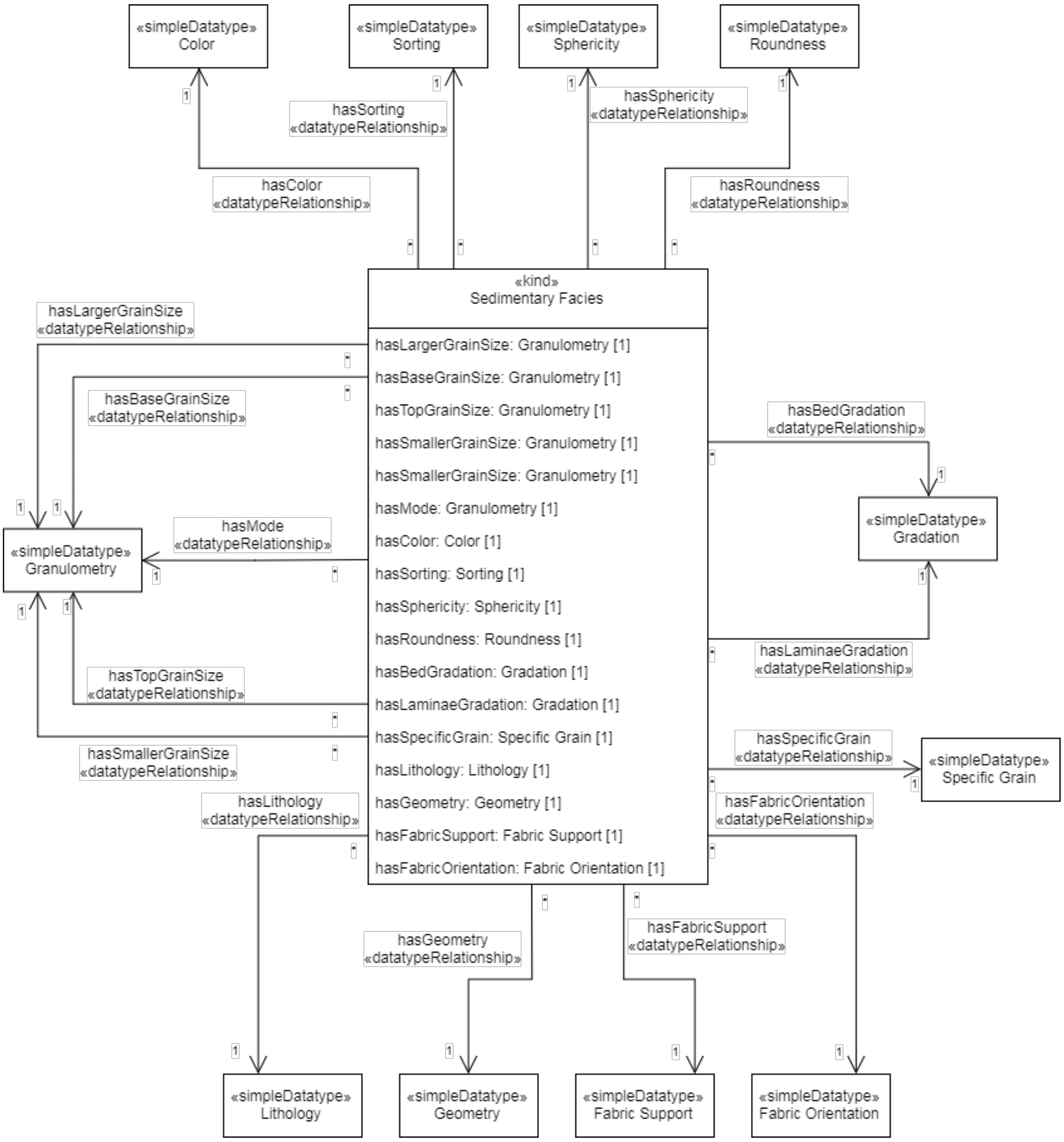


Figure 9.1: Sedimentary Facies and its attributes.

# ANNEX 2

The following plots show the log and core data from the real wells described in chapter 6. The leftmost column shows the facies described by the geologist in terms of lithology and grain size. The remaining columns show the log readings along the cored interval. These graphs were generated using the Strataledge®core description system.

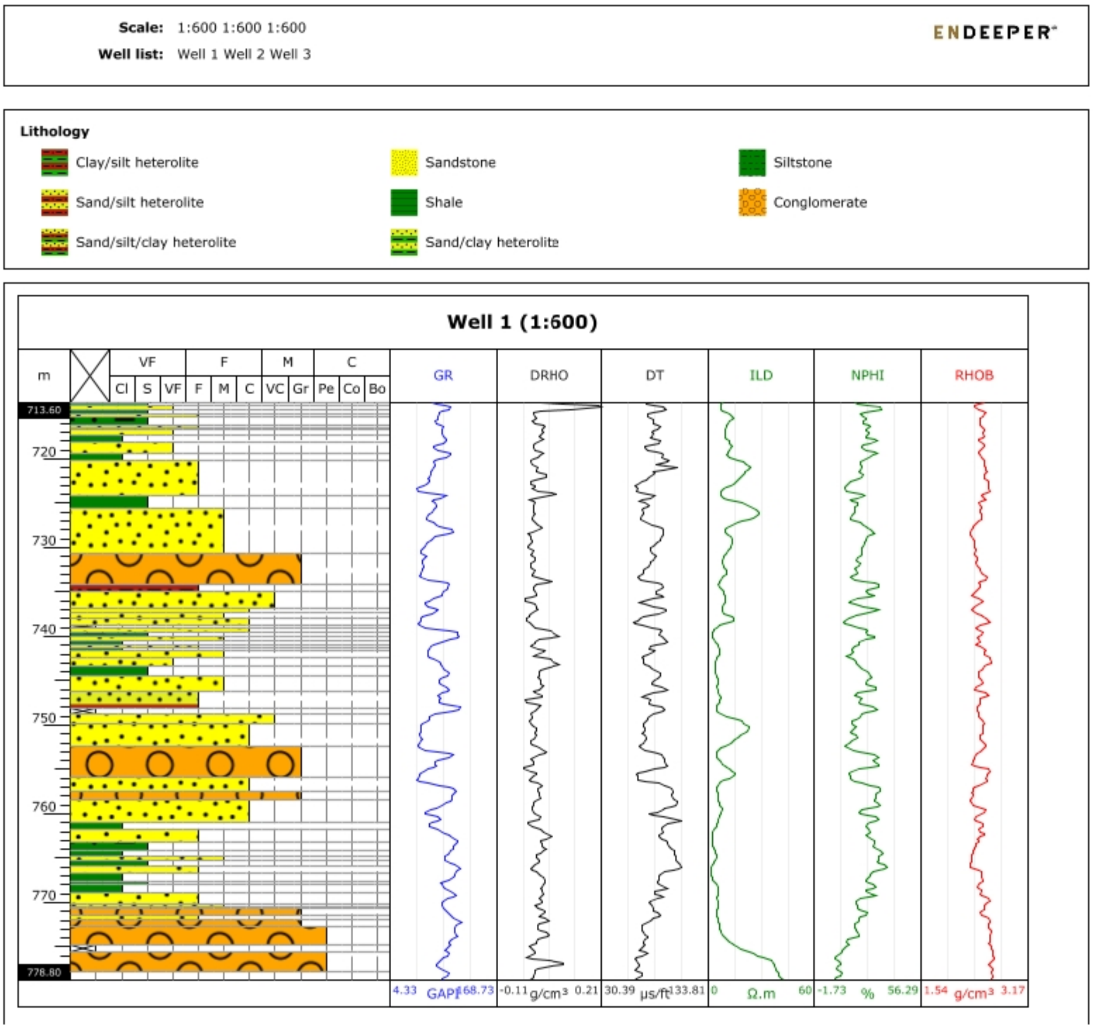


Figure 9.2: Plot of data from Well 1

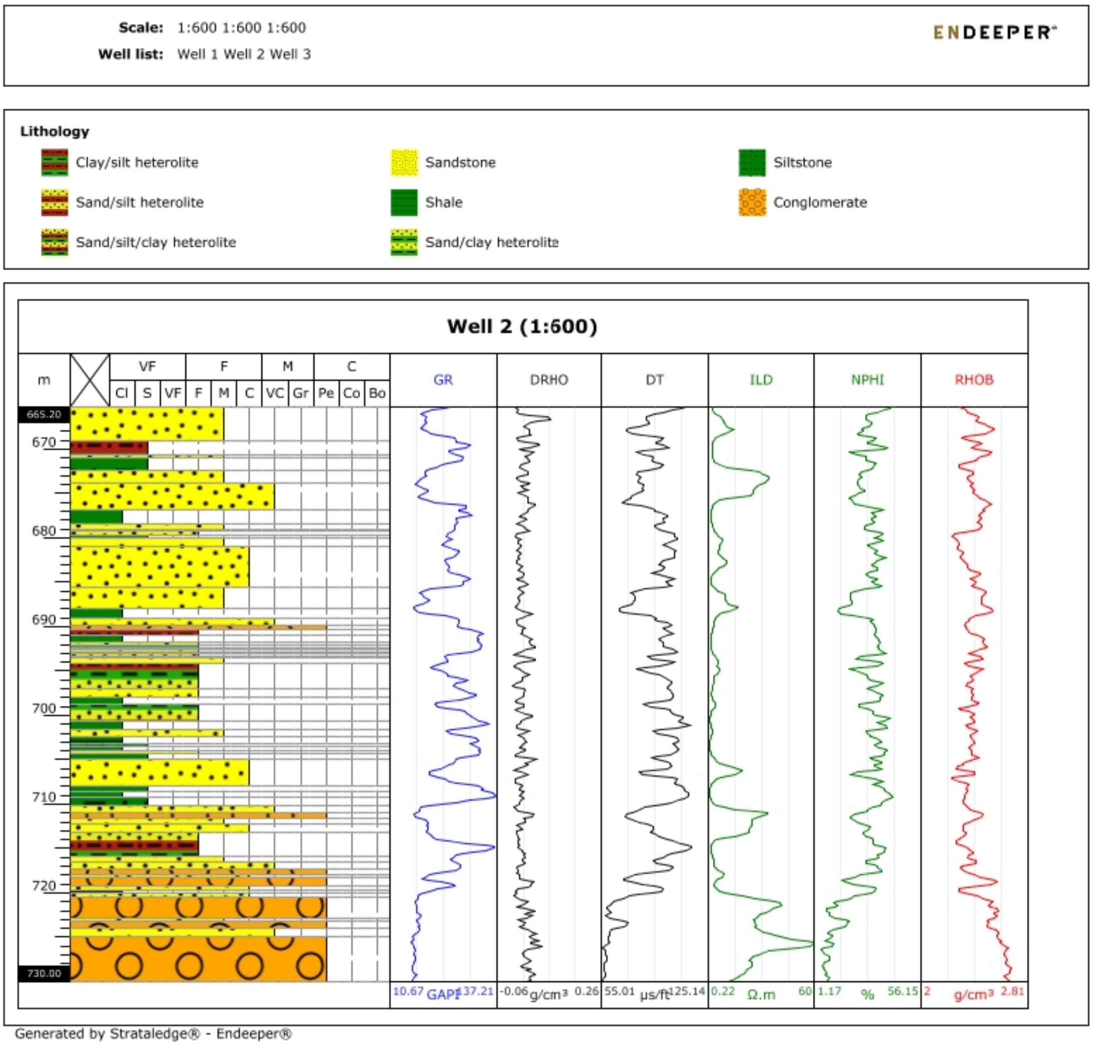
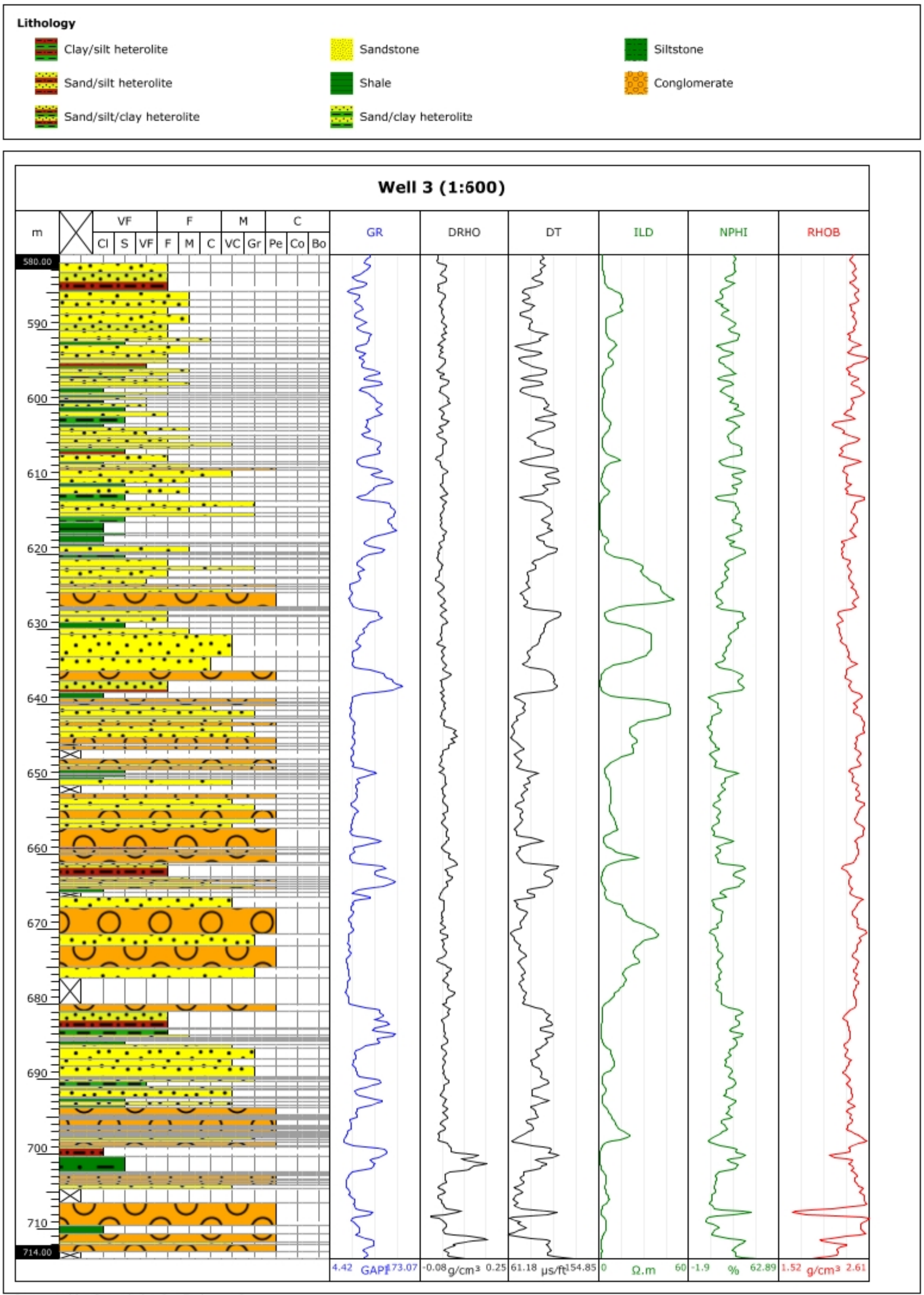


Figure 9.3: Plot of data from Well 2



Generated by Strataledge® - Endeeper®

Figure 9.4: Plot of data from Well 3