UNIVERSITY *of the*
WESTERN CAPE

The use of well log data in the creation of 3D geological maps

Charlene Omeniem Keletso Ile
3968692

Thesis submitted in fulfilment of the requirement
for the degree of Master of Science in Applied Geology
at the Department of Earth Sciences in the Faculty of Natural Sciences
for The University of the Western Cape

Supervisor: Dr M Opuwari

June 7, 2021

Declaration

I, Charlene Omeniem Keletso Ile, do hereby declare that this Masters dissertation - ***The use of well log data in the creation of 3D geological maps*** - submitted for the Masters Degree in Applied Geology for the Earth Sciences Department of the The University of the Western Cape, is entirely my own work and has not been submitted before for any degree or examination in any other university.

Signed: ____________________ Date: ____________________

Acknowledgements

Firstly, to my Lord and Saviour. Thank you for giving me strength in every aspect of my life.

To my parents. Thank you for being constant pillars of support in my life and for always pushing me to be great. I love you. Thank you.

To my sisters. Thank you for being my sounding boards and for your unwavering belief in me. I love you. Thank you.

To my supervisor - Dr Opuwari. Thank you for not only guiding me through this process, but for also giving me the room to develop as a researcher.

Abstract

Three-dimensional (3D) graphic representations of geographic environments have become commonplace in a range of fields. These representations are often an attempt to represent both geographic forms, as well as the relationships that exist between them. In contrast to other fields, the use of 3D geological models in the visualisation of the subsurface environment is relatively new. Additionally, these 3D geological models are traditionally created through the painstaking process of manual development methods. As such, the models developed are unable to fully utilise the wealth of geological data that is collected during subsurface exploration.

Therefore, the objective of this research was to create a 3D geological prototype that allowed for the visualisation of underground resource reservoirs in a faster, easier and more aesthetically appealing manner. To achieve the objectives of this research, the problem was tackled holistically by considering both the theoretical and practical components of the research. Some theoretical components that were considered are: well log – wireline log – data composition, the information that can be extracted from each well log component, geological data interpolation as well as geological visualisation. Utilisation of the theoretical component of this research facilitated the development of a programme that modelled and visualised sub-surface environments. The programme applied the information from numerous well log datasets and interpolated the various geological layers that could be found within a region.

This research used the machine perception process as the approach to develop a 3D prototype of the Bredasdorp Basin. The steps involved were made up of 5 overarching steps: data collection (acquired from the Petroleum Agency of South Africa), data preprocessing, feature extraction, data clustering and data post-processing.

As part of this research the optimal number of components to explain 95% of the data distribution was determined to be 6 components for the processed dataset. Following this 3 clusters were determined to produce the best cluster separation. The identified clusters were meant to distinguish between lithological sequences in the region, however, when examined further they did not match the expected results. A number of factors were linked to performance of the prototype, these included the distribution, consistency and imputation of the data.

Nevertheless, the research has possible implications on viewer perception in well log interpretation as well as the applicability of machine learning in the field. Following on from this research, a number of future directions can be taken with one being the incorporation of pseudo labelling in the clustering of well logs.

Contents

List of Figures

List of Tables

Nomenclature

2D	Two dimensional
3D	Three dimensional
ML	Machine Learning
GIS	Geographical Information Systems

1 Introduction

1.1 Background

Geology can be described as a descriptive science, within which scientists are consistently attempting to describe rock materials and communicate these descriptions (Peveraro, 2006). Well log data interpretation is no different. Research into the genesis of the discipline and the artistic, comprehensive, verbose and subjective illustrations that were created during the well log interpretation process, makes it apparent that well log data interpretation is inherently visual. Its dependence on visual and descriptive mediums can be credited to the fact that it is a representation of a complex, multi-dimensional subject - the geological subsurface (Jones et al., 2009).

Therefore these drawings and writings, although as detailed and descriptive as they could be, failed to fully express the core message geologists sought to portray, which is an accurate, comprehensive and recognisable representation of the subsurface and all of its complex relationships (Jones et al., 2009). Attempts to correct this propelled well log interpretation into what it is today (Peveraro, 2006). Well log interpretation focuses on measuring, recording and displaying rock material characteristics, and then deriving descriptive geological parameters from the measured quantities (Peveraro, 2006). Simply put, the practice of well log – wireline log – interpretation looks at inferring and conveying the geological characteristics (e.g. lithology) of a region by measuring the properties of rocks that surround several boreholes.

To map the geological subsurface of a region using the well log interpretation process, probe instruments (sonde) that contain multiple sensors are extended into a well, so that sedimentary property recordings can be made. As the sonde and its sensors are pulled up from the depths of the well, they transmit information about their surroundings (Luthi, 2001). Although, it is possible to post process the information gathered from the sonde's various sensors by using a range of applications both proprietary and freely available, they are often limited to two-dimensional (2D) visualisations. This is problematic because as mentioned above, relying on 2D geo-modelling solutions ultimately compromises the display of 3D data (Jones et al., 2009).

This research was, thus, carried out to answer the following question: Could more information be conveyed about the complex and highly entangled geological subsurface - through the creation of mindful 3D cartographic maps? Therefore, the objective of this investigation was to determine groups (clusters) within well log data and visualise them three-dimensionally; while also applying cartographic techniques that promoted aesthetic appeal and user comprehension.

To achieve this objective - creating appealing and recognisable 3D subsurface visualisations - certain theoretical underpinnings had to be considered and used in the development of a 3D cartographic prototype that ensures aesthetic appeal and improves understanding. Topics that were explored include the practice of well logging, machine learning (ML), geographical information systems (GIS) and the framework for cartographic design.

However, the overarching goal of this research was to contribute to the quality and versatility of well log data interpretation worldwide and, more specifically in the Bredasdorp Basin - one of the sub-basins of the Outeniqua Basin and the chosen region of study.

The Outeniqua Basin and its sub-basins (four major and one minor), all give record of a rich geological history of South Africa's south coast. Continental rifting between east and west Gondwana, extensional forces on the southern plate margins, and the subsequent thermal subsidence and late drift sedimentation in well oxygenated environments, all contributed to the formation of the enlarged basin parallel to the remnant continental shelf break – the Outeniqua basin (McMillan et al., 1997).

Since the Outeniqua Basin's discovery, countless in-depth explorations have been carried out for hydrocarbon prospecting purposes using seismic, deep borehole drilling and other geological acquisition methods (McMillan et al., 1997). Therefore, hundreds of wells have been drilled in Outeniqua's four major sub-basins and in its one of its minor sub-basin i.e. the Bredasdorp, Pletmos, Gamtoos, Algoa and Infanta sub-basins respectively (see figure 1.1).

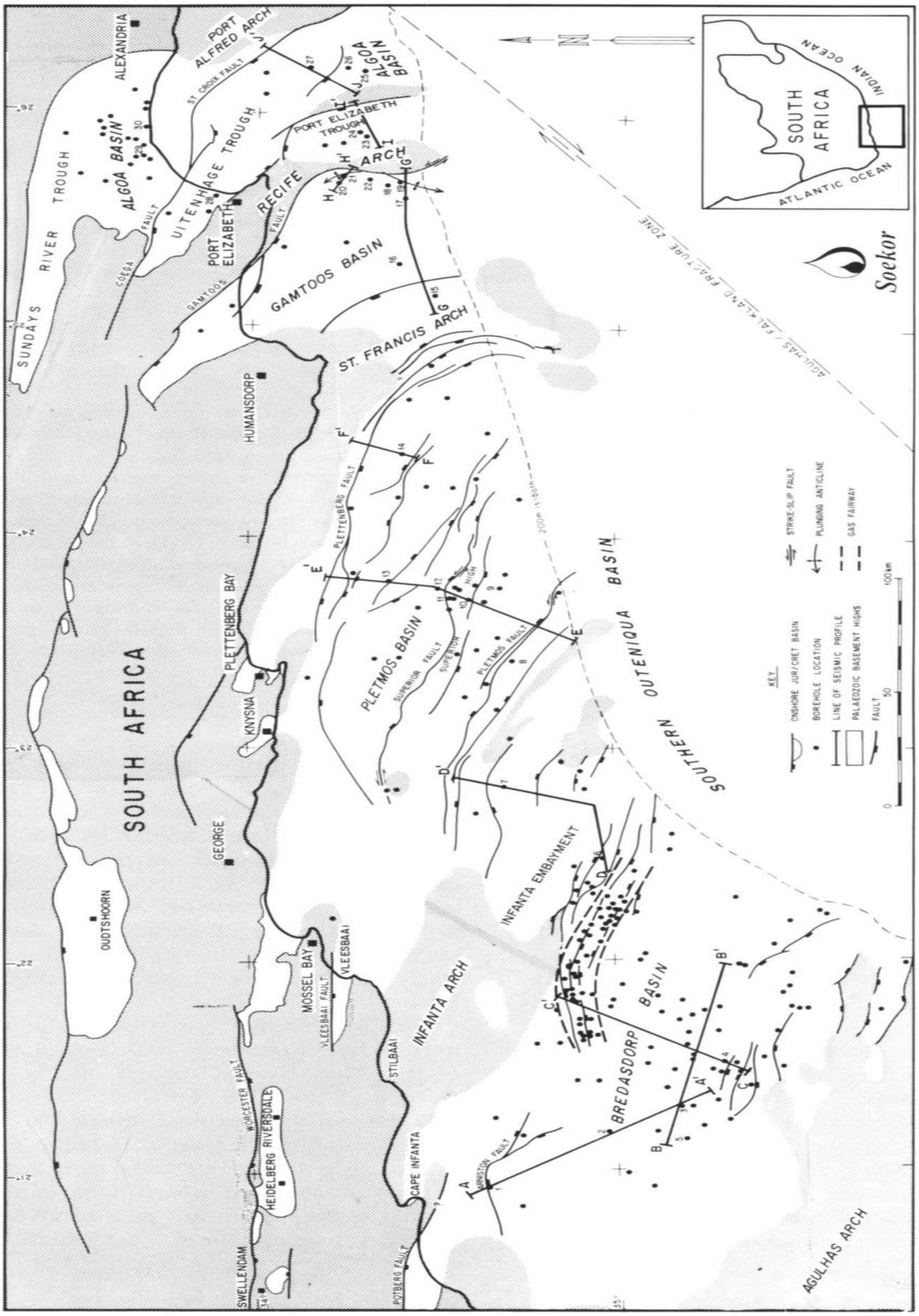


Figure 1.1: (Broad, 1990) Location map of South African southern offshore sedimentary basins. Numbered boreholes are referred to in the text.

However, due to high water depths and strong currents in the distal portions of the Outeniqua basin, most of these prospecting efforts have been focused on the Bredasdorp basin as it is the richest and most viable source of hydrocarbons (McMillan et al., 1997). Therefore, pertinent literature and theoretical frameworks were applied to well logs from the Bredasdorp basin to develop a 3D geological prototype.

1.2 Statement of the Problem

The manner in which sedimentary properties gathered from well logging are represented has a direct effect on the ability of geological users at all levels of expertise to make meaningful conclusions about the geological landscape of a region. Although most present day well log interpretation applications are suitable for their purpose, they are hindered by their dimensionality, which can lead to costly time and resource expenses. Dimensionality hinders these applications because they visualise recorded sedimentary properties in 2D space and thus do not adequately support the observation of the complex and multi-dimensional subsurface environments (Jones et al., 2009). Additionally, these representations can both be aesthetically unappealing and difficult to understand.

Hence, the driving research question behind this study was as follows:

While honouring vision and perception theory, could the complex and highly entangled geological subsurface be represented as a 3D geological prototype using data processed from well logs?

Given the above, the pertinent readings and theoretical frameworks were reviewed before applying them to assist in the development of the readable and aesthetically appealing 3D geological prototype.

1.3 Rationale of the Study

Presently, most well logging applications help maximise the value of geological subsurface data by transforming this data into visual, actionable information. However, the versatility and capacity of these applications are hindered by their two dimensionality. As the subject being displayed - the subsurface - is a three-dimensional physical environment, the need to deliver a 3D modelling solutions is of great importance because it would support confident decision making (Ford et al., 2008). Additionally, as noted by Jones et al. (2009), 3D visualisation allows for the depiction of complex geological structures, and is more inherently intuitive than standard methods.

Therefore this study explored 3D well log visualisation as a means of conveying more information about the complex geological subsurface while also ensuring for aesthetic appeal to increase the ease of use and understanding of this information. Emerging insights from this investigation will raise awareness and increase understanding about the 3D method of geo-modelling, which (unlike traditional methods) is unlimited by dimensionality (Jones et al., 2009).

1.4 Study Objectives

The objectives of this investigation were:

1. To present a geological understanding of the Bredasdorp Basin.
2. To demonstrate a clear understanding of what a well log is.
3. To determine the well logs that can be applied to 3D geological model development.
4. To identify and explain the fundamental characteristics that facilitate in user understanding and aesthetic appeal when working with cartographic representations.
5. To implement principal component analysis (PCA) and Kmeans clustering on well log data and interpret the results.
6. To develop a prototype of a 3D geological map that supports aesthetic appeal and user comprehension by adapting and combining the best practices within existing 3D modelling theory.

1.5 Concept Clarification

Python is an interpreted, high level and general purpose programming language that was first developed by Guido van Rossum in the 1980s. This programming language separates itself from its counterparts through its emphasis on code readability and multi-programming paradigm support (Van Rossum et al., 2007). Python libraries support functionalities that range from data extraction and conversion to data analysis and modelling.

Unlike traditional geological models, **three-dimensional (3D) geological models** are continuous representations of the subsurface generated from powerful modelling tools (Van der Meulen et al., 2013). These subsurface representations often include geological properties, distribution and architecture (such as lithology) (Song et al., 2019).

3D geological mapping is a multifaceted topic that deals with the three-dimensional visualistion of geological data (e.g. faults, lithology and volume) in a intuitive and choherent manner that is suitable for human perception and interpretation (Malolepszy, 2005).

Acording to Hyne (2014), **lithology** (such as sandstones, limestones, claystones and shales) is the general physical characteristics of rocks, and a common way of lithological determination is through the interpretation of well logs.

Well logs record the formation properties of an area for a given depth during the well logging process (Delfiner et al., 1987), with some geological properties captured in well logs including resistivity and porosity.

Geological basins are ovular, circular or bowl-shaped depressions in the Earth's surface, that arise from either erosion or rifting. Their low-laying nature means that they are often filled with water or sediments, thus making them good records of palaeoclimates (Rutledge et al., 2011). Three major basin types are: river drainage basins, structural basins, and oceanic basins.

1.6 Scope

Despite the fact that other concerns and issues arose from this investigation, the investigation concentrated only on matters that affected well log data comprehension and interpretation, as well as, 3D geological visualisation and prototyping. Overall, the issues that emerged in this study had an impact on the creation of an aesthetically appealing and intelligible 3D geological prototype based on well log data.

This research focused on issues, concerns and information gathered from research into well logs, ML, 3D cartographic prototype development and 3D GIS. The information gathered about the above mentioned topics were exclusively applied to the well log data gathered from the Bredasdorp geological region. Aside from topographic data for the region, no further information was considered and no additional data was gathered by any other means other than that which is stated above, including those surrounding seismic data and website development.

Three elements were constraints (limitations) in this investigation. They are time, money and the amount of information that could feasibly be gathered and effectively utilised.

1.7 Outline of Chapters

In this study there are five major chapters: the introduction, literature review, design/methodology, findings/results and conclusion. The short overview below highlights the structure and arrangement of the research conducted in this thesis.

Chapter 1

This chapter introduced the subject matter for investigation - *The use of well log data in the creation of 3D geological maps* - and provided context as to what would be presented in the thesis and why. In addition to the preliminary notes, it included

- The background of the research
- A general statement of the problem
- The rationale of the study
- The objectives of the research
- Concept clarification
- The scope of the research
- And the chapter outline for the reseacrh

Chapter 2

Relevant literature and frameworks had to be examined in order to carry out the objectives of this investigation - creation of a 3D geological model that is bolstered by cartographic comprehension and aesthetic appeal principles. Therefore, this chapter outlined some of the information that previous academics have produced, including:

- An overview of the geology of the region of interest - the Bredasdorp basin
- A narrative of well logging that covers a brief look into its history and the data that can be derived from it

- An outline of machine learning
- A review of the 3D geological cartographic design cycle and it's related components

Chapter 3

In order to provide sufficient detail about the experiment, this chapter covered the methodological aspects of the investigation. This included the type of research undertaken, the data that was collected, the tools and materials used to achieve the objectives of the research and why these methods were chosen.

Chapter 4

Chapter 4 covered the results and discussion of the investigation. The results portion of the chapter set out the key experimental results and whether the results were significant or not.

The discussion portion of the chapter examined the results in the context of the literature and established knowledge on the subject. The limitations of the research and the implications of the findings were also discussed, and the study was critically evaluated.

Chapter 5

In addition to identifying areas for future research and making recommendations, this chapter covered the critical aspects identified in the development and analysis of the geological map before concluding the overall investigation.

2 Literature Review

2.1 Introduction

Well log interpretation looks at measuring and recording rock material characteristics, and then deriving descriptive geological parameters from the measured quantities (Peveraro, 2006). In addition well logging is not only the recording and interpretation of geological quantities, but it is also the creation of meaningful visual representations. This is especially true as the data being represented is inherently visual.

However, the broader effect of well log interpretation, beyond being just a functional graphical representation, has not been extensively considered or explored. Therefore, this investigation was performed with the aim of transcending the typical and often confining two-dimensional (2D) well interpretation research that considered interpreted well logs as merely uncontextualised 2D representations, and re-orienting the discussion towards three-dimensional (3D) cartographic well representations with user appeal and comprehension.

In this chapter pertinent perspectives, literature and theoretical underpinnings on the use of unsupervised machine learning in the development of a 3D geological prototype that promotes aesthetic appeal and enhances user comprehension are discussed. In addition the geological setting of the region of interest and the wireline data that was collected from it are briefly looked at. This was done before delving into the literature surrounding the development of an unsupervised machine learning (clustering) model. Lastly, research into the concept of 3D cartography and its relation to how maps are seen and understood, was conducted.

2.2 The Geological Setting

The geological history of southern Africa, so far, spans about 3.8 billion years (Tankard et al., 2012) and gives account of years of gradual sediment accumulation and loss. Focusing on the middle to late Jurassic period gives insight into the geological and tectonic processes that resulted in the formation of the Bredasdorp Basin (Parsiegla et al., 2009), the setting for this research.

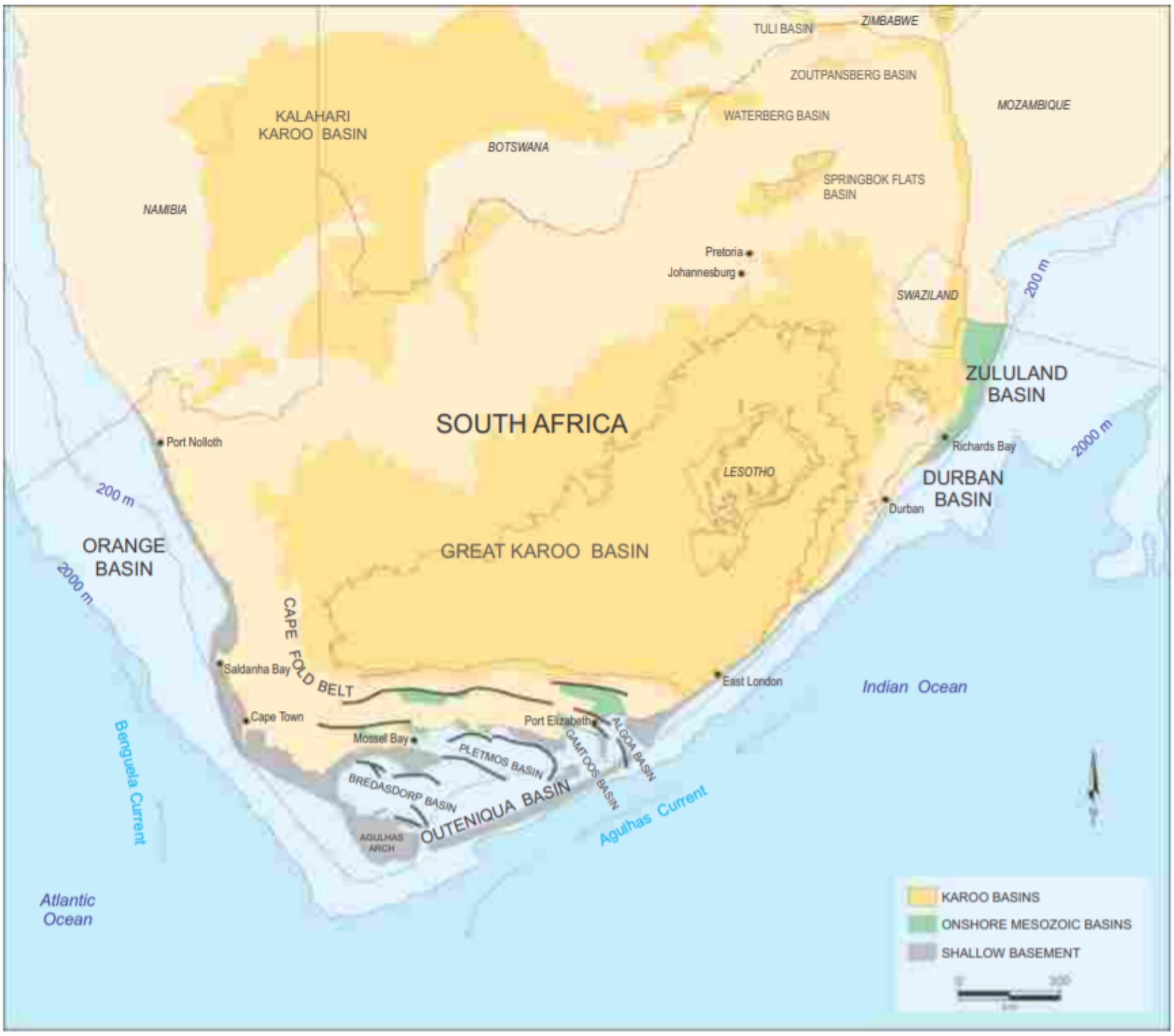


Figure 2.1: Location map of South African southern offshore sedimentary basins - including the region of interest, the Bredasdorp Basin, and its parent basin, the Outeniqua Basin (Petroleum Agency of South Africa, 2003)

The Bredasdorp Basin, a large geological repository located off South Africa's continental shelf, lies between Mossel Bay and the Cape Agulhas (see Figure 2.1) (Masindi, 2016). It is bounded by two major basement arches – the Infanta Arch to the north-east and the Agulhas Arch to the south-west. These arches are oriented parallel to the structural grain of the orogenic Cape Fold Belt and set out an elongated basin with a width of about 80 km and a length of about 200 km. This has resulted in a basin (the Bredasdorp Basin) that spans an area of around 18000 km^2. This 18000 km^2 basin is filled with sediments both from the time of continental rifting and the period after. In the basin, Upper Jurassic and Lower Cretaceous continental and marine deposits chronicle the time of continental rifting. While Cretaceous and Cenozoic divergent margin rocks tell of the sedimentation during the post-rift period (Brown et al., 1995).

To better understand the geological setting and formation of the Bredasdorp Basin, the geo-history of the basin and of its parent basin – the Outeniqua Basin – needed to be explored.

Basin Evolution: The Bredasdorp Basin

The Bredasdorp Basin is one of the four major and one minor offshore depocenters of the Outeniqua Basin. The others are the Pletmos, Gamtoos, Algoa and Infanta sub-basins (McMillan et al., 1997). The Outeniqua Basin is thus a collection of both small fault bounded and deeper sub basins located off the coast of the southern South African continental margin (Parsiegla et al., 2009). The deeper sub basins of the Outeniqua Basin are oriented closely to the Agulhas-Falkland Fracture Zone (AFFZ), a mid-ocean valley that runs from the northern edge of the Falkland Plateau to the southern edge of the African continent, and which forms the border between the continental and oceanic crusts (McMillan et al., 1997).

The Outeniqua basin for the most part consists of mid Aptian to Maastrichtian deposits on top of pre-existing rift basins and, according to McMillan et al. (1997), developed as a result of three main episodes: rift, transitional and drift episodes.

In the first episode, continental rifting occurred between the East (Antarctica-Australia-India) and West (South America) sections of Gondwana in the middle-late Jurassic to Valanginian era. Rifting between east and west Gondwana is said to have occurred along the progressively floundering rift zone between the Australasian and African Plate (Khana and Dillay, 1986). Continental rifting and the position of the Outeniqua Basin relative to the plate margin meant that it was sheared by right-lateral movements. This event was unlike the stress events that occurred at other margins in the rest of southern

Africa, where (instead of right lateral movements) extensional pull apart movements were experienced (McMillan et al., 1997).

Continental rifting was then followed up by a transitional episode (late Valanginian – early Aptian era), before the basin formation concluded with a drift episode (early Aptian to present day) (McMillan et al., 1997).

Although a post rift formation, analysis of borehole samples from the Bredasdorp basin have revealed that the region that later became the Bredasdorp basin started experiencing continental rifting around the middle-to-late Jurassic period. Extensional stress, experienced because of the breakup between the Falkland Plateau (a complex series of micro plates) and the Mozambique Ridge during continental rifting, induced normal faulting. This in turn supported the definition of elongated horsts (raised blocks of land), grabens and half grabens in the region (Brown et al., 1995). It was in these (half) graben basins – depressed blocks of land with parallel banding faults which arise from blocks of land being downthrown – that sediments were deposited. Sediments such as clastic, fluvial and shallow marine deposits were lain. According to McMillan et al. (1997) these sediments were deposited in marine and non-marine sediment successions and are made up of four main lithogenic sequences: namely a lower fluvial interval, a lower shallow marine interval, an upper fluvial interval and an upper shallow marine interval. These landward (transgressive) and seaward (regressive) sequences were primarily induced by syn-depositional normal faulting, and account for the thick successively deposited sediments visible today.

The geological history of this period is summarised in the table below:

Table 2.1: Summarised descriptions of the rift phase episodes

Rift Phase	
Interval name	Interval summary
Lower fluvial interval	The sediments in this interval were deposited by the unconsolidated detrital (alluvial) material fans and flood plains of transverse fluvial systems. Sediments such as red and minor green fine-grained clay sedimentary rocks, reddish sandstones and conglomerates (McMillan et al., 1997).
Lower shallow marine interval	The erosional regional unconformity in the lower shallow marine interval indicates the first marine incursion. As such, clean, fine-grained, well sorted, micro- and macro-fossil rich glauconitic sands overlay the red bed fluvial incursion (McMillan et al., 1997). During this period the occurrence of fault reactivated syn-depositional differential subsidence and strand barrier construction (because of the seaward growth of shore and near shore deposits) led to the coarsening and thickening of sandstones in the interval.
Upper fluvial interval	The overlying upper fluvial interval is characterised by the grain size of its deposits, which gradually decrease in size towards the top. This has been attributed to the fact that the deposits accumulated in an alluvial flood plain, with multiple wondering fluvial channels. The types of rocks in this interval (non-glauconitic sandstones, red and green claystones and siltstones) further reflect this fact (McMillan et al., 1997).

Upper shallow marine	The upper shallow marine interval is composed of variably glauconitic, quartz rich and lithic poor sandstones that developed above an unconformity that followed a second marine incursion. This second incursion is said to have been triggered by the intermittent reactivation of basin margin faulting. This is confirmed by the presence of exceptionally thick, stacked cycles of syn-sedimentary tectonic settings in the interval (McMillan et al., 1997). Although as a result of a marine incursion, the presence of micro- and macro- fossil material is rare and spatially restricted in this interval, unlike in the lower shallow marine interval (McMillan et al., 1997). According to McMillan et al. (1997) the marine transgression was then followed by a series of geologic events. Firstly, there was sea level rise and shoreline retreat (to higher ground), which caused flooding. This was followed by the seaward movement of shore sediments which caused the exposure of previously submerged sea floor. Due to their significant porosities and permeabilities, the upper shallow marine sandstones in this interval are the best gas reservoirs in the area (McMillan et al., 1997).

Following the Upper Shallow Marine interval, sedimentation continued until there was major differential subsidence and the rifting induced extensional stresses ceased. According McMillan et al. (1997) and Brown et al. (1995), the uplift and truncation of the underlying geological deposits is marked by the late Valanginian drift onset unconformity (1At1) more than 126 Ma.

After the Valanginian drift onset unconformity (1At1) marked the end of continental rifting, and before the onset of drifting, the transitional activities of thermal subsidence and reactivated faulting occurred. This is known as the transitional rift-drift phase. During this period subsidence was uniform, slow and thermally driven. Additionally, sediments and uplifted structural highs, such as horsts and bounding arches, were variably eroded because of the slower subsidence rates of the period (McMillan et al., 1997).

Sedimentation during this period occurred in deep, poorly oxygenated marine areas overlain by poorly circulating water columns. This depositional environment resulted in the deposition of mostly argillaceous (clay-rich) marine sediments, and then their transportation to deeper waters by rapid downhill currents. Also as a result of the depositional environment's unsustainable biogenic oxygen levels signs of benthonic life are rare or regionally confined (McMillan et al., 1997).

The geological history of this period is summarised in the table below:

Table 2.2: Summarised descriptions of the transitional rift-drift phase episodes

Transitional rift-drift Phase	
Interval name	Interval summary
1Atl to 5Atl (Late Valanginian to Hauterivian).	During the 1At1 to 5At1 period of the transitional phase, distal clay-rich sediments in the basin accrued in poorly oxygenated conditions. Additionally, southerly inclined submarine valleys and canyons broke up the pre-unconformity (1At1) shallow marine sandstones into distinct areas which both provided a conduit for sediment flow and pockets for gas trapping (McMillan et al., 1997).

5Atl to 13Atl (Barremian to Early Aptian)	Sediments in the 5At1 to 13At1 period transitional phase of the Bredasdorp basin formation, were deposited in a poorly circulating and poorly oxygenated environment. Also, during this period, the seaward movement of clean highly porous coastal sands was defined by both a northern margin and by the Infanta Arch (McMillan et al., 1997). During the early Barremian period (6Atl) there were 3 major channels cut into both the 6Atl sedimentary surface and the pre-1Atl rocks (i.e. upper shallow marine interval). These channels thus acted as conduits for sedimentary flow between proximal and distal portions of the basin and assisted in the formation of clay plugged gas trapping canyons (McMillan et al., 1997). Approaching the Early Aptian (13Atl) there was both an uptick in sandstone deposition, and a marked decline in the faulting subsidence rate (McMillan et al., 1997).

The end of the transitional rift drift phase occurred around the mid Albian period and was followed by a drift phase. This period is marked by two things, firstly the separation of the Falkland Plateau from Africa, and then the slow south-westerly migration of the Falkland Plateau past the coast of Southern Africa (McMillan et al., 1997).

These activities subsequently led to the establishment of a true passive margin, as well as the formation of some oblique rift half-graben sub-basins – such as the Bredasdorp basin.

The geological history of this period is summarised in the table below:

Table 2.3: Summarised descriptions of the drift phase: 13Atl to present day

Transitional rift-drift Phase	
Interval name	Interval summary
13Atl to 15Atl	The Early Aptian 13Atl unconformity ushered in a different sedimentation regimen and revealed the onset of subsidence in the central and southern portions of the Bredasdorp Basin. For the most part erosion was confined to the deeper portions of the basin, where the 13At1 unconformity was cut by an easterly-trending 5 by 50 km submarine channel (McMillan et al., 1997). Similar to the channels in the 6Atl sedimentary surface and the pre-1Atl rocks, this channelling system (the 13A channel) served as an aqueduct for the mass transport of thinly bedded turbides from the shallow to deeper reaches of the basin. However, the steep inclination of southerly regions of the basin hindered the advancement of the 13A channel into those areas. Later, the updip tributaries of the channel were clay plugged, and thus supported the accumulation and trapping of oil reservoirs (McMillan et al., 1997). The organically enriched interval overlying the 13Atl unconformity is differentiated by its northerly high-gamma organic rich claystones and westerly sand rich deposits. Although deposited in a poorly oxygenated environment, the volume of hydrocarbons that accumulated within the rock pores make this interval one of the most mature and viable oil and gas reservoir rocks (source rocks) in the basin (McMillan et al., 1997). Following the period of anoxic accumulation, there was an abrupt return of well-oxygenated seafloor conditions and improved water circulation. Evidence of this can be seen in the widespread presence of a diverse range of micro- and macro fauna (McMillan et al., 1997).

15Atl to Present Day	Sedimentation during the Cretaceous era (from the 15At1 unconformity to the L horizon) shows evidence of continental formation in the relatively thick and widespread seaward moving deposits. During this interval, although there was a decrease in subsidence rates, the abundant sandstones in the interval denote the continued high clastic input rate (McMillan et al., 1997). During the Late Cenomanian era there was a minor episode of upliftment and warping which resulted in the formation of the 15Atl unconformity. The unconformity interval was later eroded, especially in the distal eastern parts of the basin, by micro-fauna before being overlain by Early Turonian anoxic organically- and plankton-enriched sediment-starved claystone (McMillan et al., 1997). From the Turonian to mid-Coniacian age (15At1 to horizon K) there was a prograding episode where coastal sediments grow further seaward. Following this age clastic sedimentation decreased, biogenic sediments were laid down and multiple episode of truncation occurred (namely at the late Santonian-early Campanian horizon X and the Maastrichtian-Palaeocene boundary horizon L) (McMillan et al., 1997). The upliftment of the Agulhas Arch, just before horizon L, caused firstly the erosion of Late Cretaceous sediments and then the redisposition of these sedimentary debris into the Palaeocene shallow-marine glauconitic sands. Post horizon L (Tertiary to present day), the sediments deposited were almost exclusively glauconitic clays, sands and biogenic clays. Unconformities from the Late Oligocene, Miocene, Holocene and Late Pleistocene follow the Maastrichtian-Palaeocene boundary horizon, and give account of sedimentation during these periods. Although most of them, excluding the Early Miocene rocks, are thinly bedded layers (McMillan et al., 1997).

The above geo-history of the Bredasdorp Basin, a depocenter of the larger Outeniqua Basin, is summarised in the form of a chronostratigraphic correlation chart below:

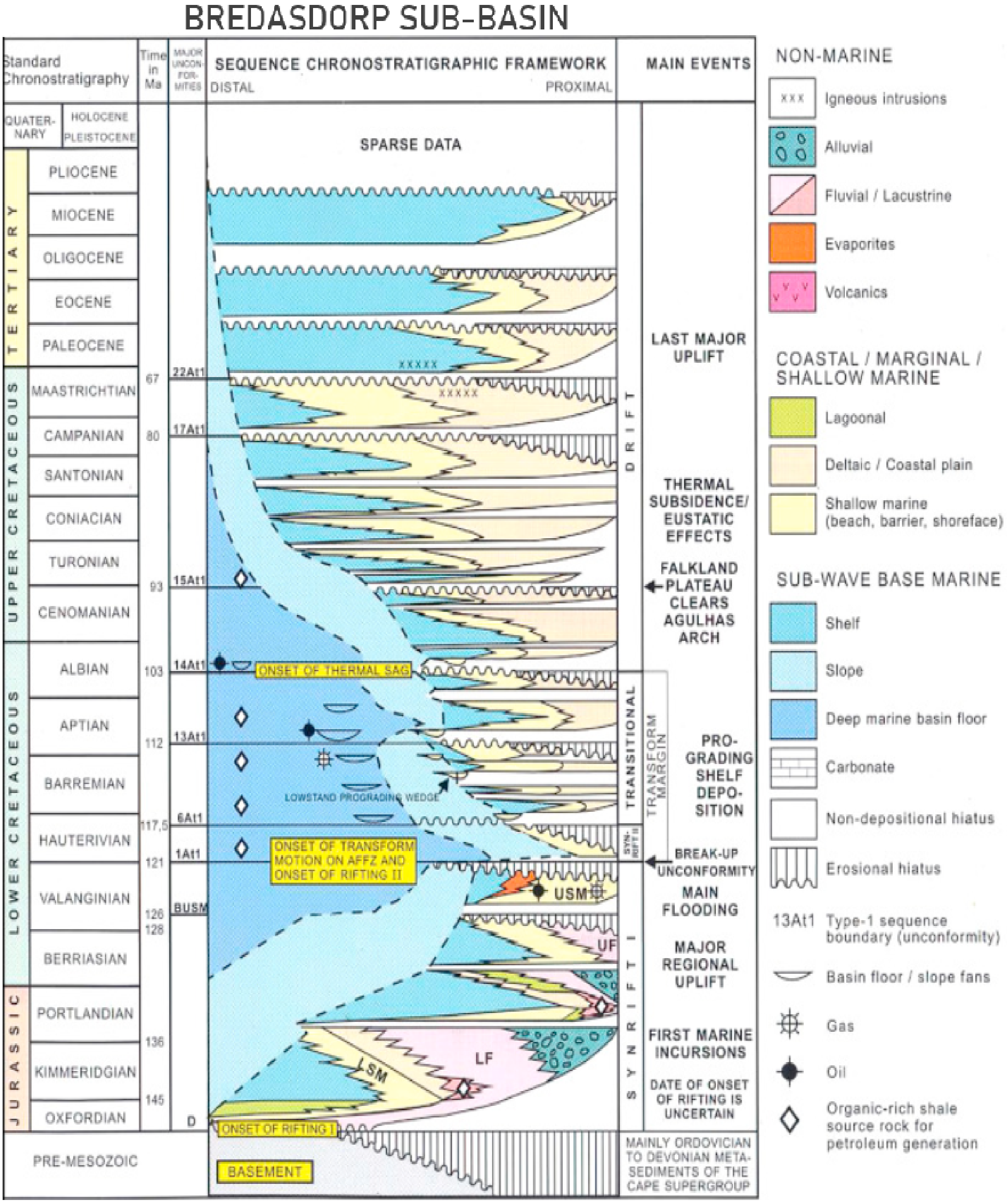


Figure 2.2: Generalised chronostratigraphy of the Bredasdorp basin (Petroleum Agency of South Africa, 2012)

2.3 Well Logs

Hydrocarbons, an invaluable non-renewable source of energy, are found in abundance in subsurface environments. However, these subsurface landscapes are tricky terrains to access and navigate. Thus, the processes of hydrocarbon exploration and exploitation often become challenging and expensive endeavors (Peveraro, 2006).

Due to a greater understanding of geological landscapes, as well as, the advent of technologically advanced machines that are used in geophysics and geo-engineering, hydrocarbon prospecting has become an almost common practice (Peveraro, 2006).

To determine areas that will yield high hydrocarbon reservoirs, accurate lithological information about subsurface environments need to be obtained from coring and drill cuttings. The cores obtained from these activities are applied to the well logging processes which allow for the determination of subsurface physical properties and lithology with respect to depth (Peveraro, 2006).

Well logging provides a cheap, quick and accurate method of obtaining subsurface petrophysical data like density, resistivity and travel time. These parameters are in turn used for hydrocarbon identification and quantification of potential pay zones and hydrocarbon reserves (Peveraro, 2006).

The hydrocarbon exploration process often begins with geological and geophysical surveys. These surveys are used to determine the types of hydrocarbons present in the subsurface by gathering information about the rock and sediment physical properties, without the expensive undertaking of tunneling or digging (Peveraro, 2006).

After surveys have been carried out wells are drilled. The drilled wells are used to confirm the existence of hydrocarbon bearing geological traps and quantify the possible pay zone by mapping of petro-physical properties against depth (Peveraro, 2006). This step in the process is generally what people in the geo-related field refer to when they speak of 'well logging'. Well logging can be carried out using one or a combination of techniques.

One technique is Measurement While Drilling (MWD), where (during the drilling process) the composition of rock samples are collected to be examined later against their depths in a specialised laboratory, (Peveraro, 2006). Another technique is Logging While Drilling (LWD), whereby sonde (probe instruments that contain multiple sensors) are used to take continuous measurements of a wells petro-physical properties against the depth of the well. These probe instruments are extended into a well using a steel cable. From the well they later transmit information about their surroundings as they are pulled up (Luthi, 2001) (see figure 2.3). The properties recorded by this method depends on the sensors and tools that are used during the well logging process e.g. resistivity tools, sonic tools, etc. (Peveraro, 2006).

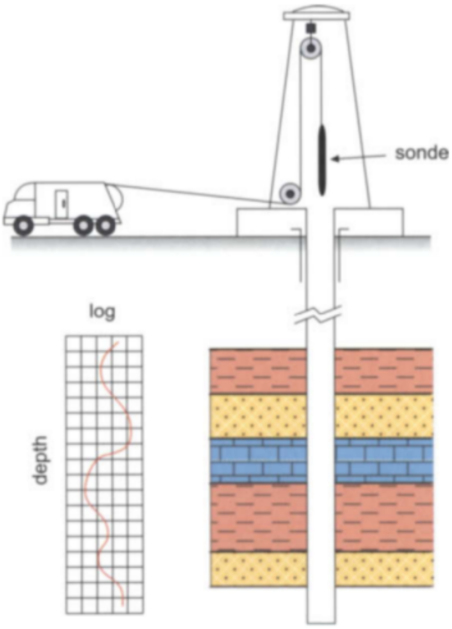


Figure 2.3: Principal of well logging (Jahn et al., 2008)

The societal importance and profitability of well logging, has meant that subsurface mapping, strata identification and the tools/methods used during these processes are topics that have been extensively covered by both academics and oil purveyors alike - with countless books, papers and charts being published on the topic. However, the topic and practice of well logging, as we know it today, can be traced back to 1837 when Professor Forbes, from The Royal Observatory Edinburgh, lowered temperature sensors into three shafts up to 24 feet (7.3 meters) deep. He did this in an effort to determine the effects of depth and time on temperature (Luthi, 2001).

Since then, those in the petro-physical field have been consumed with being able to determine sedimentary properties and fluid saturation, lithology and hence, the location of hydrocarbon bearing soils. Evidence of this can be clearly seen in the periodical publishing of books that detail the newest developments in the field of geological modelling

(Luthi, 2001). To achieve their purpose, these publications both define the state of the art in well logging field and also keep track of the technological advances through the ages. Of special interest are the advancements that have engendered more useful hydrocarbon measurement (Luthi, 2001).

The advancements in the well logging field have been, and continue to be, spurred on by many external factors. External factors including electronics and computing, drilling technology and new targets (Luthi, 2001). Electronics and computing have helped shape hydrocarbon exploration through the provision of new tools that are adaptable to lithological identification and the well logging field, as a whole. Examples of this can be seen in the high data transmission and acquisition chips, as well as the imaging and array sensors that are currently used in the field of well logging (Luthi, 2001). Advancements in the electronics and computing field have also allowed petro-physicists to make rapid and educated decisions on site because of real-time processing, quality control and advanced data visualisation (Luthi, 2001). Moreover, the development of satellites have allowed for the quick relay and display of information in near real-time, anywhere, at any time and on a range of devices. Advancements in drilling technology, specifically the development of LWD tools, have facilitated real-time data transmission, a reduction in the amount of fluid invasion during drilling as well as the prevention of borehole damage during the logging process (Luthi, 2001). Overall, the advancements in this sector have assisted in ensuring data integrity from start to finish. Lastly, as humans began to reach all corners of the globe and deplete existing resources, hydrocarbon explorative efforts had to shift to new targets. In particular, the possibilities presented by deep water targets have led to both technological advancements and new geological insights. The region of exploration challenged the logging community to develop more robust sensors as deep-water environments are geologically young, poorly consolidated, highly porous and thinly bedded. All of which can contribute to poor borehole conditions and the need for equipment that can navigate these environments (Luthi, 2001).

From the above it is clear that the general growth of humanity has led to significant progress in the logging field and made provision for a wide range of tools that can be utilised to address each logger's needs (Luthi, 2001).The relatively new tools developed from the technological advancements, as well as from the examination of new frontiers, have allowed those in the hydrocarbon and geoscience fields to carry out field explorations to an almost surgical degree.

The logs (recorded sedimentary characteristics) derived from these field explorations act as data inputs in the creation of subsurface maps. As such an understanding of these characteristics is important, if a 3D geological prototype is going to be created. However, numerous characteristics are measured during well logging. The main characteristics that will be used in the creation of the 3D subsurface maps are detailed in the following section.

2.3.1 Spontaneous Potential

The spontaneous potential (SP) curve is a measure of the potential difference between the potential of a kinetic electrode in a borehole and the potential of a static/fixed electrode at the borehole surface. During logging a borehole penetrates a permeable formation and puts two solutions of different chemical activities in contact (Peveraro, 2006). In congruence with the second law of thermodynamics, thermal agitation causes the net migration of ions from the saline rich formation water in the adjacent shale to the fresh drilling fluid in the borehole. Additionally, the negative electrical barrier created by the negative outer surface of clay mineral platelets in the shale prevents the diffusion of Cl- anions, but allows Na+ cations through. Thus, the borehole adjacent shale acts as a cation selective membrane and results in the borehole fluid becoming positive (Peveraro, 2006)(see figure 2.4).

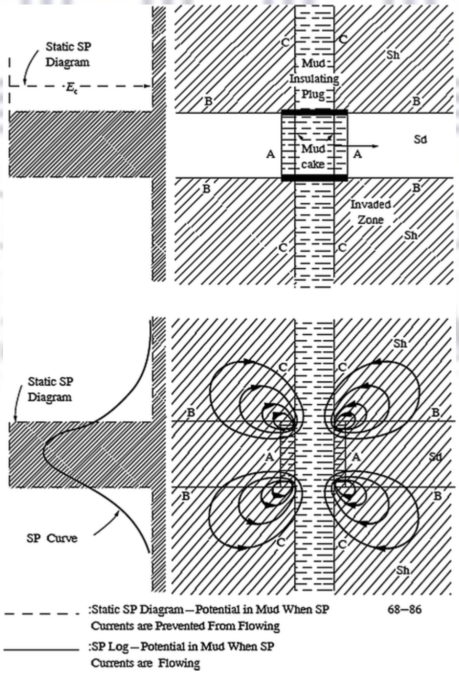

Static SP Diagram
E_c
Mud Insulating Plug
Mud cake
Invaded Zone
Sh
Sd
A
B
C
Static SP Diagram
SP Curve
Sh
Sd
A
B
C
:Static SP Diagram—Potential in Mud When SP Currents are Prevented From Flowing
68–86
:SP Log—Potential in Mud When SP Currents are Flowing

Figure 2.4: Schematic representation of potentials and current distribution in and around a permeable bed penetrated by a borehole (Peveraro, 2006)

It is important to note that SP values are not generated but are instead relative. Therefore, the shale baseline is not zero, but is the relative position from which SP deflections (and thus permeability) are measured.

During the SP measurement process no artificial currents are applied, instead the natural potential difference, in millivolts (mV), that exists between an electrode as it descends the depths of a borehole (moves downhole) and a fixed reference electrode at the surface of a borehole is recorded by a galvanometer. Each formation has its own SP, however the main objective of recording SP measurements is to allow differentiation between shale and non-shale formations. In addition to this the SP log assists formation, permeability and water resistivity determination (Peveraro, 2006).

SP is affected by a range of factors including the resistivity ratio, bed thickness, bed resistivity, borehole diameter, invasion and porous and permeable bed shaliness. Several factors influence the amplitude of an SP curve. These factors include bed thickness, bed resistivity, hole diameter, permeability and Rmf/Rw (Peveraro, 2006).

2.3.2 Gamma Ray

Most rocks have nuclei of atoms that are stable and naturally unreactive, such as clean sandstones and limestones. However, small portions of rocks are unstable and naturally reactive, and may emit their zero mass particles or photons at any time. Shales fall into this category and emit radiation from naturally occurring gamma ray sources such as the daughter elements of the Uranium-Radiam and Thorium series, as well as from radioactive potassium isotopes (40K) (Peveraro, 2006). These high energy pockets of energy (photons) emitted from excited nuclei are known as gamma rays and are the quantity measured in gamma ray (GR) logs. GR logs are captured by a scintillation detector, which records the radioactive emissions of rocks and thus assists in the lithological identification of shale and non-shale zones (Peveraro, 2006).

GR logs produced from the well logging process, measure the natural gamma radiation that originates from the radioactive elements of three main element groups, that is the thorium, uranium and potassium families. The amount of energy emitted by the radioactive elements of the aforementioned groups is usually in the spectrum of 0 – 3 million electron volts (Mev) and is generally recorded with a simple or spectral gamma ray tool. The simple, or natural, gamma ray tool takes GR readings without regard for the source of the radiation. Whereas the spectral gamma ray tool identifies the source of the radiation and (through spectral analysis determines the contribution of each element, thorium, uranium and potassium) to the overall energy spectrum (Peveraro, 2006).

Gamma radiation is a penetrating electromagnetic radiation that is progressively absorbed as it passes from one geological material to another. As such, the amount of energy emitted at a GR genesis gradually decreases as it passes from formation to formation. This effect is known as Compton scattering. Compton scattering is affected by the density of a formation with greater energy losses occurring in denser formations (Peveraro, 2006).

The GR logs derived from both simple and spectral GR tools during the well logging process, can be used to determine shale volume and lithology. However, only the spectral GR tool can determine radioactive material volume (Peveraro, 2006).

Factors that affect the radiation in rocks include age and deposition type (Peveraro, 2006). Age in particular, plays an important role in rock radioactivity; and is inversely related to the other i.e. increased gaining results in decreased radioactivity (Mennan, 2017)

2.3.3 Resistivity

Electrical resistance (R) is a substance's opposition to the flow of electrical current through it. It is this quantity that is measured in a resistivity log (which is measured in ohms). Thus, resistivity can be defined as a substances' resistance between two opposite unit cube faces at a specific temperature (Peveraro, 2006).

Resistivity logs signify the presence of fluids (like water) in rocks because rock matrices (excluding shale) are insulators, while saline fluids in their pore spaces are conductors. Resistivity is thus inversely proportional to the volume of water present in a formation. In other words, a formation with a high water content will have a low resistivity and vice versa (Peveraro, 2006).

Resistivity is useful in identifying hydrocarbons because, (in comparison to their exclusively water-bearing counterparts) the conductivity of porous rocks reduce in the presence of hydrocarbons. This fact enables the distinction between hydrocarbons and salt water in porous formations (Peveraro, 2006).

Resistivity is measured by three main methods: induction, laterlog and microresistivity logging. Microresistivity logging works by using closely oriented borehole wallmounted electrodes, while laterlog logging uses carefully constructed electrode arrangements to focus the surveying current and generate sharply focused horizontal current sheets of predetermined thicknesses. Induction logging, on the other hand, works by using high frequency alternating currents to induce concentric current loops (Peveraro, 2006).

The conductivity of a rock is a function of its porosity, the interrelation of its pores and the conductivity of the fluid in its pores (Peveraro, 2006).

2.3.4 Calipers

Boreholes are formed by rotating a circular rock-bit. Therefore, a circular borehole matching the diameter of the rock-bit is expected. However, this is often not the case. Instead the resulting hole may be circular, oval, gauged, under-gauged, over-gauged, cork-screwed or even key-holed. Gauged holes (circular and rock-bit sized) indicate the presence of hard, dense and non-shaly rocks. Under-gauged holes (rock-bit sized minus two times the thickness of mud cake infiltrates) indicate the presence of permeable, porous formations such as clays and sloughing shales. Sloughing shales can also result in over-gauged holes that are over-sized with a diameter much greater than the bit size (Peveraro, 2006).

There are several mechanical calipers that are used to determine borehole geometry. The tools fall into 6 main categories 1-arm, 2-arm, 3-arm, 4-arm, 6-arm and multi-finger tools (Peveraro, 2006). Some of these calipering devices are designed to simply measure borehole diameter while others also form an integral part in achieving the aims of the overall survey, and are therefore embedded in other tools. For example in the 3-arm caliper supports borehole diameter determination and is used as a centraliser in sonic, dipmeter and production logging tools (Peveraro, 2006).

2.3.5 Porosity

There are three logging tools that can assist with determining the porosity and thus formation mineralogy. These are density, neutron and sonic logging tools (Peveraro, 2006).

The density log measures the grams per cubic centimetre of scaled bulk density of formations. It does this by emitting a highly collimated beam of medium energy gamma rays (Peveraro, 2006). These rays collide with electrons in the formation and lose some of their energy as a result of the interaction, but continue to travel through the formation along an altered path (Compton scattering). As electron and mass density are almost identical, the inverse proportionality between the number of back-scattered gamma rays and the electron density of the formation can be used to determine the formation's mass density (Peveraro, 2006).

To achieve neutron log measurements, a chemical neutron (e.g. an Americium-Beryllium mxture) is used to barrage the formation with fast neutrons. These neutrons then collide with nuclei in the formation and slow to epithermal and then thermal neutrons. At this energy level the neutron is captured by a nucleus in the formation. To stabilise itself after the addition of the neutron the nucleus emits high energy gamma rays (Peveraro, 2006).

During this collision event the rate at which neutrons lose their energy depends on the mass similarity between the neutron and the struck nucleus. If the nucleus is of greater mass, no energy will be lost, and the neutron will bounce off elastically. However, if the nucleus and the neutron have approximately the same mass, energy will be shared and the neutron will slow (Peveraro, 2006). Hydrogen nuclei and neutrons are of almost the same mass. Therefore in a head on collision the neutron could transfer all its energy to the hydrogen nucleus. Thus, a neutron log is essentially a hydrogen log as the rate at which a neutron loses its energy by collision is directly related to the amount of hydrogen per unit volume (Hydrogen Index) present in the formation (Peveraro, 2006).

The Hydrogen Index of porous water-filled formations and shales is higher than the Hydrogen Index of formations with gas and light oil. So in addition to assisting with porosity determination, the neutron log can be used to identify shales as well as gas and light oil zones(Peveraro, 2006).

Factors that affect neutron log readings include water, clay, oil and gas i.e. essentially anything with hydrogen (Peveraro, 2006).

Sonic logs use sonic/acoustic velocity tools to determine the speed of sound in the rocks beside the borehole. Using an electrical signal, these sonic tools emit a sharp sound from an acoustic transmitter. The emitted sound moves in a spherical wave to and through the borehole wall and is refracted back before it abates. The sound refracted back to the detector is converted into an electrical signal. The amplitude of this signal is indicative of the formation's ability to carry acoustic energy and, as backed by Young's modulus, rock rigidity (Peveraro, 2006).

At the time of emission from the acoustic transmitter, the wave form is compressional.

However, on contact with the borehole wall its splits into three parts: compressional (fastest), shear and boundary wave forms (slowest). These waves retain their spacing as they travel back to the detector and the first wave arrival times are used for determining formation transit time. These transit times (along with velocity measurements) can be used to determine common rock types based on rock acoustic response observations (Peveraro, 2006)

2.4 Machine Learning

Pattern recognition is so deep-rooted in the human experience that it has become an almost subconscious activity, carried out with an ease that belies its complexity (Duda et al., 2012). Every task that makes use of this subconscious activity, often includes one or more of our senses – whether sight, smell, taste, sound or touch; and includes the ability to recognise a face, distinguish between fresh and rotten food and understand spoken words (Duda et al., 2012). According to Duda et al. (2012), pattern recognition can be defined as grouping data into patterns and making decisions based on the various pattern categories that arise.

As humans have evolved, the ability to recognise and classify patterns has passed on from generation to generation as a skill necessary for survival. With that said, and in this technological age, it is only natural for humans to seek to design machines that can carry out and improve on the application of this function. This is often referred to as machine perception and examples of this can already be seen in facial, fingerprint and automated speech recognition software (Duda et al., 2012) – with most of these design achievements stemming from the observation of how nature solves these issues.

The objectives of this investigation, log data clustering, is really no different. Consequently, the modelling techniques adopted, in order to develop a 3D subsurface prototype and achieve the objectives of the investigation, were built primarily on the machine perception principles and then more broadly on the principles of pattern classification.

In the fields of machine perception and pattern classification, machine learning is fundamental aspect. Machine learning lies at the intersection of statistics, artificial intelligence (AI) and computer science (see Figure 2.5) and involves transforming data into knowledge (Müller et al., 2016).

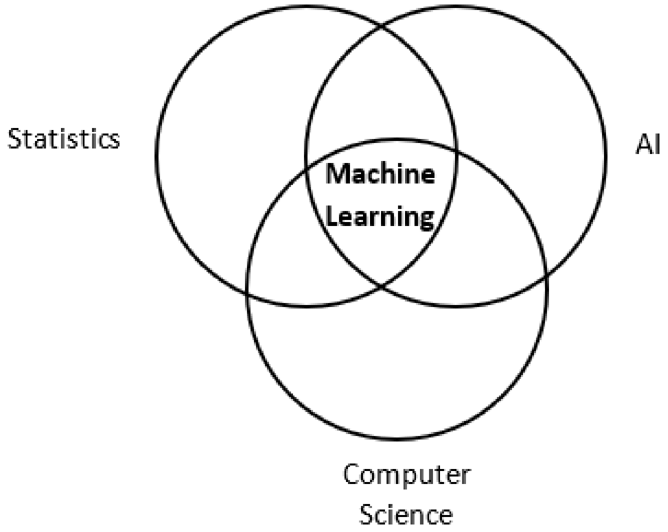


Figure 2.5: Intersection of the machine learning fields

It is this ability to extract knowledge from data that makes machine learning incredibly influential in data driven research, including machine perception and pattern classification.

In the early days of machine learning these extractions were carried out by explicitly defining conditional statements. These conditional statements spelled out decision rules that were executed depending on whether a condition was true or false. A great example of this can be seen in figure 2.6 below:

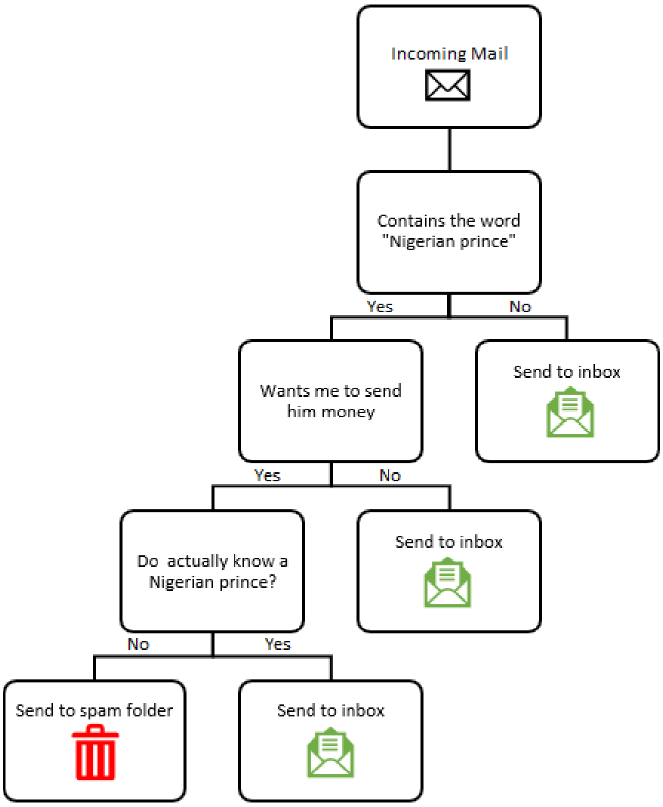

Incoming Mail
Contains the word "Nigerian prince"
Yes
No
Wants me to send him money
Send to inbox
Yes
No
Do actually know a Nigerian prince?
Send to inbox
No
Yes
Send to spam folder
Send to inbox

Figure 2.6: Decision steps in a simple spam filter (Adapted from Beyeler (2017))

Although these rules help process data and ultimately assist in the decision-making process, as stated by Beyeler (2017), they are limited by two major factors:

1. Their extreme reliance on an expert's domain knowledge, including all possible exceptions, to form decision rules.
2. Their confinement to a specific task to the point where the slightest change in the task often requires a rewrite of the entire rule system.

With the latest machine learning iterations these factors have been overcome. Therefore, when presented with a large and varied dataset machines are able to find data patterns - both hidden and apparent - without the task first having to be well defined (Beyeler, 2017).

2.4.1 Machine Learning Approaches

Most machine learning problems fall into one of three categories, that is supervised learning, unsupervised learning and reinforcement learning (Beyeler, 2017).

In supervised learning, decision making is automated by generalising from known examples (see figre 2.7). During this process the user provides input and desired output sets by labelling each data point in the dataset with a category. Using the input/output pairs as a 'teacher' the algorithm learns how to derive the output category from the input data points. Then using this knowledge base, the algorithm can then categorise an uncategorised new data point into a specific category (Beyeler, 2017).

Going back to the spam email example, to filter out spam emails a supervised learning algorithm would be provided with a large set of emails (the input) and the category of each email i.e. whether the email is a spam email or not (the output) (Müller et al., 2016). Having learnt what constitutes a spam email, the supervised learning algorithm would then be able to predict whether any future emails are spam emails (Beyeler, 2017).

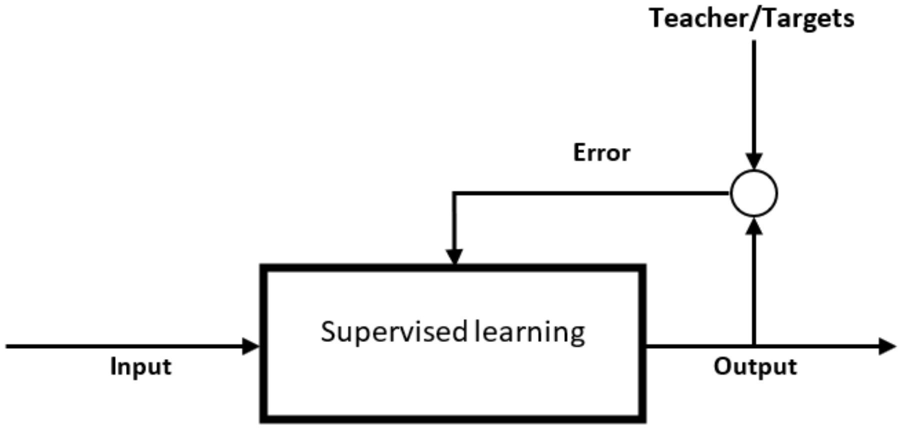


Figure 2.7: Main machine learning categories: Supervised machine learning (Reproduced from Beyeler (2017))

In unsupervised learning, the data is uncategorised and only the input data is known, so decision making is automated without a known output vector (see figure 2.8). As there is no 'teacher' to derive some knowledge from in the input dataset, the unsupervised learning algorithm organises the data into natural groups or clusters such that the points within a cluster are of great similarity while also being as disparate to other clusters as possible (Duda et al., 2012).

Often the data is simplified through a range of functions before it is clustered. These functions include dimensionality reduction, so that it can be better described and then later organised by the algorithm (Beyeler, 2017). From the simplified data the user has to hypothesise the number of clusters in the dataset before the unsupervised algorithm can assign each data point to a cluster (Duda et al., 2012).

In terms of the spam email example, knowing that there are two clusters (i.e. spam and not spam) an unsupervised learning algorithm would identify email clusters by first looking for similarities and disparities in the input data (the emails), and then grouping the data points into a cluster based on this information (the output). Although now categorised based on a cluster label, it is up to the user to interpret what each cluster means i.e. whether cluster 1 indicates spam emails and clsuster 2 non-spam emails or vice versa (Müller et al., 2016). Having learnt what properties make up a cluster, the unsupervised learning algorithm can then predict which cluster any future emails belong to.

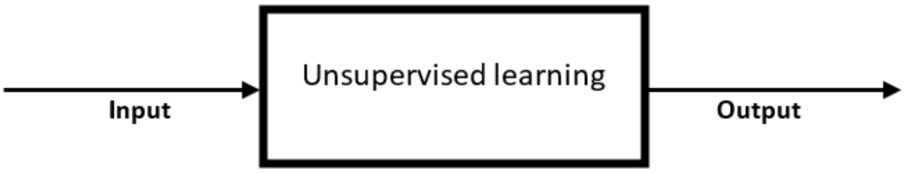


Figure 2.8: Main machine learning categories: Unsupervised machine learning (Reproduced from Beyeler (2017))

In reinforcement learning, decision making is automated by using known output vectors to strengthen the initial classification of input data (see figure 2.9) (Duda et al., 2012). During this process the data points are fed to the algorithm and then the algorithm comes to a conclusion based on this information i.e. classifies the input data. Following its conclusion, the algorithm is supplied with the feedback as to the accuracy of the classification. Using this binary right/wrong critique the reinforcement algorithm can then either maintain or modify its strategy in computing the correct category (Beyeler, 2017).

Going back to the spam email example, to filter out spam emails a reinforcement learning algorithm would be provided an email (the input). It would then draw up a tentative classification label. The algorithm will then be told whether the classification is correct or incorrect and if correct the algorithm would maintain its strategy for identifying spam emails. However, if incorrect it would alter its strategy based on the feedback (Duda et al., 2012). Having learnt what constitutes a spam email, the reinforcement learning algorithm can then predict whether any future emails are spam emails

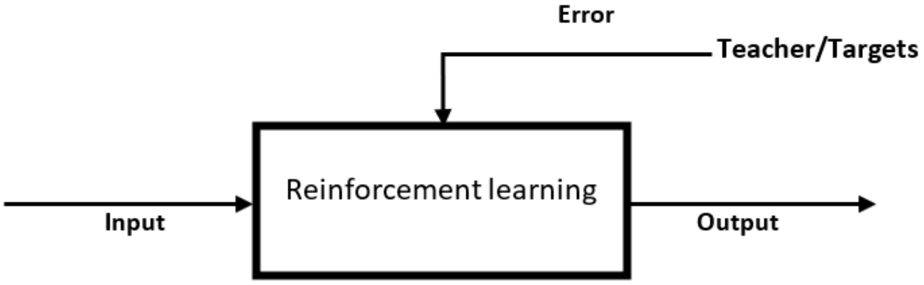


Figure 2.9: Main machine learning categories: Reinforcement machine learning (Reproduced from Beyeler (2017))

For this investigation the second family of machine learning algorithms, unsupervised machine learning, was employed. This was because the data from the study area had no ‘teacher’ (known output) to inform the learning and had instead to rely entirely on the input data to extract knowledge. Also, following the *sklearn* flow chart (see figure 2.10), as the dataset was a large sample of unlabelled data that needed to be categorised, a clustering (unsupervised learning) algorithm would have to be used.

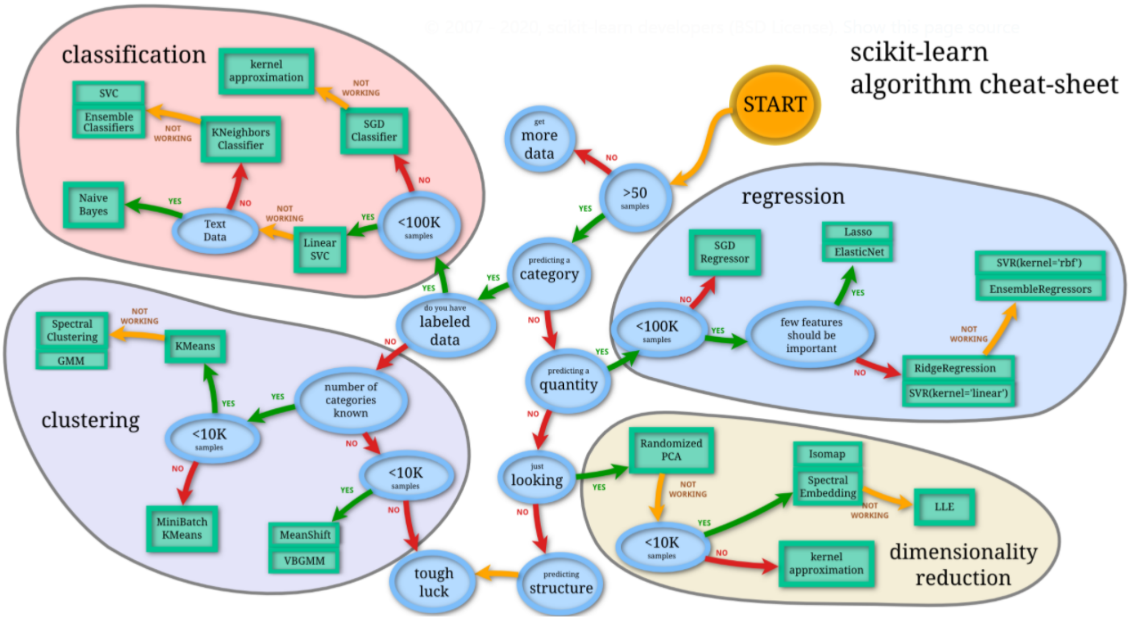


Figure 2.10: skleran flowchart on how to choose the right estimator (Developers, 2007)

2.4.2 Unsupervised Learning

Unsupervised learning comes in a multitude of forms and can be applied in numerous ways; however, the intent of its use is always to transform an input data source into a richer, more meaningful representation (Beyeler, 2017). The most common applications of machine learning are in unsupervised transformations and clustering.

Unsupervised Transformations

Unsupervised transformations use the input dataset to create new data representations that better support human or machine understanding. One such transformation is dimensionality reduction. In this transformation process, multi-feature high dimensionality data is compressed and represented as only informative essential features. Some of the most widely used dimensionality reduction algorithms include principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) (Müller et al., 2016).

In PCA the dataset is represented in a lower dimensional space by orthogonally rotating all the data points until they are aligned with the two axes that explain the most variance (Beyeler, 2017). An example of the PCA process can be seen in figure 2.11. In plot 1 of figure 2.11, component 1 is the vector that contains most of the data and which explains the direction of greatest correlation. Component 2 is a vector orthogonal to component 1 and which explains the direction of the next greatest correlation. The directions obtained from this process (components 1 and 2) are the principal components of the data and they describe the directions of greatest variance. Plot 2 in figure 2.11 shows the mean standardised data rotated to align with the axes of the first and second principal components. To reduce dimensionality of the data only certain principal components can be retained, as seen in plot 3 of figure 2.11 where only the first principal component is retained. Thus, the data is reduced from a two to one dimension dataset. Removing the rotation and applying the mean back to the data, plot 4 in figure 2.11 displays the information that was retained from the PCA process (Müller et al., 2016)..

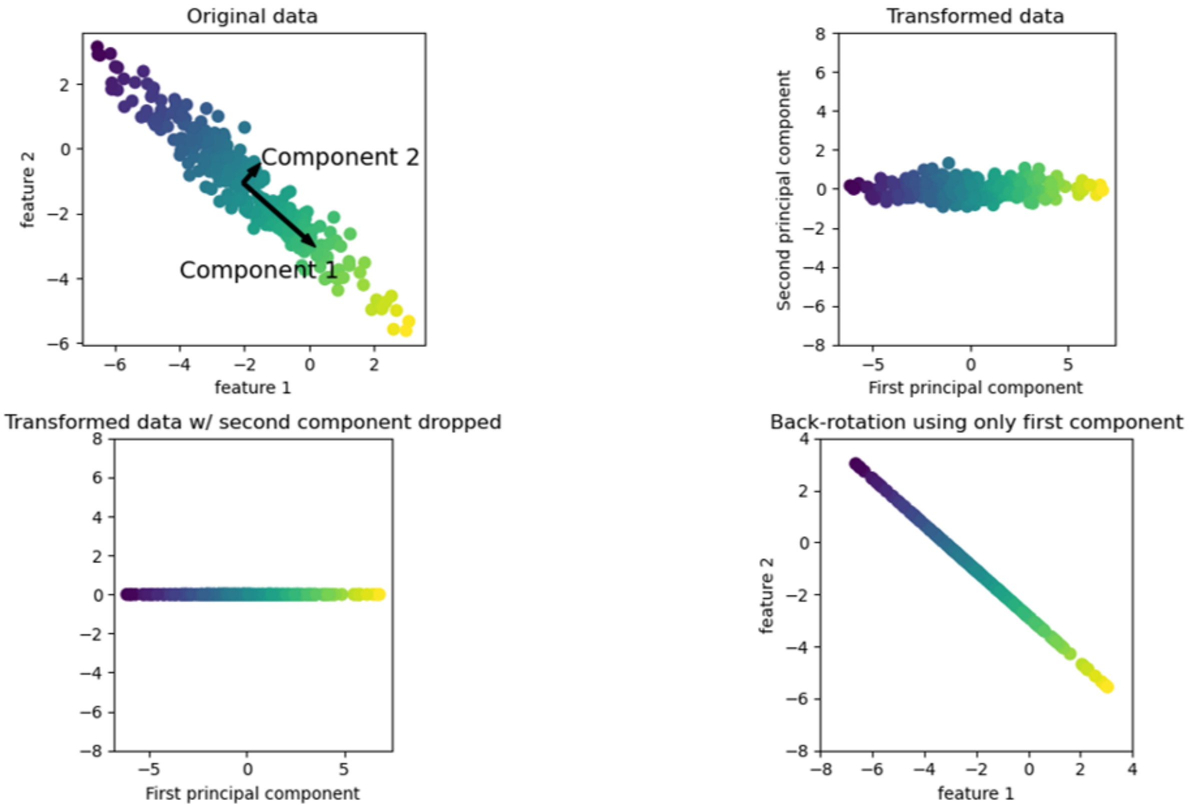


Figure 2.11: Transformation of data with PCA (Müller et al., 2016)

The next unsupervised transformation algorithm, t-SNE, starts by randomly representing the data points in two-dimensional space. Following that, the algorithm attempts to increase both the proximity of neighbouring points and the remoteness of distant points in the original feature space. Figure 2.12 below shows an example of the application of t-SNE to PCA transformed data. In the image barring a few exceptions, there is a clear separation between classes (Müller et al., 2016).

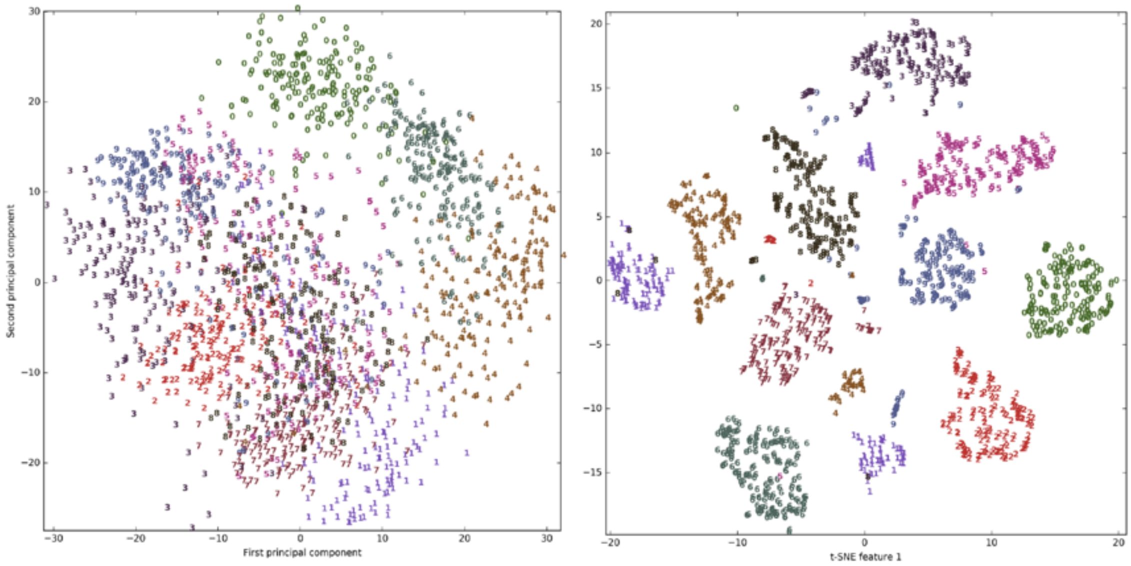


Figure 2.12: Scatter plot of the digits dataset using the first two principal components (left-hand side). Scatter plot of the digits dataset using two components found by t-SNE (right-hand side) (Müller et al., 2016)

Clustering

Clustering algorithms partition data into different classes of like objects (clusters). Similarly to t-SNE, clustering algorithms split the dataset into groups that have both great internal similarity and great external dissimilarity. There are many clustering algorithms that can achieve data partitioning, however k-means clustering is the simplest, most commonly used and best suited to the data of this investigation (based on the *sklearn* workflow, figure 2.10). K-means clustering works by finding the cluster centres of 'k' number of groups that represent the different sections of the data (Albon, 2018). According to Albon (2018), it does this by

1. Creating 'k' randomly placed cluster centers
2. Calculating the distance between each point and the cluster centers
3. Assigning each point to the group of the nearest cluster center
4. Resetting the location of cluster centers to the mean of the redetermined clusters
5. Repeating steps 2-4 until there are no more cluster membership changes

A visual representation of this process is depicted in figure 2.13 below.

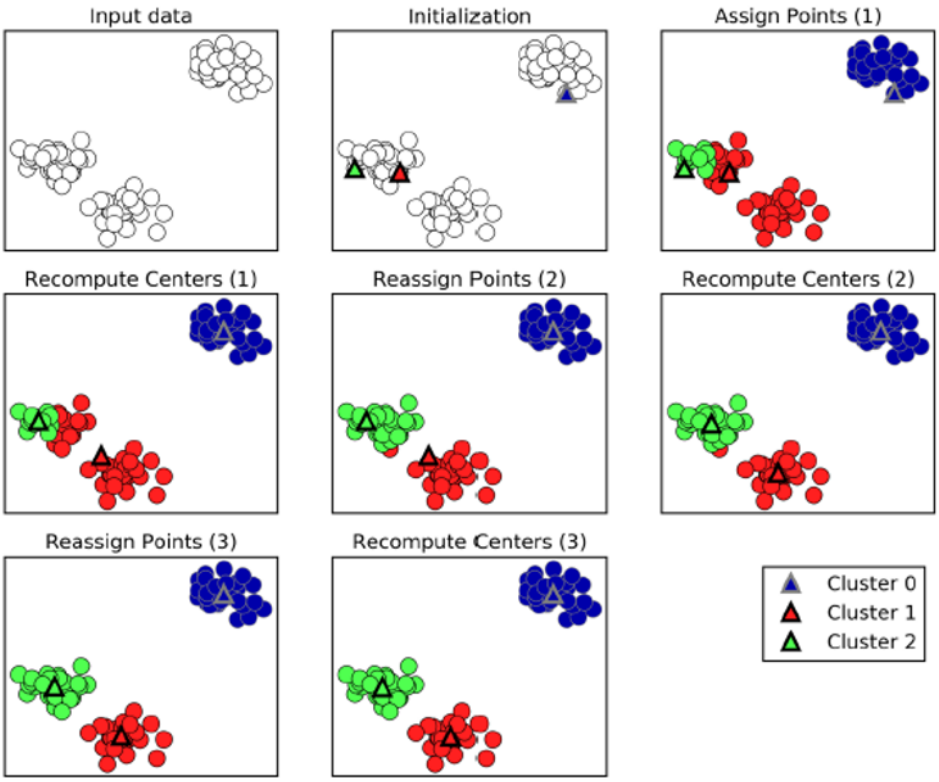


Figure 2.13: Input data and the three steps of the k-means algorithm (Müller et al., 2016)

The most important step in the k-means workflow occurs before the first step and is the definition of the number of clusters. Having to define the number of clusters beforehand can be problematic if the phenomena being modelled is complex and not fully understood. To overcome this, the elbow method and silhouette analysis can be implemented. The elbow method repeats the clustering for a range of cluster 'k' values and documents the compactness value against this 'k' value. The plotted compactness by 'k' graph resembles an arm and the 'elbow' points to smallest number of clusters that gives a very compact representation (see figure 2.14). This cluster number is what should be specified in the k-means algorithm (Beyeler, 2017). While the elbow method considers compactness, silhouette analysis takes into account the separation between clusters. By highlighting whether most of the points in a given cluster are closer to a neighbouring cluster than their own, silhouette analysis assists in cluster number selection.

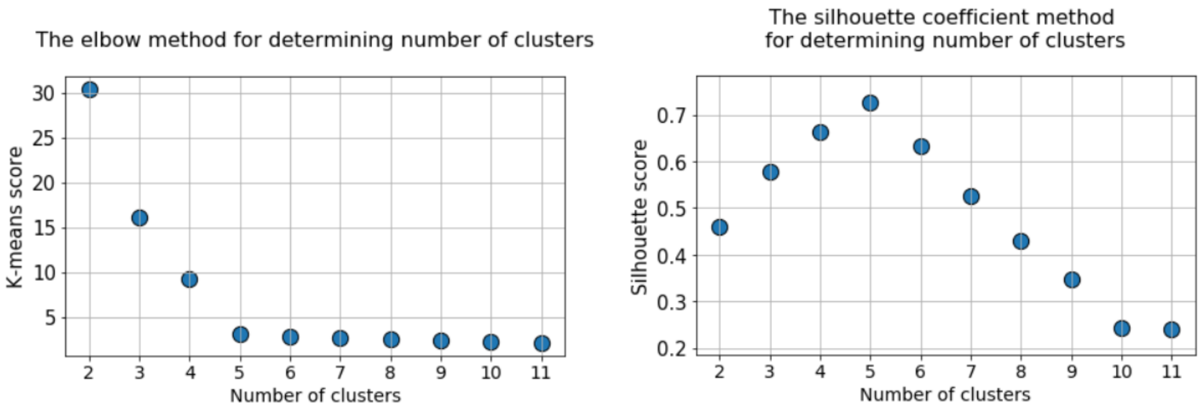


Figure 2.14: The elbow and silhouette method for determining 'k' (the number of clusters) (Adapted from Sarkar (2020))

2.5 Cartographic Design Cycle

According to Crampton and Krygier (2005) cartography the art and practice of mapping out spatial data can be traced back to the genesis of most human civilisations. This is because of humanities enduring need to visually record geographically located phenomena (Bailey and Gatrell, 1995). This recording is carried out by creating graphical representations, where image objects symbolise phenomena occurring in the real world (Rhind and Taylor, 2013). These representations, however, were often fraught with distortions due to their 2D confinement and ultimately led to maps that were unable to fully capture both the spatial and non-spatial relationships that existed within the 3D geographic environment being depicted (Monmonier, 2018).

These 2D map distortions played themselves out in all the basic map components: scale, projection and symbolisation. In addition to the inability of 2D visualisation techniques to minimise distortions, its failure to realise multi-dimensional representation, increase user coverage and step away from its paper dependence ultimately led to the development of the 3D cartographic technique. The 3D cartographic technique was developed to address the abovementioned 2D cartographic pitfalls and support the creation of multi-scale dynamic models that support user understanding and appeal (Jones et al., 2009). All these improvements thus allowed cartographic users at all experience levels to depart from a painstaking, fragile and limited visualisation technique and move towards an easily accessible and enduring method of representation (Rhind and Taylor, 2013).

Despite the move from 2D to 3D, the cartographic design cycle, a process which links a map, its maker, its user and the environment being represented, is still applicable (Stevens et al., 2012). The cartographic design cycle, as seen in figure 2.15 is a recursive process in which the outcomes of a given stage inform subsequent stages.

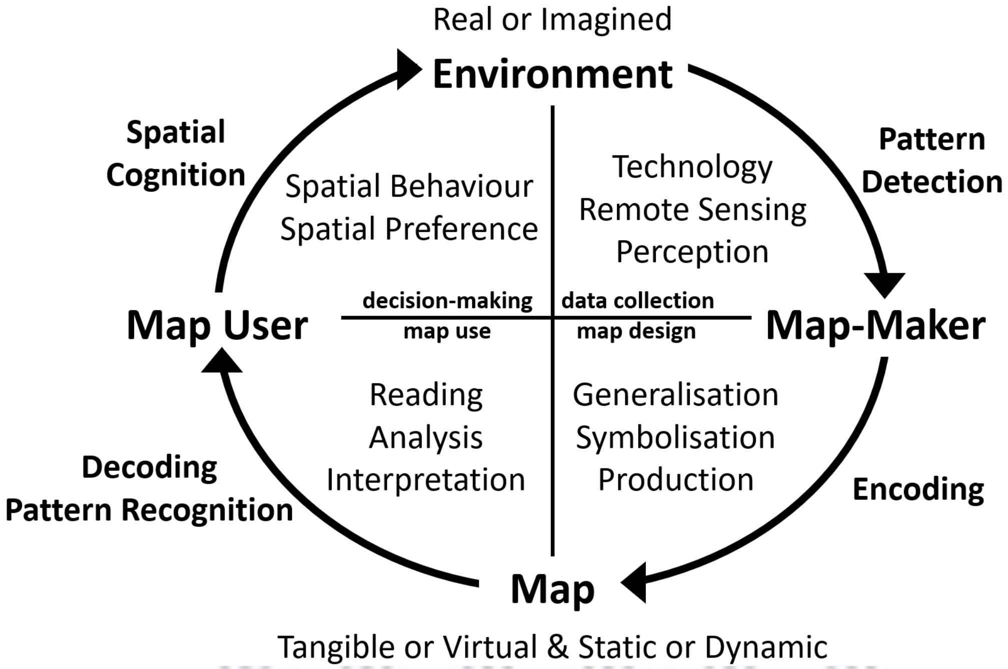


Figure 2.15: The cartographic design cycle (Reproduced from Stevens et al. (2012))

The cartographic design cycle starts with the environment being mapped. After data acquisition through both on-site and remote methods, the map-maker's perception of the physical environment (and its relationships) determines the way the data is prepared for map creation. Therefore, the patterns that exist in the raw data as well as the purpose and use of the map are all used to inform the created cartographic representation (Stevens et al., 2012).

Next, using a range of map production techniques the cartographer (map maker) attempts to visually represent the prepared data in the form of a map. This is known as encoding and the techniques used by the cartographer during this process include symbolisation and generalisation (Stevens et al., 2012).

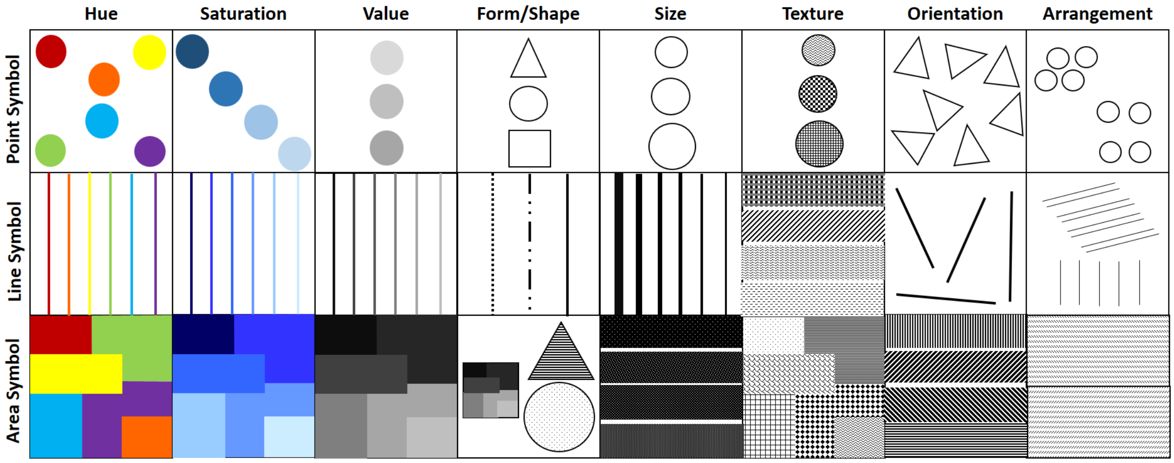


Figure 2.16: The visual variables (graphic elements) of cartographic symbols (Adapted from Tyner (2010)). First reproduced in Ile (2018).

Before deciding on the graphic elements of a map, their relationship to psychological factors have to be considered (Kraak, 1993). Some of the relationships between map graphic elements and psychological factors are detailed in the table below:

Table 2.4: The relationship between the primary graphic elements and the psychological depth cues (Kraak, 1993).

Primary Graphic Elements	Psychological Depth Cues
value	—
colour	colour
size	rectinal image size
texture	texture
orientation shape	linear perspective
—	aerial perspective
—	detail perspective
—	shades
—	obstruction/overlapping

With the necessary pyscological cues considered, symbolisation (the association of graphic elements with real world geographic objects) can be carried out. This activity improves the look and comprehension of maps (Ile, 2018). According to Haeberling (2005) there are eight essential graphic elements size, shape, spacing, orientation, arrangement, colour and brightness, texture and pattern and special graphical effects. Shown in table 2.5 below are the visual graphic elements that make up the symbolisation process.

Table 2.5: The symbolisation visual variables necessary for cartographic representation of real world graphic objects

Visual variable	Variable properties
Size	The manipulation of an object's physical proportions can either emphasise or de-emphasise some quality about it. This characteristic lends size to the effortless display of volume or amount (Tyner, 2010)
Shape	In maps, shape is used as a means to denote difference in kind (Tyner, 2010).
Colour and brightness	This variable consists of three parts - hue, value and saturation - and each plays a different role in cartographic representations. Hue is used to differentiate between objects of similar form (size and shape). Value and saturation are often used together, with value used to represent amount/quantity while saturation is used to distinguish between subcategories within a group (Tyner, 2010).
Texture	Texture is used to conjure an impression about an object and is created by amalgamating smaller elements and arranging them in a particular pattern (Tyner, 2010).
Arrangement	This variable refers to the layout of objects in a cartographic representation (Ile, 2018).
Orientation	Orientation sets the direction of objects and can thus be used as an indicator of similarity or difference (Tyner, 2010).

While symbolisation uses graphic elements to increase aesthetic appeal and information conveyance, generalisation on the other hand improves image discernment and imaging speeds through data and detail reduction.

In addition to symbolisation and generalisation, other map production techniques can be employed during the cartographic design cycle. Such production techniques include the application of lighting and environmental effects to display the relationship between geographic features, complete the representation and ensure effective visualisation of the graphic scene (Haeberling, 2005). At the end of this stage of the cartographic design cycle a coherent fit for purpose map is produced.

Due to the intelligible map designed by the cartographer, during the third stage of the cartographic design cycle, the map user can decode the symbols of the map and decipher the patterns within it. The decoded map is thus legible and available for analysis and interpretation by the user.

Lastly, information gathered during the map use informs any decisions made and actions taken. Therefore, the way maps are framed influence our spatial understanding, behavior and preferences ultimately shaping how we perceive the environment (Stevens et al., 2012).

2.5.1 Colour

The power of colour in the development of a meaningful, fit for purpose cartographic representation is often obscured by its decorative role. Therefore, study of this graphic element is required to ensure that the representation makes plain the phenomenon being mapped instead of obscuring it with flourishes.

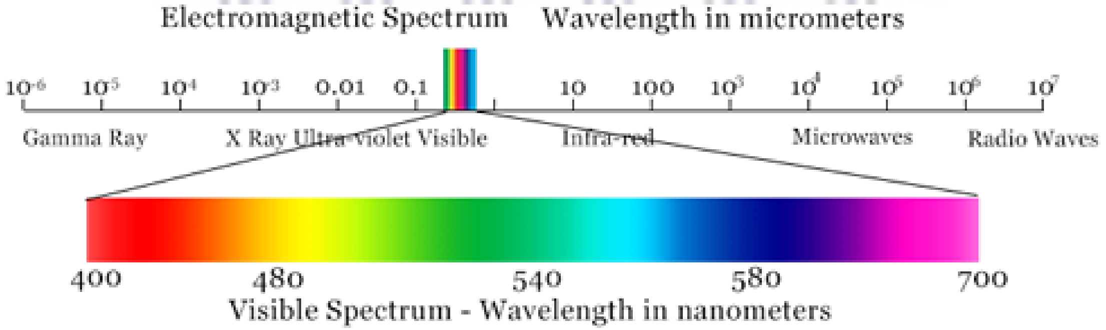


Figure 2.17: Image depicting the visible band within the electromagnetic radiation spectrum, with the wavelength for each colour band included (Monmonier, 2018)

Colour is a sensory phenomenon experienced in response to light from a narrow band of the visible electromagnetic spectrum (Hunt and Pointer, 2011). The band of visible light is between 0.4 μm and 0.7 μm (see figure 2.17), and although narrow it has been estimated that over 10 million different colours can be distinguished from this band (Judd and G, 1975). According to Hunt and Pointer (2011), this ability is only possible because of the 3 basic perceptual attributes of colour. These are brightness, colourfulness and hue and they can respectively cause and area to appear bright/dim, more/less saturated and similar to one or more portions of red, yellow, green and blue (see figure 2.18).

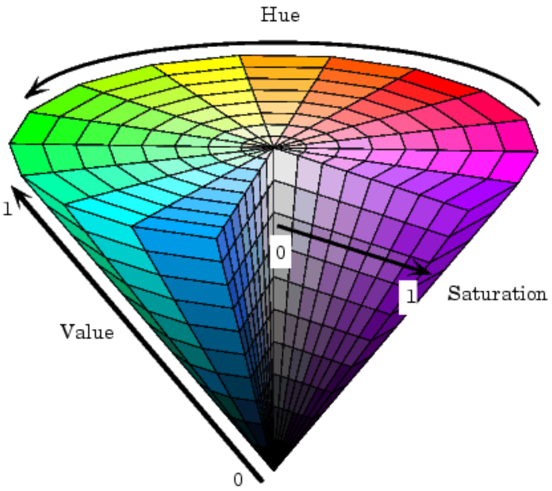


Figure 2.18: Image depicting the HSV colour space in three dimensions (hue, saturation and value). The relationship and the means of interaction of these quantities can also be seen (Monmonier, 2018)

In the figure above, hue is depicted as a colour wheel with orthogonally extending saturations centred on a value/brightness axis which ranges from black (at the bottom) to white (at the top) (Hunt and Pointer, 2011). Black and white light can be described as the absence and presence of all wavelengths from the visible band of the electromagnetic spectrum, respectively (Monmonier, 2018).

From the statements above it is clear that colour is a multifaceted tool that is able to reinforce meaning and order while also supporting the visual interest of a representation. To best ensure effective use of this colour, Lidwell et al. (2010) suggests the adherence to a few guidelines:

1. Number of colours:
    - Colour should be used conservatively and with focused intent, especially as a significant percentage of the population has a limited perception of it (Lidwell et al., 2010).
2. Colour combinations:

- Use cooler colours to mark background objects and warmer colours to distinguish foreground objects (Lidwell et al., 2010).
- Achieve aesthetic cohesion and appeal by using colours that are either combinations found in nature or that are analogous (adjacent)/complementary (opposite) colours on the colour wheel (Lidwell et al., 2010).

3. Saturation:
   - When considering saturation in a representation it is important to remember that dark colours are perceived as serious and professional, while bright colours are are seen as more friendly representations (Lidwell et al., 2010). Saturated colours, which are viewed as exciting and dynamic, are best used to indicate objects of high priority. Whereas, desaturated colours find their place in the creation of efficient and fast renditions. Above all, the use of saturated colours should be carefully considered before implementation because excessive combinations can lead to eye fatigue (Lidwell et al., 2010).
4. Symbolism:
   - The emotional and symbolic meaning of colour has to be tailored to the audience that will view the representation (Lidwell et al., 2010).

2.5.2 Eye Brain System

The output of the cartographic design cycle is a map, a visual representation of the earth. So to understand the message being conveyed by the representation a means of visual processing is required. Within humans this is satisfied by the visual system, a pathway which spans from the retina to the cortex and starts with the eyes (Hubel and Wiesel, 1979).

Without sight not only would people be unable to observe colour they would also be unable to visually process cartographic creations. Hence, it is appropriate to consider the human eye-brain system as part of this investigation.

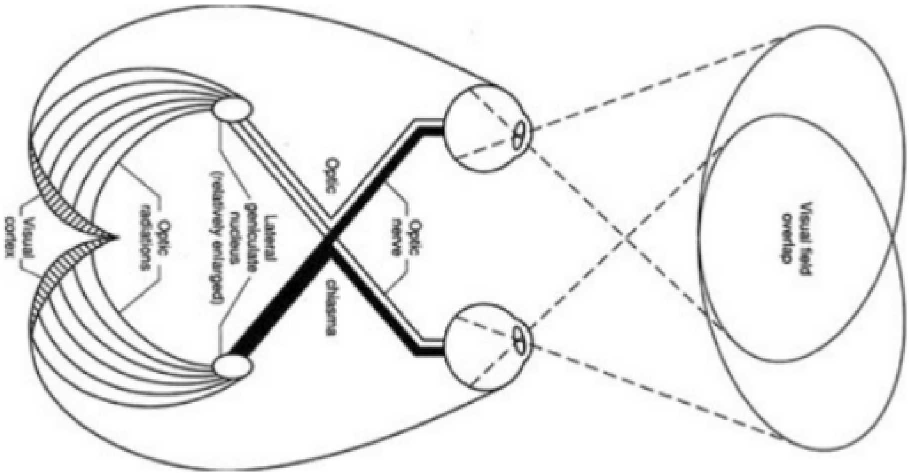


Figure 2.19: Keates (2014) schematic for the human visual system.

The iris, a structure that provides an adjustable aperture, controls the amount of light that enters the eye through the cornea (a curved transparent window within the eye). The cornea in conjunction with the cillary-adjusted lens focus light on to the retina (light sensitive cells at the back of the eye). Light then passes to the rod and cone photoreceptors (named thusly because of their shape) which are found within the retina (Snowden et al., 2012). Cones and rods, which vary in type and distribution, are sensitive to wavelengths and light respectively. Their varied type and distribution result in a selectively processed image of reality (Keates, 2014).

From the rod and cone photoreceptors, light is then passed to the retinal ganglion M (responsible for movement) and P (responsible for colour information) cells layer, before leaving the eye at the blind spot through optic nerves. Optic nerves are essentially millions of bundled blood vessels and retinal ganglion cell axons (Snowden et al., 2012).

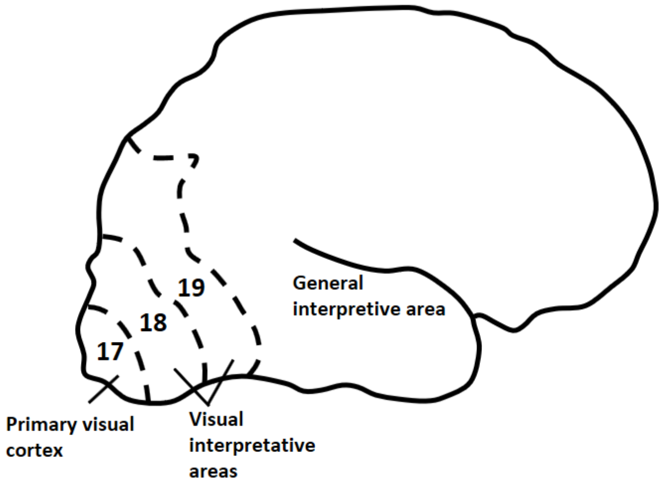


Figure 2.20: Adaptation of Keates (2014) diagram depicting the areas of the brain primarily devoted to visual perception. First reproduced in Ile (2018).

Here begins the journey to the brain, with the projection of information from the optic nerve to the lateral geniculate nuclei (LGN), a relay center for visual information. There are 6 layers in the LGN, 3 for the right eye and 3 for the left. In these layers retinotopic mapping, the orderly mapping of the visual world, is observed for the creation of a clear image. In addition to mapping the visual scene the LGN highlights information of importance by filtering out the contents of the visual field (Snowden et al., 2012). Then after traversing several synapses cells, the LGN pass their axons directly to the visual cortex. Due to the crossing of the optical nerve at the optic chiasm (the point of optic nerve conveyance), the left LGN and cortex are concerned with the visual scene from the right eye. The opposite is true for the right LGN and cortex. In a hierarchical manner simpler cells feed information from the retina to more complex cells for transformation in orientation and for the combination of retinal inputs (Hubel and Wiesel, 1979).

These transformation processes assist in the perception and comprehension of a visual scene by breaking down the graphical objects into their simplest components before regrouping it. For example in the case of a square, it would would first be split into a series of vertical and horizontal lines before being regrouped as a square. As such, the increased complexity and variability of an object necessitates more neural connections and results in increased image processing speeds as well as increased difficulty in perceiving an object (Keates, 2014).

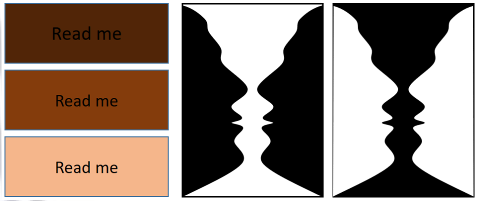
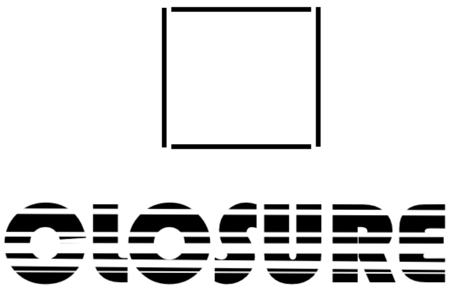
Another important understanding of visual comprehension can be gained by looking at eye movements. Unlike in the general evaluation of cartographic scenes, comprehensive map analysis (i.e. object detection, discrimination and identification) occurs by means saccadic movements instead of through the use of central (foveal) vision (Keates, 2014). During these saccadic movements the eyes go through a cycle of fixating on an object, jumping away and then refocusing on it again. As the fixation length is directly related to the complexity of an object - in terms of its size, colour, shape, texture and orientation - the definition of a simple and familiar scene is critical for promoting visual comprehension (Keates, 2014).

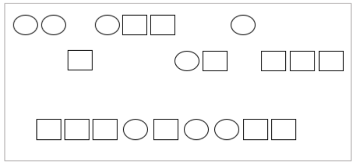
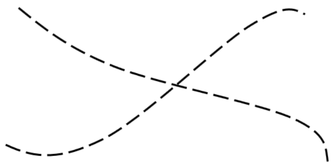
2.5.3 Gestalt

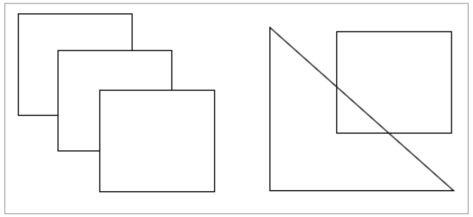
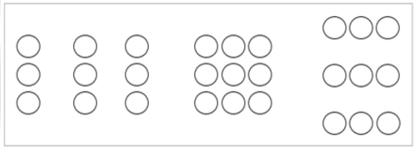
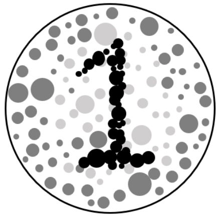
It is important for a cartographer to consider these individual cartographic elements. However, consideration of the effect of all elements on the map composition is as important. This is especially true when one considers the fact that the human eye-brain system is often incapable of observing the map elements without also observing their setting (Kent and Vujakovic, 2017). Human cognisance of the visual is, thus, dependent on relating the foreground with background. In this way the viewer's perception of the scene is not of fragmented elements, but rather of coherent, well-defined objects which are distinguishable from each other and from their environment (Roth and Bruce, 1995).

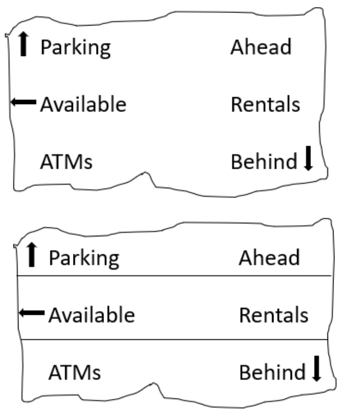
The relationships between objects and their background can be preserved through the observation of gestalt laws, which (in addition to fore- and background distinction) also aid the viewer's ability to read, analyse and interpret maps. The gestalt laws which describe how visual elements and the patterns they make are perceived (Graham, 2008), include the laws of figure/ground, closure, common fate, good continuation, law of pragnanz, proximity, similarity and uniform connectedness.

Table 2.6: Summarised descriptions of the Gestalt laws

Gestalt laws	
Description	Image
Figure/ground	
It is the separation of the stimuli within a composition into the objects of focus (figure elements) and the undifferentiated background (ground elements). Recognition and application of this law helps to ensures that the figure element is clearly seen, focused on and remembered in comparison to the ground elements. This is achieved through the creation of an unambiguous figure/ground composition. For example in figure ..., an unstable figure and ground distinction leads to object perception switching between a vase and two faces looking at each other (Lidwell et al., 2010).	 Figure 2.21: Depiction of the Gestalt law of Figure/ground as seen in both a text and image example (Adapted from Roth and Bruce (1995), and first reproduced in Ile (2018)).
Closure	
It is the perception of a group of individual elements as belonging to a single recognisable pattern object rather than existing as multiple separate entities. This is especially true when the elements are near each other and geometric in nature. Adherence to this law supports reduction in complexity and encourages the viewer participation by using smaller/fewer objects which the viewer has to complete to understand. This principle is enacted in both an almost automatic and subconscious manner. This is only possible because of humanity's preference for simplicity over complexity and order over randomness. In image ... the elements are first perceived holistically as a single object and then as individual elements (Lidwell et al., 2010).	 Figure 2.22: Depiction of the Gestalt law of Closure as seen in both an image and text example (Adapted from Graham (2008), and first reproduced in Ile (2018)).

Description	Image
Common Fate	
It refers to the perception of a group of individual elements that move together in the same direction as belonging to a related group as opposed to elements that are stationary or which move in a different direction (Lidwell et al., 2010). The advantage of this law is that it assists in the distinction between foreground and background objects by drawing the viewers gaze to related active elements (foreground) while sublimating the discernment of the stationary elements (background). For example n figure ... viewers will tend to group the shapes in the top row by their common fate and group the bottom row, of shapes based on their type (i.e. in the bottom row adjacent squares/circles will be grouped together (Lidwell et al., 2010).	 Figure 2.23: Depiction of the Gestalt law of Common Fate (Adapted from Lidwell et al. (2010)).
Good Continuation	
It is the perception of aligned elements as a single group/set that is more related than unaligned elements (Lidwell et al., 2010). This law facilitates the distinction between related and unrelated elements by aligning elements and guiding the viewer's gaze along the objects form with as little disruption as possible.'For example in figure..., the curved lines are still perceived as being part of a cohesive object despite there being gaps. This is because the eye percieves the path behind the structure (Lidwell et al., 2010).	 Figure 2.24: Depiction of the Gestalt law of Continuation (Adapted from Roth and Bruce (1995), and first reproduced in Ile (2018)).

Description	Image
Law of Pragnanz	
This law refers to the interpretation of ambiguous elements in the simplest way possible as opposed to interpreting it more complexly (Lidwell et al., 2010). The application of this law improves perception and recall by minimising the number of elements in a composition. This law thus supports the viewer's use of his/her own cognitive resources to translate compositions into their simplest designs. For example, the shapes in image ... are perceived simply as complete squares and triangles instead of as more complex geometries (Lidwell et al., 2010).	 Figure 2.25: Depiction of the Gestalt law of Pragnanz (Adapted from Lidwell et al. (2010)).
Proximity	
This principle deals with the perception of elements located near each other as a related set in contrast to elements located further away (Lidwell et al., 2010). Consideration of this law is a powerful tool for indicating similarities and differences while also reducing complexity. This law is applied by locating elements in the same group near each other and unrelated elements farther away. For example in image ... the proximity between the circles influences whether they are grouped as a square, columns or rows (Lidwell et al., 2010).	 Figure 2.26: Depiction of the Gestalt law of Proximity (Adapted from Lidwell et al. (2010)).
Similarity	
Similarity identifies the perception of individual elements with similar traits as belonging to the same group as opposed to dissimilar elements (Lidwell et al., 2010). Three main traits to consider when working with similarity are: colour, size and shape. The application of this law ensures that the similarity between elements is recognised by showing them as related to each other with respect to shape, colour and size (Lidwell et al., 2010).For example in figure ... although there is a see or dots, the number 1 is distinguishable because the dots of which the number consists are the same colour.	 Figure 2.27: Depiction of the Gestalt law of Similarity (Adapted from Graham (2008), and first reproduced in Ile (2018)).

Gestalt laws	
Description	Image
Uniform Connectedness	
This law describes how individual elements connected to each other by uniform visual properties are perceived as belonging to the same or a related set as opposed to the elements connected to varying visual properties or elements which are completely unconnected (Lidwell et al., 2010). Using this law improves relatedness as well as understanding of a composition by drawing the viewer's gaze to the related regions and highlighting their sequence. For example in figure ... the grouping of related words ensure that there is no room for misconception (Lidwell et al., 2010).	 Parking Ahead Available Rentals ATMs Behind Parking Ahead Available Rentals ATMs Behind Figure 2.28: Depiction of the Gestalt law of Uniform Connectedness (Adapted from Lidwell et al. (2010)).

2.5.4 Thematic Maps

For cartographers, a driving force in map creation is the consideration of the map's purpose.

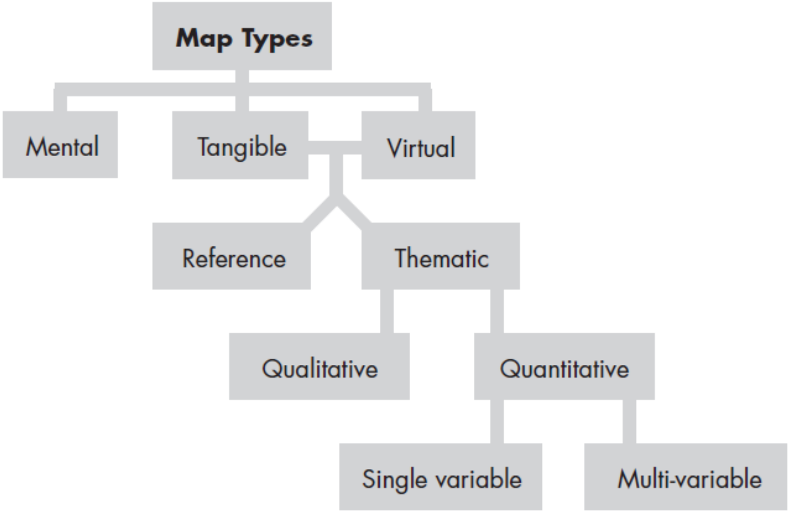


Figure 2.29: Dent et al. (2009) schematic for the different map classifications.

There are two main objectives in map creation: the display of a variety of features on a location focused map or the display of the structural characteristics of a geographical feature (Dent et al., 2009). Seeking to achieve the former objective results in the production of a reference map, whereas seeking to achieve the latter objective results in the production of a thematic map.

As the purpose of this investigation was to illustrate the characteristics of a geographical feature an understanding of thematic maps (created through the manipulation of graphic variables and with a single purpose in mind (Muehlenhaus, 2013)) had to be gained.

The success of thematic maps hinge on the type of data being used, either categorical (qualitative) or numerical (quantitative) data. Categorical data are data that can be assigned to discrete, non-numerical classes. The different classes can be distinguished by a range of graphic elements including shape, size and colour. Numerical data, however, are concerned with the representation of rank/magnitude within a data set. Therefore, qualitative maps show the spatial distribution or location of features while quantitative maps are more focused on showing feature quantities (Dent et al., 2009). Regardless of the data type thematic maps highlight intent, assist with the display of spatial distribution and ease the decision-making process.

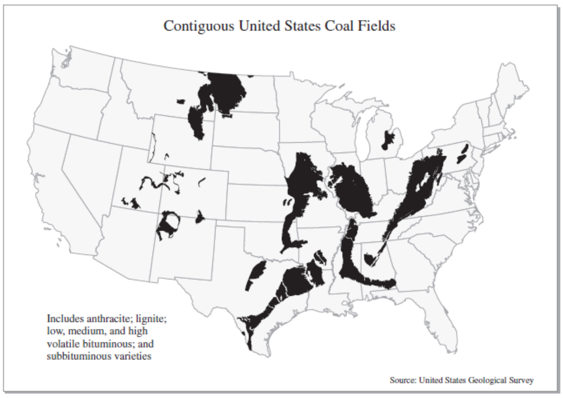

Contiguous United States Coal Fields
Includes anthracite; lignite; low, medium, and high volatile bituminous; and subbituminous varieties
Source: United States Geological Survey

Figure 2.30: Qualitative thematic map example (Dent et al., 2009)

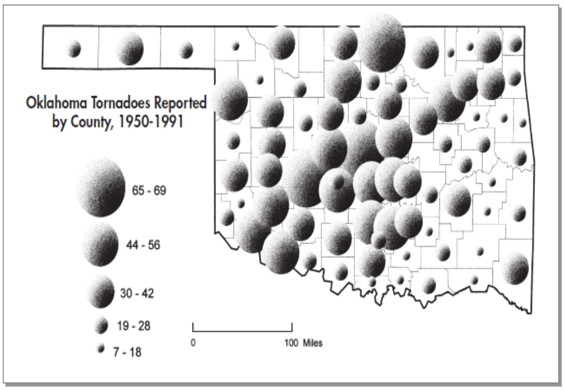


Figure 2.31: Quantitative thematic map example (Dent et al., 2009)

Thematic maps are composed of three cartographic units a *basemap* which provides spatial context to a *thematic overlay* which sets the purpose of the map and, lastly, *auxiliary map elements*.

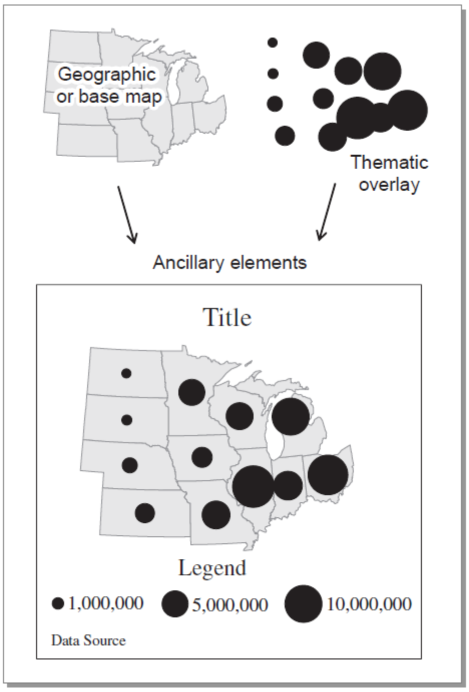


Figure 2.32: The components of a thematic representation (Dent et al., 2009)

It has become very common for thematic maps to be produced using geographical information systems (GIS) therefore it is a tool that requires some consideration (Dent et al., 2009).

2.5.5 Geographical Information Systems

3D geological modelling is a hot topic in a range of fields including the geosciences. This is because of its ability to define the boundaries between geological strata, enhance visibility and improve the accuracy of geological analysis (Zhu et al., 2012). To achieve the creation of such a model, various professionals in the geographical field - from urban developers to geologists - often look to GIS. GIS portray a simplified view of a complex reality and encompass the interaction of people and machines for the collection, storage, modelling, manipulation, management and dissemination of geographic information (Worboys and Duckham, 2004).

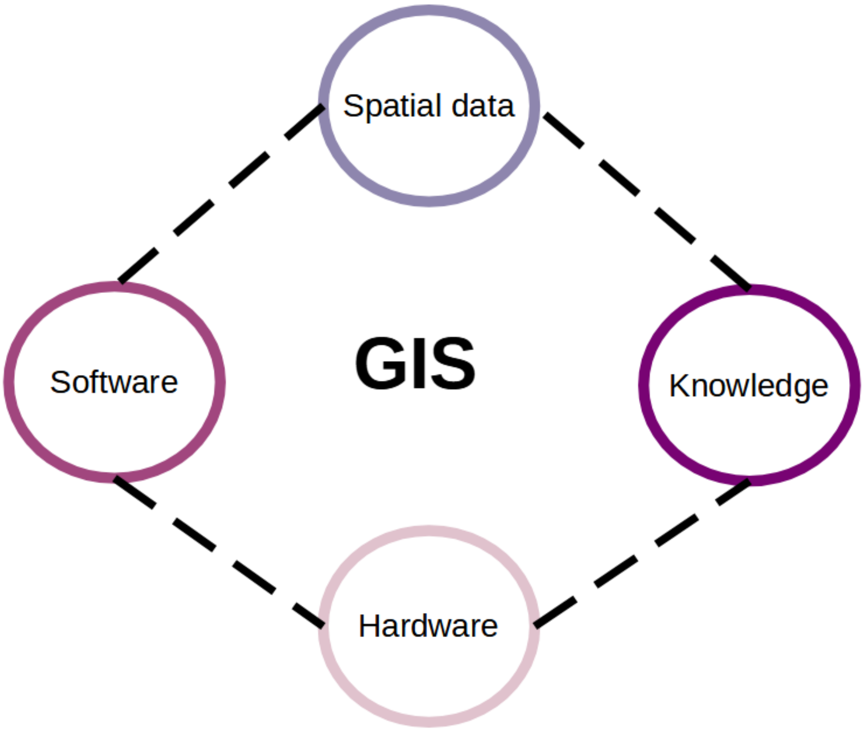


Figure 2.33: Visual depicting a simplified diagrammatic representation of GIS

GIS relies on four components: geographic data, human knowledge and experience, hardware and software. The first is geographic data i.e. having data that describes some phenomenon (Grinderud, 2009). This data can either be spatial (a geographic location e.g. residential address) or non-spatial (descriptive information about a geographic location e.g. residence owner) (Jovanović, 2016). With the data in hand, user knowledge and experience determines the degree to which the available technology is exploited. Here familiarity with both the system and the phenomena being represented yields the best outcomes (Grinderud, 2009). Next, geographic information technologies hardware and software are used to map, explore, process, interpret, share and store the spatial and non-spatial data.

Of the entire system, the most important technological unit is undoubtedly the data store (database). The database, a data container organised based on a data model and used for the storage and retrieval of data (Worboys and Duckham, 2004), lies at the heart of all GIS. Databases support various manipulation techniques including generalisation and transformation, where the data is smoothed, projected and scaled. Therefore, the development of a sound data model is key. To glean real spatial insights, analytical techniques (such as volume, area and overlay operations) are applied to the manipulated data set. These insights can then be displayed as maps, graphs, tables, reports and other such presentation formats. It is thus, this technology (GIS) that will be leveraged to display the outputs of this investigation – 3D cartographic maps.

At present, the tools available to geo-related professionals are confined to two-dimensional (2D) spatial visualisation, which not only cause difficulty in displaying complex real-world objects but also in processing and manipulating it. It was the inability of 2D systems to achieve successful and true representations of the 3D world, that led to the development of the 3D modelling technique – a technique used to directly transfer reality into a 3D digital model. The success of 3D modelling is in its freedom of a fixed viewing position.

3 Method

3.1 Introduction

This chapter covers both the approach used to develop a 3D geological prototype and describes the facets associated with it. As such, pertinent workflows and diagrams are included to aid explanation. Overall, both qualitative and quantitative data and a stepwise methodology were used to develop the algorithm that achieved the objectives of the investigation (the development of an aesthetically appealing user comprehensible 3D geological prototype). However, while the derived subsurface prototype was developed by instantiating an algorithm that makes its own inferences from unclassified/unlabeled input data; the adopted methodology took some geological understandings into account. This was especially true during the feature imputation, normalisation, engineering and selection stages of the data pre-processing.

3.2 Research Approach

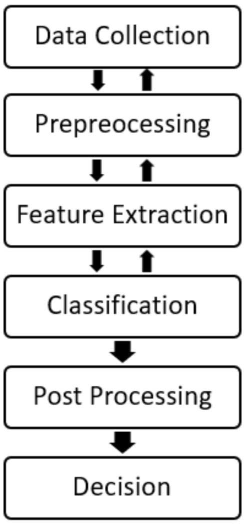


Figure 3.1: Image showing the approach used in the development of a 3D prototype geomodel (adapted from Chopra et al. (2019)).

The machine perception process used is summarised in Figure 3.1 and has both a 'bottom-up' and 'top-down' flow to enable the response incorporation of later levels. However, traditionally the process starts with the phenomena being observed. During this stage the phenomena of interest is studied and observations are recorded. These recordings often hinge on three characteristics: the nature of the phenomena, environmental factors and sensor response settings and characteristics (Bychkovskiy et al., 2003). Consequently, these characteristics have great impact on the breadth and quality of the data recorded. As the data used in this investigation was accessed and not collected, these characteristics played a major role in the results obtained as well as in the way the data was processed and used.

After data collection, and in adherence with the machine perception approach, data exploration was the next step undertaken. During this step, knowledge was gained about the data by learning about the information collected. As such, the input features (i.e. well logs) available for predicting the target variable (i.e. the groups – that the wells represented) were inspected. Besides this the data was explored to gain a better understanding of its structure and quality by looking at the data format (i.e. numerical/categorial, organized/unorganised), completeness and distribution.

With data exploration completed, the next step carried out was data pre-processing. In pre-processing the data was cleaned, scaled and formatted to support noise reduction as well as the development of the best prototype possible (Duda et al., 2012). Although there are numerous other pre-processing activities, such as data encoding and binarising, this investigation only focused on pre-processing activities that could be applied to numerical data. This was because the data almost entirety consisted of only quantitative data.

After the cleaned training data was split into two sets (a training and testing set) it was then available for use in dimensionality reduction, visualisation and clustering machine learning algorithms. In the application of these models the line between over and under fitting had to be tread very carefully. This endeavour was carried out to ensure that the prototype captured the complexity of the data as best as possible, without making it unadaptable to new datasets.

The penultimate step of this process was post-processing and involved analysing and evaluating the prototype. Here, the developed prototype was assessed by looking at factors such as bias, run time and input feature variable importance. Although not always the case, during this step GIS visualisation was implemented to facilitate the decision-making process and highlight possible areas of geological exploration.

The process ended with the decision step, where conclusions were drawn based on the output of the machine perception workflow.

3.3 Data Desrciption

The data used as part of this investigation was accessed and not collected. Thus, instead of detailing the data collection process, the data source and the data itself will be described.

The geological logs were obtained courtesy of the Petroleum Agency of South Africa (PASA), an agency which promotes on- and off-shore oil and gas exploration and development on behalf of South Africa's government (Petroleum Agency of South Africa, 2013). PASA provided the geological logs for 3 wells surveyed over a period of 16 years (2000 – 2016) and located off the Bredasdorp coast: F-04, F-06 and F-08 (see figure 3.2).

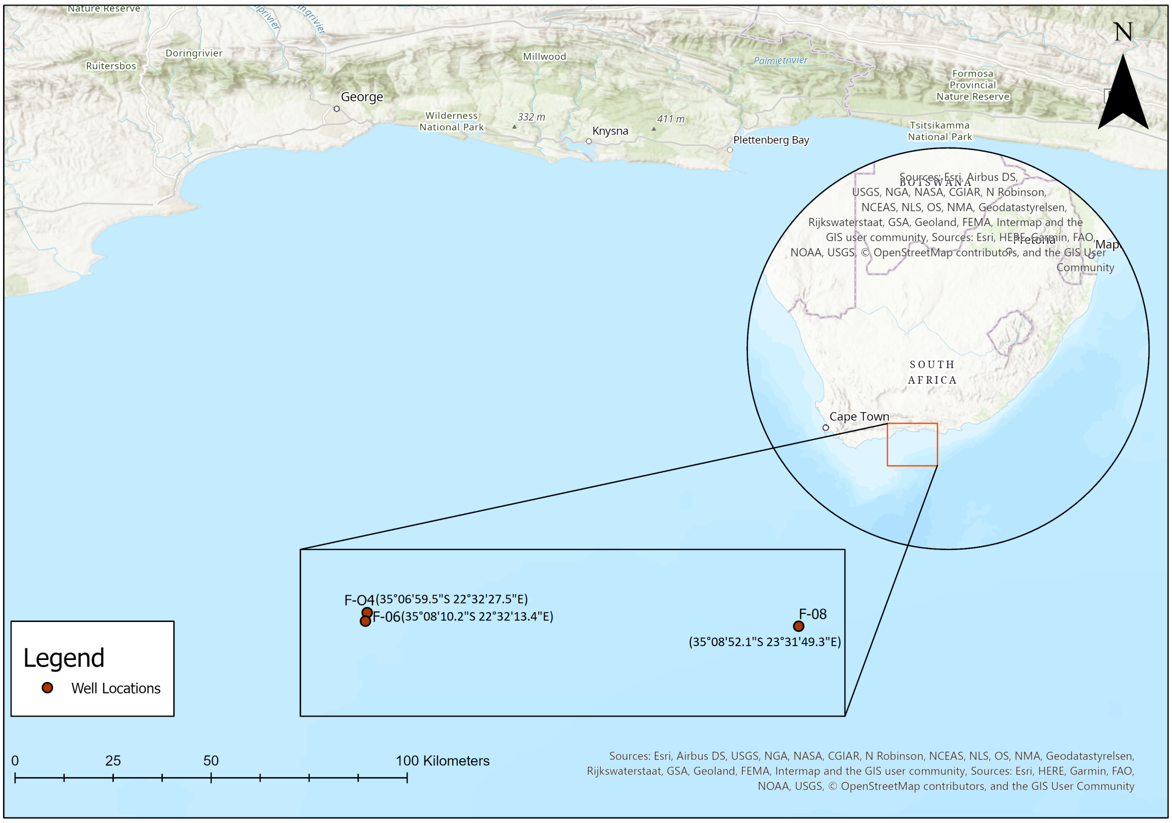

N
George
Knysna
Plettenberg Bay
Wilderness National Park
Tsitsikamma National Park
Formosa Provincial Nature Reserve
Ruitersbos
Doringrivier
Millwood
BOTSWANA
Pretoria
SOUTH AFRICA
Cape Town
F-O4(35°06'59.5"S 22°32'27.5"E)
F-06(35°08'10.2"S 22°32'13.4"E)
F-08
(35°08'52.1"S 23°31'49.3"E)
Legend
● Well Locations
0
25
50
100 Kilometers
Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Figure 3.2: Well locations

The information captured from the wells were commonly used logs (curves) such as caliper logs, spontaneous potential logs, resistivity logs, and many other wireline logs. The information also detailed well and parametric information, to provide environmental context.

In total 439 .las files held the curves for all three wells; with 314 , 61 and 64 logs belonging to wells F-04, F-06 and F-08 respectively. Despite there being over 140 curves in some files only 15 of them (DEPT, GR, TNPH, NPHI, RHOB, LLD, MSFL, MRES, MTEM, SP, CALS, BS, DT, DTLN and ITT) were used to develop the prototype. These 15 were chosen because they are some of the most common, descriptive and useful features that can be used in well log interpretation.

3.4 Data Pre-Processing

3.4.1 Data Conversion

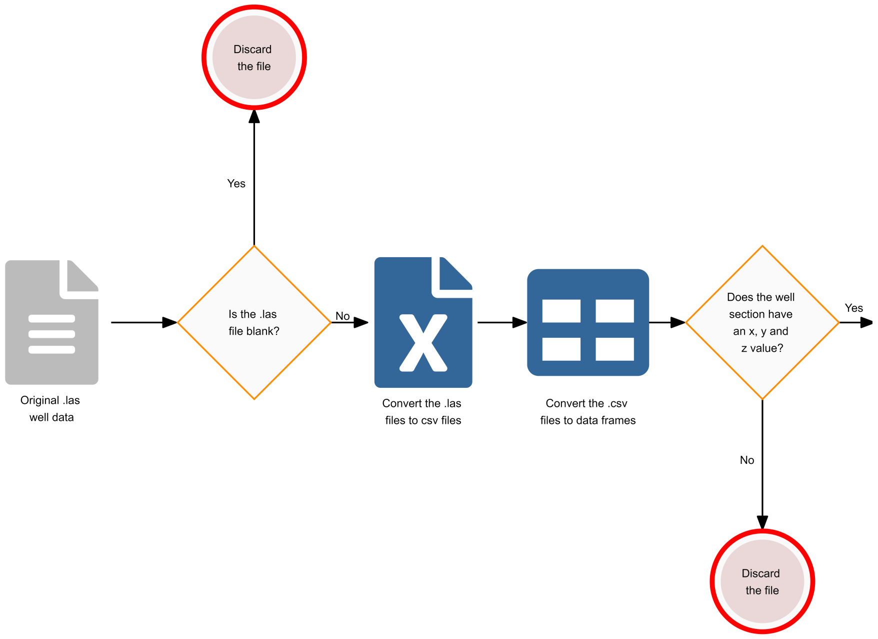


Figure 3.3: Workflow for exploring the data

The raw data for this machine learning endeavour was stored in las file format over multiple files. Therefore, the first step in preprocessing the data involved getting the raw data into the development environment in a readable and manageable format. This was best achieved by reading all the las files for each well and then converting them into comma-separated value (csv) files (as seen in Figure 3.4).

The converted csv files were then read into the development environment before being converted into data frames. This format conversion was done because the tabular (row and column based) data frames structured the data, promoted intuitive and versatile data use and also supported data wrangling i.e. the transformation, cleaning and organising of raw data (Albon, 2018).

When performing the format conversion from .las to .csv files, a number of wells either did not have well coordinates or were entirely blank (i.e. datasets that had header information but were otherwise completely empty). The presence of these quantities was necessary for prototype development, therefore these incomplete files were discarded as their use in the project would not have meaningfully added to the research but would have detracted from it instead.

3.4.2 Data Exploration

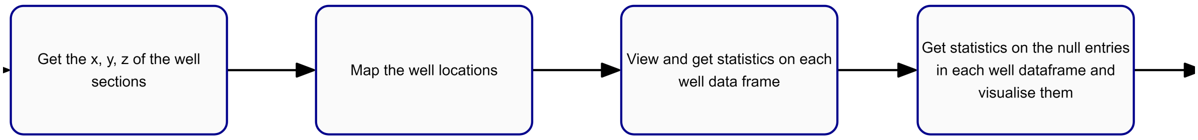

Get the x, y, z of the well sections
Map the well locations
View and get statistics on each well data frame
Get statistics on the null entries in each well dataframe and visualise them

Figure 3.4: Workflow for exploring the data

After the data was loaded in the correct format it was explored to garner a better understanding of data structure and content. First, the wells were mapped to gain a spatial understanding of their locations. Although technically a step in the data transformation process, the well locations were obtained by converting the degree-minute-second coordinates (dms) to decimal degrees (dd) and then displayed on a map.

Following the spatial exploration, samples from the data set were displayed. Samples, instead of the entire data set, were viewed because the data set was reasonably large and only a quick exploration of a few records was necessary to understand the data scheme. As such, for each data frame a view of the first 5 rows was created and the dimensions (number of columns and rows) were extracted. Also, as part of the data exploration process, descriptive and summary statistics for all the numerical columns were obtained. It was from these two data exploration processes that it was gathered that each column corresponds to one well log while each row corresponds to one observation.

As missing values are ubiquitous in almost all machine learning problems, the data exploration process wrapped up with visualisation of the null entries in the data.

3.4.3 Feature Selection

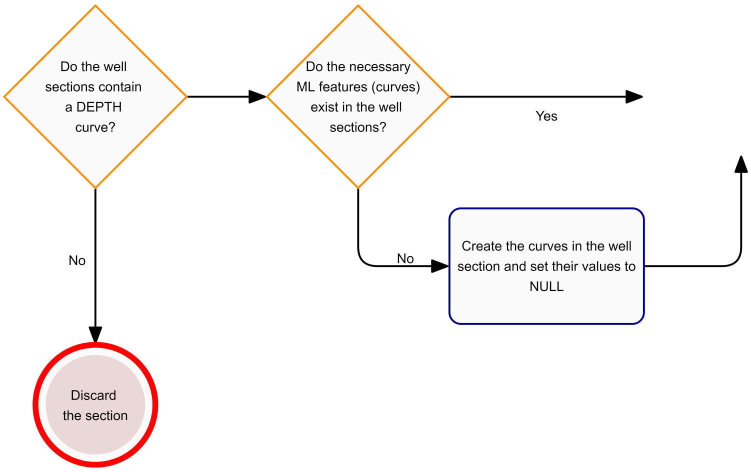

Do the well sections contain a DEPTH curve?
Do the necessary ML features (curves) exist in the well sections?
Yes
No
No
Create the curves in the well section and set their values to NULL
Discard the section

Figure 3.5: Workflow for selecting features

Between the data integration activities (see subsection 3.4.4 below), the data was recomposed to only reflect features (well logs/curves) that were important in well log clustering. This was done by identifying and extracting conventional well log interpretation curves from the data.

In this vein, only well logs that had depth (DEPT) curves were extracted and then assessed for the presence of other pertinent well logs. As mentioned in section 3.3, aside from DEPT the other curves extracted were GR, TNPH, NPHI, RHOB, LLD, MSFL, MRES, MTEM, SP, CALS, BS, DT, DTLN and ITT. These logs were selected because they are the most commonly captured curves during the well logging process and would, therefore, be consistent measures to which to apply machine learning algorithms.

However, if a curve (excluding DEPT) did not exist for a record, an empty entry was created. This allowed for data imputation later in the preprocessing workflow (see subsection 3.4.5).

3.4.4 Data Integration

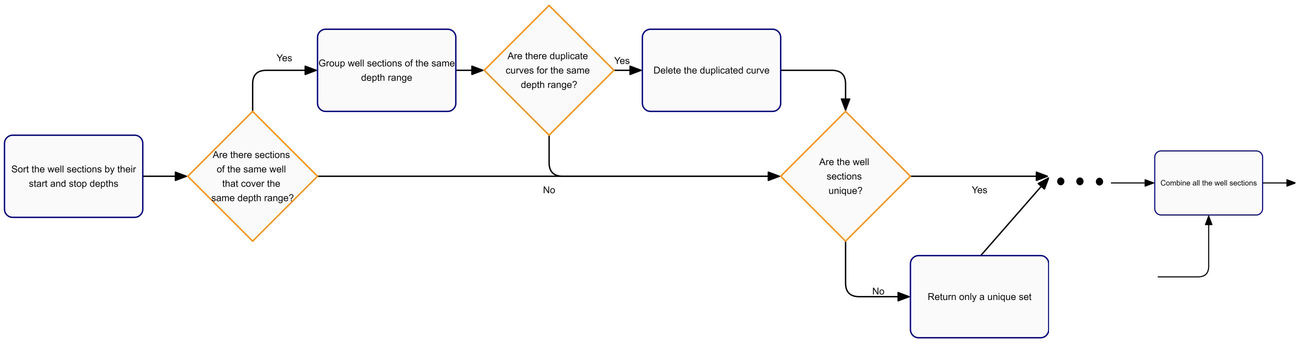

Sort the well sections by their start and stop depths
Are there sections of the same well that cover the same depth range?
Yes
Group well sections of the same depth range
Are there duplicate curves for the same depth range?
Yes
Delete the duplicated curve
No
Are the well sections unique?
Yes
No
Return only a unique set
Combine all the well sections

Figure 3.6: Workflow for integrating the data sets

After a brief look at the condition of the data, and some pertinent descriptive and summary statistics, the next steps involved concatenating the separate data frames into a single unit. Concatenating the data supported meaningful information extraction and analysis from a unified structure.

To achieve the desired data configuration, the data frames were put through a few processes. The first involved sorting all the well sections by their start and stop depths, and then separating the data into unique and non-unique depth ranges. If there were multiple data frames with the same depth range, these files were grouped into one set.

The grouped data were then evaluated to ensure that each column (well log) was unique. If the feature wasn't unique, then one of the duplicated columns were dropped so that curve singularity could be established.

Penultimately, the uniqueness of each depth range was reconfirmed before the unique depth range data frames were combined into one data frame warehouse. After aggregating the data into one unified structure the next step in pre-processing (data cleaning) could be carried out.

3.4.5 Data Cleaning

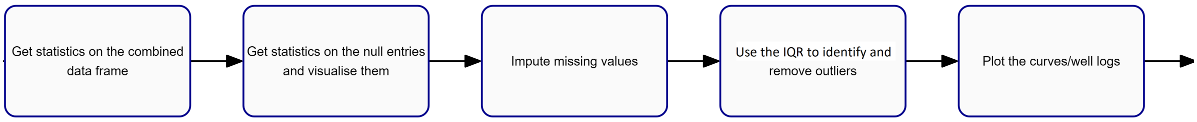

Get statistics on the combined data frame
Get statistics on the null entries and visualise them
Impute missing values
Use the IQR to identify and remove outliers
Plot the curves/well logs

Figure 3.7: Workflow for cleaning the data

Before cleaning the data, a fresh set of visual and descriptive statistics were obtained for the aggregated data frame. This included the location, spread and quantity of null entries in the data frame.

Following this, the main activities of data cleaning - missing data and outlier management - were executed. These data inconsistencies had to be identified and addressed because a consistent and complete data set was crucial for carrying out clustering and making predictions.

There are two primary ways for handling missing values removing the data or imputing it. In the data removal process rows or columns are deleted based on the percentage of missing values they have. The deletion threshold is mutable but usually lies between 70-75% (Beyeler, 2017)). The following equation was used determine which features fell above this threshold:

$$missing\ entries\ (\%) = \frac{n_{missing}}{n}$$

In this equation, $n_{missing}$ is the number of missing entries and n is the total number of entries.

The advantage of removing null entries is that it ensures a usable data set and can be carried out both quickly and easily. Despite these advantages, it can also result in the loss of important features.

Therefore, to balance the data loss, data imputation was also carried out. With this approach statistical methods (e.g. mean, median, mode) or a modelled approach is used to derive missing values.

For this investigation the modelled imputation approach was used because the values in the data set correlate with each other and a modelled approach considers missing entries as functions of all the other entries in a record (row) (Pedregosa et al., 2011).

With the missing values taken care of, the next activity in data cleaning - outlier management - was tackled. Here, points that were very large or small with respect to the data distribution were removed. For each column (well log), this was achieved by

1. Sorting the data
2. Plotting a boxplot
3. Getting the upper and lower fences as well as the interquartile range (IQR)
4. Removing the data points that were either below Q1 – (1.5 x IQR) or above Q3 + (1.5 x IQR) (see figure 3.8)

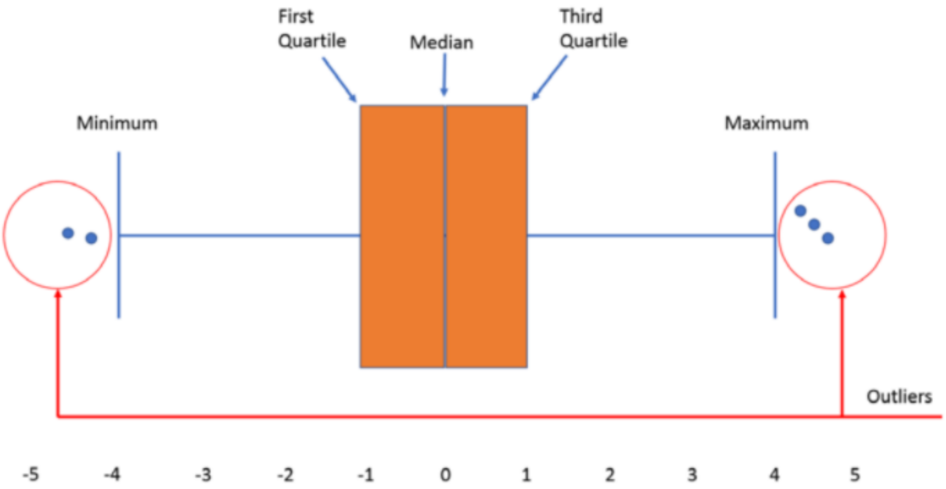


Figure 3.8: Sample of outliers in a boxplot

Before the final step in the pre-processing workflow the cleaned data (up to this point) was visualised. The visualisation carried out was a log (formation parameter) by depth plot, which is the standard well log method of display. To the trained eye, these plots can provide immediate lithological identification, porous and non-porous rock distinction and potential pay zone recognition (Peveraro, 2006).

3.4.6 Data Scaling

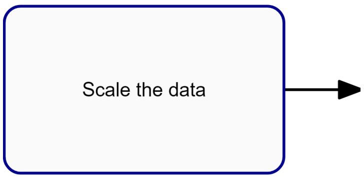


Figure 3.9: Workflow for scaling the data

The final step in the pre-processing workflow (data scaling) involved transforming the features so that they used the same scale, magnitude and range. This was an essential step in preparing the data for clustering because machine learning algorithms weigh a feature's importance based on its magnitude and therefore non-normalised readings could greatly affect any predictions made.

The scalar used to transform the data was a Standard Scalar. This scalar was used because it scales the data into a uniform unit over the entire data range, and in this manner ensures that the appropriate effect of each feature is considered (Beyeler, 2017). The Standard Scalar module works by first subtracting each value(x) from the mean (μ) of all the data and then dividing it by the variance of the data (σ) i.e.

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

At this stage the data was in a format suitable for the data sensitive machine learning algorithms and could used in the next stage of the machine learning workflow - clustering.

3.5 Clustering

3.5.1 Train/Test Split

As mentioned in section 2.4.2 unsupervised machine learning is made up of two main activities: unsupervised transformations and clustering.

However, before the pre-processed data could be applied to machine learning algorithms, the data had to be separated into a testing and training set (see figure 3.11). The training set was created to provide the machine learning algorithms with data knowledge, upon which predictions (on the test set) could be made.

The training and test set were created using the 80:20 train/test split ratio, where 80 percent of the data was used for training and the 20 percent was used for testing (see figure 3.10) .

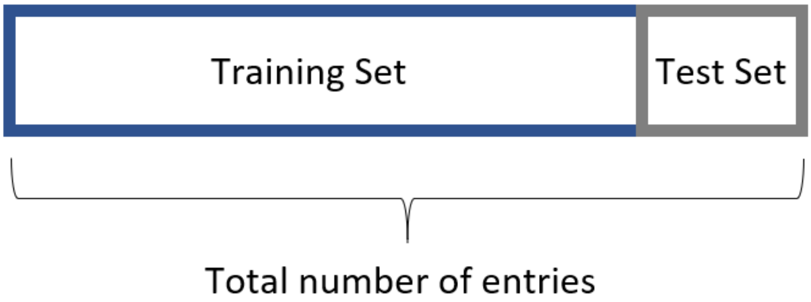
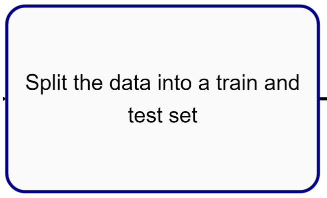


Figure 3.10: Train/Test split (Bronshtein, 2017)



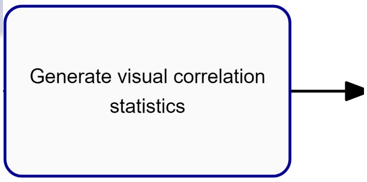


Figure 3.11: Workflow step for splitting the data into a train and test the data

Figure 3.12: Workflow step for carrying visual data correlation

3.5.2 Data Correlation

Next, using the training set, two visual correlation matrices were created as a means to better understand the linear relationships between each variable. The first correlation plot was created by linearly relating all the features in the training set, getting the correlation value between them, associating these values with a set colour range and then plotting the colour associated correlations.

The second correlation matrix was created by plotting the features under investigation in 2D space. The manner of representation was altered depending on whether it was on the upper triangle, lower triangle or diagonal. On the lower triangle the features relationships were represented as scatter plots i.e points on a 2D axis where the value of a feature determined its location. On the upper triangle features relationships were represented as hexbins i.e. colour associated quantity counts in a binned 2D feature space. Lastly, the diagonal represented feature relationships as histograms i.e. quantity counts in grouped ranges (see figure 3.13).

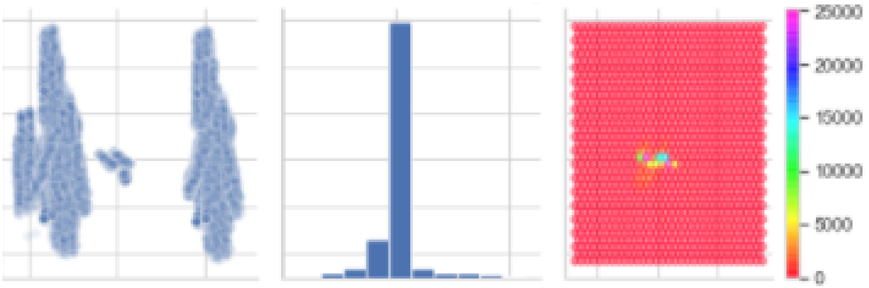


Figure 3.13: From left to right to left examples of a scatter plot, histogram and hexbin plot

3.5.3 Principle Component Analysis

With the preliminary activities completed, the first unsupervised exercise - PCA - was conducted.

To determine the optimal number of dimensions to apply to the PCA algorithm, the PCA algorithm was run iteratively for the total number of features (columns) in the data frame, which is to say that the algorithm was run for a *2 component* feature space up to *n component* spaces. Running the algorithm this away allowed the minimum number of components necessary for 95% data variance to be ascertained.

Using this value, the number of components that explain 95% of the data variance, the PCA was run again and applied to the training data set. Application of the PCA algorithm to the training data set reduced its dimensionality while maintaining as much information as possible

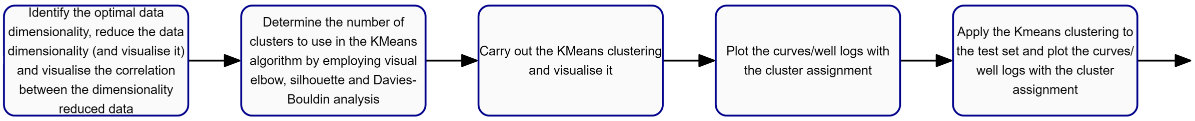

Identify the optimal data dimensionality, reduce the data dimensionality (and visualise it) and visualise the correlation between the dimensionality reduced data
Determine the number of clusters to use in the KMeans algorithm by employing visual elbow, silhouette and Davies-Bouldin analysis
Carry out the KMeans clustering and visualise it
Plot the curves/well logs with the cluster assignment
Apply the Kmeans clustering to the test set and plot the curves/ well logs with the cluster assignment

Figure 3.14: Workflow for performing PCA and KMeans Clustering

3.5.4 KMeans clustering

Similarly to the process undergone during PCA, KMeans clustering involved working out optimal algorithm parameters, running the algorithm for the optimal algorithm parameters and applying the KMeans algorithm to the PCA-reduced data set.

In this case, the optimal algorithm parameters (the number of clusters) were determined by using three evaluation metrics: the elbow method, the silhouette coefficient and the Davies-Bouldin score. Assessment of these three metrics showed coincidence and pointed to the number of clusters that should be used in the KMeans clustering algorithm. Therefore the KMeans clustering was re-run for the optimal number of clusters and applied to the PCA reduced data set.

The output of the KMeans process, a cluster label for each point, was then visualised as a frequency count before it was mapped against both the PCA reduced data and the entire training data set.

Next the KMeans algorithm was programmed to make cluster label predictions with the test set as the input data set. These cluster assignments along with those for the test set were then visualised alongside the log (formation parameter) by depth plots to wrap up the well log interpretation process.

3.6 Data Post-Processing

3.6.1 Data Conversion

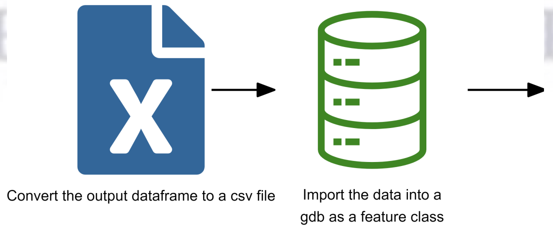

Convert the output dataframe to a csv file
Import the data into a gdb as a feature class

Figure 3.15: Workflow for converting the data for 3D modelling

To visualise the clustered well logs as a prototyped 3D geomodel, the results of the KMeans clustering had to be extracted and converted into a csv file - a GIS readable format. This csv file was then read into two visualisation environments, one programmatically created through python and the other created via a GIS application. In the GIS application the csv file was read in as a feature class - a geographical layer - and stored in a file geodatabase. It was on this geographical layer that an interpolated 3D subsurface prototype was created.

During the conversion of the csv file into a feature class, the desired spatial reference was set to a projected coordinated system as the interpolation tool requires this projection type.

3.6.2 GIS prototyping

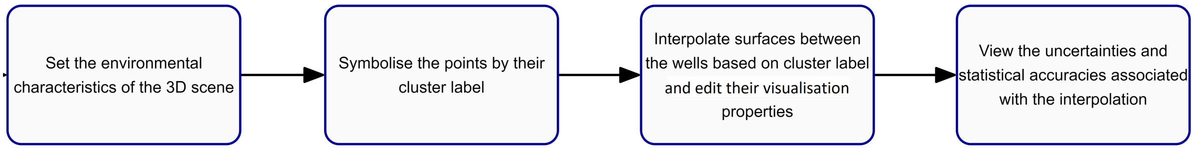

Set the environmental characteristics of the 3D scene
Symbolise the points by their cluster label
Interpolate surfaces between the wells based on cluster label and edit their visualisation properties
View the uncertainties and statistical accuracies associated with the interpolation

Figure 3.16: Workflow for 3D GIS prototyping

The final step in the development of a 3D geological prototype, prototyping of the GIS model, consisted of four steps:

1. Setting up the appearance of the scene
2. Setting up the appearance of the features
3. Interpolating a surface between the points in the feature class
4. Viewing the uncertainties and accuracies around the interpolation

To kick off 3D rendering, topographical points for the southern tip of South Africa were extracted from Google Earth and converted into readable text files. These points were then imported into the 3D scene which was created with vedo (a pythonic library that supports the visualisation of 3D objects). Before being scaled and coloured to approximate relaity, the imported points were interpolated to get a topographical surface for the region. With this completed, the classified wells were imported into the scene and their appearances were set. The symbology of the features were set to graduated colours for each of the unique cluster labels.

To get a geo-statistical understanding of the uncertainty and accuracy of a surface interpolated between the class predicted wells, 3D surface interpolation was carried out using the 3D Empirical Bayesian Kriging geo-statistical method in ArcGIS Pro (a powerful GIS desktop application developed by Esri). To set up the 3D scene in the application, a topobathy (a combination topographic and bathymetric) basemap was imported and the vertical exaggeration of the surface was set to 10. This surface was also given a transparent surface colour so that subsurface visualisation would be possible. Next the projection of the scene was set to match that of the imported feature class before the appearance of the feature class was set. Next the symbology of the feature class was set to graduated colours for each of the unique cluster labels.

The optimal parameters for the surface interpolation were determined through the use of the geo-statistical wizard, which allowed for parameter tuning and result simulation before final application. Overall, the use of ArcGIS Pro allowed for regional class values to be set between the wells based on the limited known class values.

Data post-processing concluded with the assessment of the automatically generated geo-statistical uncertainty and accuracy of the interpolated surface.

3.7 Methodological Framework

The entire methodology used in the creation of the 3D geological prototype can be seen in the image below:

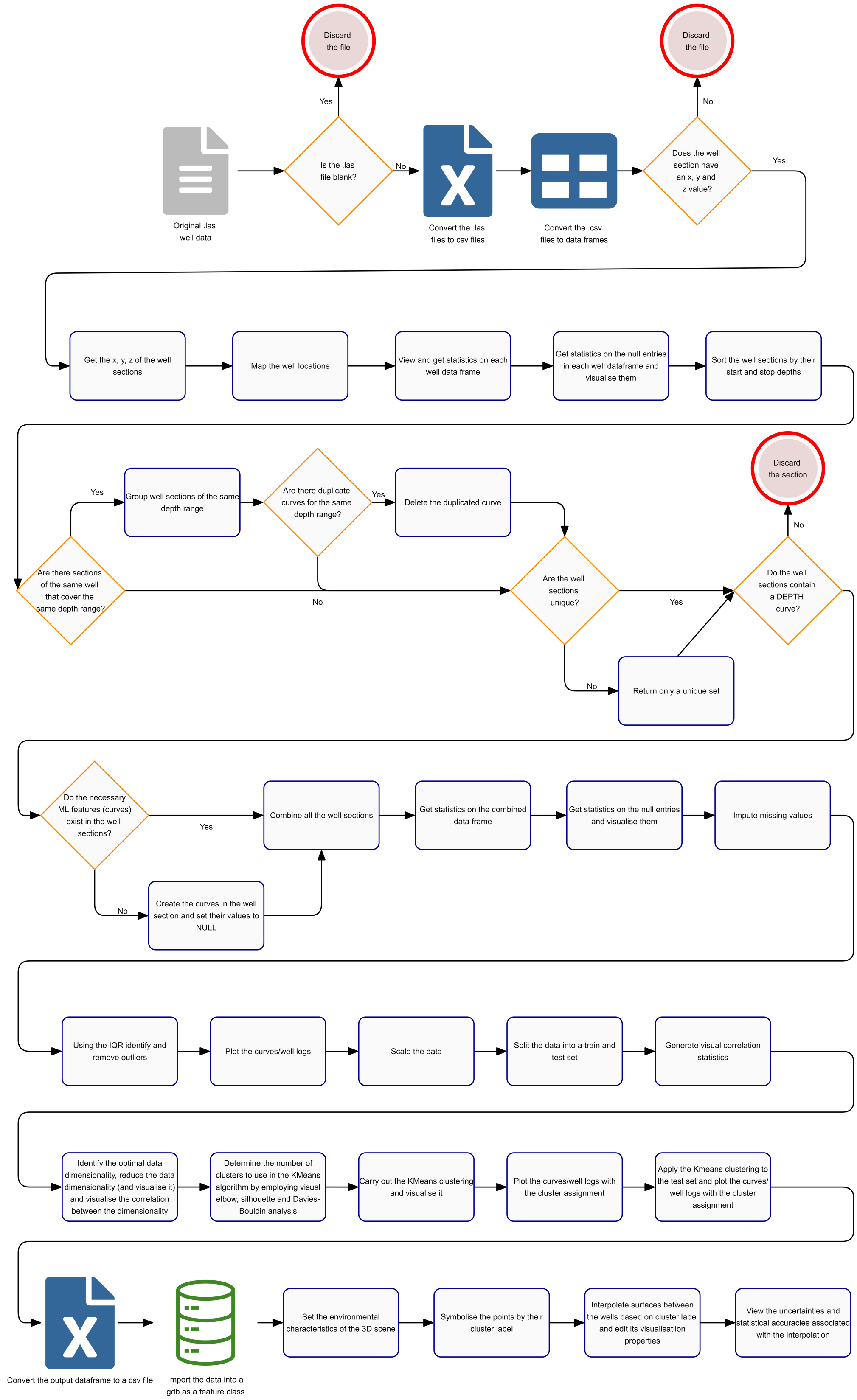

Discard the file
Yes
Discard the file
No
Original .las well data
Is the .las file blank?
No
Convert the .las files to csv files
Convert the .csv files to data frames
Does the well section have an x, y and z value?
Yes
Get the x, y, z of the well sections
Map the well locations
View and get statistics on each well data frame
Get statistics on the null entries in each well dataframe and visualise them
Sort the well sections by their start and stop depths
Discard the section
Yes
Group well sections of the same depth range
Are there duplicate curves for the same depth range?
Yes
Delete the duplicated curve
No
Are there sections of the same well that cover the same depth range?
No
Are the well sections unique?
Yes
Do the well sections contain a DEPTH curve?
No
Return only a unique set
Do the necessary ML features (curves) exist in the well sections?
Yes
Combine all the well sections
Get statistics on the combined data frame
Get statistics on the null entries and visualise them
Impute missing values
No
Create the curves in the well section and set their values to NULL
Using the IQR identify and remove outliers
Plot the curves/well logs
Scale the data
Split the data into a train and test set
Generate visual correlation statistics
Identify the optimal data dimensionality, reduce the data dimensionality (and visualise it) and visualise the correlation between the dimensionality
Determine the number of clusters to use in the KMeans algorithm by employing visual elbow, silhouette and Davies-Bouldin analysis
Carry out the KMeans clustering and visualise it
Plot the curves/well logs with the cluster assignment
Apply the Kmeans clustering to the test set and plot the curves/ well logs with the cluster assignment
Convert the output dataframe to a csv file
Import the data into a gdb as a feature class
Set the environmental characteristics of the 3D scene
Symbolise the points by their cluster label
Interpolate surfaces between the wells based on cluster label and edit its visualisatiion properties
View the uncertainties and statistical accuracies associated with the interpolation

Figure 3.17: Complete methodology for the creation of a 3D geological prototype

4 Results and Analysis

4.1 Introduction

Four of the six objectives of the investigation (listed below) were covered in the literature review.

1. To present a geological understanding of the Bredasdorp Basin.
2. To demonstrate a clear understanding of what a well log is.
3. To determine the well logs that can be applied to 3D geological model development.
4. To identify and explain the fundamental characteristics that facilitate in user understanding and aesthetic appeal when working with cartographic representations.

This chapter presents the results of the 3D geological prototype creation process for wells in the Bredasdorp Basin. Therefore, the chapter contains the results obtained to achieve the last two objectives of the investigation (listed below).

1. To implement principal component analysis (PCA) and Kmeans clustering on well log data and interpret the results.
2. To develop a prototype of a 3D geological map that supports aesthetic appeal and user comprehension by adapting and combining the best practices within existing 3D modelling theory.

As such this chapter covers the results of data optimisation, data reduction (through PCA), cluster selection (through unsupervised machine learning metrics) and geological prototype analysis. The results are presented in the form of data frames, box plots, correlation plots, unsupervised machine learning metric plots, and log plots. From these results measures were extracted and comparisons were made to give an indication of how the 3D geological prototyping fared.

4.2 3D Geological Prototype Process Outline

To achieve the objectives set out in this investigation, and develop a robust solution, both the theoretical and practical facets of the problem had to be considered. The factors considered are outlined in Figure 4.1:

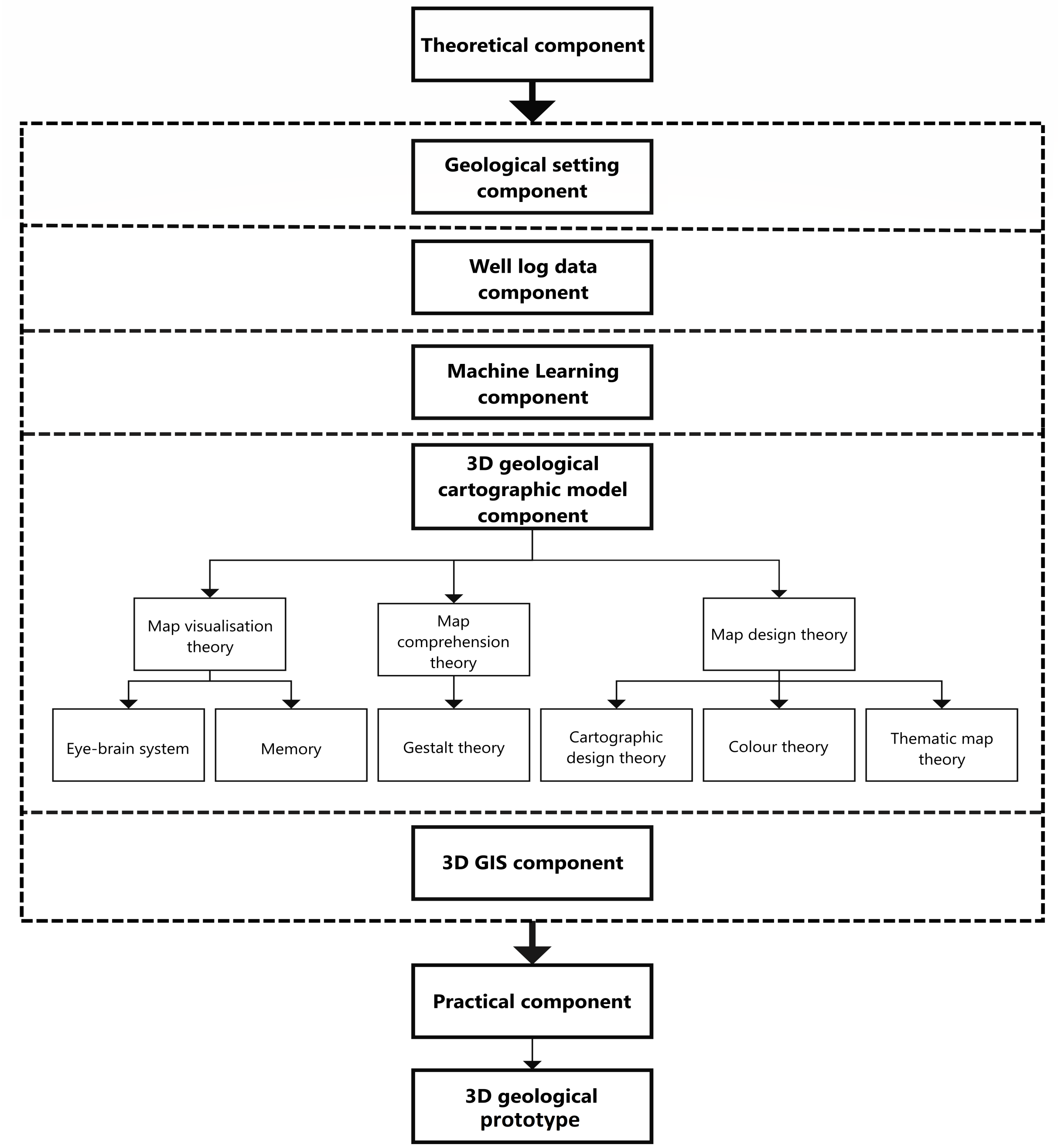
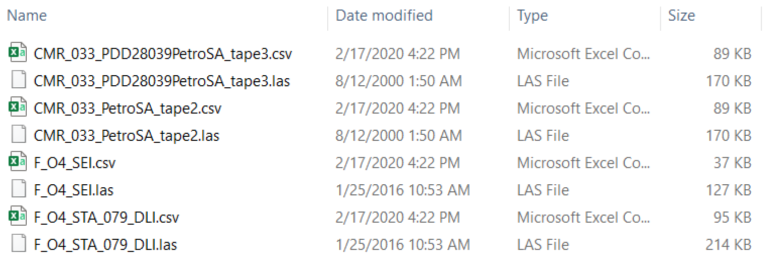


Figure 4.1: Holistic approach applied in the development of a 3D geological model based on well log data.

4.3 Data Optimisation Results

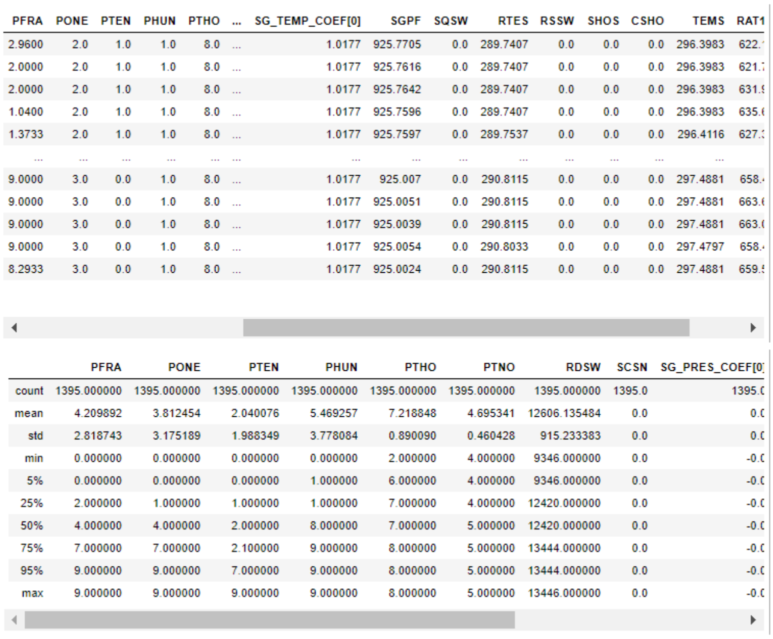
A total of 439 .las files were read into the development environment and coverted into .csv files (see figure 4.2).



Name	Date modified	Type	Size
CMR_033_PDD28039PetroSA_tape3.csv	2/17/2020 4:22 PM	Microsoft Excel Co...	89 KB
CMR_033_PDD28039PetroSA_tape3.las	8/12/2000 1:50 AM	LAS File	170 KB
CMR_033_PetroSA_tape2.csv	2/17/2020 4:22 PM	Microsoft Excel Co...	89 KB
CMR_033_PetroSA_tape2.las	8/12/2000 1:50 AM	LAS File	170 KB
F_O4_SEI.csv	2/17/2020 4:22 PM	Microsoft Excel Co...	37 KB
F_O4_SEI.las	1/25/2016 10:53 AM	LAS File	127 KB
F_O4_STA_079_DLI.csv	2/17/2020 4:22 PM	Microsoft Excel Co...	95 KB
F_O4_STA_079_DLI.las	1/25/2016 10:53 AM	LAS File	214 KB

Figure 4.2: File coversion: .las to .csv

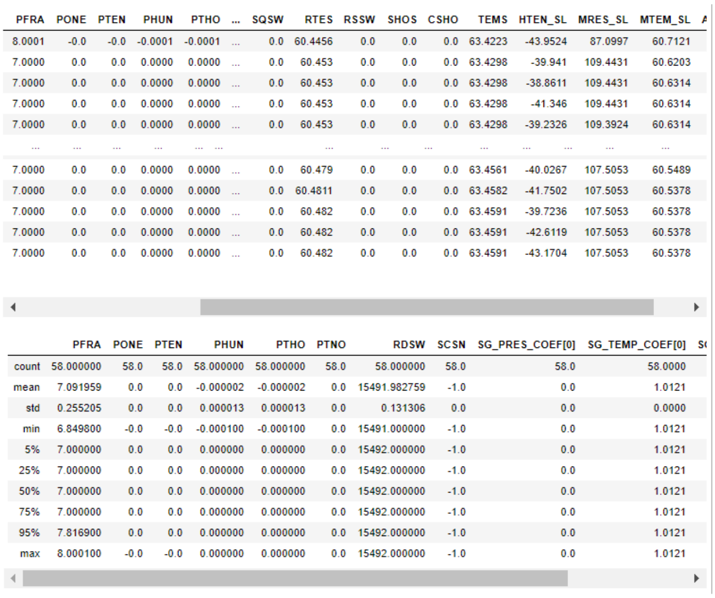
However, only 280 of them were not blank and were populated with x, y and z coordinate information. Therefore 159 las files had to be discarded as unsuitable for clustering. The kept data frames varied in size, features (well logs/curves) and thus statistics (see figure 4.3 and 4.4).



PFRA	PONE	PTEN	PHUN	PTHO	...	SG_TEMP_COEF[0]	SGPF	SQSW	RTES	RSSW	SHOS	CSHO	TEMS	RAT1
2.9600	2.0	1.0	1.0	8.0	...	1.0177	925.7705	0.0	289.7407	0.0	0.0	0.0	296.3983	622.
2.0000	2.0	1.0	1.0	8.0	...	1.0177	925.7616	0.0	289.7407	0.0	0.0	0.0	296.3983	621.7
2.0000	2.0	1.0	1.0	8.0	...	1.0177	925.7642	0.0	289.7407	0.0	0.0	0.0	296.3983	631.
1.0400	2.0	1.0	1.0	8.0	...	1.0177	925.7596	0.0	289.7407	0.0	0.0	0.0	296.3983	635.
1.3733	2.0	1.0	1.0	8.0	...	1.0177	925.7597	0.0	289.7537	0.0	0.0	0.0	296.4116	627.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9.0000	3.0	0.0	1.0	8.0	...	1.0177	925.007	0.0	290.8115	0.0	0.0	0.0	297.4881	658.
9.0000	3.0	0.0	1.0	8.0	...	1.0177	925.0051	0.0	290.8115	0.0	0.0	0.0	297.4881	663.
9.0000	3.0	0.0	1.0	8.0	...	1.0177	925.0039	0.0	290.8115	0.0	0.0	0.0	297.4881	663.
9.0000	3.0	0.0	1.0	8.0	...	1.0177	925.0054	0.0	290.8033	0.0	0.0	0.0	297.4797	658.
8.2933	3.0	0.0	1.0	8.0	...	1.0177	925.0024	0.0	290.8115	0.0	0.0	0.0	297.4881	659.

	PFRA	PONE	PTEN	PHUN	PTHO	PTNO	RDSW	SCSN	SG_PRES_COEF[0]
count	1395.000000	1395.000000	1395.000000	1395.000000	1395.000000	1395.000000	1395.000000	1395.0	1395.0
mean	4.209892	3.812454	2.040076	5.469257	7.218848	4.695341	12606.135484	0.0	0.0
std	2.818743	3.175189	1.988349	3.778084	0.890090	0.460428	915.233383	0.0	0.0
min	0.000000	0.000000	0.000000	0.000000	2.000000	4.000000	9346.000000	0.0	-0.0
5%	0.000000	0.000000	0.000000	1.000000	6.000000	4.000000	9346.000000	0.0	-0.0
25%	2.000000	1.000000	1.000000	1.000000	7.000000	4.000000	12420.000000	0.0	-0.0
50%	4.000000	4.000000	2.000000	8.000000	7.000000	5.000000	12420.000000	0.0	-0.0
75%	7.000000	7.000000	2.100000	9.000000	8.000000	5.000000	13444.000000	0.0	-0.0
95%	9.000000	9.000000	7.000000	9.000000	8.000000	5.000000	13444.000000	0.0	-0.0
max	9.000000	9.000000	9.000000	9.000000	8.000000	5.000000	13446.000000	0.0	-0.0

Figure 4.3: Overview of the log data for a large data set with 1395 records



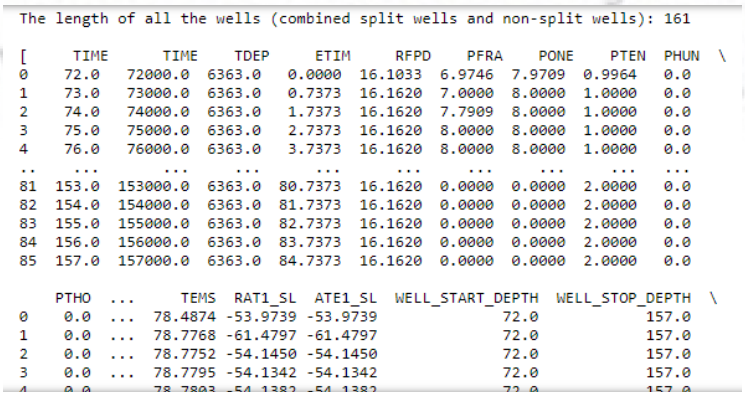
PFRA	PONE	PTEN	PHUN	PTHO	...	SQSW	RTES	RSSW	SHOS	CSHO	TEMS	HTEN_SL	MRES_SL	MTEM_SL	A
8.0001	-0.0	-0.0	-0.0001	-0.0001	...	0.0	60.4456	0.0	0.0	0.0	63.4223	-43.9524	87.0997	60.7121	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.453	0.0	0.0	0.0	63.4298	-39.941	109.4431	60.6203	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.453	0.0	0.0	0.0	63.4298	-38.8611	109.4431	60.6314	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.453	0.0	0.0	0.0	63.4298	-41.346	109.4431	60.6314	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.453	0.0	0.0	0.0	63.4298	-39.2326	109.3924	60.6314	
...	...	...	...	...	...		...	...	...		...	...	...	...	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.479	0.0	0.0	0.0	63.4561	-40.0267	107.5053	60.5489	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.4811	0.0	0.0	0.0	63.4582	-41.7502	107.5053	60.5378	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.482	0.0	0.0	0.0	63.4591	-39.7236	107.5053	60.5378	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.482	0.0	0.0	0.0	63.4591	-42.6119	107.5053	60.5378	
7.0000	0.0	0.0	0.0000	0.0000	...	0.0	60.482	0.0	0.0	0.0	63.4591	-43.1704	107.5053	60.5378	

	PFRA	PONE	PTEN	PHUN	PTHO	PTNO	RDSW	SCSN	SG_PRES_COEF[0]	SG_TEMP_COEF[0]	S
count	58.000000	58.0	58.0	58.000000	58.000000	58.0	58.000000	58.0	58.0	58.0000	
mean	7.091959	0.0	0.0	-0.000002	-0.000002	0.0	15491.982759	-1.0	0.0	1.0121	
std	0.255205	0.0	0.0	0.000013	0.000013	0.0	0.131306	0.0	0.0	0.0000	
min	6.849800	-0.0	-0.0	-0.000100	-0.000100	0.0	15491.000000	-1.0	0.0	1.0121	
5%	7.000000	0.0	0.0	0.000000	0.000000	0.0	15492.000000	-1.0	0.0	1.0121	
25%	7.000000	0.0	0.0	0.000000	0.000000	0.0	15492.000000	-1.0	0.0	1.0121	
50%	7.000000	0.0	0.0	0.000000	0.000000	0.0	15492.000000	-1.0	0.0	1.0121	
75%	7.000000	0.0	0.0	0.000000	0.000000	0.0	15492.000000	-1.0	0.0	1.0121	
95%	7.816900	0.0	0.0	0.000000	0.000000	0.0	15492.000000	-1.0	0.0	1.0121	
max	8.000100	-0.0	-0.0	0.000000	0.000000	0.0	15492.000000	-1.0	0.0	1.0121	

Figure 4.4: Overview of the log data for a small data set with 58 records

Exploration of data provided a quick understanding of the logs (features) available for use during the classification as well as the statistics (count, mean, standard deviation, minimum, maximum and percentile ranks) of each feature.

From the retained wells, 198 data frames had a depth range that matched another, while 82 wells had a unique depth ranges. The data frames with the repeated depth ranges can be explained as observations split over multiple files. Therefore, these data frames were aggregated to form 161 unique depth ranges (including the 82 already unique depth ranges) (see figure 4.5).



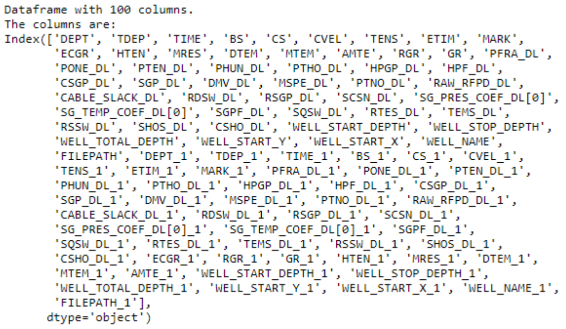
```
The length of all the wells (combined split wells and non-split wells): 161

[     TIME       TIME    TDEP     ETIM     RFPD    PFRA    PONE    PTEN   PHUN  \
0    72.0    72000.0  6363.0   0.0000  16.1033  6.9746  7.9709  0.9964   0.0
1    73.0    73000.0  6363.0   0.7373  16.1620  7.0000  8.0000  1.0000   0.0
2    74.0    74000.0  6363.0   1.7373  16.1620  7.7909  8.0000  1.0000   0.0
3    75.0    75000.0  6363.0   2.7373  16.1620  8.0000  8.0000  1.0000   0.0
4    76.0    76000.0  6363.0   3.7373  16.1620  8.0000  8.0000  1.0000   0.0
..    ...        ...     ...      ...      ...     ...     ...     ...    ...
81  153.0   153000.0  6363.0  80.7373  16.1620  0.0000  0.0000  2.0000   0.0
82  154.0   154000.0  6363.0  81.7373  16.1620  0.0000  0.0000  2.0000   0.0
83  155.0   155000.0  6363.0  82.7373  16.1620  0.0000  0.0000  2.0000   0.0
84  156.0   156000.0  6363.0  83.7373  16.1620  0.0000  0.0000  2.0000   0.0
85  157.0   157000.0  6363.0  84.7373  16.1620  0.0000  0.0000  2.0000   0.0

    PTHO  ...     TEMS  RAT1_SL  ATE1_SL  WELL_START_DEPTH  WELL_STOP_DEPTH  \
0    0.0  ...  78.4874 -53.9739 -53.9739              72.0            157.0
1    0.0  ...  78.7768 -61.4797 -61.4797              72.0            157.0
2    0.0  ...  78.7752 -54.1450 -54.1450              72.0            157.0
3    0.0  ...  78.7795 -54.1342 -54.1342              72.0            157.0
4    0.0     78.7803 -54.1382 -54.1382              72.0            157.0
```

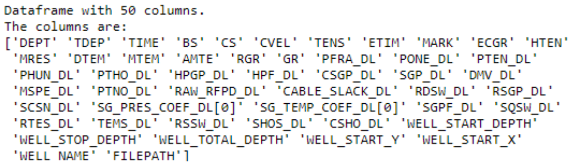
Figure 4.5: Header of the 1st data frame (out of 161 data frames) that has a unique depth range

Grouping the observations that had been split over multiple data frames, into 1 data frame, resulted in duplicate columns. Therefore, for each data frame a unique set of features (columns) was obtained. In some cases this action drastically reduced the dimensionality of the data frame (see figures 4.6 and 4.7).



```
Dataframe with 100 columns.
The columns are:
Index(['DEPT', 'TDEP', 'TIME', 'BS', 'CS', 'CVEL', 'TENS', 'ETIM', 'MARK',
       'ECGR', 'HTEN', 'MRES', 'DTEM', 'MTEM', 'AMTE', 'RGR', 'GR', 'PFRA_DL',
       'PONE_DL', 'PTEN_DL', 'PHUN_DL', 'PTHO_DL', 'HPGP_DL', 'HPF_DL',
       'CSGP_DL', 'SGP_DL', 'DMV_DL', 'MSPE_DL', 'PTNO_DL', 'RAW_RFPD_DL',
       'CABLE_SLACK_DL', 'RDSW_DL', 'RSGP_DL', 'SCSN_DL', 'SG_PRES_COEF_DL[0]',
       'SG_TEMP_COEF_DL[0]', 'SGPF_DL', 'SQSW_DL', 'RTES_DL', 'TEMS_DL',
       'RSSW_DL', 'SHOS_DL', 'CSHO_DL', 'WELL_START_DEPTH', 'WELL_STOP_DEPTH',
       'WELL_TOTAL_DEPTH', 'WELL_START_Y', 'WELL_START_X', 'WELL_NAME',
       'FILEPATH', 'DEPT_1', 'TDEP_1', 'TIME_1', 'BS_1', 'CS_1', 'CVEL_1',
       'TENS_1', 'ETIM_1', 'MARK_1', 'PFRA_DL_1', 'PONE_DL_1', 'PTEN_DL_1',
       'PHUN_DL_1', 'PTHO_DL_1', 'HPGP_DL_1', 'HPF_DL_1', 'CSGP_DL_1',
       'SGP_DL_1', 'DMV_DL_1', 'MSPE_DL_1', 'PTNO_DL_1', 'RAW_RFPD_DL_1',
       'CABLE_SLACK_DL_1', 'RDSW_DL_1', 'RSGP_DL_1', 'SCSN_DL_1',
       'SG_PRES_COEF_DL[0]_1', 'SG_TEMP_COEF_DL[0]_1', 'SGPF_DL_1',
       'SQSW_DL_1', 'RTES_DL_1', 'TEMS_DL_1', 'RSSW_DL_1', 'SHOS_DL_1',
       'CSHO_DL_1', 'ECGR_1', 'RGR_1', 'GR_1', 'HTEN_1', 'MRES_1', 'DTEM_1',
       'MTEM_1', 'AMTE_1', 'WELL_START_DEPTH_1', 'WELL_STOP_DEPTH_1',
       'WELL_TOTAL_DEPTH_1', 'WELL_START_Y_1', 'WELL_START_X_1', 'WELL_NAME_1',
       'FILEPATH_1'],
      dtype='object')
```

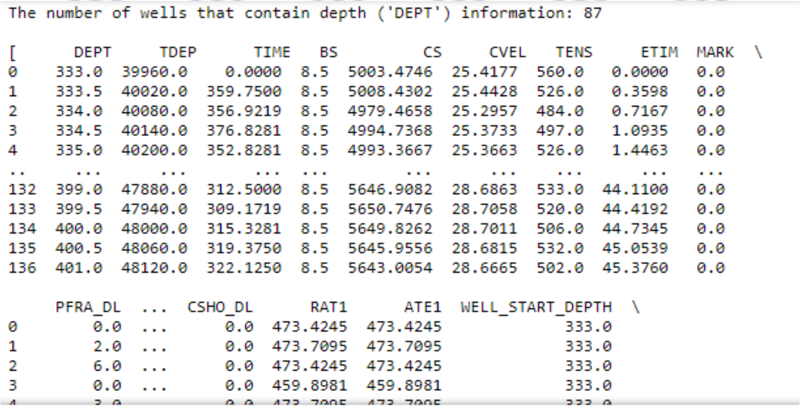
Figure 4.6: Data frame with duplicate columns



```
Dataframe with 50 columns.
The columns are:
['DEPT' 'TDEP' 'TIME' 'BS' 'CS' 'CVEL' 'TENS' 'ETIM' 'MARK' 'ECGR' 'HTEN'
 'MRES' 'DTEM' 'MTEM' 'AMTE' 'RGR' 'GR' 'PFRA_DL' 'PONE_DL' 'PTEN_DL'
 'PHUN_DL' 'PTHO_DL' 'HPGP_DL' 'HPF_DL' 'CSGP_DL' 'SGP_DL' 'DMV_DL'
 'MSPE_DL' 'PTNO_DL' 'RAW_RFPD_DL' 'CABLE_SLACK_DL' 'RDSW_DL' 'RSGP_DL'
 'SCSN_DL' 'SG_PRES_COEF_DL[0]' 'SG_TEMP_COEF_DL[0]' 'SGPF_DL' 'SQSW_DL'
 'RTES_DL' 'TEMS_DL' 'RSSW_DL' 'SHOS_DL' 'CSHO_DL' 'WELL_START_DEPTH'
 'WELL_STOP_DEPTH' 'WELL_TOTAL_DEPTH' 'WELL_START_Y' 'WELL_START_X'
 'WELL NAME' 'FILEPATH']
```

Figure 4.7: Data frame with unique columns

After examining the 161 unique depth range data frames for the presence of DEPT logs, only 87 data frames passed and were kept, while the rest were removed (see figure 4.8).



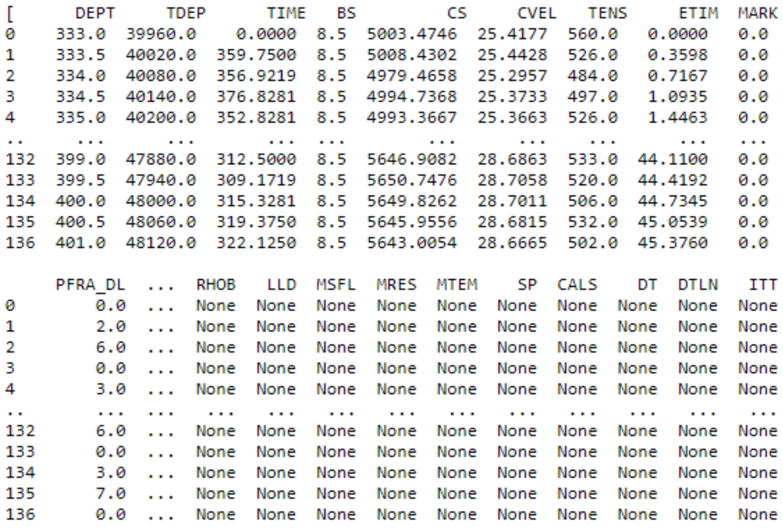
```
The number of wells that contain depth ('DEPT') information: 87

[      DEPT     TDEP      TIME   BS         CS     CVEL   TENS     ETIM  MARK  \
0    333.0  39960.0    0.0000  8.5  5003.4746  25.4177  560.0   0.0000   0.0
1    333.5  40020.0  359.7500  8.5  5008.4302  25.4428  526.0   0.3598   0.0
2    334.0  40080.0  356.9219  8.5  4979.4658  25.2957  484.0   0.7167   0.0
3    334.5  40140.0  376.8281  8.5  4994.7368  25.3733  497.0   1.0935   0.0
4    335.0  40200.0  352.8281  8.5  4993.3667  25.3663  526.0   1.4463   0.0
..     ...      ...       ...  ...        ...      ...    ...      ...   ...
132  399.0  47880.0  312.5000  8.5  5646.9082  28.6863  533.0  44.1100   0.0
133  399.5  47940.0  309.1719  8.5  5650.7476  28.7058  520.0  44.4192   0.0
134  400.0  48000.0  315.3281  8.5  5649.8262  28.7011  506.0  44.7345   0.0
135  400.5  48060.0  319.3750  8.5  5645.9556  28.6815  532.0  45.0539   0.0
136  401.0  48120.0  322.1250  8.5  5643.0054  28.6665  502.0  45.3760   0.0

     PFRA_DL  ...  CSHO_DL      RAT1      ATE1  WELL_START_DEPTH  \
0        0.0  ...      0.0  473.4245  473.4245             333.0
1        2.0  ...      0.0  473.7095  473.7095             333.0
2        6.0  ...      0.0  473.4245  473.4245             333.0
3        0.0  ...      0.0  459.8981  459.8981             333.0
4        3.0           0.0  473.7095  473.7095             333.0
```

Figure 4.8: Header of the 1st data frame (out of 87 data frames) that had a DEPT curve

The data frames with DEPT information, were each assessed for the presence of the logs necessary for clustering (i.e. GR, TNPH, NPHI, RHOB, LLD, etc.). If these logs were not present, they were added as empty values (see figure 4.9). The number of necessary (cluster) logs present in each data frame varied from three to twelve logs. Therefore, at least three empty logs were added for each data frame. This action ensured that the data frames shared the same column dimension and could thus be concatenated into one data frame based on their depths.



```
[      DEPT     TDEP      TIME   BS          CS     CVEL   TENS      ETIM  MARK
0     333.0  39960.0    0.0000  8.5   5003.4746  25.4177  560.0    0.0000   0.0
1     333.5  40020.0  359.7500  8.5   5008.4302  25.4428  526.0    0.3598   0.0
2     334.0  40080.0  356.9219  8.5   4979.4658  25.2957  484.0    0.7167   0.0
3     334.5  40140.0  376.8281  8.5   4994.7368  25.3733  497.0    1.0935   0.0
4     335.0  40200.0  352.8281  8.5   4993.3667  25.3663  526.0    1.4463   0.0
..      ...      ...       ...  ...         ...      ...    ...       ...   ...
132   399.0  47880.0  312.5000  8.5   5646.9082  28.6863  533.0   44.1100   0.0
133   399.5  47940.0  309.1719  8.5   5650.7476  28.7058  520.0   44.4192   0.0
134   400.0  48000.0  315.3281  8.5   5649.8262  28.7011  506.0   44.7345   0.0
135   400.5  48060.0  319.3750  8.5   5645.9556  28.6815  532.0   45.0539   0.0
136   401.0  48120.0  322.1250  8.5   5643.0054  28.6665  502.0   45.3760   0.0

      PFRA_DL  ...  RHOB   LLD  MSFL  MRES  MTEM    SP  CALS    DT  DTLN   ITT
0         0.0  ...  None  None  None  None  None  None  None  None  None  None
1         2.0  ...  None  None  None  None  None  None  None  None  None  None
2         6.0  ...  None  None  None  None  None  None  None  None  None  None
3         0.0  ...  None  None  None  None  None  None  None  None  None  None
4         3.0  ...  None  None  None  None  None  None  None  None  None  None
..        ...  ...   ...   ...   ...   ...   ...   ...   ...   ...   ...   ...
132       6.0  ...  None  None  None  None  None  None  None  None  None  None
133       0.0  ...  None  None  None  None  None  None  None  None  None  None
134       3.0  ...  None  None  None  None  None  None  None  None  None  None
135       7.0  ...  None  None  None  None  None  None  None  None  None  None
136       0.0  ...  None  None  None  None  None  None  None  None  None  None
```

Figure 4.9: Header and footer of the 1st data frame (out of 87 data frames) that has had empty logs added to it

As stated above, dimension and feature matching (during the concatenation process) resulted in empty (Null) values in the combined data frame (see figure 4.11). Therefore only statistics on columns that were completely populated could be obtained (see figure 4.10)

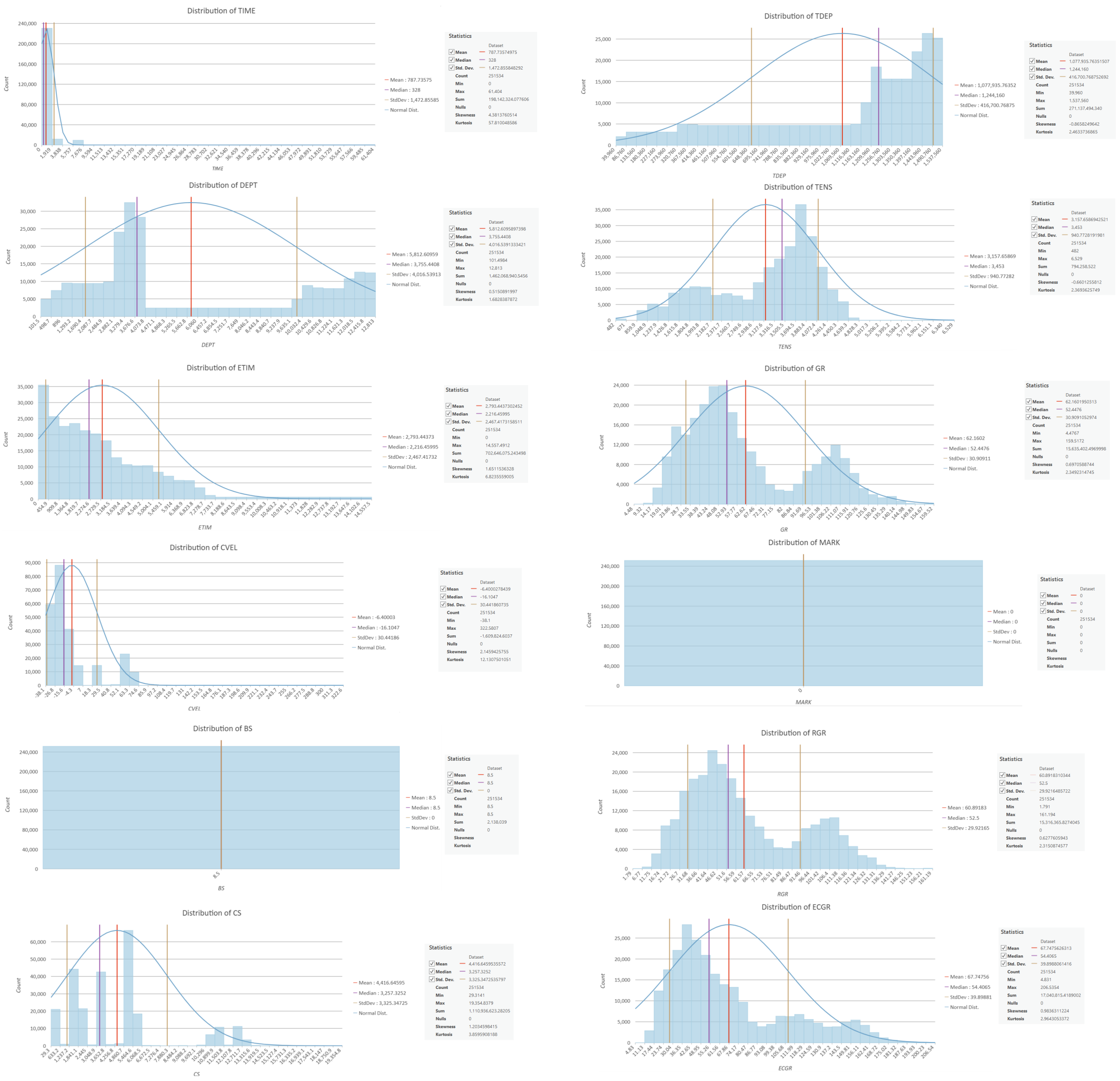
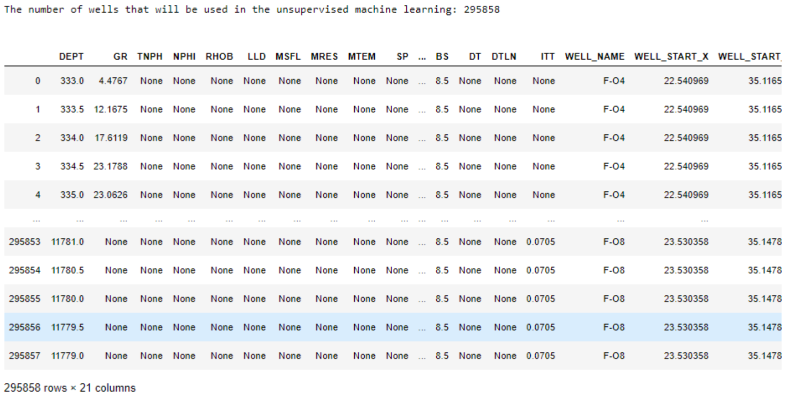


Figure 4.10: Statistics of the non null features

Followinng this, the dataframes were filtered to only reflect the logs necessary for clustering.



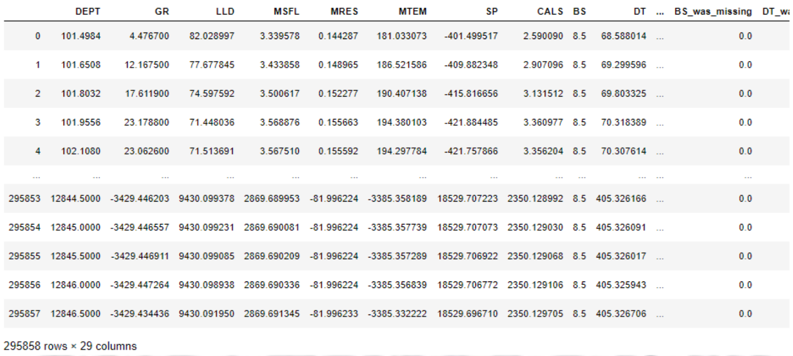
```
The number of wells that will be used in the unsupervised machine learning: 295858
```

	DEPT	GR	TNPH	NPHI	RHOB	LLD	MSFL	MRES	MTEM	SP	...	BS	DT	DTLN	ITT	WELL_NAME	WELL_START_X	WELL_START
0	333.0	4.4767	None	None	None	None	None	None	None	None	...	8.5	None	None	None	F-O4	22.540969	35.1165
1	333.5	12.1675	None	None	None	None	None	None	None	None	...	8.5	None	None	None	F-O4	22.540969	35.1165
2	334.0	17.6119	None	None	None	None	None	None	None	None	...	8.5	None	None	None	F-O4	22.540969	35.1165
3	334.5	23.1788	None	None	None	None	None	None	None	None	...	8.5	None	None	None	F-O4	22.540969	35.1165
4	335.0	23.0626	None	None	None	None	None	None	None	None	...	8.5	None	None	None	F-O4	22.540969	35.1165
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
295853	11781.0	None	None	None	None	None	None	None	None	None	...	8.5	None	None	0.0705	F-O8	23.530358	35.1478
295854	11780.5	None	None	None	None	None	None	None	None	None	...	8.5	None	None	0.0705	F-O8	23.530358	35.1478
295855	11780.0	None	None	None	None	None	None	None	None	None	...	8.5	None	None	0.0705	F-O8	23.530358	35.1478
295856	11779.5	None	None	None	None	None	None	None	None	None	...	8.5	None	None	0.0705	F-O8	23.530358	35.1478
295857	11779.0	None	None	None	None	None	None	None	None	None	...	8.5	None	None	0.0705	F-O8	23.530358	35.1478

295858 rows × 21 columns

Figure 4.11: Header and footer of the concatenated data frame

The missing values were addressed by running an imputation method (see figure 4.12) before the data set was assessed for outliers (see figure 4.13). All the logs (except the DEPT log) had values that fell outside of the range specified by the IQR rule. Therefore these values were removed from the data set to ensure that the clustering process was not biased by points unrepresentative of the data distribution (see figures 4.13 and 4.14).

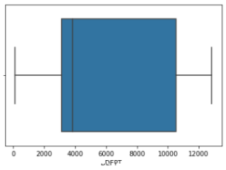


	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	DT_wa
0	101.4984	4.476700	82.028997	3.339578	0.144287	181.033073	-401.499517	2.590090	8.5	68.588014	...	0.0	
1	101.6508	12.167500	77.677845	3.433858	0.148965	186.521586	-409.882348	2.907096	8.5	69.299596	...	0.0	
2	101.8032	17.611900	74.597592	3.500617	0.152277	190.407138	-415.816656	3.131512	8.5	69.803325	...	0.0	
3	101.9556	23.178800	71.448036	3.568876	0.155663	194.380103	-421.884485	3.360977	8.5	70.318389	...	0.0	
4	102.1080	23.062600	71.513691	3.567510	0.155592	194.297784	-421.757866	3.356204	8.5	70.307614	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.5	405.326166	...	0.0	
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.5	405.326091	...	0.0	
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.5	405.326017	...	0.0	
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.5	405.325943	...	0.0	
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.5	405.326706	...	0.0	

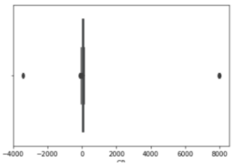
295858 rows × 29 columns

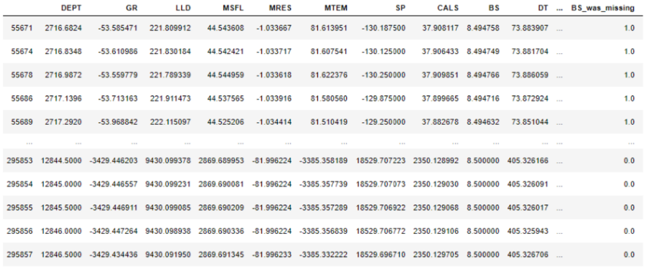
Figure 4.12: Data frame with imputed values

DEPT



GR

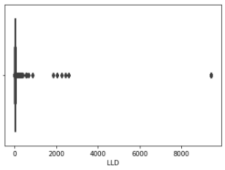


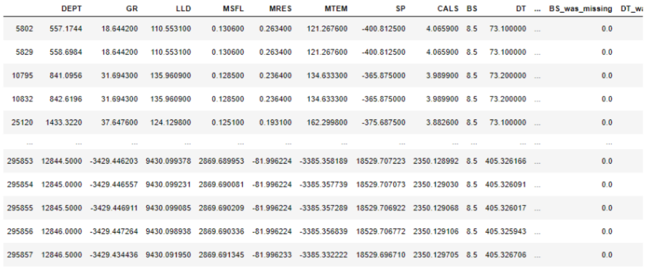


	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing
55671	2716.6824	-53.585471	221.809912	44.543608	-1.033667	81.613951	-130.187500	37.908117	8.494758	73.883907	...	1.0
55674	2716.8348	-53.610986	221.830184	44.542421	-1.033717	81.607541	-130.125000	37.906433	8.494749	73.881704	...	1.0
55678	2716.9872	-53.559779	221.789339	44.544959	-1.033618	81.622376	-130.250000	37.909851	8.494766	73.886059	...	1.0
55686	2717.1396	-53.713163	221.911473	44.537565	-1.033916	81.580560	-129.875000	37.899665	8.494716	73.872924	...	1.0
55689	2717.2920	-53.968842	222.115097	44.525206	-1.034414	81.510419	-129.250000	37.882678	8.494632	73.851044	...	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.500000	405.326166	...	0.0
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.500000	405.326091	...	0.0
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.500000	405.326017	...	0.0
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.500000	405.325943	...	0.0
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.500000	405.326706	...	0.0

26374 rows × 29 columns

LLD

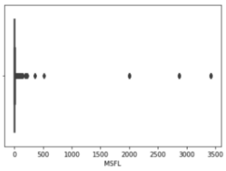


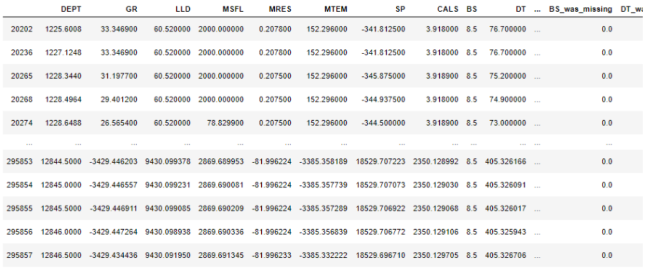


	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	DT_wa
5802	557.1744	18.644200	110.553100	0.130600	0.263400	121.267600	-400.812500	4.065900	8.5	73.100000	...	0.0	
5829	558.6984	18.644200	110.553100	0.130600	0.263400	121.267600	-400.812500	4.065900	8.5	73.100000	...	0.0	
10795	841.0956	31.694300	135.960900	0.128500	0.236400	134.633300	-365.875000	3.989900	8.5	73.200000	...	0.0	
10832	842.6196	31.694300	135.960900	0.128500	0.236400	134.633300	-365.875000	3.989900	8.5	73.200000	...	0.0	
25120	1433.3220	37.647600	124.129800	0.125100	0.193100	162.299800	-375.687500	3.882600	8.5	73.100000	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.5	405.326166	...	0.0	
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.5	405.326091	...	0.0	
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.5	405.326017	...	0.0	
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.5	405.325943	...	0.0	
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.5	405.326706	...	0.0	

44625 rows × 29 columns

MSFL

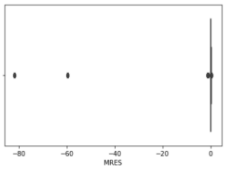


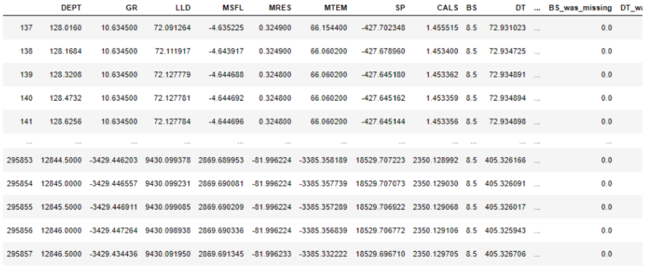


	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	DT_wa
20202	1225.6008	33.346900	60.520000	2000.000000	0.207800	152.296000	-341.812500	3.918000	8.5	76.700000	...	0.0	
20236	1227.1248	33.346900	60.520000	2000.000000	0.207800	152.296000	-341.812500	3.918000	8.5	76.700000	...	0.0	
20265	1228.3440	31.197700	60.520000	2000.000000	0.207500	152.296000	-345.875000	3.918900	8.5	75.200000	...	0.0	
20268	1228.4964	29.401200	60.520000	2000.000000	0.207500	152.296000	-344.937500	3.918000	8.5	74.900000	...	0.0	
20274	1228.6488	26.565400	60.520000	78.829900	0.207500	152.296000	-344.500000	3.918900	8.5	73.000000	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.5	405.326166	...	0.0	
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.5	405.326091	...	0.0	
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.5	405.326017	...	0.0	
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.5	405.325943	...	0.0	
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.5	405.326706	...	0.0	

47012 rows × 29 columns

MRES

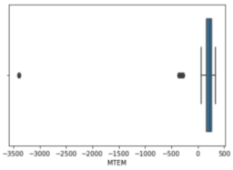


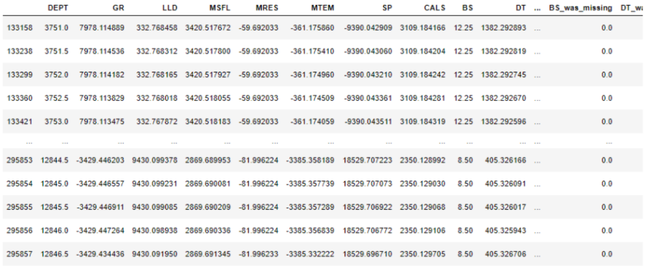


	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	DT_wa
137	128.0160	10.634500	72.091264	-4.635225	0.324900	66.154400	-427.702348	1.455515	8.5	72.931023	...	0.0	
138	128.1684	10.634500	72.111917	-4.643917	0.324900	66.060200	-427.678960	1.453400	8.5	72.934725	...	0.0	
139	128.3208	10.634500	72.127779	-4.644688	0.324800	66.060200	-427.645180	1.453362	8.5	72.934891	...	0.0	
140	128.4732	10.634500	72.127781	-4.644692	0.324800	66.060200	-427.645162	1.453359	8.5	72.934894	...	0.0	
141	128.6256	10.634500	72.127784	-4.644696	0.324800	66.060200	-427.645144	1.453356	8.5	72.934898	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.5	405.326166	...	0.0	
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.5	405.326091	...	0.0	
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.5	405.326017	...	0.0	
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.5	405.325943	...	0.0	
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.5	405.326706	...	0.0	

57618 rows × 29 columns

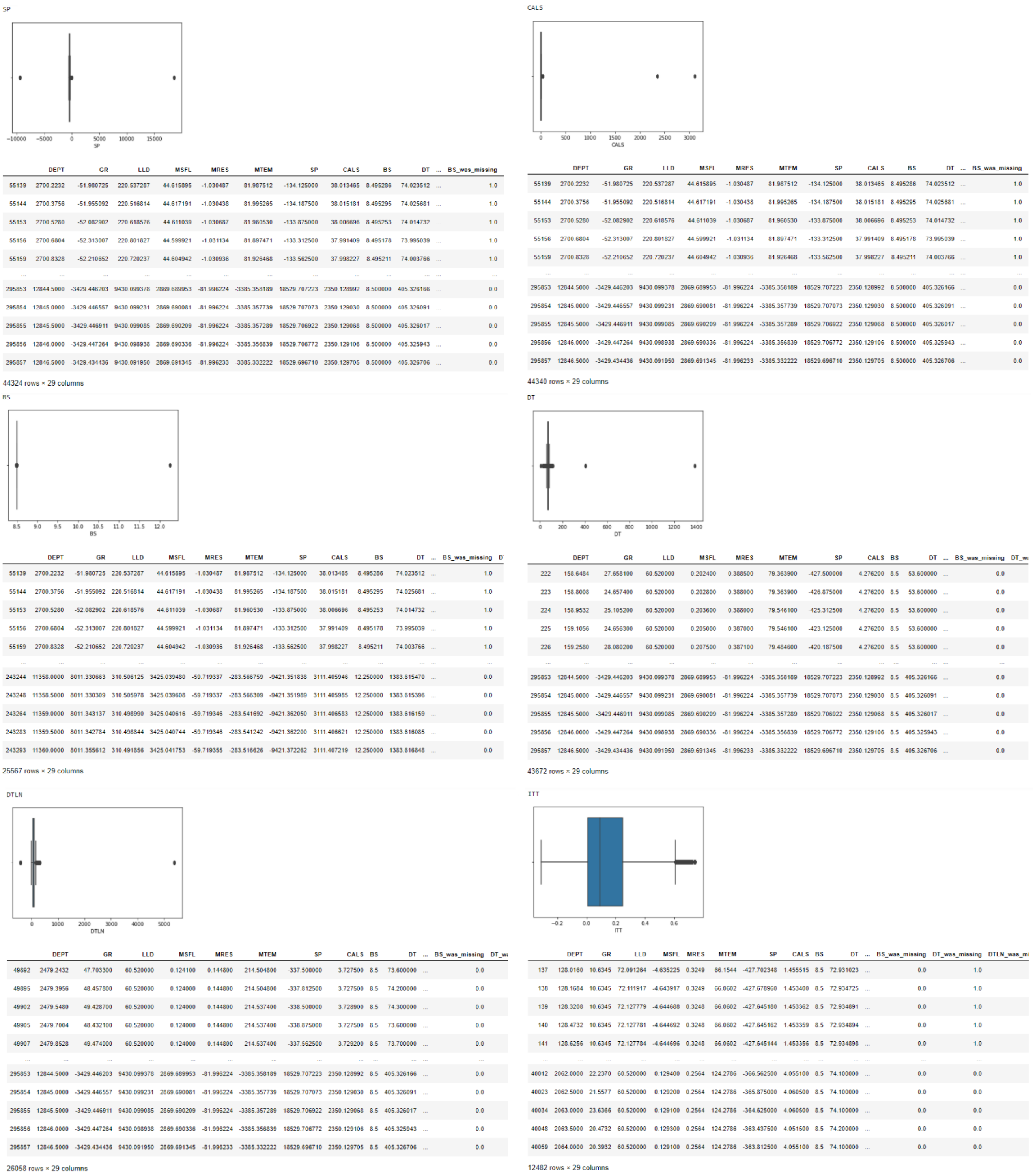
MTEM





	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	DT_wa
133158	3751.0	7978.114889	332.768458	3420.517672	-59.692033	-361.175860	-9390.042909	3109.184166	12.25	1382.292893	...	0.0	
133238	3751.5	7978.114536	332.768312	3420.517800	-59.692033	-361.175410	-9390.043060	3109.184204	12.25	1382.292819	...	0.0	
133299	3752.0	7978.114182	332.768165	3420.517927	-59.692033	-361.174960	-9390.043210	3109.184242	12.25	1382.292745	...	0.0	
133360	3752.5	7978.113829	332.768018	3420.518055	-59.692033	-361.174509	-9390.043361	3109.184281	12.25	1382.292670	...	0.0	
133421	3753.0	7978.113475	332.767872	3420.518183	-59.692033	-361.174059	-9390.043511	3109.184319	12.25	1382.292596	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
295853	12844.5	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.50	405.326166	...	0.0	
295854	12845.0	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.50	405.326091	...	0.0	
295855	12845.5	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.50	405.326017	...	0.0	
295856	12846.0	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.50	405.325943	...	0.0	
295857	12846.5	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.50	405.326706	...	0.0	

18012 rows × 29 columns



	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing
55139	2700.2232	-51.980725	220.537287	44.615895	-1.030487	81.987512	-134.125000	38.013465	8.495286	74.023512	...	1.0
55144	2700.3756	-51.955092	220.516814	44.617191	-1.030438	81.995265	-134.187500	38.015181	8.495295	74.025681	...	1.0
55153	2700.5280	-52.082902	220.618576	44.611039	-1.030687	81.960530	-133.875000	38.006696	8.495253	74.014732	...	1.0
55156	2700.6804	-52.313007	220.801827	44.599921	-1.031134	81.897471	-133.312500	37.991409	8.495178	73.995039	...	1.0
55159	2700.8328	-52.210652	220.720237	44.604942	-1.030936	81.926468	-133.562500	37.998227	8.495211	74.003766	...	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.500000	405.326166	...	0.0
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.500000	405.326091	...	0.0
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.500000	405.326017	...	0.0
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.500000	405.325943	...	0.0
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.500000	405.326706	...	0.0

44324 rows × 29 columns

	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing
55139	2700.2232	-51.980725	220.537287	44.615895	-1.030487	81.987512	-134.125000	38.013465	8.495286	74.023512	...	1.0
55144	2700.3756	-51.955092	220.516814	44.617191	-1.030438	81.995265	-134.187500	38.015181	8.495295	74.025681	...	1.0
55153	2700.5280	-52.082902	220.618576	44.611039	-1.030687	81.960530	-133.875000	38.006696	8.495253	74.014732	...	1.0
55156	2700.6804	-52.313007	220.801827	44.599921	-1.031134	81.897471	-133.312500	37.991409	8.495178	73.995039	...	1.0
55159	2700.8328	-52.210652	220.720237	44.604942	-1.030936	81.926468	-133.562500	37.998227	8.495211	74.003766	...	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.500000	405.326166	...	0.0
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.500000	405.326091	...	0.0
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.500000	405.326017	...	0.0
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.500000	405.325943	...	0.0
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.500000	405.326706	...	0.0

44340 rows × 29 columns

	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	D
55139	2700.2232	-51.980725	220.537287	44.615895	-1.030487	81.987512	-134.125000	38.013465	8.495286	74.023512	...	1.0	
55144	2700.3756	-51.955092	220.516814	44.617191	-1.030438	81.995265	-134.187500	38.015181	8.495295	74.025681	...	1.0	
55153	2700.5280	-52.082902	220.618576	44.611039	-1.030687	81.960530	-133.875000	38.006696	8.495253	74.014732	...	1.0	
55156	2700.6804	-52.313007	220.801827	44.599921	-1.031134	81.897471	-133.312500	37.991409	8.495178	73.995039	...	1.0	
55159	2700.8328	-52.210652	220.720237	44.604942	-1.030936	81.926468	-133.562500	37.998227	8.495211	74.003766	...	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
243244	11358.0000	8011.330663	310.506125	3425.039480	-59.719337	-283.566759	-9421.351838	3111.405946	12.250000	1383.615470	...	0.0	
243248	11358.5000	8011.330309	310.505978	3425.039608	-59.719337	-283.566309	-9421.351989	3111.405985	12.250000	1383.615396	...	0.0	
243264	11359.0000	8011.343137	310.498990	3425.040616	-59.719346	-283.541692	-9421.362050	3111.406583	12.250000	1383.616159	...	0.0	
243283	11359.5000	8011.342784	310.498844	3425.040744	-59.719346	-283.541242	-9421.362200	3111.406621	12.250000	1383.616085	...	0.0	
243293	11360.0000	8011.355612	310.491856	3425.041753	-59.719355	-283.516626	-9421.372262	3111.407219	12.250000	1383.616848	...	0.0	

25567 rows × 29 columns

	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	DT_wa
222	158.6484	27.658100	60.520000	0.202400	0.388500	79.363900	-427.500000	4.276200	8.5	53.600000	...	0.0	
223	158.8008	24.657400	60.520000	0.202800	0.388000	79.363900	-426.875000	4.276200	8.5	53.600000	...	0.0	
224	158.9532	25.105200	60.520000	0.203600	0.388000	79.546100	-425.312500	4.276200	8.5	53.600000	...	0.0	
225	159.1056	24.656300	60.520000	0.205000	0.387000	79.546100	-423.125000	4.276200	8.5	53.600000	...	0.0	
226	159.2580	28.080200	60.520000	0.207500	0.387100	79.484600	-420.187500	4.276200	8.5	53.600000	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.5	405.326166	...	0.0	
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.5	405.326091	...	0.0	
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.5	405.326017	...	0.0	
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.5	405.325943	...	0.0	
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.5	405.326706	...	0.0	

43672 rows × 29 columns

	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	DT_wa
49892	2479.2432	47.703300	60.520000	0.124100	0.144800	214.504800	-337.500000	3.727500	8.5	73.600000	...	0.0	
49895	2479.3956	48.457800	60.520000	0.124000	0.144800	214.504800	-337.812500	3.727500	8.5	74.200000	...	0.0	
49902	2479.5480	49.428700	60.520000	0.124000	0.144800	214.537400	-338.500000	3.728900	8.5	74.300000	...	0.0	
49905	2479.7004	48.432100	60.520000	0.124000	0.144800	214.537400	-338.875000	3.727500	8.5	73.600000	...	0.0	
49907	2479.8528	49.474000	60.520000	0.124000	0.144800	214.537400	-337.562500	3.729200	8.5	73.700000	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
295853	12844.5000	-3429.446203	9430.099378	2869.689953	-81.996224	-3385.358189	18529.707223	2350.128992	8.5	405.326166	...	0.0	
295854	12845.0000	-3429.446557	9430.099231	2869.690081	-81.996224	-3385.357739	18529.707073	2350.129030	8.5	405.326091	...	0.0	
295855	12845.5000	-3429.446911	9430.099085	2869.690209	-81.996224	-3385.357289	18529.706922	2350.129068	8.5	405.326017	...	0.0	
295856	12846.0000	-3429.447264	9430.098938	2869.690336	-81.996224	-3385.356839	18529.706772	2350.129106	8.5	405.325943	...	0.0	
295857	12846.5000	-3429.434436	9430.091950	2869.691345	-81.996233	-3385.332222	18529.696710	2350.129705	8.5	405.326706	...	0.0	

26058 rows × 29 columns

	DEPT	GR	LLD	MSFL	MRES	MTEM	SP	CALS	BS	DT	...	BS_was_missing	DT_was_missing	DTLN_was_mi
137	128.0160	10.6345	72.091264	-4.635225	0.3249	66.1544	-427.702348	1.455515	8.5	72.931023	...	0.0	1.0	
138	128.1684	10.6345	72.111917	-4.643917	0.3249	66.0602	-427.678960	1.453400	8.5	72.934725	...	0.0	1.0	
139	128.3208	10.6345	72.127779	-4.644688	0.3248	66.0602	-427.645180	1.453362	8.5	72.934891	...	0.0	1.0	
140	128.4732	10.6345	72.127781	-4.644692	0.3248	66.0602	-427.645162	1.453359	8.5	72.934894	...	0.0	1.0	
141	128.6256	10.6345	72.127784	-4.644696	0.3248	66.0602	-427.645144	1.453356	8.5	72.934898	...	0.0	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
40012	2062.0000	22.2370	60.520000	0.129400	0.2564	124.2786	-366.562500	4.055100	8.5	74.100000	...	0.0	0.0	
40023	2062.5000	21.5577	60.520000	0.129200	0.2564	124.2786	-365.875000	4.060500	8.5	74.100000	...	0.0	0.0	
40034	2063.0000	23.6366	60.520000	0.129100	0.2564	124.2786	-364.625000	4.060500	8.5	74.100000	...	0.0	0.0	
40048	2063.5000	20.4732	60.520000	0.129300	0.2564	124.2786	-363.437500	4.051500	8.5	74.200000	...	0.0	0.0	
40059	2064.0000	20.3932	60.520000	0.129100	0.2564	124.2786	-363.812500	4.055100	8.5	74.100000	...	0.0	0.0	

12482 rows × 29 columns

Figure 4.13: Box plots depicting the distribution of the data, with outlier records summarised below the plot

Following the outlier removal process, the distribution of the data was once again visualised to get a sense of the data's new spread (see figures 4.14).

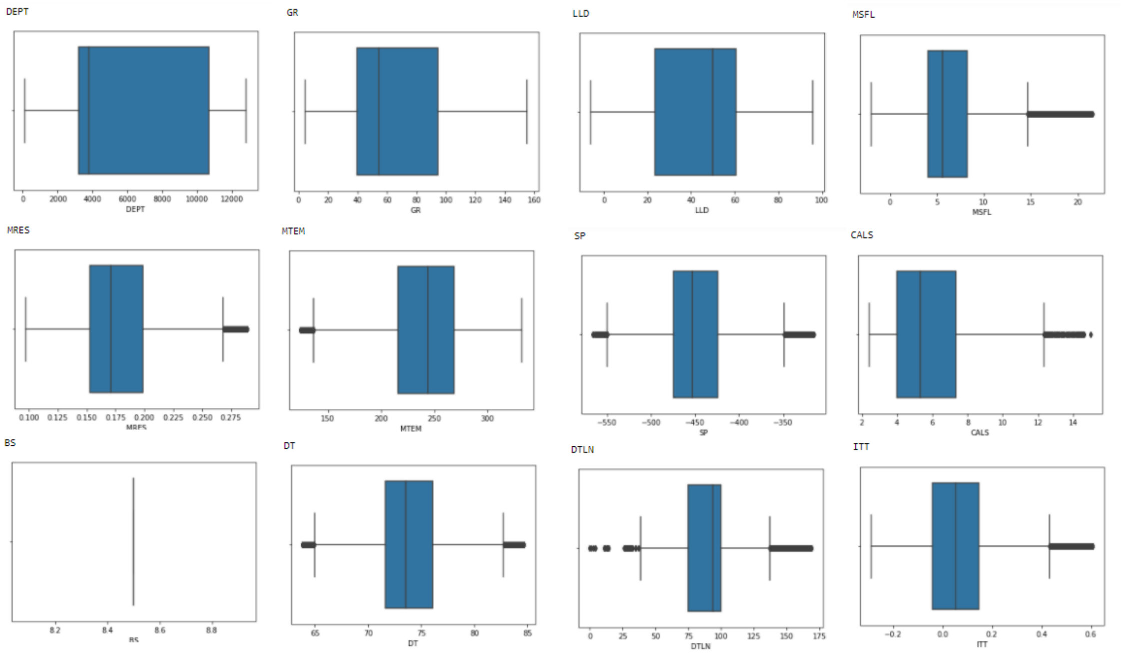


Figure 4.14: Box plot depicting the distribution of the data, with the outliers removed

4.4 PCA Optimal Parameter Derivation

One of the key objectives in this research was deriving the best parameters possible for the dimensionality reduction algorithm. This was of importance because this value had a significant impact on the results of the clustering; and hence prototype creation.

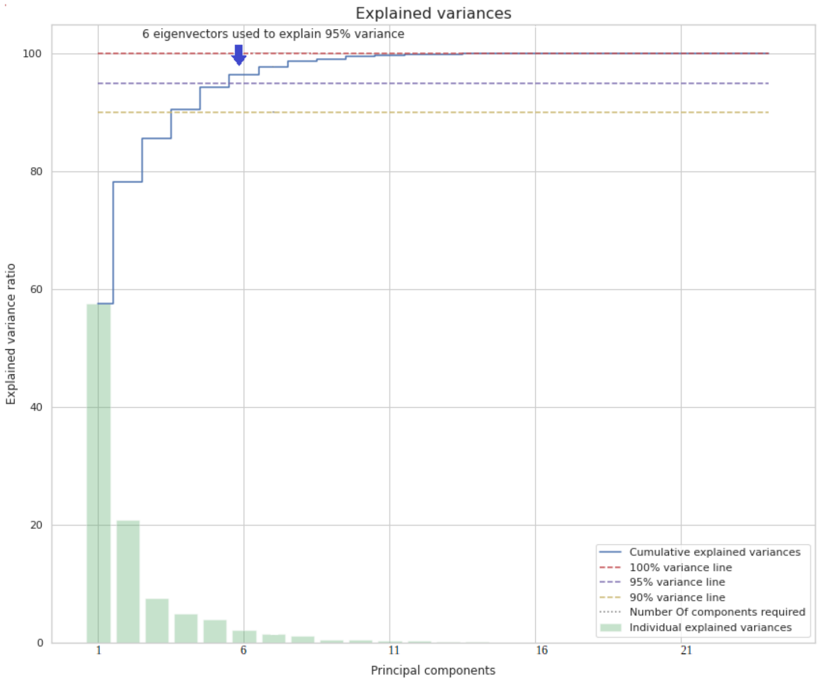


Figure 4.15: Principal Component Analysis (PCA) number of components

Figure 4.15 depicts the optimal number of components for the dimensionality reduction i.e. 6 components. This number was obtained from an iteratively run PCA process and indicates that 6 components explain 95% of the data distribution. Which means that the first 6 components describe the greatest variances within the data and can be used to reconstruct a majority of it (thus making the remaining components redundant). Therefore the optimal stretch and rotation from the 13 dimensional well log data set to a 6 dimensional PCA data space has been found.

Here it is also important to note that the exercise of reducing the dimensionality of the data (to that of only its necessary components), also automatically filtered out any random noise that might have been embedded within the data. Also, standardising the data points and removing and outliers was critical in determing the optimal number of components as the dimensionality reduction algorithm is sensitive to these qualities.

Looking at 4.15, the explained varience ratio in the first principal component is at a 58% variance because it is a linear combination of all the features such that it accounts for as much of the variance in the data as possible. Similar to the first component, the second principal component is a linear combination of features such that as much of the remaining variation as possible is accounted for. Thus bringing the total variation at the end of the second component to a 79% variance. The remaining four principal components adhere to this same property, that is they are linear combinations that account for as much of the remaining variation as possible. Therefore principal components 3, 4, 5 and 6 with variances of 6%, 6%, 3% and 2% respectively, bring the total variance in the data to 96% - which is just over the set PCA threshold of 95%.

4.5 Principal Component Interpretation

The PCA process ensured that only the most descriptive and relevant portions of the data was used to identify clusters, in addition to facilitating with faster visualisation because of the reduced feature space.

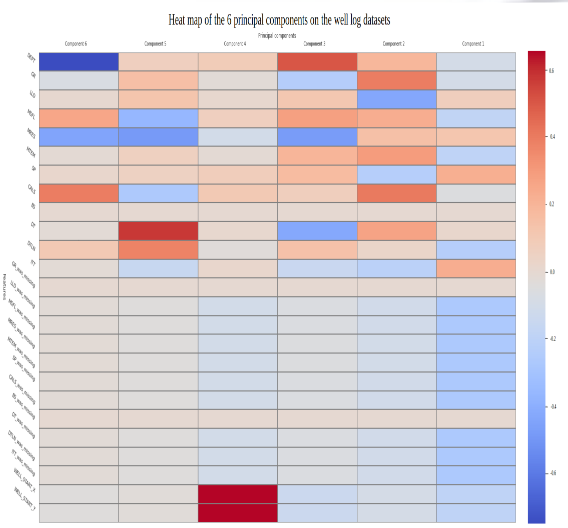
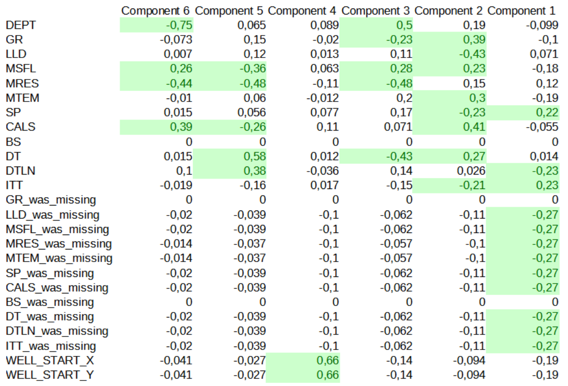


Figure 4.16: Heatmap showing the correlations between the principal components and the original variables.



	Component 6	Component 5	Component 4	Component 3	Component 2	Component 1
DEPT	-0,75	0,065	0,089	0,5	0,19	-0,099
GR	-0,073	0,15	-0,02	-0,23	0,39	-0,1
LLD	0,007	0,12	0,013	0,11	-0,43	0,071
MSFL	0,26	-0,36	0,063	0,28	0,23	-0,18
MRES	-0,44	-0,48	-0,11	-0,48	0,15	0,12
MTEM	-0,01	0,06	-0,012	0,2	0,3	-0,19
SP	0,015	0,056	0,077	0,17	-0,23	0,22
CALS	0,39	-0,26	0,11	0,071	0,41	-0,055
BS	0	0	0	0	0	0
DT	0,015	0,58	0,012	-0,43	0,27	0,014
DTLN	0,1	0,38	-0,036	0,14	0,026	-0,23
ITT	-0,019	-0,16	0,017	-0,15	-0,21	0,23
GR_was_missing	0	0	0	0	0	0
LLD_was_missing	-0,02	-0,039	-0,1	-0,062	-0,11	-0,27
MSFL_was_missing	-0,02	-0,039	-0,1	-0,062	-0,11	-0,27
MRES_was_missing	-0,014	-0,037	-0,1	-0,057	-0,1	-0,27
MTEM_was_missing	-0,014	-0,037	-0,1	-0,057	-0,1	-0,27
SP_was_missing	-0,02	-0,039	-0,1	-0,062	-0,11	-0,27
CALS_was_missing	-0,02	-0,039	-0,1	-0,062	-0,11	-0,27
BS_was_missing	0	0	0	0	0	0
DT_was_missing	-0,02	-0,039	-0,1	-0,062	-0,11	-0,27
DTLN_was_missing	-0,02	-0,039	-0,1	-0,062	-0,11	-0,27
ITT_was_missing	-0,02	-0,039	-0,1	-0,062	-0,11	-0,27
WELL_START_X	-0,041	-0,027	0,66	-0,14	-0,094	-0,19
WELL_START_Y	-0,041	-0,027	0,66	-0,14	-0,094	-0,19

Figure 4.17: Table showing the correlations between the principal components and the original variables with the significant correlations highlighted in green.

To understand the features (well logs) described by each component, the principal components were plotted against each of the origional features (see figure 4.16) . This action allowed for features that were strongly correlated (either positively or negatively) with each component to be seen and extracted. Before significant features for each component could be extracted a correlation cutoff magnitude had to be set. The correlation cutoff magnitude for this operation was set at values above or below 0.2. This value was chosen by considering the range of the data and then selecting a correlation value that allowed every feature to be described by a component at least once.

Before delving further into the component interpretation, an important observation can be made: The only features that weren't described in any of the principal components were BS and BS_was_missing. This is because BS and BS_was_missing returned a correlation value of 0 for all the components. Therefore for this dataset, BS and BS_was_missing (and by extension borehole shape and size) did not have a relationship with any of the components.

First Principal Component

The first principal component is correlated with twelve of the 25 features used in the well log interpolation process. The twelve features are: SP, DTLN, ITT, LLD_was_missing, MSFL_was_missing, MRES_was_missing, MTEM_was_missing, SP_was_missing, CALS_was_missing, DT_was_missing, DTLN_was_missing and ITT_was_missing. The first principal component is thus a measure of well log data completeness, formation transit time and shale presence.

The component is an inverse measure of data completeness because it describes the most significant values for all the non-zero imputed features (i.e. LLD_was_missing, MSFL_was_missing, MRES_was_missing, MTEM_was_missing, SP_was_missing, CALS_was_missing, DT_was_missing, DTLN_was_missing and ITT_was_missing). The correlation magnitude for all these features is -0.27, which means that a decrease in these features results in an increase in the value of the first componenet. Additionally, as these features vary together by the same amount, a change in one feature will cause the other features in this set to change by an equivalent value.

The first component is also a formation transit measure because it gives record of two of the three sonic log (DTLN and ITT) used in the investigation. Although, these logs have an equal correlation magnitude they act in different directions, with DTLN having a correlation value of -0.23 and ITT having a correlation value of 0.23. Therefore, a decrease in DTLN will result in an increase in the first component value while the opposite is true for the ITT log.

SP accounts for the shale presence measure in component 1. With a value of 0.22 it increases as the component increases. Along with ITT, SP varies positively while the other features in the component vary in the opposite direction.

Second Principal Component

In the second principal component, five of its features (GR, MSFL, MTEM, CALS and DT) vary together positively, with correlation values of 0.39, 0.23, 0.3, 0.41 and 0.27 respectively. While three features (LLD, SP and ITT) vary together in the opposite direction. These negatively correlated features have correlation values of -0.43, -0.23 and -0.21 respectively.

Most interestingly this component fully represents two feature above the set threshold. These features are LLD (a resistivity measure) and MTEM (a measure of mud temperature).

Third Principal Component

This component can be viewed as a measure of the quality of formation gamma radiation, resistivity and transit time, as well as borehole depth. This is because DEPT, GR, MSFL, MRES and DT have the highest above threshold values for the component. The magnitude of each of these features is 0.5, -0.23, 0.28, -0.48 and -0.43 respectively.

GR, DT and MRES vary together, decreasing as the componenet increases, while an increase in DEPT and MSFL causes the component to increase. The two resistivity measures (MSFL and MRES) are inversely related to the componet. As such, the third component increases as MSFL increases, but decreases as MRES increases.

Fourth Principal Component

The fourth principal component has a strong positive correlation with two of the origional features i.e. the xy location of the well (i.e. WELL_START_X AND WELL_START_Y). Therefore, the fourth principal component increases as they increase and can be viewed as a locational measure.

For the investigation dataset, these two features vary together and their equal correlation magnitudes (0.66 for both WELL_START_X AND WELL_START_Y) suggests that a change in either of the features would produce an identically proportional change in the principal component.

Fifth Principal Component

This component is described by five features: MSFL, MRES, CALS, DT and DTLN. Although MSFL, DT and DTLN are entities in other components, the highest correlation values for these features are seen in this component (regardless of the direction). The opposite, however, is true for CALS, where the component indicates the smallest above threshold value for the feature.

In the case of MRES, the feature is equally represented in both this and the third component, with a value of -048. This correlation value is indicative of the negative relationship between a formation's resistivity and the component's value. So, as MRES decreases the value of the component will increase. This pattern, of inverse proportionality is also displayed by MSFL (another resistivity measure) with a value of -0.36, and CALS (a borehole geometry measure) with a value of -0.26.

The only features that increases as the fifth principal component increases are the time measures, with DT and DTLN having correlation values of 0.58 and 0.38 respectively.

Seeing that all the features in this component have already been accounted for by other components, their addition adds redundancy to the results.

Sixth Principal Component

The sixth principal component is strongly correlated with four of the original well log features. However, the features that make up this component vary in opposite directions. That is for descreasing DEPT and MRES values the component increases. While, for MSFL and CALS, their positive component correlation means that the value of the component increases as they increase.

This component can be viewed as a measure of a formations resistivity as well as a boreholes geometry and depth.

Furthermore, we see that the sixth principal component correlates most strongly with the DEPT. In fact, it could be said that based on a correlation value of -0.75 that this principal component is primarily a measure of DEPT.

4.6 Derivation and Interpretation of the Optimal KMeans Parameters

Unlike in supervised learning, unsupervised machine learning algorithms do not have have a 'teacher' to learn from. And without domain knowledge, specifying the number of clusters to partition the data into can be problematic. To overcome this downstream modelling, where the response of the KMeans model to a given number of clusters, was employed. In this approach, the effect of ’k’ clusters on the performance of a model was assessed by specifying and testing the KMeans model on a range of ’k’ clusters.

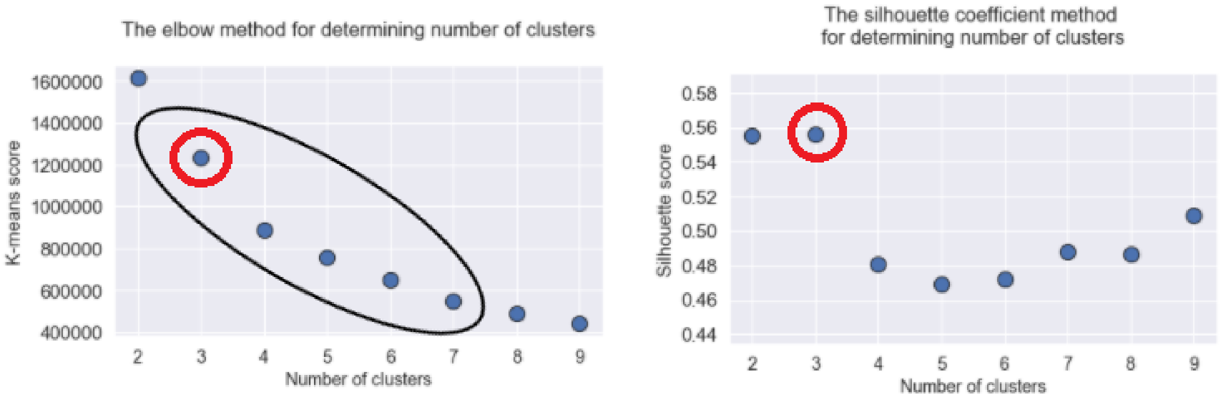


Figure 4.18: KMeans number omf clusters

Two downstream evaluation metrics were used to assess the performance of ’k’ clusters on the KMeans model. The first metric used was the elbow method, as seen in plot 1 of figure 4.18. Using the sum squared distance (SSE) between the data points and their assigned cluster centres, the elbow method indicated that the ideal number of clusters could have been anywhere between 3 and 7 clusters.

To clarify this uncertainty, the second metric silhouette analysis was used. This metric was calculated by getting the coefficient between the mean intra-cluster distance and the mean nearest cluster distance. As the value for the silhouette coefficient ranges from 1 to -1, 1 essentially indicates correct cluster assignment (a great distance between clusters), -1 an incorrect cluster assignment (an incredibly small distance between clusters) and 0 a debatable cluster assignment (a small distance between clusters). The values returned from this metric fell between 0.4 and 0.6, with the values peaking at around 0.56 before falling again (as seen in the second plot of figure 4.18).

These values indicated a decent cluster assignment for all of the cluster numbers tested (2-9), but that the best cluster separation would be achieved with 3 clusters with a silhouette coefficient of about 0.56. Figure 4.19 is a visual depiction of the silhouette metric and for 3 clusters shows that the about 60-70 percent of the data is clustered into the second cluster (labelled 1), while the other 30-40 % is split between clusters 1 (labelled 0) and 3 (labelled 2) (see figure 4.20).

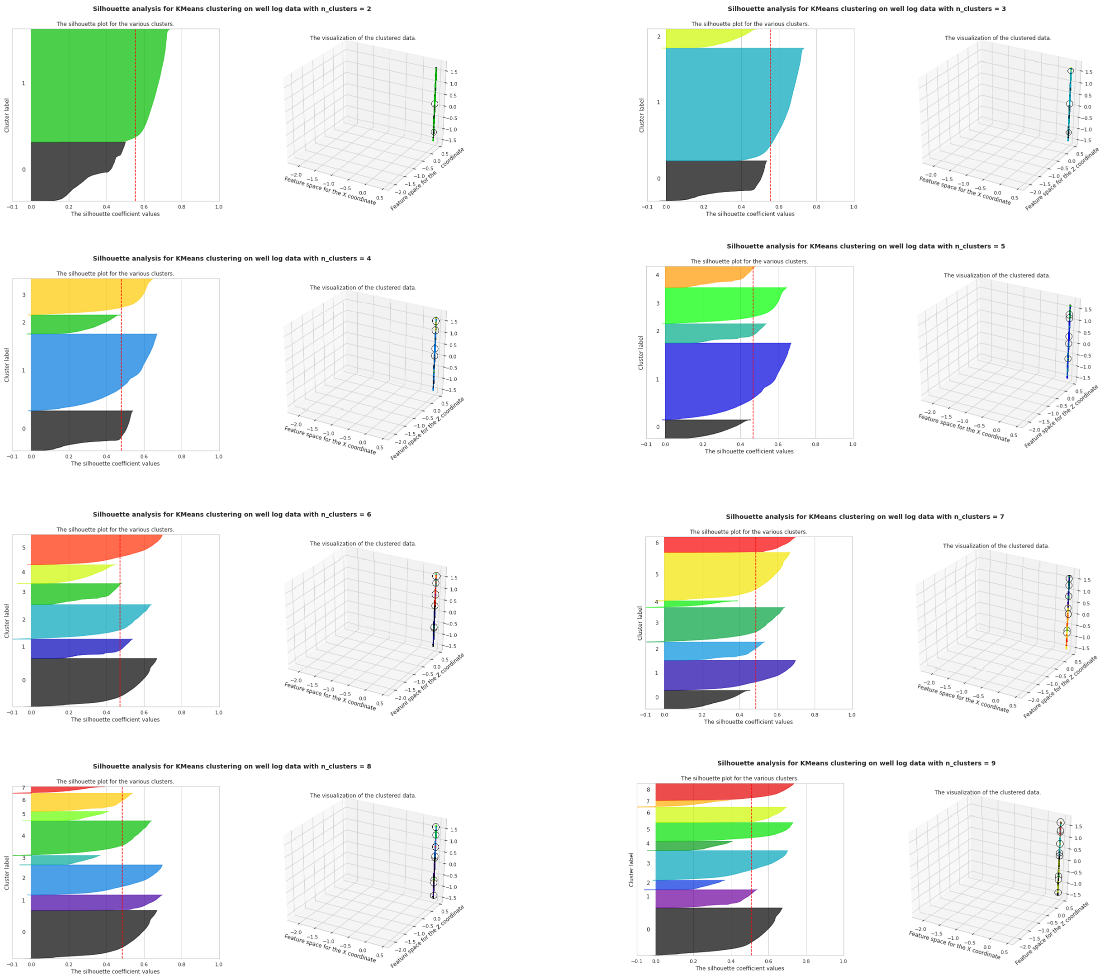


Figure 4.19: Visual silhouette analysis

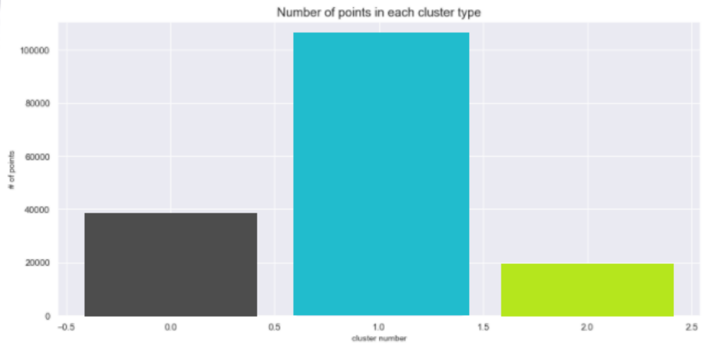


Figure 4.20: KMeans Cluster frequency

The log plots, with the associated cluster label (exluding the first plot), are presented for the concated well data.

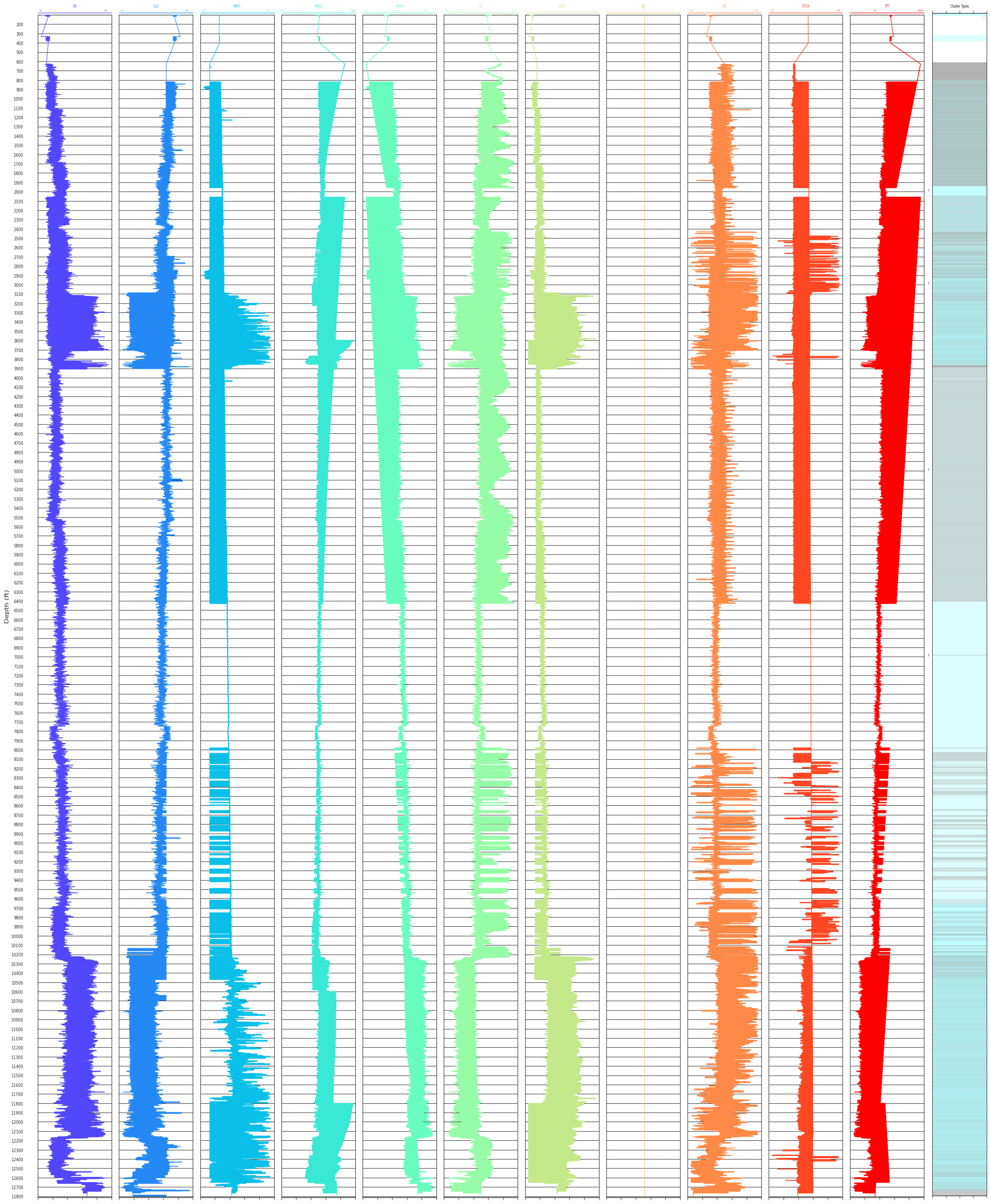


Figure 4.21: Well log plot for the dataset set

Focusing on the 5000 - 12000 depth range for the GR plot, the distinction between sandstones and shales can be seen. Additionally, the reservoir seals - impermeable rocks that form barriers above and below reservoir sections - are identifiable (see figure 4.22). However, the cluster plot for the interval does not match the lithological types indeicated by the GR plot. The factors that most-likely affected the prototypes performance are detailed in the section below.

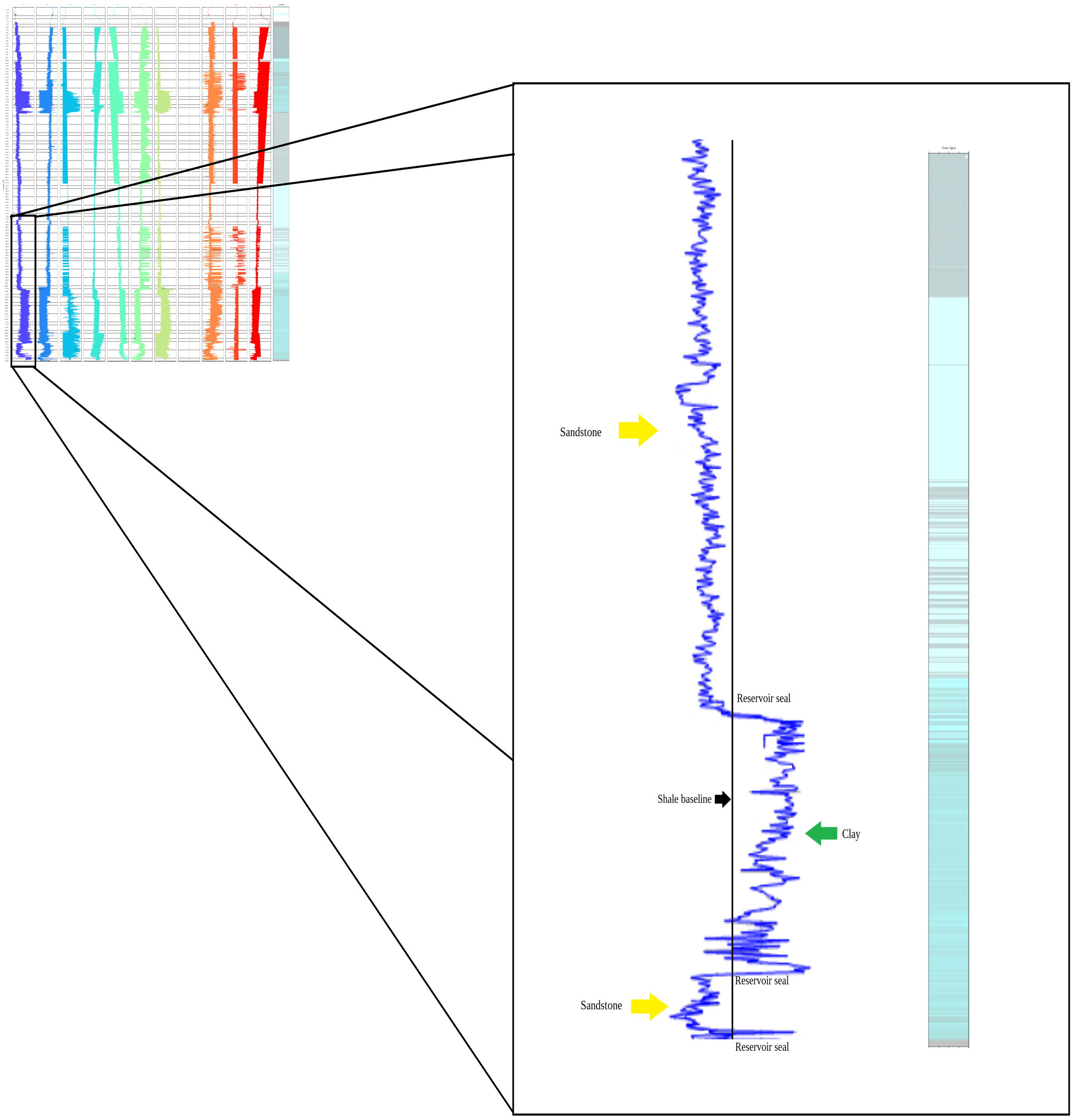


Figure 4.22: Image depicting the intervals of interset and sealing points the bottom depth of the concatenated dataset.

4.7 Geological Prototype Analysis

To interpolate a surface between the well logs, based on their assigned clusters, and to complete prototype development the data set was imported into both a pythonic and GIS environment. Visualisation of the points in 3D showed that the Data optimisation had reduced the data set to only points in well F-04. This can be attributed to three main reasons:

1. The number of data points wells F-06 and F-08 contributed to the combined data set were minimal in comparison to the number contributed to by well F-04. This unequal distribution could have skewed what the model determined as outliers and removed the values for wells F-06 and F-08.

2. Most of the logs used in the creation of the model (i.e. all of the logs besides DEPTH, TNPH, SP, BS, and ITT) were not present in wells F-06 and F-08. Therefore, these logs (GR, NPHI, RHOB, LLD, MSFL, MRES, MTEM, CALS, DT and DTLN) had to be added as empty values (see figures 4.23, 4.24 and 4.25). This resulted in wells F-06 and F-08 contributing to most of the null values in the data set (see figures 4.23, 4.24 and 4.25).

3. In addition to pertinent logs being missing, their values had to be imputed to get pseudo data for those missing entries. These imputed, model approximated values, could have been calculated as values lower or higher than they should have been and were thus removed - by the model - because of the its classification of them as outliers.

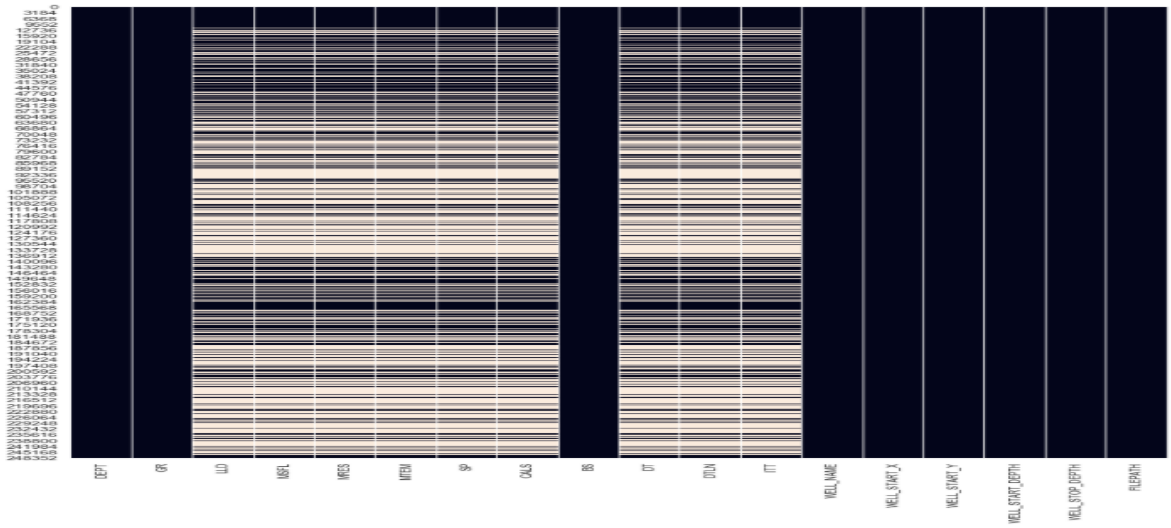


Figure 4.23: Visual representation of the null values in well F-04

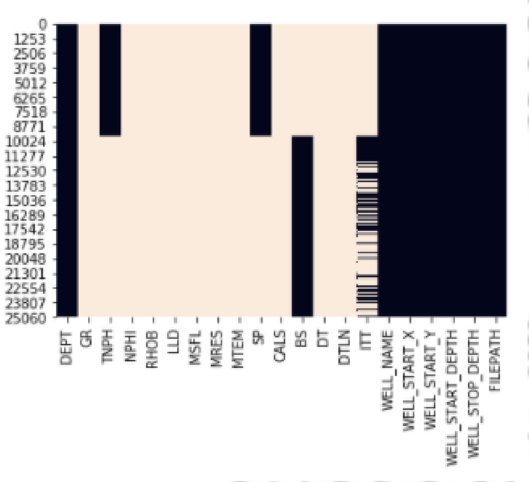


Figure 4.24: Visual representation of the null values in well F-06

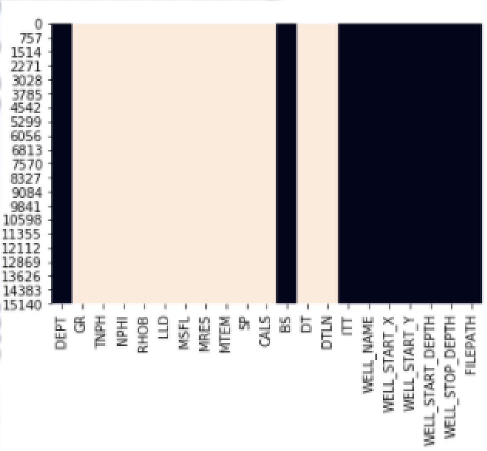


Figure 4.25: Visual representation of the null values in well F-08

Since a surface could not be interpolated from only well F-04, pseudo data had to be generated for what well F-06 and F-08 would have been in order to generate a 3D geological prototype. This process was achieved by duplicating the results for well F-04 and offsetting them by variable amounts.

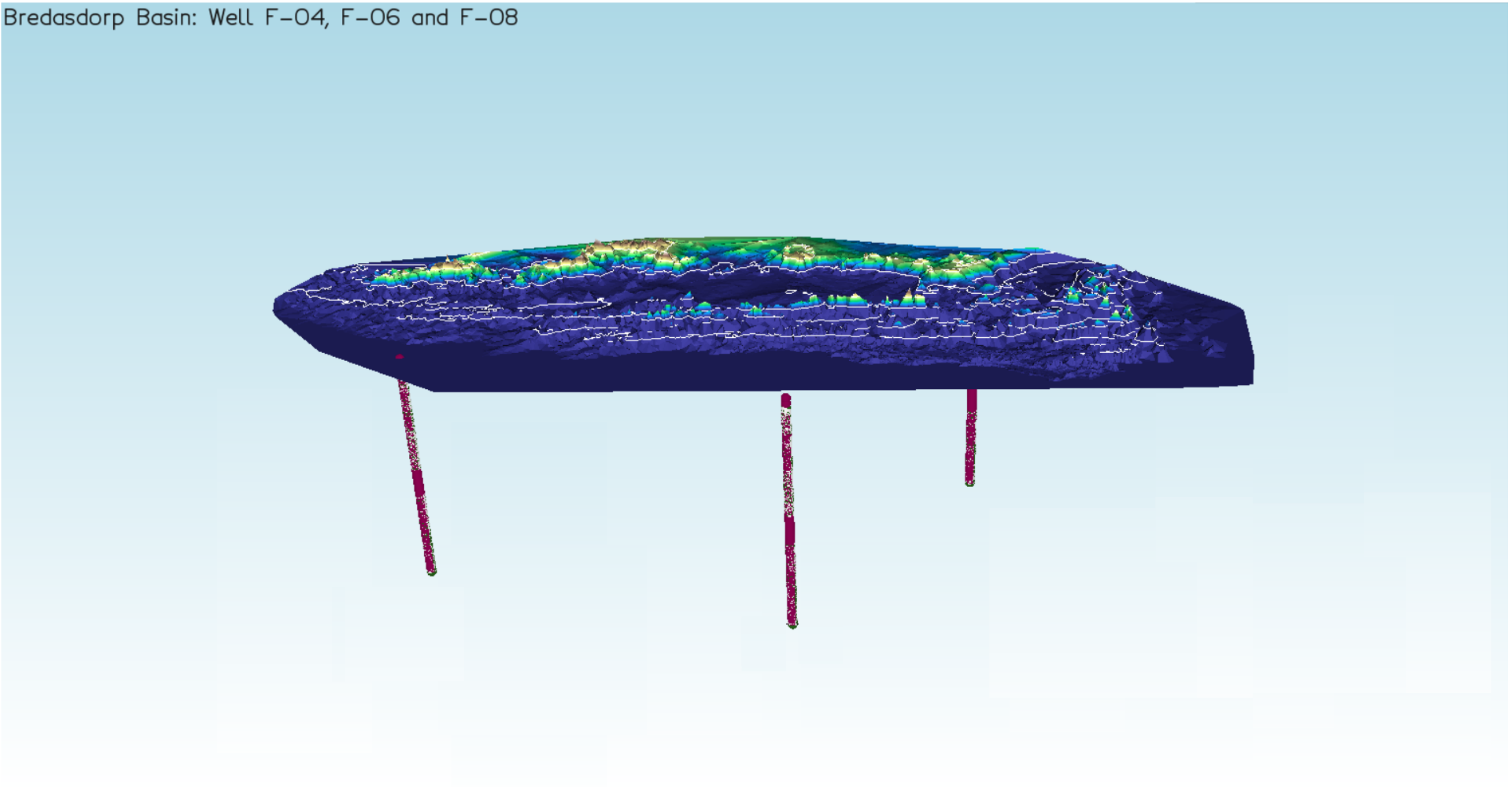


Figure 4.26: 3D geological prototype with well log coloured by cluster name.

The prototype geomodel developed (see figure 4.26 and fig: predictions) had the uncertainties depicted in 4.28 at the same level. These uncertainties vary but generally increase the further away the interpolated surface gets from the well.

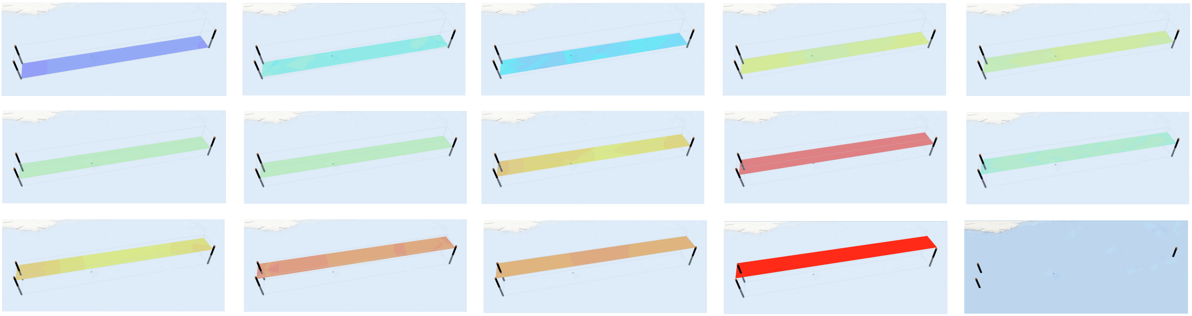


Figure 4.27: Interpolated surface based on the cluster labels of the input data set (slices taken every 1000 m)

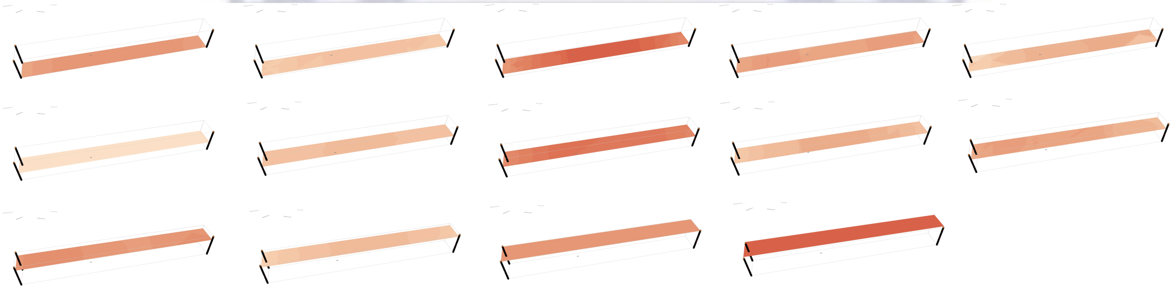
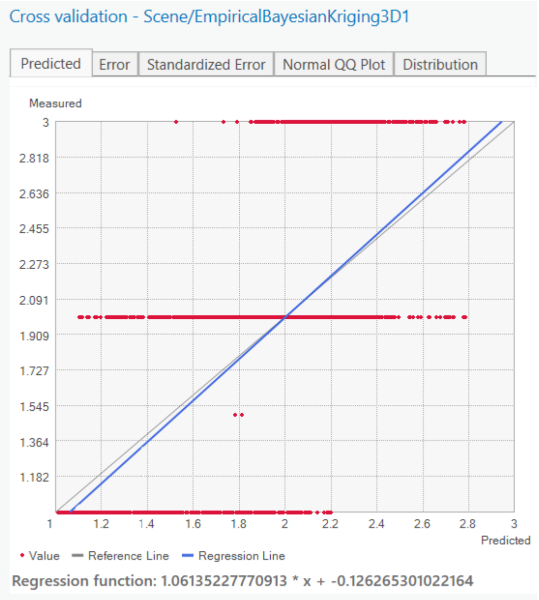


Figure 4.28: Uncertainties associated with the interpolated surface (slices taken every 1000 m)

Delving deeper into how well the interpolated surface is predicted, the inclination of the prediction plot shows there is a high correlation between the points. This can be seen in how the line fitted through the data (seen in blue and described by the equation on the bottom of the image) is close to the 1:1 auto-correlation grey line.

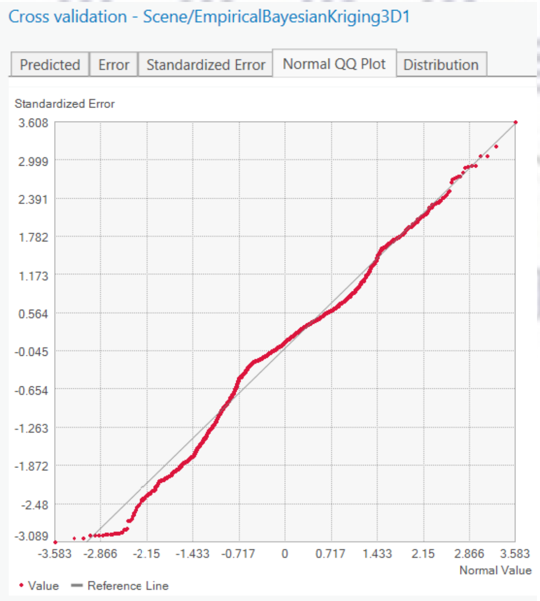


Cross validation - Scene/EmpiricalBayesianKriging3D1

Regression function: 1.06135227770913 * x + -0.126265301022164

Figure 4.29: Predicted vs Measured plot

The quantiles of the difference between the predicted and measured values from a standard normal distribution can be seen in the Normal QQ Plot graph below. Here the close correation beweten the data points and the grey line show that the errors of the predictions, from their true values, are mostly normally distributed.

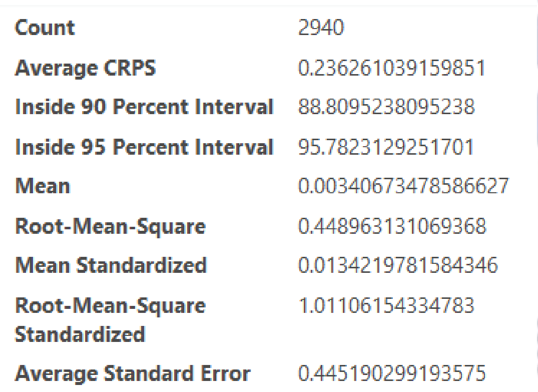


Cross validation - Scene/EmpiricalBayesianKriging3D1

Figure 4.30: Normal plot

From the statistical outputs depicted in the image below (see figure 4.31), a couple inferences can be made about the interpolated surface:

- As the root mean squared standardised errors are greater than 1, the variability in the predictions are being underestimated. This is also confirmed by the average standard errors being less than the root mean squared prediction errors.
- The inside 90% interval, shows that about 89 percent of points fall within a 90 percent cross validation confidence interval.
- The inside 95% interval, shows that about 96 percent of points can be found within a 95 percent cross validation confidence interval.
- The average Continuous Ranked Probability Score (CRPS) of all points, at about 0.2 shows that there is a deviation between the predictive cumulative distribution function and each observed data value



Count	2940
Average CRPS	0.236261039159851
Inside 90 Percent Interval	88.8095238095238
Inside 95 Percent Interval	95.7823129251701
Mean	0.00340673478586627
Root-Mean-Square	0.448963131069368
Mean Standardized	0.0134219781584346
Root-Mean-Square Standardized	1.01106154334783
Average Standard Error	0.445190299193575

Figure 4.31: Interpolated surface summary statistics

5 Conclusion

5.1 Introduction

3D visualisation in well log interpretation has only been around for a comparatively short period of time in the history of the field, and it was only through major technological advancements that it became a possibility. The relative infancy of the domain means that it is a quickly becoming a burgeoning field of research with endless possibilities. To contribute to knowledge in the field, this dissertation set out to build a 3D geological prototype from well logs. The developed prototype had to apply well log interpretation theory as well as vision and perception theory to promote user understanding and aesthetic appeal. These outcomes were achieved by investigating the study area, well logs, unsupervised learning as well as GIS and the cartographic design process.

5.2 Application of the Research

The findings of this investigation detail the versatility and practicality of 3D well log interpretation, which are tools that can be transferred to other fields and industries that require data processing, classification and visualisation.

5.3 Implications of the Research

- The viewer focused prototype development (in terms of understanding and appeal) could promote greater examination of the influence of design on perception and appeal.

- The outcomes of this investigation highlights the applicability of machine learning in geological data processing and visualisation.

5.4 Recommendations and Future Work

Future research directions may focus on the following:

- The application of pseudo-labelling in the clustering of well logs. By applying this recommendation, the accuracy of the geological prototype would be improved upon as the build is based on pseudo-labels of a certain degree of confidence.
- The incorporation of a web component to both the processing and visualisation of subsurface environments. This web component could consist of an interactive website that supports users in carrying out machine learning processes on their own data before displaying the rendered geological maps.
- The development of this investigation's machine learning workflow into a tool that allows for parameter tuning in addition to well log and machine learning algorithm selection.
- Comparison of the different machine learning algorithms (as well as their hyper parameters) on the clustering and visualisation of well logs.
- The extension of the research to other subsurface datasets in the development of a 3D geological prototype. Datasets such as seismic, fault and temperature isosurface recordings.

Bibliography

Albon, C. (2018), *Machine learning with python cookbook: Practical solutions from preprocessing to deep learning*, " O'Reilly Media, Inc.".

Bailey, T. C. and Gatrell, A. C. (1995), *Interactive spatial data analysis*, Vol. 413, Longman Scientific & Technical Essex.

Beyeler, M. (2017), *Machine Learning for OpenCV*, Packt Publishing Ltd.

Broad, D. (1990), Petroleum geology of gamtoos and algoa basins, *in* 'Geological Society of South Africa', pp. 60–63.

Bronshtein, A. (2017), 'Train/test split and cross validation in python', *Understanding Machine Learning* .

Brown, L. F. et al. (1995), *Sequence Stratigraphy in Offshore South African Divergent Basins: An Atlas on Exploration for Cretaceous Lowstand Traps by Soekor (Pty) Ltd, AAPG Studies in Geology 41*, AAPG.

Bychkovskiy, V., Megerian, S., Estrin, D. and Potkonjak, M. (2003), A collaborative approach to in-place sensor calibration, *in* 'Information processing in sensor networks', Springer, pp. 301–316.

Chopra, R., England, A. and Alaudeen, M. (2019), *Data Science with Python: Combine Python with machine learning principles to discover hidden patterns in raw data*, Packt Publishing.
URL: *https://books.google.co.za/books?id=RYmkDwAAQBAJ*

Crampton, J. W. and Krygier, J. (2005), 'An introduction to critical cartography', *ACME: An International Journal for Critical Geographies* **4**(1), 11–33.

Delfiner, P., Peyret, O., Serra, O. et al. (1987), 'Automatic determination of lithology from well logs', *SPE formation evaluation* **2**(03), 303–310.

Dent, B., Torguson, J. and Hodler, T. (2009), *Cartography: Thematic Map Design*, McGraw-Hill Higher Education.
URL: *https://books.google.co.za/books?id=HGounQAACAAJ*

Developers, S. L. (2007), 'Choosing the right estimator', *Available at: https://scikit-learn. org/stable/tutorial/machine_learning_map* .

Duda, R. O., Hart, P. E. and Stork, D. G. (2012), *Pattern classification*, John Wiley & Sons.

Ford, J., Burke, H., Royse, K. and Mathers, S. (2008), 'The 3d geology of london and the thames gateway: a modern approach to geological surveying and its relevance in the urban environment'.

Graham, L. (2008), 'Gestalt theory in interactive media design', *Journal of Humanities & Social Sciences* **2**(1).

Grinderud, K. (2009), *GIS: The geographic language of our age*, Tapir Academic Press.

Haeberling, C. (2005), Cartographic design principles for 3d maps–a contribution to cartographic theory, *in* 'Proceedings of ICA Congress Mapping Approaches into a Changing World'.

Hubel, D. H. and Wiesel, T. N. (1979), 'Brain mechanisms of vision', *Scientific American* **241**(3), 150–163.

Hunt, R. W. G. and Pointer, M. R. (2011), *Measuring colour*, John Wiley & Sons.

Hyne, N. (2014), *Dictionary of petroleum exploration, drilling & production*, PennWell Corporation.

Ile, C. O. K. (2018), Cartographic designs for 3d maps: Enhancing affect. [Unpublished honours dissertation].

Jahn, F., Cook, M. and Graham, M. (2008), *Hydrocarbon exploration and production*, Elsevier.

Jones, R., McCaffrey, K., Clegg, P., Wilson, R., Holliman, N. S., Holdsworth, R., Imber, J. and Waggott, S. (2009), 'Integration of regional to outcrop digital data: 3d visualisation of multi-scale geological models', *Computers & Geosciences* **35**(1), 4–18.

Jovanović, V. (2016), 'The application of gis and its components in tourism', *Yugoslav Journal of Operations Research* **18**(2).

Judd, D. and G, W. (1975), *Color in Business, Science, and Industry*, John Wiley and Sons.

Keates, J. S. (2014), *Understanding maps*, Routledge.

Kent, A. J. and Vujakovic, P. (2017), *The Routledge Handbook of Mapping and Cartography*, Routledge.

Khana, S. and Dillay, G. (1986), 'Seychelles: Petroleum potential of this indian ocean paradise', *Oil Gas J.;(United States)* **84**(12).

Kraak, M.-J. (1993), 'Three-dimensional map design', *The Cartographic Journal* **30**(2), 188–194.

Lidwell, W., Holden, K. and Butler, J. (2010), *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*, Rockport Pub.

Luthi, S. (2001), *Geological well logs: Their use in reservoir modeling*, Springer Science & Business Media.

Malolepszy, Z. (2005), Three-dimensional geological maps, *in* 'The Current Role of Geological Mapping in Geosciences', Springer, pp. 215–224.

Masindi, R. (2016), 'A review of a small production gas field in central bredasdorp basin, based on new seismic, integrated with core and log data.'.

McMillan, I., Brink, G., Broad, D. and Maier, J. (1997), Late mesozoic sedimentary basins off the south coast of south africa, *in* 'Sedimentary Basins of the World', Vol. 3, Elsevier, pp. 319–376.

Mennan, A. (2017), 'Well log interpretation and 3d reservoir property modeling of maui-b field, taranaki basin, new zealand'.

Monmonier, M. (2018), *How to lie with maps*, University of Chicago Press.

Muehlenhaus, I. (2013), *Web cartography: map design for interactive and mobile devices*, CRC Press.

Müller, A. C., Guido, S. et al. (2016), *Introduction to machine learning with Python: a guide for data scientists*, " O'Reilly Media, Inc.".

Parsiegla, N., Stankiewicz, J., Gohl, K., Ryberg, T. and Uenzelmann-Neben, G. (2009), 'Southern african continental margin: Dynamic processes of a transform margin', *Geochemistry, Geophysics, Geosystems* **10**(3).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Petroleum Agency of South Africa, P. (2003), *South African exploration opportunities*, Information brochure. South African Agency for Promotion of Petroleum.

Petroleum Agency of South Africa, P. (2012), *South African exploration opportunities*, Information brochure. South African Agency for Promotion of Petroleum.

Petroleum Agency of South Africa, P. (2013), 'What is petroleum agency sa?', *Available at: https://www.petroleumagencysa.com/* .

Peveraro, R. (2006), 'Well log interpretation', *Petroskills-OGCI, Course Notes, Tulsa, OK* .

Rhind, D. W. and Taylor, D. F. (2013), *Cartography Past, Present and Future: A Festschrift for FJ Ormeling*, Elsevier.

Roth, I. and Bruce, V. (1995), *Perception and representation: Current issues*, Sociology and Social Change.

Rutledge, K., Ramroop, T., Boudreau, D., McDaniel, M., Teng, S., Sprout, E., Costa, H., Hall, H. and Hunt, J. (2011), 'basin'.
URL: *https://www.nationalgeographic.org/encyclopedia/basin/*

Sarkar, T. (2020), 'Clustering metrics better than the elbow method - kdnuggets'.
URL: *https://www.kdnuggets.com/2019/10/clustering-metrics-better-elbow-method.html*

Snowden, R., Snowden, R. J., Thompson, P. and Troscianko, T. (2012), *Basic vision: an introduction to visual perception*, Oxford University Press.

Song, R., Qin, X., Tao, Y., Wang, X., Yin, B., Wang, Y. and Li, W. (2019), 'A semi-automatic method for 3d modeling and visualizing complex geological bodies', *Bulletin of Engineering Geology and the Environment* **78**(3), 1371–1383.

Stevens, J., Smith, J. and Bianchetti, R. (2012), 'Mapping our changing world', *MacEachren AM, Peuquet DJ. Department of Geography, The Pennsylvania State University, University Park* .

Tankard, A. J., Martin, M., Eriksson, K., Hobday, D., Hunter, D. and Minter, W. (2012), *Crustal evolution of southern Africa: 3.8 billion years of earth history*, Springer Science & Business Media.

Tyner, J. A. (2010), 'Principles of map design. new york'.

Van der Meulen, M., Doornenbal, J., Gunnink, J., Stafleu, J., Schokker, J., Vernes, R., Van Geer, F., Van Gessel, S., Van Heteren, S., Van Leeuwen, R. et al. (2013), '3d geology in a 2d country: perspectives for geological surveying in the netherlands', *Netherlands Journal of Geosciences* **92**(4), 217–241.

Van Rossum, G. et al. (2007), Python programming language., *in* 'USENIX annual technical conference', Vol. 41, p. 36.

Worboys, M. and Duckham, M. (2004), *GIS: A Computing Perspective, Second Edition*, Taylor & Francis.
URL: *https://books.google.co.za/books?id=x4e2IVV0u9gC*

Zhu, L., Zhang, C., Li, M., Pan, X. and Sun, J. (2012), 'Building 3d solid models of sedimentary stratigraphic systems from borehole data: an automatic method and case studies', *Engineering Geology* **127**, 1–13.