TECHNICAL UNIVERSITY OF CRETE

Analysis of Well Log Data using Time Series Models and Geostatistical Methods

A thesis submitted in fulfilment of the requirements for the Diploma degree in the

SCHOOL OF MINERAL RESOURCES ENGINEERING

Anastasia Xenaki

Advisory Committe:

Professor D. T. Hristopulos (supervisor), School of Mineral Resources Engineering, Technical University of Crete

Professor A. Vafeidis, School of Mineral Resources Engineering, Technical University of Crete

Professor N. Varotsis, School of Mineral Resources Engineering, Technical University of Crete

December 13, 2019

Abstract

This thesis focuses on the statistical analysis of well log data from two hydrocarbon reservoirs located in Labrador Island, Western Newfoundland (Canada). The data is obtained from two onshore wells (Finnegan and Seamus). We focus on the analysis of four logs (spontaneous potential, Gamma radiation and two induction logs) from six different formations. The thesis has three main objectives: (i) to estimate the probability distributions and spatial correlations in data obtained from the same well, (ii) to evaluate cross-correlations between logs across the two different wells, and (iii) to explore methods for the reconstruction of missing well log data.

With respect to the first objective, the exploratory statistical analysis indicates that the majority of the respective properties do not follow the Gaussian distribution. However, after removing an empirically determined trend function, the residuals are closer to the Gaussian distribution. The spontaneous potential and Gamma radiation indicators can be described by Cauchy and Gumbel distributions, while the induction indicators by means of the Gamma and Weibull distributions. Variogram analysis suggests that spontaneous potential and Gamma Radiation conform to the same type of theoretical variogram model with similar sill and range values.

In reference to the second objective, the statistical analysis indicates weak cross correlations between log data measured at the two different wells. The Gamma radiation logs show both positive and negative cross correlations which are overall higher (in magnitude) than for the respective correlations for the other three logs.

Regarding the third objective, the comparison of the performance of different imputation, interpolation and time series algorithms for gap filling indicates that linear interpolation, linear weighted moving average and less often the Kalman-ARIMA methods are the top-performing algorithms for well log gap filling.

Περίληψη

Η Γεωστατιστική παρέχει εργαλεία στατιστικής ανάλυσης για την μελέτη χωρικών ή χωροχρονικών φυσικών μεταβλητών. Η προσέγγιση μιας χωροχρονικής γεωστατιστικής ανάλυσης σε δεδομένα υδρογονανθράκων τα όποια χαρακτηρίζονται από ανομοιογένεια αποτελεί αντικείμενο έρευνας. Ωστόσο, μέθοδοι ανάλυσης χρονοσειρών που περιλαμβάνουν την επεξεργασία μονοδιάστατων χρονικών δεδομένων, μπορούν κάλλιστα να εφαρμοστούν σε προβλήματα που αφορούν μονοδιάστατα χωρικά δεδομένα, όπως για παράδειγμα δεδομένα διαγραφιών από γεωτρήσεις. Τόσο η επιστήμη της Γεωστατιστικής όσο και η ανάλυση Χρονοσειρών είναι αντικείμενα με ερευνητικό ενδιαφέρον και εφαρμογές σε πολλά επιστημονικά πεδία, όπως στον τομέα της μεταλλευτικής, της μηχανικής πετρελαίου, σε τομείς περιβαλλοντικών επιστημών, στη τηλεπισκόπηση και την επιστήμη των υλικών.

Σκοπός της εργασίας είναι η εφαρμογή γεωστατιστικών μεθόδων και μοντέλων ανάλυσης χρονοσειρών με στόχο την ανάλυση των χωρικών συσχετίσεων διαγραφιών γεώτρησης, όπως επίσης και την πρόβλεψη (αποκατάσταση) κενών δεδομένων (missing data). Τα κενά δεδομένων προέρχονται είτε από αστοχίες εξοπλισμού, είτε από δυσλειτουργία των αισθητήρων, ή ακόμη από σφάλματα στα συστήματα "αποστολής/ανάκτησης" δεδομένων. Τα προβλήματα αυτά δυσχεραίνουν την διαδικασία εκτίμησης των γεωλογικών σχηματισμών από τις διαγραφίες. Η παρούσα εργασία γράφτηκε με την προοπτική προσέγγισης των δύο παραπάνω θεμάτων, δηλαδή της ανάλυσης των χωρικών συσχετίσεων και της αποκατάσταση ς κενών δεδομένων.

Η παρούσα εργασία επικεντρώνεται στην ανάλυση διαθέσιμων δεδομένων από διαγραφίες ταμιευτήρων υδρογονανθράκων που βρίσκονται στην νήσο Labrador στο Δυτικό Newfoundland (Καναδάς). Η μελέτη βασίζεται στην ανάλυση των ακόλουθων τεσσάρων επιλεγμένων διαγραφιών: φυσικό δυναμικό, δείκτης ακτινοβολίας γάμμα και ηλεκτρομαγνητικής επαγωγής από έξι σχηματισμούς που βρίσκονται σε δύο επάκτιες γεωτρήσεις, ονόματι Finnegan και Seamus. Οι δύο γεωτρήσεις βρίσκονται σε απόσταση 14.5 χλμ μεταξύ τους. Αναλυτικότερα, οι υπό μελέτη σχηματισμοί της γεώτρησης Finnegan είναι οι ακόλουθοι: Goose (American) Tickle με 1422 δεδομένα, Table Point με 725 δεδομένα, Aguathuna με 250 δεδομένα, Catoche με 624 δεδομένα, Boat Harbour με 599 δεδομένα, και Watts Bight με 349 δεδομένα. Αντίστοιχα, οι υπό μελέτη σχηματισμοί της γεώτρησης Seamus είναι οι ακόλουθοι: Goose (American) Tickle με 1700 δεδομένα, Table Point με 871 δεδομένα,

Aguathuna με 347 δεδομένα, Catoche με 721 δεδομένα, Boat Harbour με 819 δεδομένα, και Watts Bight με 406 δεδομένα.

Η συγκεκριμένη μελέτη εστιάζει σε τρία κύρια σημεία ενδιαφέροντος: (i) εκτίμηση χωρικών συσχετίσεων από διαγραφίες που λαμβάνονται από την ίδια γεώτρηση, (ii) εκτίμηση της ετεροσυσχέτισης μεταξύ των μετρήσεων των ίδιων φυσικών ιδιοτήτων από δύο διαφορετικές γεωτρήσεις, (iii) διερεύνηση μεθόδων για την ανακατασκευή κενών δεδομένων.

Για τον πρώτο στόχο, προσδιορίσαμε τους διάφορους γεωλογικούς σχηματισμούς στα σημεία των γεωτρήσεων. Εφαρμόσαμε διερευνητική στατιστική ανάλυση δεδομένων για να προσδιοριστούν οι κατανομές πιθανότητας κάθε διαγραφίας ανά σχηματισμό, όπως επίσης και για να προσδιοριστούν τα βέλτιστα μοντέλα βαριογραμμάτων ανά διαγραφία.

Για τον δεύτερο στόχο χρησιμοποιήσαμε μεθόδους παρεμβολής με σκοπό να δημιουργήσουμε δύο ευθυγραμμισμένα σύνολα δεδομένων με κοινό βήμα δειγματοληψίας. Αυτή η διαδικασία κρίθηκε απαραίτητη διότι διαφορετικοί σχηματισμοί βρίσκονται σε διαφορετικά βάθη κατά μήκος της κάθε γεώτρησης, ενώ το βήμα δειγματοληψίας διαφέρει μεταξύ των γεωτρήσεων. Εν συνεχεία συγκρίναμε τα αποτελέσματα των τιμών των διάφορων μεθόδων παρεμβολής που εφαρμόστηκαν για τον υπολογισμό της ετεροσυσχέτισης των γεωτρήσεων. Οι μέθοδοι αυτοί περιλαμβάνουν την γραμμική παρεμβολή, την κυβική παρεμβολή, την παρεμβολή σφηνοειδών συναρτήσεων (splines) όπως επίσης και την τεχνική του κοντινότερου γείτονα.

Ο τρίτος στόχος αποσκοπεί στην διερεύνηση των εφαρμογών της ανάλυσης χρονοσειρών στην εκτίμηση κενών δεδομένων διαγραφιών. Τα μοντέλα αντικατάστασης (imputation) και παρεμβολής χρησιμοποιούνται συνήθως για να πληρώσουν κενά δεδομένων σε διαγραφίες γεωτρήσεων. Για τον υπολογισμό της απόδοσης των διάφορων μεθόδων, οι καταγεγραμμένες διαγραφίες διαχωρίζονται σε δύο διακριτά σύνολα: το σύνολο εκπαίδευσης (οι τιμές των διαγραφιών σε αυτό το σύνολο θεωρούνται γνωστές) και το σύνολο ελέγχου (όπου οι τιμές των διαγραφιών θεωρούνται άγνωστες). Η συμπλήρωση των κενών έγινε με την χρήση των μεθόδων αντικατάστασης, παρεμβολής και χρονοσειρών. Οι μέθοδοι περιλαμβάνουν την Kalman Arima, την μέθοδο μέσο όρου, την γραμμική παρεμβολή και την παρεμβολή σφηνοειδών συναρτήσεων (Spline) όπως επίσης και τον απλός κινούμενο μέσο όρο, και τον ζυγισμένο κινητό μέσο όρο. Η ακρίβεια της πρόβλεψης βασίστηκε στην απόσταση ανάμεσα στα αυθεντικά δεδομένα και στις εκτιμήσεις μέσω των μεθόδων αντικατάστασης, παρεμβολής ή χρονοσειρών.

Σχετικά με τον πρώτο στόχο της εργασίας, η διερευνητική ανάλυση έδειξε ότι η πλειοψηφία των μετρούμενων ιδιοτήτων δεν ακολουθεί την Γκαουσσιανή κατανομή πιθανότητας. Σε αυτές τις περιπτώσεις, κατόπιν αφαίρεσης μιας συνάρτησης τάσης τα στατιστικά υπόλοιπα βρίσκονται πιο κοντά στην Γκαουσιανή κατανομή. Τα αποτελέσματα της μελέτης αποδεικνύουν ότι οι διαγραφίες φυσικού δυναμικού και διάταξης ακτινοβολίας γάμμα περιγράφονται καλύτερα από κατανομές πιθανότητας που ορίζονται σε όλο το πεδίο των πραγματικών τιμών, όπως οι κατανομές Cauchy και Gumbel. Αντίθετα, οι διαγραφίες επαγωγής περιγράφονται καλύτερα από κατανομές που ορίζονται στο σύνολο των θετικών αριθμών όπως οι κατανομές Γάμμα και Weibull. Τα αποτελέσματα της ανάλυσης των βαριογραμμάτων έδειξαν ότι οι διαγραφίες φυσικού δυναμικού και διάταξης ακτινοβολίας γάμμα, πολύ συχνά προσαρμόζονται στο ίδιο θεωρητικό μοντέλο βαριογράμματος και αυτό οφείλεται στο γεγονός ότι οι τιμές της οροφής και της ζώνη επιρροής είναι παρόμοιες. Η ανάλυση βαριογράμματος επιβεβαιώνει την υψηλή χωρική ετερογένεια που χαρακτηρίζει τις καταγραφές διαγραφιών.

Όσον αφορά στο δεύτερο στόχο της εργασίας, τα αριθμητικά αποτελέσματα της στατιστικής ανάλυσης έδειξαν ασθενή ετεροσυσχέτιση μεταξύ των ιδιοτήτων που μετρήθηκαν στις δύο γεωτρήσεις. Η ετεροσυσχέτιση εξετάστηκε μέσω του υπολογισμού μέτρων στατιστικής εξάρτησης όπως η συχέτιση Pearson και η συσχέτιση Spearman. Όπως αναφέρθηκε ανωτέρω, για τον υπολογισμό των ετεροσυσχετίσεων χρησιμοποιήθηκαν μεθόδοι παρεμβολής με σκοπό την ομογενοποίηση του βήματος δειγματοληψίας. Όλες οι μέθοδοι οδήγησαν σε παρόμοιες εκτιμήσεις των συσχετίσεων. Οι διαγραφίες διάταξης ακτίνων γάμμα απέδωσαν το πιο ενδιαφέροντα αποτελέσματα σε σχέση με τις υπόλοιπες διαγραφίες, παρουσιάζοντας τόσο θετικές όσο και αρνητικές τιμές συσχετίσεων. Οι τιμές των θετικών συντελεστών συσχέτισης κυμαίνονται από 0.001 έως 0.483, ενώ οι τιμές των αρνητικών συντελεστών συσχέτισης εκτείνονται από -0.142 έως -0.001.

Αναφορικά με τον τρίτο στόχο της εργασίας, η σύγκριση και η ποσοτικοποίηση της απόδοσης των αλγορίθμων αντικατάστασης και παρεμβολής έδειξε ότι η γραμμική παρεμβολή, ο ζυγισμένος μέσος όρος, και λιγότερο συχνά η μέθοδος Kalman-Arima, είναι τα πιο αποδοτικά μοντέλα αποκατάστασης κενών δεδομένων.

Acknowledgements

First of all, I would like to thank the members of my thesis committee. Professor D. T. Hristopulos, for tirelessly answering all of my questions and intrigued me right from the beginning. Thank you for sharing your expertise and insights at critical moments. I have very much enjoyed working with you. I would also like to thank Professor A. Vafeidis and Professor N. Varotsis for their useful comments.

I would like to express my sincerest and heartfelt gratitude to the members of the Geostatistics Laboratory in the School of Mineral Resources Engineering at the Technical University of Crete, Mr. Manolis Petrakis, Dr. Andrew Pavlides and Mrs. Vasiliki Agou. I really appreciate all of the questions that you have asked me throughout the process and all of the feedback you have provided me with, which surely has greatly contributed to the quality of this thesis.

I would also like to thank my dear friends, of whom some I have met during my studies at TUC, for their motivation, support and great laughs!

Last but not least, I owe a debt of gratitude to my parents Dimitrios and Virginia, who have always believed in me, for their unconditional love and support, and Dimitris for his patience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Geostatistics provides tools for the statistical analysis of spatial or spatiotemporal data. On the other hand, time series analysis provides tools for processing temporal data; such tools can also be applied to one-dimensional spatial data such as well log data. Both geostatistics and time series analysis can be applied to mining and petroleum engineering data, as well as ground-based environmental and remote sensing data. Well logs were developed and used to perform geothermal ([90]; [79]; [87]; [68]), geotechnical ([75]; [19]; [82]) and environmental studies ([85]; [59]; [44]). Observations about how deep a formation is and what type of lithologies are expected to be found while the borehole is being drilled deeper and deeper, can be obtained and be further analyzed.

For example, in geology, well logging readings are a considerable source of information that can be used to create a preliminary geological map that is necessary for surface exploration ([66]; [42]; [76]), in petrophysics, they provide a unbias evaluation of the potential production of hydrocarbon reservoirs ([99]; [88]; [48]); in geophysics, collecting and assessing high precision well-log data is the first step of seismic analysis and real-time evaluation of the formation's and fluid's properties ([63]; [91]; [94]); in petroleum engineering well-log data is used to estimate parameters for numerical simulations ([28]; [65]; [86]). In this thesis we apply geostatistical methods and time series methods to analyze well-log correlations and to predict missing data (missing data reconstruction).

In this thesis we apply geostatistical methods and time series models to analyze well-log correlations and to predict (reconstruct) missing data. The missing data problem occurs due to instrument failures, sensor malfunctions and data "send/retrieval" problems that hinder the evaluation of geological formations. Correlations reflect spatial continuity and provide information needed for prediction methods that are used to reconstruct gaps in the data.

The study focuses on the analysis of available well logs from two hydrocarbon reservoirs located in Labrador Island, Western Newfoundland (Canada). The data is obtained from two onshore wells (Finnegan and Seamus). The distance between the Finnegan and Seamus well is 14.7km. We will focus on the analysis of four select logs (Spontaneous Potential, Gamma Radiation and two Induction logs) from six formations. More explicitly, the examined formations from Finnegan well are the following: Goose (American) Tickle, comprising 1422 data points; Table Point comprising 725 data points; Aguathuna with 250 data points; Catoche with 624 data points; Boat Harbour with 599 data points, and Watts Bight with 349 data points. Accordingly, the formations probed in the Seamus well are: Goose (American) Tickle with 1700 data points; Table Point with 871 data points; Aguathuna with 347 data points; Catoche with 721 data points; Boat Harbour with 819 data points, and finally Watts Bight with 406 data points.

This thesis has three main objectives: (i) to estimate spatial correlations in well log data obtained from the same well, (ii) to evaluate cross-correlations between measurements of the same property across the two different wells, and (iii) to explore methods for the reconstruction of missing data.

To address the first objective, we identify the different geological formations at the locations of the two wells. We apply exploratory data analysis to determine the relevant probability distributions and summary statistics, as well as variogram analysis to identify spatial correlations within each formation and to determine the optimal variogram model for each measured property.

In order to address the second objective, we use several interpolation methods to create two data sets with formation alignment and common sampling step. This processing is necessary since different formations are found at different depths along each drill hole, and the data in each drill hole have unequal sampling steps. We compare the different interpolation methods used in terms of the resulting values of well-to-well log correlations; these methods comprise: linear interpolation, cubic and spline interpolation, and nearest neighbor interpolation.

The third objective is to investigate the potential of time series methods for estimating missing data in the well log data series. Missing data imputation and interpolation models are typically used to improve missing well log data quality. To assess the performance of different methods, full data records are split into two disjoint sets: a training set (where the log values are assumed to be known) and a testing set (where the log values are assumed to be missing). We use different imputation, interpolation, and time series methods for filling the gaps (testing set values); these methods comprise: Kalman ARIMA, mean imputation, linear and spline interpolation, as well as linear weighted and simple moving average methods. The prediction accuracy (which measures the agreement between the original testing set values and the imputed or interpolated values) is used to quantify the performance of the gap-filling methods.

With respect to the first objective of the study, the exploratory analysis indicates that the majority of the respective properties do not follow the Gaussian distribution. However, after removing a trend function, the residuals are closer to the Gaussian distribution. Results demonstrate that the Spontaneous potential and Gamma radiation indicators can be most often described by Cauchy and Gumbel distributions. In contrast, the Induction indicators can be most often described means of the Gamma and Weibull distributions. The results of the variogram analysis indicate that Spontaneous potential and Gamma Radiation indicators are mostly fitted to the same type of theoretical variogram model, with similar sill and range values. The variogram analysis confirmed that high spatial heterogeneity characterizes the entire span of the logging records.

With respect to the second objective, the statistical analysis indicates a weak correlation between the respective properties measured at the two different wells. The association between the data at the neighboring wells is examined by means of statistical dependence measures such as the Pearson's linear correlation coefficient and Spearman's rank correlation coefficient. The cross correlations calculated from the processed data using different interpolation models lead to similar values. The Gamma radiation logs show both positive and negative correlation which are overall higher (in magnitude) than for the other three logs. The values of the positive correlation coefficients range from 0.001 to 0.483, while the values of the negative correlation coefficients range from -0.142 to -0.001.

Regarding the third objective of the study, the comparison of the performance of different imputation, interpolation and time series algorithms for gap filling indicates that linear interpolation, linear weighted moving average and less often the Kalman-ARIMA methods are the top-performing algorithms.

Chapter 2

Formation Evaluation : Well Logs

Using well logs in oil and gas exploration is like "hunting on a game preserve"

George R. Pickett
Colorado School of Mines

In situ measurements taken by running logs can give answers to whether a geological structure of a potential oil or gas reservoir exists. Additional information includes the finding of the reservoir's location in the geological strata, the productivity of the upstream, midstream and downstream industry and the inductive inferences of evidence of a near reservoir.

Interpretation of well-logs delineates the properties related to geology and petrophysics, such as the determination of rock and reservoir fluids composition, which are usually deduced from examinations of outcrops, cores and cuttings. Any other useful information can be obtained by measuring the natural gamma ray radiation, bulk density, sonic transit time etc. Consequently, log data often constitutes the *signature* of the rock. Those well logging techniques would be further examined in this section.

2.1 Basic Log Types

The *while drilling* evaluation techniques are beneficial and allow a real-time characterization of the drilled formations. These techniques require expensive high-technology sensors to be inserted in the bottomhole assembly, while performing high resolution records ([24]; [62]; [18]; [81]). For this reason, a brief introduction of the main types of logs will be presented.

Logging While Drilling (LWD)

The Logging-While-Drilling (LWD) formation evaluation sensors acquire downhole data while drilling, collecting mainly petrophysical data. The measuring elements are part of the instrumented *Bottom Hole Assembly*, also called BHA, the drilling collars; pulses of the signals are transmitted to the surface via the mud column. The advantages of LWD are:

1. Access to real time information.
2. Mud invasion does not have an effect on measurements.
3. The LWD tools is more serviceable for collecting data from tough structural environments, such as deviated wells, horizontal wells or an unstable borehole.
4. The LWD sensor provides information about the well's placement and stability while minimizing the risk of a stuck pipe, thus a safer and more efficient hole is drilled.

However, there are factors restricting the LWD tool's efficiency and those are mentioned below:

1. Data transmission/recording may be affected by the speed's telemetry or by the existence of pumped mud into the drill string.
2. Limited memory size.
3. Most LWD tools are powered by batteries with limited battery life that fluctuates from 40 to 90 hours depending on the tool.

4. LWD tool's placement in the bit have to be taken into consideration due to some technical limitations. For instance, ROP's productiveness and sufficiency can possible be influenced by the location of the tool in the drill string.

Measurements While Drilling (MWD)

The Measurement-While-Drilling formation evaluation technique measures data which is near the bit, without interrupting the standard drilling operations. The recorded information reaches the surface by the exact mechanism of transmission of the LWD tool (mud pressure pulses). The advantages of MWD are:

1. Real time directional drilling operations monitoring.
2. Advantageous use in wellbore completion.
3. Estimation of drilling formation properties and drilling parameters, such as the bottom hole pressure, the torque and the weight on the bit, in the interest of optimizing the drilling process.

2.2 Well Logging Methods

Drilling and geophysical techniques are more often used in modern exploration and evaluation of a formation. Well logging data acquisition and interpretation is of the utmost interest of geoscientists. The measurements made with logging tools provide accurate and reliable information of both the rock and its fluid content. Several significant advances have been developed in order to make the acquisition of the data a credible process, including the interpretation of well log data in various rock formations. This practice is considered rather biased than to extract information given from a scattered core analysis sample. Therefore gives the advantage of an objective visualization of the formations at the specific scale plus a representative and more detailed description of the formations. These developments can ensue a precise, even if errors are present, well log data interpretation and reinterpretation and a quickly data obtainment, whereas reduce the total well cost. It is of considerable importance to cite that a wireline logging cost is usually ranging from 5 to 10% of the the total well cost covering approximately 90% of the total geological information which is illustrated in figure 2.1 .

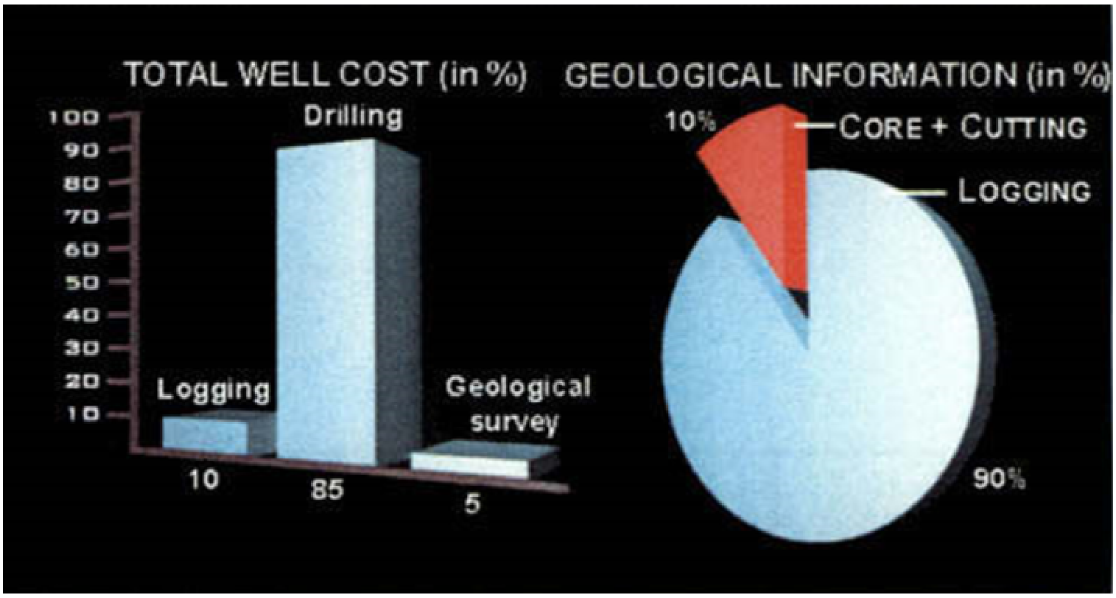


Figure 2.1: Average logging cost represents the 10% of the total well cost which is providing the 90% of the total geological information. Figure retrieved from [81]

We classify the various well logging measurements into two board categories according to their properties ([80]). The first group includes *natural* or *spontaneous* phenomena. The basic equipment employs a single detector to acquire data from the wellbore (*passive system*). The second group includes *induced* phenomena. The basic equipment requires a sources or an emitter to appropriately stimulate a response in the formation, annexed to a detection system to track down the presence of electromagnetic waves and radioactivity.

The categories of the logging measurements that arise from *natural* or *spontaneous* phenomena are: Natural gamma radioactivity, Spontaneous potential (SP), Temperature of formation, Hole-diameter (caliper log) and the Hole inclination (deviation log).

The categories of the logging measurements that arise from *induction* phenomena are: Electrical (resistivity, conductivity, dielectric constant), Nuclear (density, photoelectric absorption, hydrogen index, macroscopic thermal neutron capture cross-section, elemental composition, proton spin relaxation time) and Acoustic measurements (acoustic velocity, acoustic-signal amplitude, well seismics).

2.2.1 The Spontaneous Potential Log

Spontaneous Potential is proved to be a considerable useful tool that permits the efficient collection of a substantial data set. Readings of spontaneous potential can give strong and significant evidence about the indication of lithology, porosity and permeability of the different drilled formations. Conclusions about the location, the formation water salinity and hence the formation oil saturation along the drilled hole are made. The drilling operations are performed in order to find the pay zone rich in hydrocarbons. Readings and core analysis samples are inspected and analyzed in order to make correlations between the well to generally characterize the constituting rock properties ([8]).

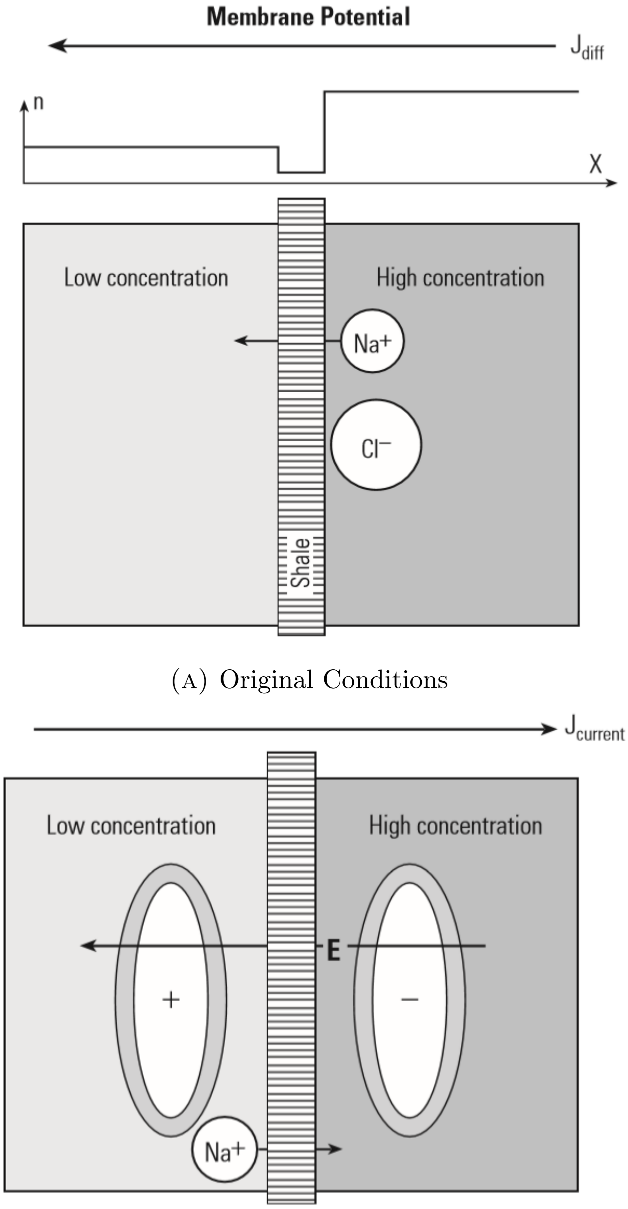
Principle

Continuous recordings of the spontaneous potential include the electrochemical potential difference, measured with a voltmeter, between a single electrode in the borehole and a ground referenced electrode placed at the surface. Electrochemical potentials of interest are the *liquid junction potential* and the *membrane potential.*

Liquid junction potential: Let's consider two sodium chloride solutions and a membrane barrier separating the two different concentrations. Then, the higher concentration solution's ion will tend to drift to the less concentrated solution, since the Na^+ alacrity is slower than the Cl^- ions, thus creating a *liquid junction potential.* The maximum liquid junction potential will be measured if the salinity between the mud filter (less concentrated solution) and the formation water (more concentrated solution), is great.

Membrane potential: Created in molecular constructions between shale and sandstone beds. In figure 2.2 a semipermeable shale barrier acts like an ionic sieve and separates the two different salinities solution. The less mobile Na^+ ions are travelling through the membrane more rapidly that the Cl^- ions since the shale barrier is more permeable to Na^+ ions than to Cl^- ions. In figure 2.2a, the current density of the diffusing particles is J_{diff} and n is the particles concentration. At this point, the negative charge causes no movement of the Na^+ and Cl^- ions in the region. In figure 2.2b, a charge separation occurs when an electric field is applied. The magnitude of the ionic current $J_{current}$ increases and the Na anions are passing to the right region while the Cl cations are slowing down to the left until the anions and cations reach an equilibrium, thus creating a *membrane potential.*

Factors affecting the measurements



(A) Original Conditions

(B) Dynamic Conditions

FIGURE 2.2: Generation of the membrane potential. Figure retrieved from [32]

Some typical responses of a Spontaneous Potential log are illustrated in the figure 2.3. Correspondence of spontaneous potential measurements depend mainly on the following addressing factors:

1. Thickness of the permeable bed; when the SP curve is narrowed then it requires a bed thickness correction.

2. Bed resistivity; high resistivity levels reduce the reflection of the SP curve

3. Shale content; reduces the SP deflection

4. Hydrocarbon content; reduces the SP deflection

5. Mud and water resistivity; oil-based mud can not be used when SP is recorded since electrical conductive paths through the mud are blocked.

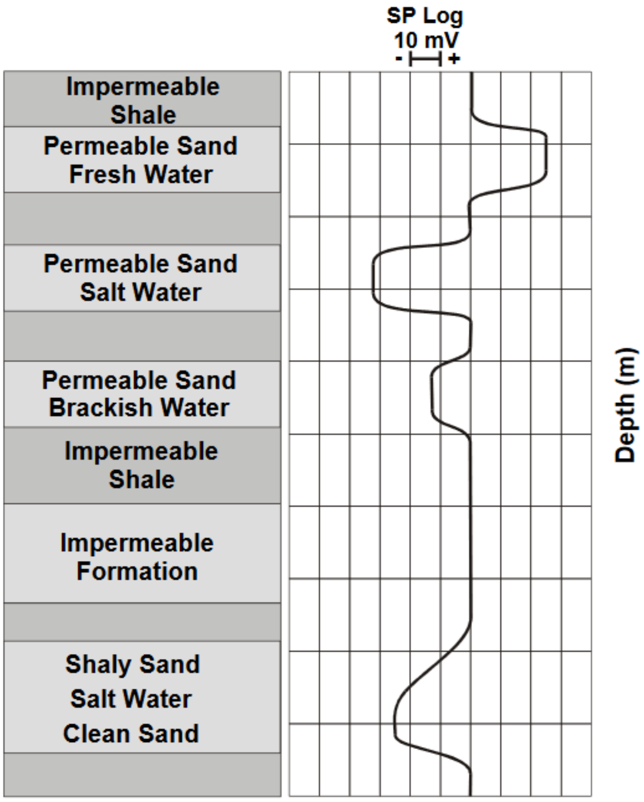


FIGURE 2.3: Common responses of a Spontaneous Potential log. Adapted from [36]

2.2.2 The Gamma Ray Log

Gamma Ray measurements are practically used for three main reasons. Readings of gamma ray result in evaluation of the shale content of a formation or a shale reservoir. In other applications, it can be used for analysis of the lithology and mineralogy of the drilled formation. Moreover, it can be used for stratigraphic correlations. Those correlations are based on shale distributions in the studied geological area and the age of shale. When correlations are made, we need to take into account the contamination from non-shale radioactive sources ([7]).

Principle

The *gamma ray* log records the total natural gamma radiation emitted from isotopes of three main source elements: ^{40}K(potassium), ^{232}Th(thorium), ^{283}U(uranium).
The gamma rays emitted from an isotope in the formation gradually discrete in energy. Hence, the gamma ray intensity that the log measures is a function of: (a) the initial gamma ray emission ; and (b) the Compton scattering in the formation that the gamma rays encounter between the gamma emission and the detector. In figure 2.4, an illustration of the Gamma Ray log in comparison with the Spontaneous Potential and Caliper log is presented. On average, a shale contains 6 ppm uranium, 12 ppm thorium, and 2 ppm potassium. The magnitude of gamma ray measurement in clean limestone, salts, coal, anydrite, shaly sand and dolomite is usually small while in case of shale is relatively large.

Factors affecting the measurements

The dependency of the measurements responds mainly to the concentration of K, Th, U occur in the formation. Other minor factors including :

- Interfering peaks close to the principle peaks in each window of energy band.
- Two "escape peaks" related to each principle high energy peak, resulting in Th interference in the U window, and Th and U interference in the K window.
- The bore-hole size, tool position (centering/eccentricity).
- Mud weight, casing size and weight and cement thickness.

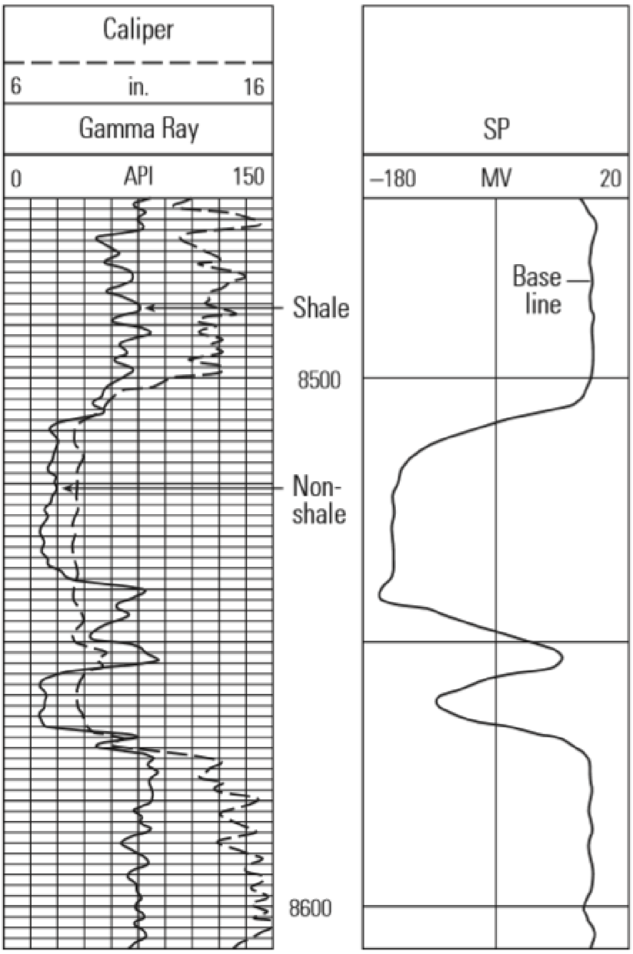


Figure 2.4: Caliper and gamma ray curve in comparison with the spontaneous potential curve. The studied formation is referred to clean and shale zones. Figure retrieved from [32]

2.2.3 Induction Log

Induction logs are a type of Resistivity log. Induction logging devices are recommended when the drilling fluid is oil-based, air or gas-based mud that do not conduct electricity. Induction logging tools measure the formation's resistivity and conductivity for saturation estimates (differentiate the water-bearing zones from the hydrocarbon-bearing ones) when induced by a focused magnetic field.

Principle

The *Array Induction Log (AIL)* tool includes of a multiple transmitting coil and eight groups of receiving coils, spacing from 6 inches to 6 ft at three or one frequency. Each array consists of a single transmitter coil and two receivers. The tool measures the conductivity of the formation by corresponding to multi-frequency and multi-coil pairs.

Different resistivity curved with three vertical resolutions $1, 2$ and 4 ft are obtained at different investigation depths, $10, 20, 30, 60$ and 90 inches. Other induction tools produce two types of signals; the inphase (R-signal) and the quadrature (X-signal) induction signal. The inphase signal is presented on standard dual induction—SFL log presentations while both the inphase and quadrature signal are combined during advanced processing in the logging tool itself to run real time corrections for environmental and geological conditions. Modern induction logs include several sets of coils with focused currents. Thus, the effects of the borehole and surrounding formations are minimized. Most modern resistivity log suites use different depths of investigation with various combinations of measurements ([4], [54]).

Factors affecting the measurements

Correspondence of conductivity measurements depend mainly on the following addressing factors:

1. Mud inside the borehole; recommended when the drilling fluid is oil-based, air or gas-based mud
2. Bed thickness; is not recommended in resistive and compact formations since the signal level is low.
3. Formation resistivities; dramatic increase of the difference between apparent and true formation resistivity.

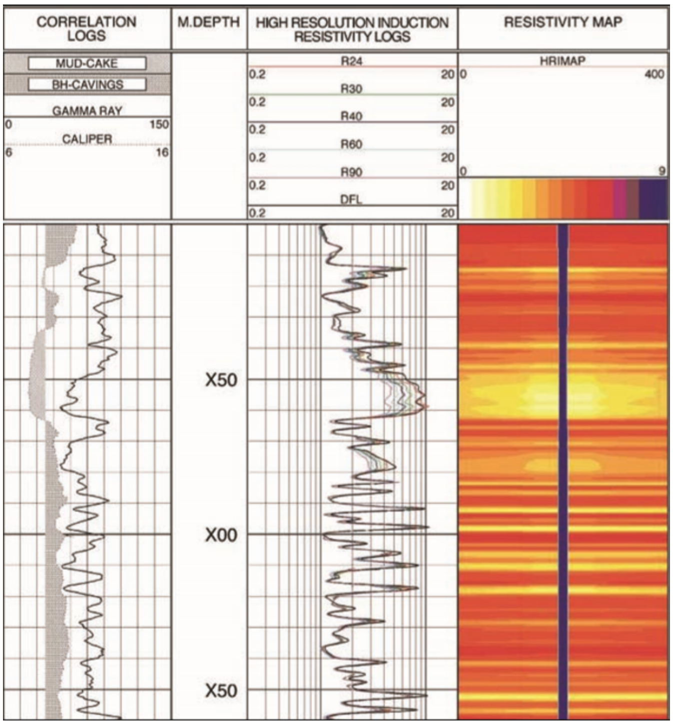


Figure 2.5: Example of high resistivity induction log from Halliburton (Oil field service company). Figure retrieved from [8]

2.3 Well log quality

The key data points that we can collect from a drilled formation are the measurements of a hydrocarbon-bearing productive zones; the definition of the reservoir type and thickness; the distinct prediction of porosity and permeability of the prospective zones; the determination of the fluid type, flow and migration through the pores of the complex geological environment.

In a process of planning and conducting a well log operation, well log quality control is a subject of major interest. Acquisition problems including skips, nose, spikes and missing data result in data misinterpretation. The majority of well logs include systematic error and environmental corrections. Those corrections are not able to completely eliminate the occurred errors. Nevertheless, the measurements' correctness becomes crucial when observation points are very close to the decision making threshold. Additionally, we need to clarify that by increasing the frequency of logging may not be a guarantee of increased knowledge of information and by no means does reduce the overall logging costs.

Chapter 3

An Introduction to Time Series

3.1 Stochastic Processes and Time Series

Time series are sets of observations taken sequentially at a specified time vector $\mathbf{t} = (t_1, \ldots, t_N)^\top$. Observations that contain data points taken continuously over some time interval are referred to as **continuous**-time series, while observations that consist of individual data points separated by time intervals are referred to as **discrete**-time series (e.g. seismic imaging data) ([100], [61]). An example of a discrete time series is illustrated in figure 3.1. In this thesis the term "data" will always refer to acquired observations as a discrete sequence at uniform intervals. *Time series analysis* is the statistical methodology pertained to the analysis of such sequence of data.

The sequence of variables $Y_1, Y_2, \ldots, Y_N$ or (Y_t), at times $t = 1, 2, \ldots, N$, is called a time series, where N is the number of observations of the time series Y_t. The study of a time series requires the collection of a large number of observations taken by a specific time frequency. We often use time series analysis to understand the past and therefore to predict the future. One basic feature of a time series analysis is the interpolation of the observed correlation between two successive values.

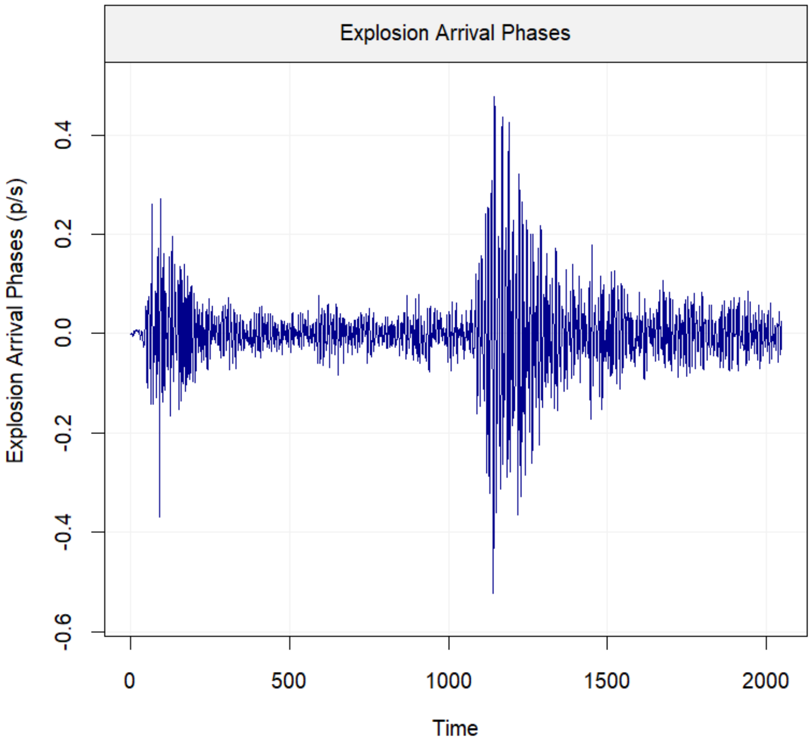


FIGURE 3.1: Example of displacement recorded during an explosion. Data retrieved from "astsa" package in R statistical computing environment.

3.2 Fundamental Concepts

In this section an introduction on the principal points of the statistical moments will be made. Statistical moments are functions expressed explicitly by an analytical expression[1] and they are often used to express statistical characteristics of a random field. There are four moments of a probability distribution that are briefly overviewed. The first moment is the *mean*, the second moment is the *variance*, the third moment is the *skewness* and the fourth moment is the *kurtosis*. The first two statistical moments provide information about the appearance of a distribution, whereas the third and the fourth moments provide information about the symmetry and shape of the distribution ([39]).

[1] Also defined as deterministic functions.

3.2.1 Moments

Let us consider a stochastic process $\{Y_t$:t= $0, \pm 1, \pm 2, \pm 3, ...\}$. Then we can define the **mean** function as:

$$\mu_t = E(Y_t) \tag{3.1}$$

for $t = 0, \pm 1, \pm 2, \pm 3, \dots$.

Generally, μ_t may differ at each time point t.

The **variance** of a random variable X can be determined as:

$$Var(X) = \sigma_X^2 = E[X - E(X)]^2 \tag{3.2}$$

or

$$Var(X) = \sigma_X^2 = E(X^2) - [E(X)]^2 \tag{3.3}$$

We call the **standard deviation**(σ_x) the positive square root of the variance of X.

The **standardized** version of X is described as:

$$X^* = \frac{(X - \mu_X)}{\sigma_X} \tag{3.4}$$

The **covariance** of X and Y is defined as:

$$Cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{3.5}$$

In time series analysis, the same function is called **autocovariance** and it is given by:

$$\gamma_k = Cov(Y_t, Y_{t+k}) = E[(Y_t - \mu_t) \cdot (Y_{t+k} - \mu_{t+k})] = E(Y_t Y_{t+k}) - \mu_t \mu_{t+k} \tag{3.6}$$

The **autocorrelation** function is defined as:

$$\rho_{t,t+k} = \text{Corr}(Y_t, Y_{t+k}) \tag{3.7}$$

for $t = 0, \pm 1, \pm 2, \pm 3, \dots$, where the *correlation coefficient* of Y_t and Y_{t+k}, is defined by:

$$\rho_{t,t+k} = \text{Corr}(Y_t, Y_{t+k}) = \frac{Cov(Y_t, Y_{t+k})}{\sqrt{\sigma_{Yt}^2 \sigma_{Yt+k}^2}}. \tag{3.8}$$

Note that the correlation coefficient satisfies

$$-1 \leq \mathrm{Corr}(Y_t, Y_{t+k}) \leq 1. \tag{3.9}$$

In the case of standardized variables Y_t^* and Y_{t+k}^*, then $\rho = E(Y_t^* Y_{t+k}^*)$.

Values of $\rho_{t,t+k}$ near ±1 signify strong *linear* dependence between random variables, whereas values below 0.2 signify low linear dependence. Values of Y_t,Y_{t+k} are *uncorrelated* if the autocorrelation function is equal to zero.

In order to examine the covariance function properties of every possible time series models, lets consider x_1,x_2,...,x_m and y_1,y_2,...,y_n are constants while t_1,t_2,...,t_m and s_1,s_2,...,s_n are time points, then:

$$Cov\left[\sum_{i=1}^{m} x_i Y_{t_i}, \sum_{j=1}^{n} y_i Y_{s_i}]\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_i Cov(Y_{t_i}, Y_{s_j}) \tag{3.10}$$

The **skewness** measures the asymmetry of the distribution and is defined as ([96]):

$$s_t = \frac{E[(X-\mu)^3]}{\sigma_x^3}. \tag{3.11}$$

This is estimated from a sample $(x_1, \ldots x_n)$ by means of the average

$$\hat{s}_t = \sum_{i=1}^{n} \frac{(x_i - \mu)^3}{n\,\hat{\sigma}_x^3},$$

where n is the sample size and $\hat{\sigma}_x$ is the sample estimate of the standard deviation.

If

- $s_t \in [-1, 1]$ then the distribution is *highly skewed*
- $s_t \in [-1, -0.5]$ or $s_t \in [0.5, 1]$ then the distribution is *moderately skewed*
- $s_t \in [-0.5, 0.5]$ then the distribution is approximately *symmetric*

The **kurtosis** measures the heaviness or the lightness of the tail of the distribution relative to the normal distribution of the same variance and is defined as:

$$k_t = \frac{E[(X-\mu)^4]}{\sigma_x^4}. \tag{3.12}$$

This is estimated from a sample $(x_1, \dots x_n)$ by means of the average

$$\hat{k}_t = \sum_{i=1}^{n} \frac{(x_i - \mu)^4}{n\,\hat{\sigma}_x^4},$$

where n is the sample size and $\hat{\sigma}_x$ is the sample estimate of the standard deviation.

If

- $k_t = 3$ then the distribution is *Gaussian*
- $k_t \in [3, \infty]$ then the distribution is *leptokurtic*
- $k_t \in [-\infty, 3]$ then the distribution is *platykurtic*

3.3 Time Series Analysis

There is a distinguished remark that we need to take under consideration when dealing with time series data. The fact that there is the profound relationship between the impute current values to that of its preceding or later data points that affect the parameter we are interested in ([20], [61]).

In figure 3.2 the main time series components are illustrated.

1. Trend - the increasing or decreasing overall direction of the value in the series, over time.
2. Seasonality - repeating variations or short-term cycles in the series caused by re-occurring events.
3. Random component - random shifts in the time series that may be ascribed to noise or other unsystematic events.

Other time series components include:

1. Outliers (Special events) - abnormal observations due to random or special events. Special attention needs to be taken when analyzing or interpreting the outliers in order to be effectively characterized.

2. Level shifts - sudden fluctuations on the mean time series level.

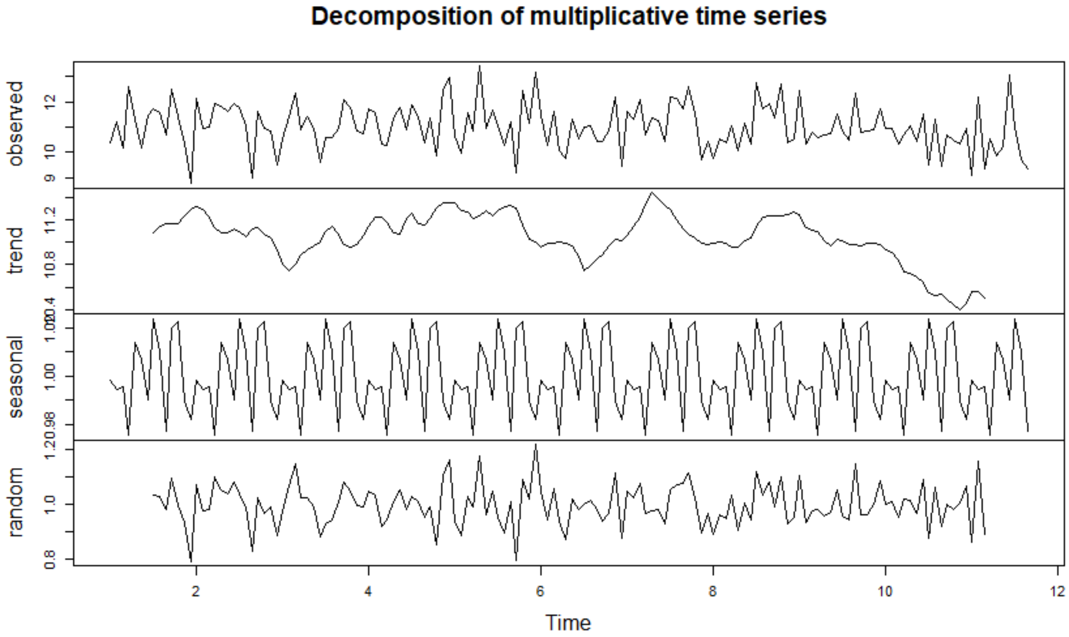


FIGURE 3.2: Decomposition of multiplicative time series. The number of observations is equal to 150 and the number of observations per unit of time is equal to 14.

At this point we need to elucidate the difference between the three dominant types of time series, stationary, additive and multiplicative. Their composition is considered as follows:

The stationary model's main assertion is that the mean, variance and autocorrelation (see section 4.2) are constant through the course of time.

stationary = seasonality and/or noise

The main characteristic of the additive model is that all components are independent to each other and are implemented in the same attributed unit of measurement.

additive = trend + seasonality + noise

On contrast, in the multiplicative model, only the trend component has the same attributed unit of measurement of the observed time series, while all the other components are independent to this same unit.

*multiplicative = trend * seasonality * noise*

In the additive model case, the trend component doesn't affect the seasonality when calculating the values of the series. This assumption can be verified, especially when analyzing natural phenomena.

3.3.1 Trend Removal

We need to take into consideration that the data gathered through logging is liable to sampling or measurement error. That is, real data often exhibit more complicated trend models. By means of simplicity, the trend function u_x will be modelled by low-order polynomials of the coordinators of the series' data points in order to ensure consistency of interpretation of the spatial direction in the data, and on the other hand, to examine under which possible circumstances the effect of a trend on a variogram (see section 5.3) might be bypassed to allow a sufficient analysis of the data. In Table 3.1 some common 1D trend models are shown. The selection of the best trend model is done by means of Least Squares Errors (LSE). An indicative plot appears in Figure 3.3.

Model	**Trend Function (1D)**
Mean	$u_x = a_0$
Linear	$u_x = a_0 + a_1 x$
Quadratic	$u_x = a_0 + a_1 x + a_2 x^2$
Cubic	$u_x = a_0 + a_1 x + a_2 x^2 + a_3 x^3$
Quartic	$u_x = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$

Table 3.1: Common trend functions

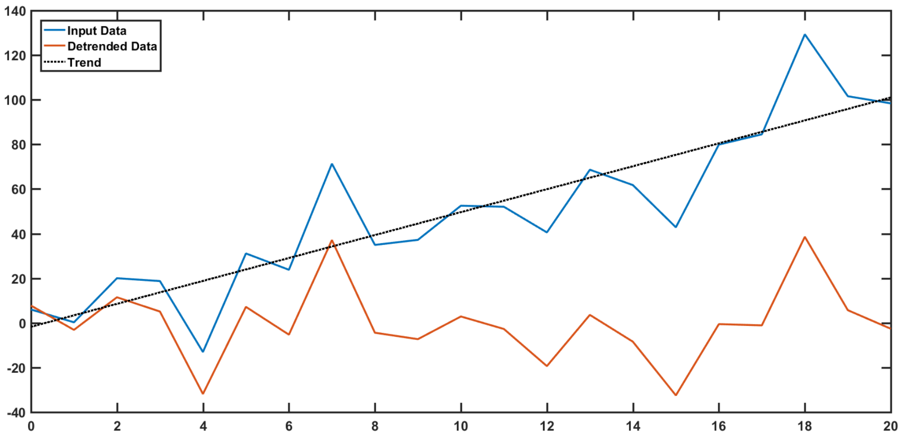


Figure 3.3: Time series with linear trend (blue line) and residuals (red line) after removing the trend (black dotted line).

3.3.2 Stationarity

The term **stationary process** implies that the properties of the process do not change over time ([40], [53]). A process Y_t is said to be *strictly stationary* if the joint distribution of Y_{t_1},Y_{t_2},...,Y_{t_n} is equal to the joint distribution of Y_{t_1-k},Y_{t_2-k},...,Y_{t_n-k}, at time points t_1,t_2,...,t_n and all possible time lag k. Thus, it follows that $E(Y_t) = E(Y_{t-k})$ and $Var(Y_t)$=$Var(Y_{t-k})$ for all t and k so that the mean function is constant and the variance is also constant over time. On the contrary, a process Y_t, is said to be *weakly stationary* if it has the same mean value, μ, at all time points; it has the same variance, γ_0, at all time points; and $Cov(Y_t, Y_{t+k}) = \gamma_k$ for all lags k, meaning that the covariance of the values at any two time points, t,$t-k$ depend only on lag k.

Stationarity Tests

Fitting a model in time series usually implies the prerequisite that the time series are stationary. It is possible to check the stationary behaviour by using a variety of tools. Those include the Ljung-Box test; the Augmented Dickey-Fuller (ADF) t-test; the Kwiatkowski-Philips-Schmidt-Shin (KPSS) test; the Wavelet Spectrum Test and the Priestley-Subba Rao (PSR) test ([70]). In this thesis, the stationary or nonstationary behaviour will be determined by the Kwiatkowski-Philips-Schmidt-Shin (KPSS) test.

In the KPSS test the model is based on linear regression and the represented component are the sum of three parts: the deterministic trend (t), a random walk process (r_t), and a stationary error (ϵ_t) of the first equation, estimated by the ordinary least squares regression (OLS) and by stationary assumption. The model is being described by the following equation:

$$y_t = at + r_t + \epsilon_t \tag{3.13}$$

when $r_t = r_{t-1} + u_t$. The component u_t is the error term of the second equation, by assumption of an i.i.d. series. If $a = 0$, then y_t is stationary around r_t, alternatively, if $a \neq 0$, the y_t is stationary around a linear trend ([89]). The data is usually log-transformed in order to eliminate any exponential trends and transform them into linear ones.

KPSS test often erroneously rejects the hypothesis that the data can be modeled as a stationary time series. This type of error can be prevented by combining the results of the KPSS and the ADF tests.

Chapter 4

Non-stationary Time Series

4.1 Simple Time Domain Models

The main purpose of time series analysis is to develop a mathematical model that provides plausible definitions for source data of a relevant stochastic process ([84]). One possible way to define the stochastic process is to determine the *joint probability density function* of the sequence of variables $\{Y_1,Y_2,\ldots,Y_N\}$ that can be described as

$$f(Y_1, Y_2, \ldots, Y_N) \tag{4.1}$$

If the probability density function were specified, then a future value point of the time series could be easily assessed at a particular probability. However, it is impossible to completely identify those multivariate distributions due to high number of parameters that they contain.

In this section one group of simple time domain models will be analyzed. Those models are used to produce more advanced models.

4.1.1 Independently and Identically Distributed

The term "independently and identically distributed" (iid) implies that the random variables of the sequence $\{Y_1,Y_2,\ldots,Y_N\}$ have the same distribution and also are mutually independent.

Let us suppose that $\{Y_1,Y_2,...,Y_N\}$ are *i.i.d.* with the same distribution as a random variable Y, then the probability distribution of this stochastic process is equal to the product of single probability density functions:

$$f(Y_1, Y_2, .., Y_N) = f(Y_1) \cdot f(Y_2) \cdot ... \cdot f(Y_N) \tag{4.2}$$

thus,

$$E(Y_1 + Y_2 + ... + Y_N) = N \cdot E(Y) \tag{4.3}$$

$$Var(Y_1 + Y_2 + ... + Y_N) = N \cdot Var(Y) \tag{4.4}$$

4.1.2 White Noise

A *white noise* time series is an example of a stationary process. We assume that the $\{e_t\}$ so-called *white noise* process has zero mean and denote σ^2 variance for all t, respectively:

$$\{e_t\} \sim WN(0, \sigma^2) \tag{4.5}$$

where W denotes the white independent noise, thus we write $W \sim i.i.d.(0, \sigma^2_W)$,([16]).

In the case of a *white noise*, the previous values of a time series cannot be properly processed in order to predict a future value, thus forecasting is impossible. The residuals of a typical regression describe an example of *white noise* whereas define *random errors*, *stochastic shocks* or other *innovations*. White noise can be used for synthetic data simulation.

As mention in the 4.1.1 section , an *i.i.d.* process is a case of *white noise*. A sequence of random variables $\{e_t\}$ is *i.i.d.* if:

$$E(e_t) = \mu, constant \tag{4.6}$$

$$\gamma_0 = E(e_s^2) = \sigma^2_{e_s}, \forall t \tag{4.7}$$

where e_t is independent of e_s for all t and s , and $t \neq s$. If the values of a time series w_t follow a standard normal distribution:

$$w_t \sim N(0, \sigma^2) \tag{4.8}$$

then the series is known as *Gaussian white noise* (Figure 4.1).

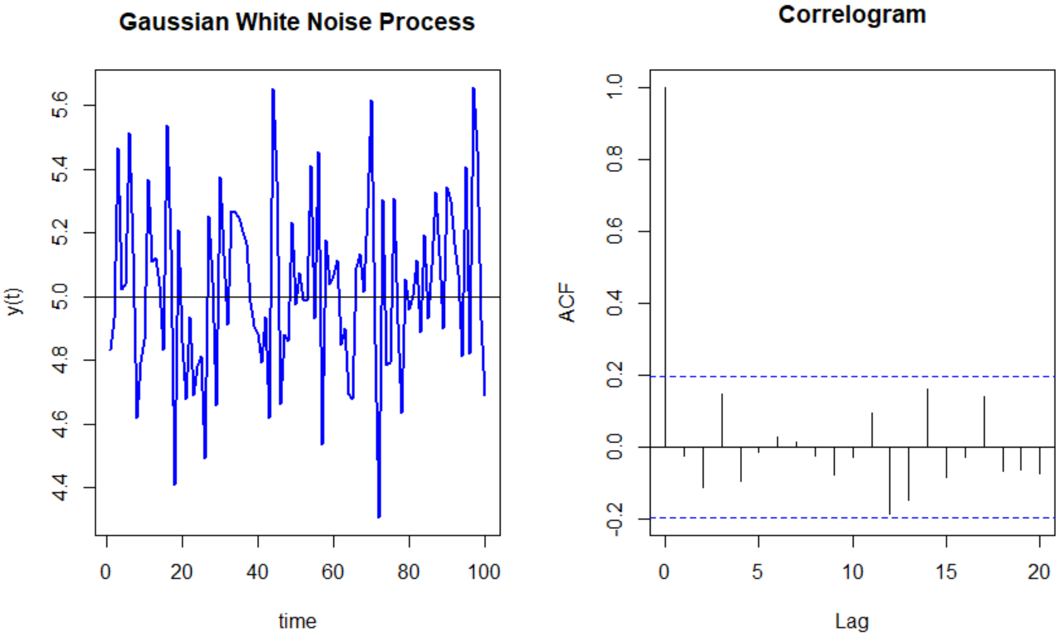


FIGURE 4.1: Simulation of N=100 random values of a Gaussian white noise (w_t) with mean $\mu = 5$ and standard deviation σ_{w_t}=0.3 (left). The auto-correlations of a simulated white noise series are all zero except at zero lag (see ACF plot).

4.1.3 Random Walk Model

A *random walk model* describes a series in which the change from one time point t to another time point $t + 1$ are random. It is defined as the time series Y_t that results when a completely random displacement ε_t is added to the previous Y_{t-1} according to equation (4.9):

$$Y_t = Y_{t-1} + \varepsilon_t, \tag{4.9}$$

for $t = 1, 2, 3...n$, with Y_0=0 and ε_t $i.i.d$N(0,σ^2_ε) variables.

A simple example of a random walk model is described in figure 4.2 where we can assert that an individual is walking into a path. The probability of taking a step back, forward or move in the right or left direction is equal.

Starting from t=1 and then using the successive substitution method in 4.9 we may rewrite it as follows:

$$Y_1 = Y_0 + \varepsilon_1 \tag{4.10}$$

$$Y_2 = Y_1 + \varepsilon_2 = Y_0 + \varepsilon_1 + \varepsilon_2 \tag{4.11}$$

$$Y_t = Y_0 + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_t = Y_0 + \sum_{i=1}^{t} \varepsilon_i \tag{4.12}$$

A *random walk plus drift* model is given by (Eq.4.13):

$$Y_t = \delta + Y_{t-1} + \epsilon_t \tag{4.13}$$

for t=1,2,3...,n, with Y_0=0 and ε_t $i.i.d$ $N(0, \sigma_\varepsilon^2)$ are the white noise innovation terms.

The general solution of the equation 4.13 results from the same method of successive substitution implemented in equation 4.9. Thus,

$$Y_t = Y_0 + \delta t + \sum_{i=1}^{t} \epsilon_i \tag{4.14}$$

for t=1,2,3...,n, with ϵ_t $i.i.d$ normally distributed $N(0, \sigma_\epsilon^2)$ innovation terms.

The random walk with drift is not stationary, which can be seen by calculating the mean $E(Y_t)$ and the variance γ_0 that are functions of time t ([33]):

$$E(Y_{t+n}) = Y_t + na + \sum_{i=1}^{n} E(\varepsilon_{t-i}) = Y_t + na \tag{4.15}$$

The first moment indicates that the process is not mean stationary:

$$Var(Y_{t+n}) = \sum_{i=1}^{n} Var(\varepsilon_{t-i}) = \sum \sigma_{\varepsilon_t}^2 = n\sigma_{\varepsilon_t}^2 \tag{4.16}$$

The second moment indicates that the process in not variance stationary and the variance changes depending on time t. A summarizing image of a simulated random walk with and without drift appears at figure 4.3.

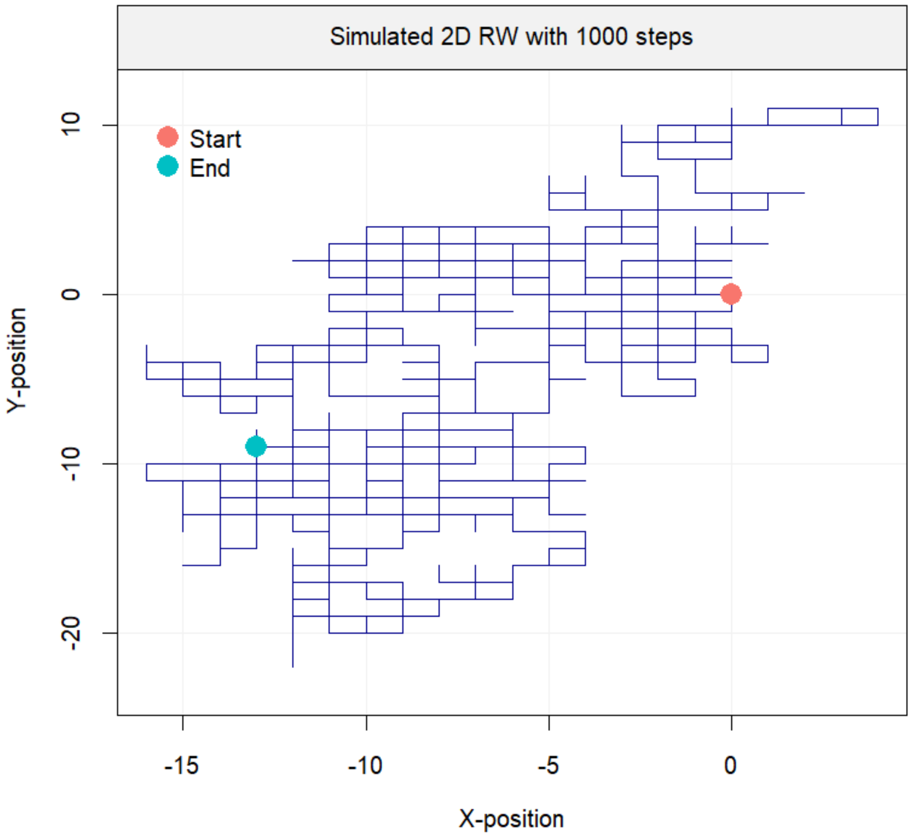


FIGURE 4.2: Realization of a simulated 2D random walk with 1000 steps.

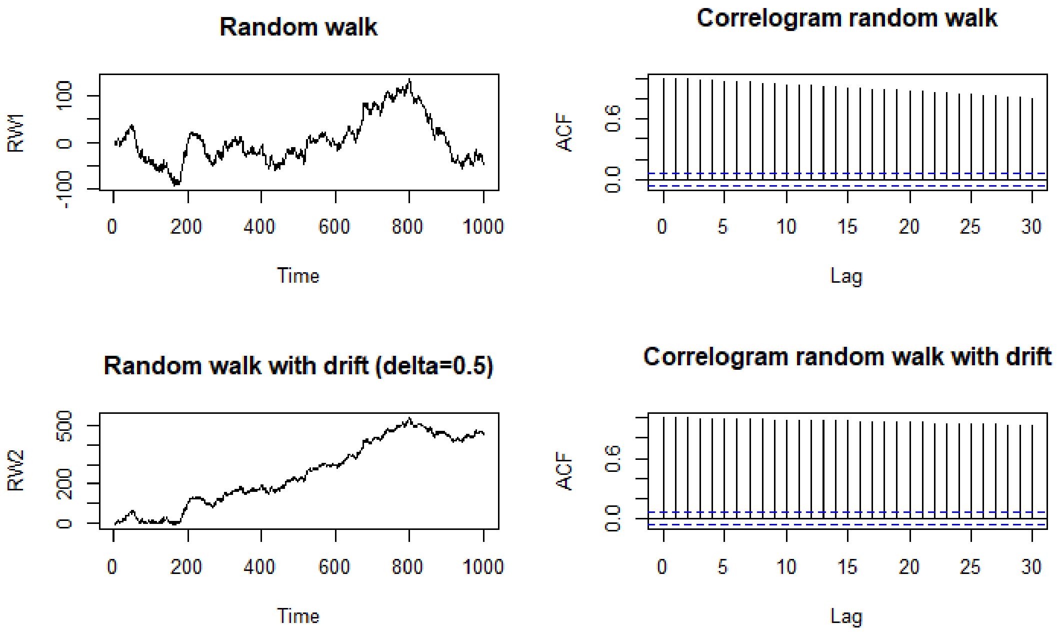


FIGURE 4.3: Top left: Simulated random walk of 1000 random values. Bottom left: Simulated random walk of 1000 random values with drift $\delta = 0.5$. The decline of the respective autocorrelation functions (ACF) (top right and bottom right plots) progresses slowly in both cases. The decline of the ACF for the random walk with drift progresses more slowly than the ACF of the random walk without drift.

4.2 Autocovariance and Autocorrelation Function

As previously referred in chapter 3, the mathematical definition of the sample covariance between two stochastic variables $x = x_t$, $y = y_t$ is set as follows:

$$c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{4.17}$$

Then the *autocovariance* at lag k is given by the following expression:

$$\gamma_k = Cov(y_t, y_{t+k}) = E[(y_t - \bar{y})(y_{t+k} - \bar{y})] \tag{4.18}$$

where $k = 0, 1, 2, ...$ and y_t and y_{t+k} are values of the time series at different times. The variance of the time series corresponds to the autocovariance at lag $k = 0$. For a stationary time series the variance is constant.

The *autocorrelation coefficient* at lag k is computed by means of the equation:

$$\rho_k = \frac{E[(y_t - \bar{y})(y_{t+k} - \bar{y})]}{\sqrt{E[(y_t - \bar{y})^2]E[(y_{t+k} - \bar{y})^2]}} = \frac{Cov(y_t, y_{t+k})}{Var(y_t)} \tag{4.19}$$

The cross-correlation statistics for positive values of lag k between the two variables is defined by ([67])

$$c_{xy} = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y}) \tag{4.20}$$

Thus, the cross-correlation function is given by the expression:

$$\rho_{xy} = \frac{c_{xy}}{\sqrt{c_{xx}\, c_{yy}}}. \tag{4.21}$$

4.3 Non-Stationarity

A time series is called *nonstationary* if the mean and/or variance change over time. A simple example of nonstationarity is the random walk with or without a drift. Non-stationary behaviour is common in nature, especially in the fields of economics ([5]) or signal analysis ([38]).

Well log data is often corrupted by the different types of noise that results in nonlinear and nonstationary characteristic behaviour, causing abrupt discontinuities in the series, whereas making the recognition of the formation boundaries a difficult and ambiguous process. Additional signal analysis techniques must be applied so as to take care of nonstationary well log signals.

4.4 Forecasting Models

The second part of this thesis frames on predicting the values for the continuous gaps in the well logs, acquired through sensing tools, providing better quality of information for the following steps in the interpretative formation evaluation techniques.

Many study-cases have chosen a variety of well logs and used varied algorithms and methods to identify the correlations between logs. The most commonly tested techniques are the ones of *generalised linear models* - Ordinary least squares (OLS), Bayesian Ridge Regression(BRR), and Random Sample Consensus (RANSAC) - and *non-linear models* - Artificial Neural Net-works (ANN), Random Forests (RF), and Gradient Tree Boosting (GB) ([52] ; [74]; [14]; [25]). There are other techniques for grid filling when the data set is incomplete using the Maximum entropy spectrum analysis, minimum curvature or natural neighbor shorting ([46]; [101]).

However, the science community considers the rapid growth of application of time series forecasting methods to be of great utilitarian value in order to fill missing data under specific mathematical statements. Nevertheless, this practice is still in its incipient stages due to some complex conditions concerning the analysis of data. For instance, a major issue includes the unravelling and extraction of the convoluted trend and seasonality of the well log data, which are often stymied by their high complexity.

4.4.1 Auto-regression

The general equation for linear regression is defined as:

$$y = \alpha + \beta x + \epsilon \tag{4.22}$$

where α is the intercept, β is the linear co-efficient, x is the independent variable and ϵ is the random noise. In most cases, there are more than one parameters that affect the study and thus multiple regression is preferably used. The following equation describes the above:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{4.23}$$

Generally, we may assert that the auto-regression (AR) model predicts the next point from the use of the previous value points of the data. This is defined by the equation:

$$X_t = c + \beta_1 * x_{t-1} + \beta_2 * x_{t-2} \ldots \beta_n * x_{t-n} + \epsilon \tag{4.24}$$

where c is a constant which is, in some cases, zero and the mean of the time series, x_{t-n} are the independent previous value points, β_n are the parameters of the model and ϵ is the error term which is also called *the innovation term*. A white noise model is typically used to describe the innovation term.

4.4.2 Moving Average

It's possible to replace the white noise series w_t by a moving average to smooth out fluctuations, trends and cycles of the time series ([83]). Suppose that V_t is the moving average of span N at time period t, then :

$$V_t = \frac{Y_t + Y_{t-1} + ... + Y_{t-N+1}}{N} = \frac{1}{N} \sum_{i=t-N+1}^{t} Y_i \tag{4.25}$$

where $Y_t, Y_{t-1}, ..., Y_{t-N+1}$ are the most recent N observations with weight zero to all other observations ([61]). An illustration of two moving average models of the white noise process shown in figure 4.1 appear in the graphs of figure 4.4.

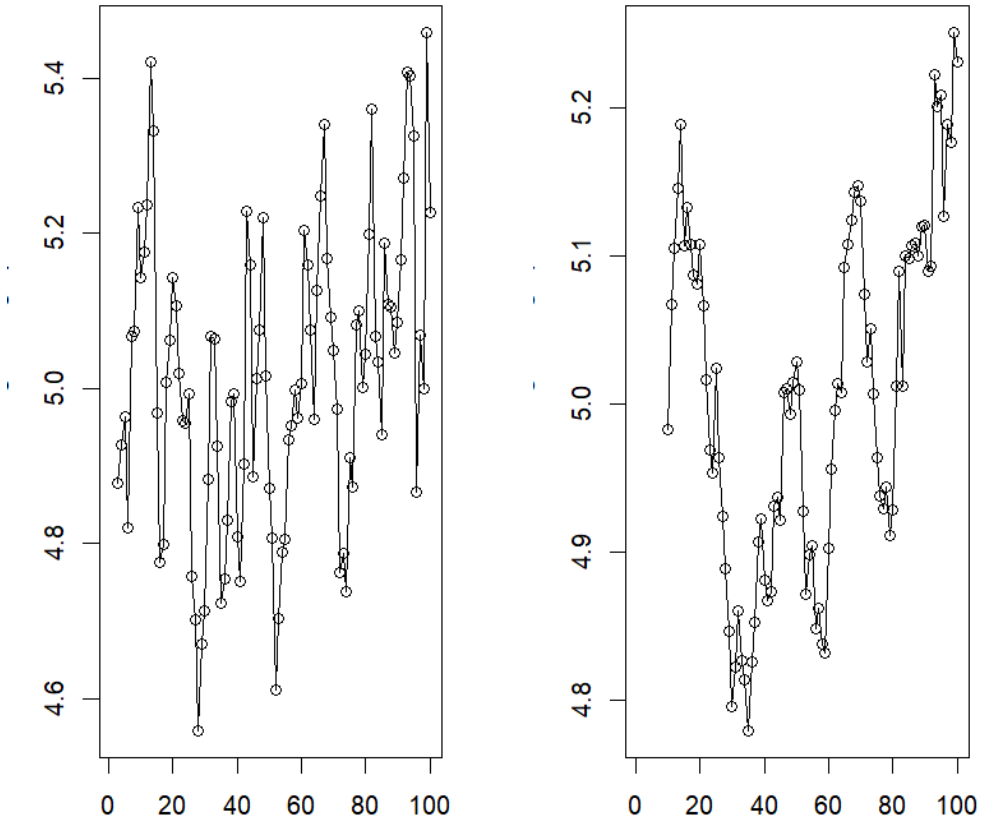


FIGURE 4.4: Moving averages of the white noise process shown in figure 4.1. Left: 3-point moving average. Right: 10-point moving average.

4.4.3 Autoregressive Integrated Moving Average (ARIMA)

The combination of a d-degree differencing with autoregresssion and a moving average model is called an **ARIMA**(p, d, q) model (Table 4.1, Table 4.2) and can be written as follows :

$$Y_t = \delta + \varphi_1 Y_{t-1} + ... + \varphi_{p+d} Y_{t-p-d} + ... + \varepsilon_t - ... - \theta_1 \varepsilon_{t-1} - ... - \theta_q \varepsilon_{t-q} \qquad (4.26)$$

where ϕ are the parameters of the $AR(p)$ component model and θ are the parameters of the $MA(d)$. The model implies the assumption of a stationary times series, without trend and a constant variance and mean throughout the series. In reality, thought, this is rarely the case. In order to model a non-stationary time series we initially remove the trend next we transform the data into stationary data set, we perform the model on the adapted data and finally the trend aspect is added back into the main series ([60]).

The model denotes the dependency of a number of previous values Y_{t-j}, $j = 1, \dots, p+d$, of a current random error ε_t and a number of previous errors ε_{t-j},$j = 1, \dots, q$.

p	Order of the auto-regressive part
d	Degree of first differencing involved
q	Order of the moving average part

Table 4.1: ARIMA model of order p,d,q ([40]).

White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Moving Average	ARIMA(0,0,q)

Table 4.2: Basic ARIMA models ([40]).

4.4.3.1 ARIMA Parameter Selection

Based on the aforementioned statements, it is important to carefully choose the optimum order of the ARIMA model. This selection process includes the determination of the Auto-correlation function (ACF), referred in equation 4.19, and Partial-autocorrelation function (PACF) in order to find the values p, d, q that optimize the metric of interest.

Auto-correlation is defined as the degree of correlation between the current observed data point and its previous or future point, thus is a measurement of the order of the dependence. The interval between the observed data point and its previous values used in measure the correlation is called the *lag*. Partial-autocorrelation is a measurement of correlation between observations' residuals with the next lag k value.

If we consider $m = p + q + P + Q$, where p, q are the non-seasonal components of an ARIMA model and P, Q are the seasonal components of an ARIMA model (referred as SARIMA) then the optimum chosen components are the ones that minimize the Akaike Information Criterion (AIC) (see section 4.6).

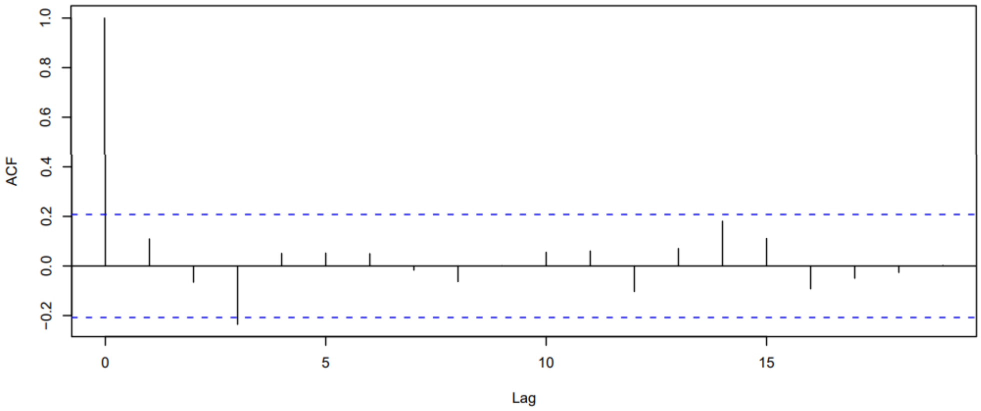


FIGURE 4.5: Correlogram estimate of the auto-correlation function.

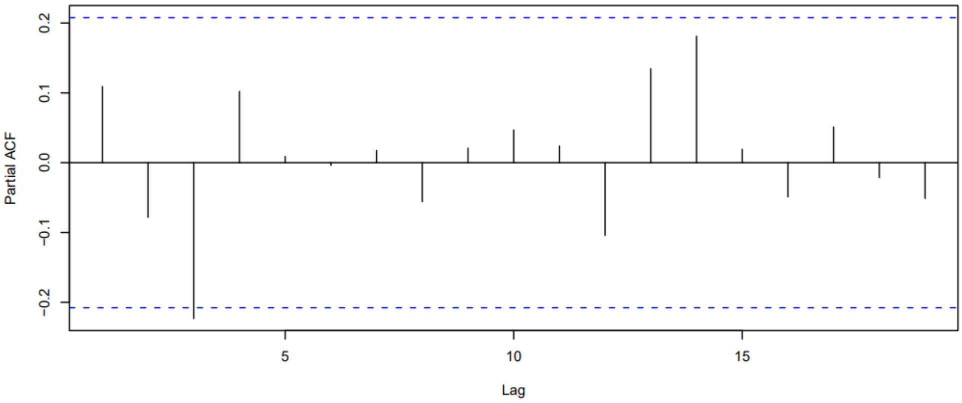


FIGURE 4.6: Correlogram estimate of the partial-autocorrelation function

4.4.4 Model Selection

The following step in data analysis is to obtain a model with good prediction accuracy. This statistical technique of evaluation and model selection is called *Cross-Validation (CV)*. It is crucial to identify and include all the important factors and interaction and at the same time omit the unimportant ones. In practise, the data set will be equally partitioned into two segments: one used for *training* and the other used for *testing*. Various procedures of different validation methods are proposed in order to estimate accuracy. The least biased accuracy types of cross-validation is the *regual cross-validation*, the *leave-one-out Cross-Validation (LOOCV)*, the *leave-p-out Cross-Validation (LpOCV) and the* k-fold Cross Validation (k-fold CV) ([27], [45]). In this analysis we are concerned with the common validation measures

Validation Measures

A number of validation measurement formulas provide an evaluation of association between each model. The tested equations are the Mean Absolute Error (MAE), the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), the Pearson's correlation coefficient (r_P) and the Spearman's correlation coefficient (r_S). The regression error metrics are useful for evaluating the model's precision. Pearson's and Spearman's correlation coefficients benchmark linear and monotonic relationships between the predicted and estimated variables, respectively. Let y be the observed values, $\hat{y_t}$ be the predictive values and n be the number of observations. Writing the formulas explicitly, we have:

$$\mathbf{MAE} = \frac{1}{n}\sum_{t=1}^{n}|\hat{y_t} - y| \tag{4.27}$$

The MAE is used to measure the prediction's closeness and accuracy. MAE gives more weight to the average magnitude of errors between predicted and the corresponding observations.

$$\mathbf{MSE} = \frac{1}{n}\sum_{t=1}^{n}[\hat{y_t} - y]^2 \tag{4.28}$$

The MSE refers to the sum of squared bias and variance and is a useful metric for providing indirect mathematical insight about the behaviour of the natural processes.

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}[\hat{y_t} - y]^2} \tag{4.29}$$

The RMSE is used to calculate the prediction's closeness and accuracy while measuring the quadratic average magnitude of relatively large errors.

$$\mathbf{MAPE} = \frac{100}{n}\sum_{i=t}^{n}\frac{|\hat{y_t} - y_t|}{|y_t|} \tag{4.30}$$

The MAPE expresses the mean absolute percentage error and has the advantage of being scale independent. Very high actual values will result in extremely low error and vise versa.

$$r_P = \frac{\sum_{t=1}^{n} \left(\hat{y}_t - \bar{\hat{y}}\right)\left(y_t - \bar{y}\right)}{\sqrt{\sum_{t=1}^{n} \left(\hat{y}_t - \bar{\hat{y}}\right)^2 \sum_{t=1}^{n} \left(y_t - \bar{y}\right)^2}} \tag{4.31}$$

The Pearson's correlation is frequently used to measure the degree to which two variables are correlated, thus querying their linear dependency. The closer the Pearson's product (r_P) is to 1 or -1, the more accurate the linear fit is.

$$r_S = 1 - \frac{-\sum_{t=1}^{n}(d_t)^2}{n(n^2-1)} \tag{4.32}$$

The Spearman's correlation measures the strength of association between two sets of continuous variables $\hat{y}_t$ and y_t where d_i indicates the differences between the ranks of $\hat{y}_t$ and y_t.

4.5 Correlation Techniques

K-Nearest Neighbor(KNN)

The K-Nearest Neighbor (KNN) technique will be used in order to correlate two discrete physical property logs of two distinguished oil and gas wells that have different depth steps resolutions and are located in the same studied area. This technique uses a *k-dimensional tree* (also called k-d tree) to store and organize spatial data in a k-dimensional space.

Given a set Ω of points n, we need to rapidly find the closest point in the metric space (k-neighbor), more simply, we need to find the k objects nearest to the query point q ([3]).

First, the calculation of the *Euclidean* distance between the numeric values of the data points is implemented. The algorithm will compute the distance between each data point and the test data. Finally, the data points that have the highest probability in being similar to the tested data are classified. The mathematical formula of the Euclidean distance is shown below.

$$\begin{aligned} d(p,q) = d(q,p) &= \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_i-p_i)^2} \\ &= \sqrt{\sum\nolimits_{i=1}^{n} (q_i-p_i)^2} \end{aligned} \tag{4.33}$$

The KNN is a non-parametric technique, which means that no additional assumptions about the data sets needs to be taken. Yet, certain caution must be taken when using this algorithm, especially when data points are in boundary which can lead to misclassification.

Classification measures for missing values

Missing data on model induction is a rather major drawback concerning the fields of Machine Learning (ML); Data Mining (DM) and other correlated areas. Some good reference in the area are : [9], [56], [17], [55], [37].

In the scientific field of applied geophysics the majority of data sets are obtained from measurements of natural (or spontaneous) phenomena and induced phenomena at prediction time. The conditions under which the open-hole and cased-hole measurements are made often cause the data to have several gaps. Classification measures in such cases is useful so as to classify the unknown values.

Cost function

When performing multiple well logs correlation a practical solution for optimizing a reasonable computational cost is of high significance for computer implementation. In his study, [50], proposes the *dynamic depth warping method* where a pair of well logs A(n) and B(m) with the n-th and m-th point in the well A and B, respectively, the cost function is a difference metric $d(n,m)$ of matching points between the two well logs, $|A(n)-B(m)|$. The cost function is defined as:

$$d(n,m) = \frac{\sqrt{\sum_{i=1}^{k} |A_i(n)-B_i(m)|^2\, W(k)}}{k} \tag{4.34}$$

for $i = 1, 2, ..., k$ logs and $W(k)$ the weighting coefficient for the k-th log.

More resent studies emphasize in the intrinsic importance of utilizing deliberated training models of neural networks. A more comprehensive description of this practise can be found in [31], [10].

4.6 Fitting Criteria

Maximum Likelihood Estimation

The Maximum likelihood estimation method (MLE) is an indispensable tool used for parameter estimation and is preferred for a variety of mathematical modelling techniques when the data is non-normal. Suppose that x_i are $i.i.d$, then the likelihood is defined as:

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

The $L(\theta)$ signifies the observing probability f the given data as a function of θ. In order to maximize the product of the previous function, we maximize the log likelihood, using the fact that the logarithm is an increasing function:

$$l(\theta) = \sum_{i=2}^{n} \log(f(x_i \mid \theta))$$

This method can be performed on data so as to extract as much information as possible.

Information criteria are useful for model selection. In this thesis the AIC and BIC criteria are used to determine which distribution model is most appropiate for a given set of stochastic variables. The mathematical expressions of these criteria are written below:

AIC : Akaike Information Criterion

The AIC approach aims to clarify the best fitted model of the observed data via the principles of MLE and negative entropy [1].

$$AIC = -2 \log L(\theta) + 2k \tag{4.35}$$

[1] Measure of diveregence of normality ([15])

BIC : Bayesian Information Criterion

The BIC approach aims to identify the best fitted model of the observed data by comparing probabilities, under the consideration that each of the candidate models is the true model.

$$BIC = -2 \log L(\theta) + k \log(n) \tag{4.36}$$

Concisely, both criteria can be used in order to reassure the robustness of a model's fitness. These criteria are giving optimal model selection results under defined conditions, whereas fail to fully describe the complexity of a real model problem. Hence, the understanding of the nature of the problem is necessary.

Goodness-of-fit statistics

The measurements of goodness of fit of a statistical model is an important step on data analysis in order to examine if the initial hypotheses about the observation process fit a model adequately as well as if we can consider it consistent with those hypotheses. The following tests can be used for such a reason are the *Kolmogorov–Smirnov test*; the *Cramér–von Mises criterion*; the *Anderson–Darling test*. In this thesis the Anderson–Darling test would be used for the analysis. All distributions tested for this particular thesis are fully specified in chapter 5.2.

Chapter 5

Data Analysis Processes

When it comes to data processing for interpretation, there is not a standard procedure for every data set. Usually the investigator follows a sequence of operations to result in correct conclusions. In time series analysis and forecasting there are some general accepted steps performed, including the evaluation and trend model removal and then residuals diagnostic processes. In the following section, a brief description of each procedure is presented.

5.1 Preliminary and Exploratory Analysis

Preliminary data analysis aims to provide summary statistics for all data and examine if there are issues that can affect the modeling processes. Univariate analysis refers to the analysis of data that contain only one variable. Multivariate analysis is the analysis which examines the relationship between two or more variables. The primary analysis includes both univariate and multivariate analysis ([21]).

Exploratory data analysis aims to provide information about the various characteristics of a data-set by displaying several graphical techniques and tools. The following tools are going to be used in this thesis.

Histograms

A histogram is a graphical display that forms the shape of a probability distribution function by plotting a number of observations from a distribution. We can define a histogram as a function that calculates a number of intervals n_i and then divides them into variable values. The calculated density histogram is a discrete function with values

$$\frac{f_i}{n(c_{R,i} - c_{L,i})}, \quad i = 1, \ldots, N_b \tag{5.1}$$

where f_i is the frequency of the data for each histogram class (bin) i, $[c_{L,i}, c_{R,i}]$ is the width of each bin (note that $c_{L,i+1} = c_{R,i}$), n is the number of samples and N_b is the number of bins.

QQ plots

A commonly used technique for informally calculating goodness-of-fit as well as estimating the scale and location for a family of distributions F, is called QQ plot or quantile-quantile plot ([2]). The scale parameter defines the heaviness of the tail. In some cases, is hard to judge the normality from a histogram. A normal QQ plot graphs the shape of the empirical distribution (y-axis) against the shape of a normal distribution (x-axis) thus provides a visual check in order to examine whether or not the points are close to a straight line. For the thesis's purposes we will adapt this method to the problem of detecting the lack-of-fit at the distribution tails.

PP plots

Probability-probability plots (also known as PP plots) are a graphical tool for interpreting CDFs of a family of distributions against one another ([97]). PP plots are well suited to compare probability distributions that are not overlapping. Notably, the PP-plot is sensitive to differences in the middle of a distribution, in comparison with those in the case of the QQ plot.

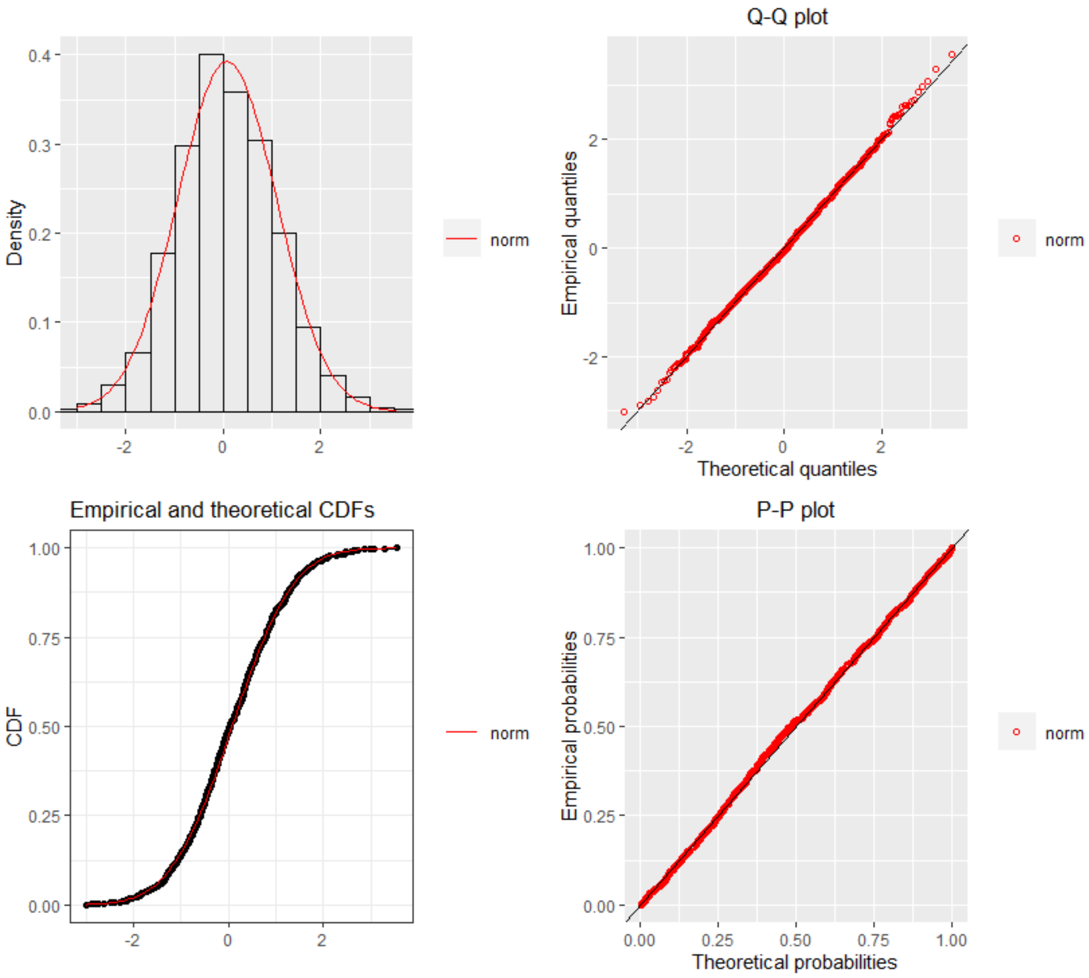


FIGURE 5.1: Example of 1000 generated random values from the standard normal distribution, with zero mean and standard deviation equal to one.

5.2 Probability Distributions

Many geophysical processes are usually modeled and based on the distributions[1] described in this section. Those distributions refer to stochastic processes. In this study, we will use the term *stochastic* instead of random to describe a non-random evolution of the natural process. Estimating the parameters of a distribution is a challenge. Those parameters are usually complicated functions that depend on the geophysical parameters of interest ([26]). The data is subject to a great degree of uncertainty that we wish to describe, in a simple and effective way. Thus, we use the *probability theory*.

While geophysical data-sets obtained by formation evaluation tools can compute natural properties or describe natural phenomena, we have to do some simplifications, as time discretization at the annual time scale ([47]), so that we can perform a classical statistics implementation of our data.

The typical elements of any distribution are variables included in the *probability distribution function* (PDF) and the *cumulative distribution function* (CDF). Even different order moments can be regarded as parameters that can make inferences about the location, scale and shape of the distribution. Common discrete distributions include the Binomial, Geometric, Logarithmic, Poison, Zipf and more, while common continuous distributions include Cauchy, Laplace, Gaussian (or Normal), Beta, Gamma, Student - t, Weibull, Pareto, Exponential, Gumpel and many more ([58]). Only a few of them will be analyzed for the propose of this thesis.

Cumulative Distribution Function

A $F_X(x)$ function is said to be a Cumulative Distribution Function (CDF) if it has the following characteristics:

1. $\frac{dF_X(x)}{dx} \geq 0$
2. $F_X(-\infty) = 0$
3. $F_X(+\infty) = 1$

The CDF function gives the corresponding probability of a set of random variables x that occur below a specific value and is expressed by the mathematical formula of :

[1] Primary statistical tool for analysing and illustrating raw data.

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{5.2}$$

Probability Density Function

A $f_X(x)$ function is called a Probability Density Function (PDF), for a sample area X, if it has the following properties:

1. $f_X(x) > 0$
2. $\int_X f_X(x)dx = 1$

The PDF function gives the probability of both continuous and discrete distributions within a specific range of values and is expressed by the equation of:

$$Pr[b \geq X \geq \alpha] = \int_{\alpha}^{b} f_X(x)dx \tag{5.3}$$

5.2.1 Probability distribution models

Normal Distribution

We call X a normal random variable with elements $\mu \in \mathbb{R}$, $\sigma^2 > 0$ and we can rewrite it as $X \sim N(\mu, \sigma^2)$. The Normal distribution is also known as Gaussian distribution and in non-technical literature is called the bell curve.

Probability density function :

$$f(x;\sigma,\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.4}$$

where $x \in \mathbb{R}$. We can imply that $f_Y(\mu - y) = f_Y(\mu + y), y \in \mathbb{R}$, which means that f_Y is symmetrical to the μ parameter. There is maximum point at $y_0 = \mu$ that is the only local (and absolute) maximum. The inflection points are $y_1 = \mu - \sigma$ and $y_2 = \mu + \sigma$.

Weibull Distribution

The Weibull distribution can be considered as the generalization of the exponential distribution.

Probability density function :

$$f(x;\lambda,\alpha) = \alpha\lambda^{\alpha}x^{\alpha-1}e^{\lambda x^{\alpha}} \tag{5.5}$$

where $x \in \mathbb{R}$, $\lambda > 0$, $\alpha \geq 0$. The α parameter is called *shape parameter.* When α increases, the curve narrows. The λ parameter is called the scale parameter. Weibull distributions with $\alpha < 1$ have a decreasing failure rate, whereas Weibull distributions with $a > 1$ have an increasing failure rate.

Gamma Distribution

The Gamma distribution can be considered as the generalization of the exponential distribution.

Probability density function :

$$f(x;\lambda,\alpha) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} \tag{5.6}$$

where $x \geq 0$. The gamma function is defined as : $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx$, $a > 0$. The Gamma distribution is right-skewed.

Logistic Distribution

Probability density function :

$$f(x;\sigma,w) = \frac{e^{-w}}{[\sigma(1+e^{-w})]^2} \tag{5.7}$$

for any $\alpha \in \mathbb{R}$, $\beta > 0$ and $x \in \mathbb{R}$ while w is defined as : $w = \frac{x-\alpha}{\beta}$. A The $f(x)$ is symmetric about $x = \alpha$ plus the f increases on $(-\infty,\alpha)$ and decreases on (a,∞).

Log-logistic Distribution

Probability density function :

$$f(x;k,z) = \frac{kz^{k-1}}{(1+z^k)^2} \tag{5.8}$$

for any $k \in (0,\infty)$ while k and z are defined as : $k = \frac{\beta}{\alpha}$, $z = \frac{x}{\alpha}$, $x > 0, \alpha > 0, \beta > 0$. The α and β elements denote the scale and the shape parameters, respectively.

Chauchy Distribution

We call X a cauchy random variable with elements $\alpha \in \mathbb{R}$, $\gamma > 0$.

Probability density function :

$$f(x;\alpha,\gamma) = \frac{1}{\pi\gamma}[1 + (\frac{x-\alpha}{\gamma})^2]^{-1} \tag{5.9}$$

for $x \in \mathbb{R}$. The $f(x)$ is symmetric about $x = \alpha$, it increases and then decreases, when the mode is $x = \alpha$. As x approaches ∞ or $-\infty$ then $f(x) \to 0$.

Gumbel Distribution

We call X a gumbel random variable with elements $\mu \in \mathbb{R}$ and $\beta > 0$ and is a particular case of the class of extreme-value distributions. A Gumbel distribution is right-skewed.

Probability density function :

$$f(x;\mu,\beta) = \frac{1}{\beta}e^{-(z+e^{-1})} \tag{5.10}$$

where z is defined as : $z = \frac{x-\mu}{\beta}$, $x \in \mathbb{R}$. At the location of the mode ($x = \mu$), the density $f(x) = e^{-1}$ is approximately 0.37, regardless of the value of β.

5.3 Spatial Modelling : Estimation of Spatial Correlation

This section refers to the well-log data correlations and the importance of investigation of spatial correlation of the heterogeneity and variability of physical properties the geological strata. This can be done with variography, which establishes the rate of similarity between sample points as a function of a distinct seperation distance h. Their visual representation is displayed in Figure 5.2.

5.3.1 Experimental Variogram

The computational method of the experimental variogram is the Matheron's method of moments (MoM) estimator ([64]):

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{N(h)} [z(u) - z(u+h)]^2 \tag{5.11}$$

where $N(h)$ is the number of comparison pairs for lag h, $z(u)$ and $z(u+h)$ are the observed values of z at locations u and $u+h$. In other words, the variogram is defined as equal to one half of the average of squared differences between the field values.

Nugget

The existence of *nugget effect* (or *nugget*) is related to the fluctuation of the short range variability in the data. The nugget is equal to the intersection of the variogram with the y-axis of the graph. If the nugget is larger in comparison with the sill then that indicates too much noise and really small spatial correlation. Notice that below the intersection point no information can be obtained for interpretation ([77]) .

Sill

The *sill* of a variogram is the inflection point of the curve at which levels off and represents the variance of the variables. Positive or negative spatial correlation occurs when the data points are below or above the sill, respectively. The existence of trends in the data can be indicated by the behavior of the variogram curve based on the sill. In that case, the trends have to be proceed accordingly ([41]).

Range

The distance at which the variogram's value points level off to the sill is known as the *range* and is a maximum correlation length estimation between two sampling points at separation distance h. One remark is that spatial correlation can be calculated if the point distances are greater than the range, but is practically zero. ([34]).

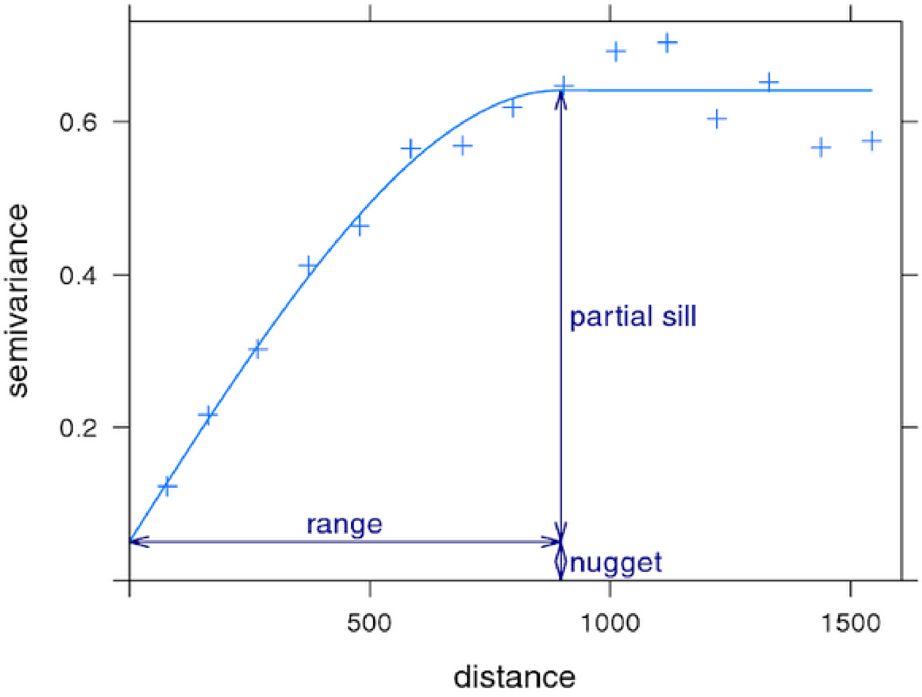


Figure 5.2: The three principal parameters of the variogram from [11]

5.3.2 Theoretical Variogram model

Bochner's Theorem

The *covariance* is defined as a deterministic function between two points and denotes the interdependence of those points on a field Z on $\mathbb{R}^d$. However, it is not correct to consider that a deterministic function can be defined as a covariance function. The covariance functions cannot be any functions unless they meet under some conditions. Those conditions must be determined, as they represent several theoretical models. Therefore, the experimental spatial correlation adjust to a defined fitting model. The conditions that define the permissible covariance functions are provided by the Bochner's theorem.

Theorem 5.1 (Bochner's Theorem)**.** *A function $\tilde{c}_X$ is a permissible covariance function, if the following conditions hold:*

1. *The integral $\tilde{c}_X(\mathbf{k}) = \int c_X(r)\, e^{i\mathbf{k}\cdot\mathbf{r}} d\mathbf{r}$ exists and is symmetric, i.e., $\tilde{c}_X(\mathbf{k}) = \tilde{c}_X(-\mathbf{k})$,*
2. *it is non negative for all frequencies $\mathbf{k}$, an*
3. *is bounded for all frequencies $\mathbf{k}$.*

Variogram Models

Fitting a variogram model to the empirical variogram is necessary for two main reasons ([71]):

1. Spatial prediction algorithm (Kriging) requires spatial continuity of the data.
2. A variogram model can ensure a positive definite model of spatial variability.

The most common are the spherical, exponential, gaussian, and power functions. These models ensure mathematical stability during calculations ([41]) and are known to be positive defined. Examples of experimental and theoritical variogram models used in the present thesis are :

Nugget Effect Model

The nugget effect model represents a constant value for all distances greater to zero. It is described by the function:

$$\gamma(h) = \begin{cases} 0 & \text{if h} = 0 \\ 1 & \text{if h} < 0 \end{cases} \tag{5.12}$$

In order to optimize any process under study is essential to understand the nature of the nugget effect since the model describes the spatially uncorrelated range of the observed values.

Spherical Model

Represented by quadratic modified equation. It is described by the function:

$$\gamma(h) = \begin{cases} C_0 + C_1 \left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3\right] & \text{if } 0 < h \leq a, \\ C_0 + C_1 & \text{if } h \geq a. \end{cases} \tag{5.13}$$

C_0 denotes the nugget variance and C_1 refers to the variance of the spatially correlated component. The quantity α is a distance parameter and indicates the spatial dependence.

Exponential Model

Similar to the spherical model in variability with distance reaching the sill asymptotically. It is described by the function:

$$\gamma(h) = \begin{cases} C_0 + C_1 \left[1 - e^{\frac{-h}{a}}\right] & h > 0, \\ 0 & h = 0. \end{cases} \tag{5.14}$$

The parameter a denotes the range of the spatial dependence, also referred to as correlation length.

Gaussian Model

The Gaussian model uses the normal distribution curve, thus has a parabolic shape in short distances where phenomena identical. It is described by the function:

$$\gamma(h) = C_0 + C_1(1 - e^{(\frac{-h}{a})^2}) \tag{5.15}$$

The Gaussian model has smoother spatial changes than other experimental variogram models.

Matérn Model

The Matérn model can be characterized as a generalization of various variogram model functions. The Exponential model for $\nu = 0.5$, the Whittle's model for $\nu = 1$, as well as the Gaussian model for $\nu = \infty$ are some of the cases. The parameter v is referred to the literature as the *smoothness* parameter. Different behaviors of the model can be described due to great flexibility of the number of the parameter ν. The corresponding variogram function with $\nu = \infty$ denotes a smooth behavior. Alternatively, if the $\nu \approx 0$, then is related to a very rough behavior. The model is described by the function:

$$\gamma(h) = C_0 + C_1[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}(\frac{h}{a})^{\nu} K_{\nu}(\frac{h}{a})] \tag{5.16}$$

where C_0 is the nugget effect. The sill is the sum of C_0 and C_1. The K_ν denotes the *Bessel function*:

$$K_{\nu}(t) = \frac{\Gamma(a)}{2}(\frac{t}{2})^{-\nu} \tag{5.17}$$

and the $\Gamma(\nu)$ denotes the *Gamma* function :

$$\Gamma_{\nu} = \int_0^{\infty} e^{-t}u^{\nu-1}dt \tag{5.18}$$

The non-negative parameter of the covariance is the component ν.

Power Model

A spacial case of the power model is the *Linear* model, where $a = 1$ and h describes the slope. It is described by the function:

$$\gamma(h) = C_0 h^a \tag{5.19}$$

for $0 < a < 2$. The parameter a describes the variation's intensity while the parameter $2H$ describes the curvature. There is no sill for the power-law variogram. Thus, allow for infinite variance.

Pentaspherical Model

Represents a five-dimensional analogue of the spherical model. The formula is described by the function:

$$\gamma(h) = \begin{cases} C_1[\frac{15h}{8a} - \frac{5}{4}(\frac{h}{a})^3 + \frac{3}{8}(\frac{h}{a})^5] & h \leq a \\ C_1 & h > a \end{cases} \tag{5.20}$$

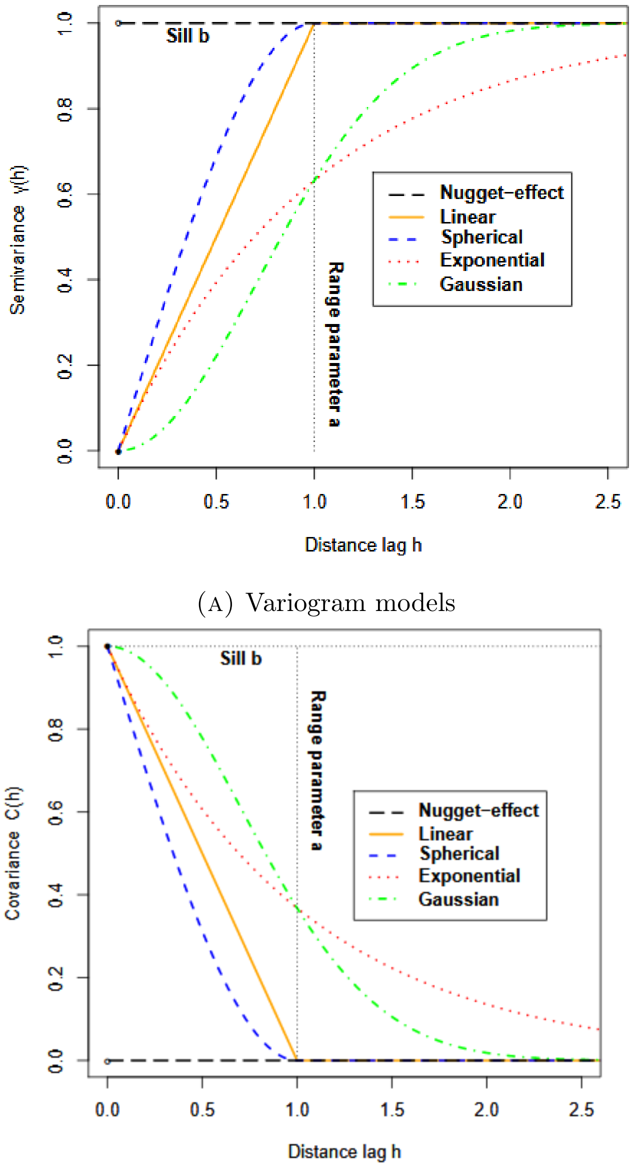
The parameter a denotes the range and the C_1 is the sill. One remark is that the curve rises gradually in comparison with the spherical model, with gradient $15C_1/8a$.

Circular Model

The formual is describe by the function:

$$\gamma(h) = \begin{cases} C_1[1 - \frac{2}{\pi}cos^{-1}\frac{h}{a} + \frac{2h}{\pi a}\sqrt{1 - \frac{h^2}{a^2}}] & h \leq a \\ C_1 & h > a \end{cases} \tag{5.21}$$

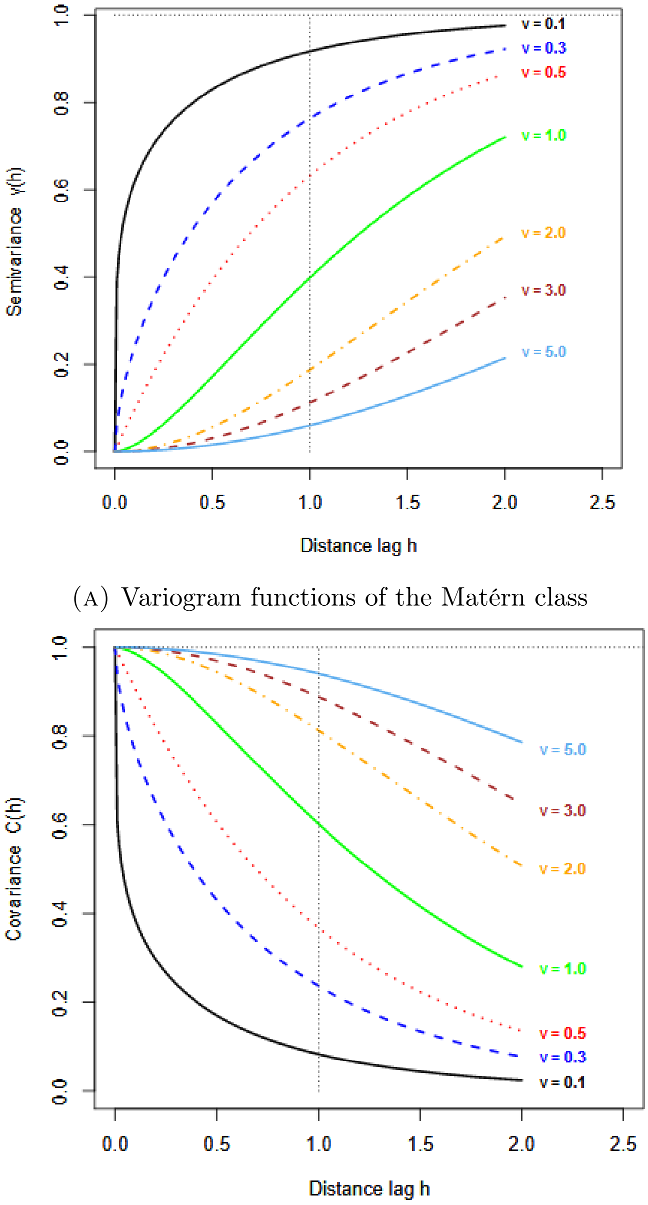
The parameter a denotes the range and the C_1 is the sill. The model's curve rises rightly and reaches the range with gradient $4C_1/a\pi$.



(A) Variogram models

(B) Permissible covariance models

FIGURE 5.3: The used parameters are range parameter $a = 1$ and sill $b = 1$. Image retrieved from [49].



(A) Variogram functions of the Matérn class

(B) Covariance functions of the Matérn class

FIGURE 5.4: The used parameters are range parameter $a = 1$ and sill $b = 1$ and varying ν smoothness parameter. Image retrieved from [49].

5.4 Variogram Fitting Methods

The purpose of fitting a theoritical variogram model to the calculated experimental variogram model is to estimate the optimum variogram parameters. The smoothing parameter of the variogram is defined by the number of lags k, yet there is no established rule for selecting the optimum number of lags. Some proposed methods for choosing the optimum fitted model, based on the leave-one-out cross-validation (LOOCV) and the Akaike information criterion (AIC) are the Ordinary Least Squared (OLS) and the Weighted Least Squares (WLS) ([43]). Some good studies in the area are the ones of [57] and [69] with applications in the geophysical study field.

Least Squares

In the Ordinary Least Squares (OLS) method we attemp to estimate the parameter vector θ of the theoritical variogram $\gamma(h)$ fitted in the experimental variogram $\hat{\gamma}(h)$, hence minimize the sum of square differences $R(\theta)$ given by the following equation :
For $i = 1, 2, ..., k$

$$R(\theta) = \sum_{i=1}^{k} w_i^2 [\hat{\gamma}(h_i) - \gamma(h_i;\theta)]^2 \tag{5.22}$$

In the case of OLS the weights w_i are equal to 1. The OLS method assumes that all differences resulted from the optimization process are normally distributed and independent.

In the case of Weighted Least Squares (WLS) the weights w_i are dependent upon the weighting method and $w_i^2 = 1/Var(\hat{\gamma}(h_i))$. One method of weighting is described by [22] and is given by the formula:

$$R(\theta) = \frac{1}{2} \sum_{i=1}^{k} N_i [\frac{\hat{\gamma}(h_i)}{\gamma(h_i;\theta)} - 1]^2 \tag{5.23}$$

where N_i are the number of pairs for the lag i. WLS fitting is more accurate for short distances, while on the hand the OLS performs an overall best fit at all distances considering constant variance. In R environment this could be done by using the argument *fit.method*. For the purpose of the thesis, the *fit.method=1* with weights N_j from the experimental variogram and *fit.method=7* with weights N_j/h_j^2 from the experimental variogram, are going to be used. Those weights depend on fitting parameters.

Cross-Validation Method

With the cross-validation method it is possible to calculate the error between the real and the predicted value for a number of data points with known values. Therefore, it allows us to compute the goodness of the performance of each interpolation algorithm. The main selected methods of CV for the purpose of this thesis have been analytically extensively adverted in the subsection 4.4.4

Chapter 6

Well Logs Correlation

In this section our research aims at finding a solution for the problem of correlation between available well logging data. This section seeks to address the following concepts:

- Implement geostatistical preliminary and exploratory analysis in order to demonstrate spatial dependencies.
- Investigate methods of improving accuracy of well to well log correlations. The solution to the problem is based on interpolation methods.

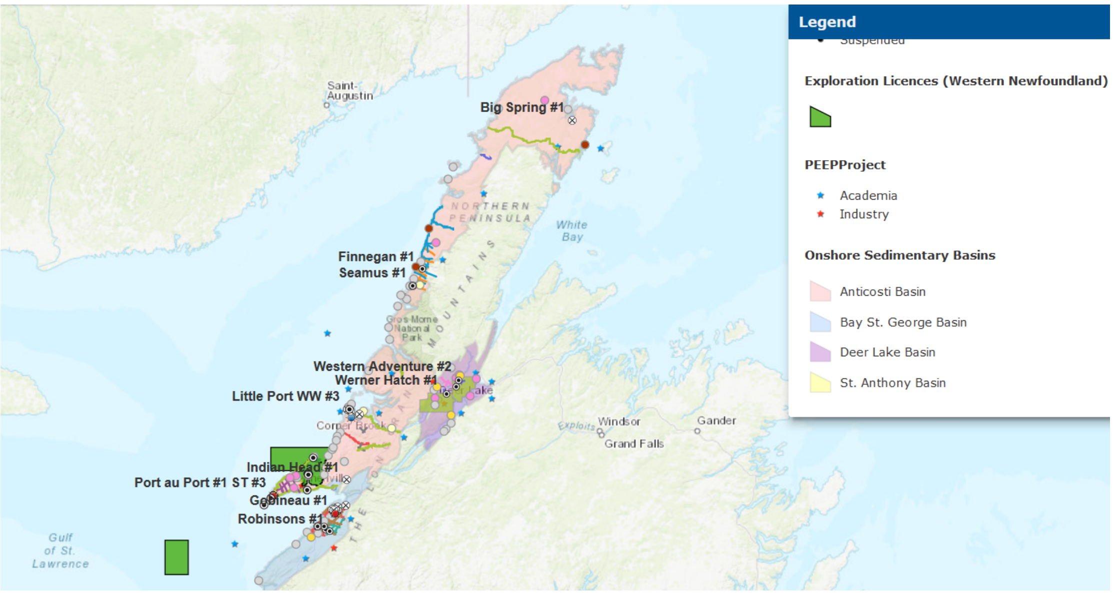


Figure 6.1: Finnegan and Seamus hydrocarbon wells location in Western Newfoundland and Labrador Island. Image retrieved from the Government of Newfoundland and Labrador website, section; Onshore Maps and Data.

6.1 Information about the Hydrocarbon Gas Reservoir

In this thesis we use well logs from two onshore natural gas reservoirs in Western Newfoundland and Labrador Island (Canada). The onshore exploration wells Seamus and Finnegan were drilled by Nalcor Energy and partners in 2010/2011 and had good gas shows. However, both were suspended since the natural gas encountered in the wells was non-commercial. The Seamus well was drilled to a total final depth of $3,160m$ while the Finnegan well reached an onshore depth of $3,130m$. The data-sets gathered as a result of drilling, testing and seismic analysis of the wells can be integrated for (i) the better understanding of the Western Newfoundland and Labrador Island petroleum geology, (ii) onshore studies of the regional stratigraphy and correlation into offshore blocks, (iii) extrapolation to various offshore exploration licences in the area.

The two drilled well-bores are located within the Cambrian Ordovician-Anticosti basin. The Anticosti basin is the largest Paleozoic basin of Western Newfoundland and Labrador Island with both offshore and onshore covered areas. The geological model of the basin contains rock sequences from Lower Cambrian to Devonian evolution period of the northern Appalachian orogen including a sliver of overlying carboniferous clastics that are associated with multiple tectonic events. Good oil and gas production reservoirs are presented in the Lower Ordovician and Mid-Upper Ordovician (HTD), the Carbonate thrust slice, and the Lower Devonian sandstone. Dolomitized carbonate rocks and sandstones are the predominant reservoir rocks in the Anticosti basin.

The studied group of Goose Tickle includes the Goose (American) Tickle formation, the Table Head group includes the Table Point formation while the St.George group includes the Aguathuna, Catoche, Boat Harbour and Watts Bight formation. All those formations are present to both studied wells. Reservoir potential is recognized within the stratigraphic unit of Goose (American) Tickle and Table Point formations while main reservoirs are recognized within the St.George group.

These hydrocarbons occurrences associated with the Paleozoic basin are considered of high geological risk with regard to hydrocarbon mitigation, oil biodegradation, and lateral seal. The use of high quality acquisitive data can lead to the direct detection of porosity and fluid type and thus minimize the specified risk.

The recorded data includes the GR (Gamma Ray), SP (Spontaneous Potential), A10 (Array Induction Two Foot Resistivity, Depth of Investigation: 10in) and A20 (Array Induction Two Foot Resistivity, Depth of Investigation: 20in) well logs.

All the information and data used for this thesis has been retrieved from the Final well reports of Seamus and Finnagan wells found in [1].

Figures 6.2, 6.3, 6.4, 6.5 illustrate the available well log data of the selected formations.

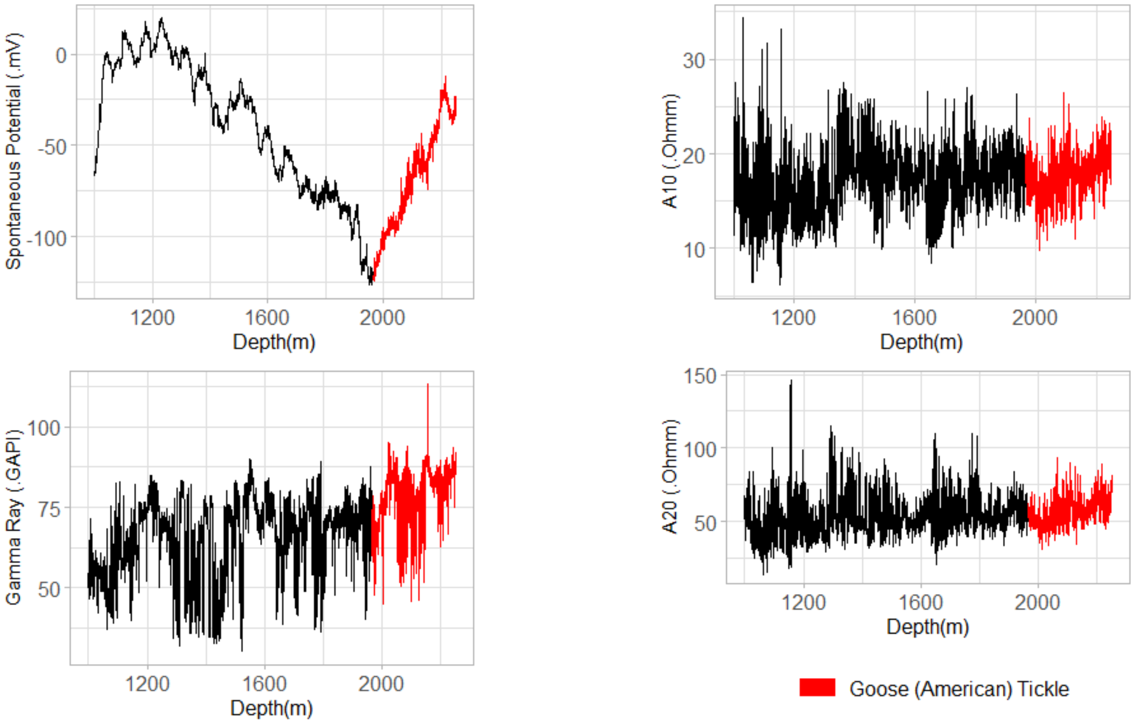


FIGURE 6.2: Finnegan well logs of 311mm hole section. The spontaneous and gamma-ray logs are displayed on the left side of the log. The induction resistivity logs are on the right.

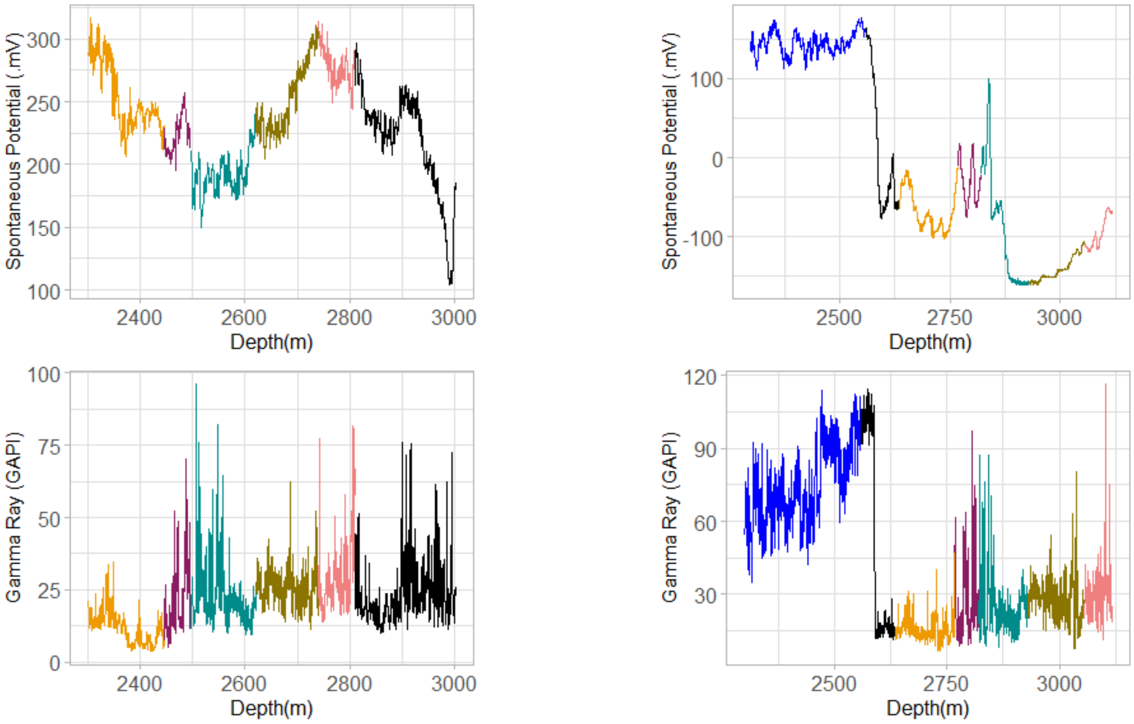


FIGURE 6.3: The spontaneous and gamma-ray logs of Finnegan 216mm hole section and Seamus 216mm hole section are displayed on the left and right side of the log, respectively.

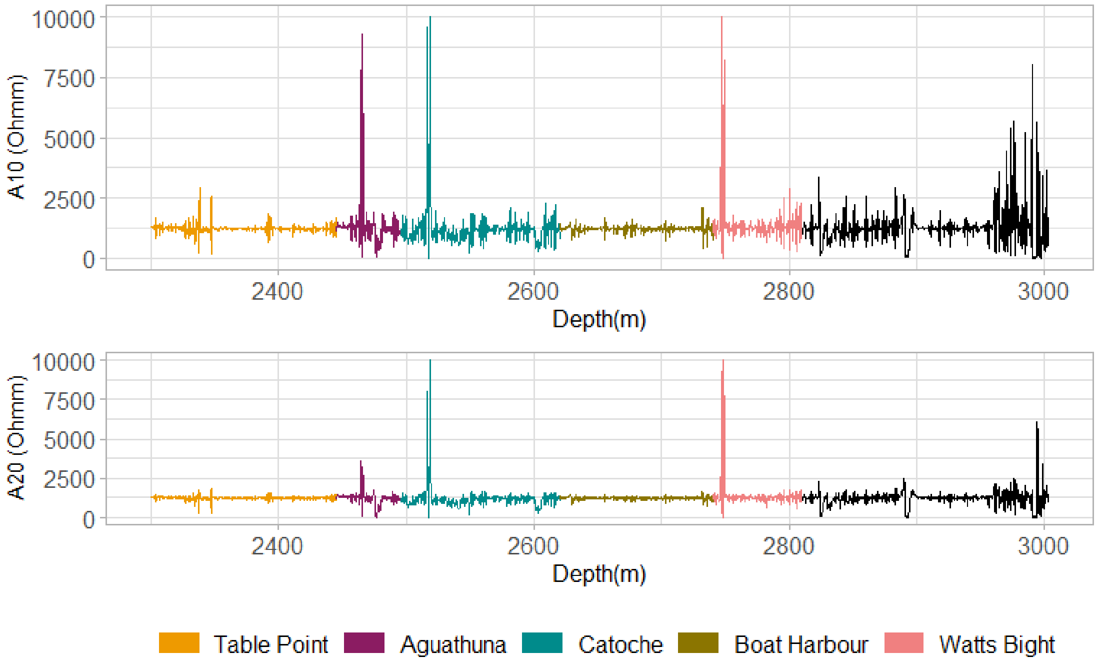


FIGURE 6.4: Finnegan well logs of the induction resistivity logs of 216mm hole section.

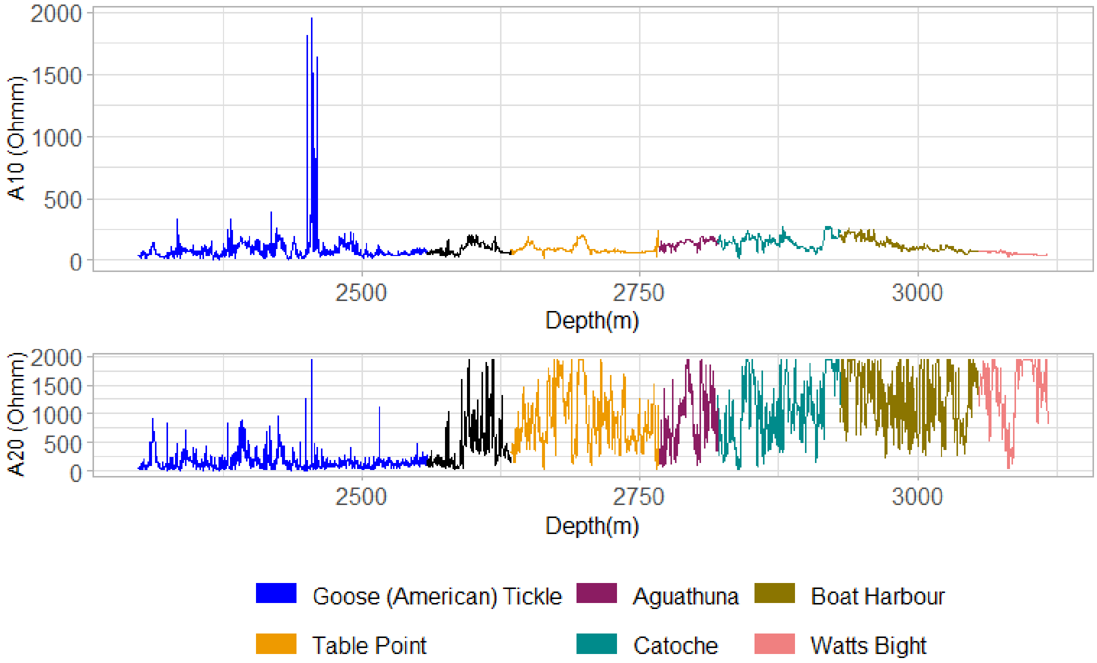


FIGURE 6.5: Seamus well logs of the induction resistivity logs of 216mm hole section.

6.1.1 Notions and Assumptions for the Data

In table 6.1 a summary of the selected formations' thickness and data consistency is represented. Indications of complex, non-stationary behavior can be visualized by the figures 6.2, 6.3, 6.4, 6.5, as the means and variances are not constant over depth.

	Formations	**Formation thickness (m)**	**Observed data (n)**
Finnegan	Goose (American) Tickle	284.4	1422
	Table Point	145	725
	Aguathuna	50	250
	Catoche	124.8	624
	Boat Harbour	119.8	599
	Watts Bight	69.8	349
Seamus	Goose (American) Tickle	259.1	1700
	Table Point	132.6	871
	Aguathuna	52.7	347
	Catoche	109.8	721
	Boat Harbour	124.8	819
	Watts Bight	61.8	406

Table 6.1: Summary of studied data-sets.

The analysis is progressed by removing potential trends in the datasets; in order to ensure consistency of interpretation of the spatial direction in the data and on the other hand, to examine under which possibly circumstances the effect of a trend on a semivariogram might be bypassed to allow a sufficient analysis of the data.

For a quick-well log interpretation the principle data sources have been used in order to locate and identify the different geological formations. Firstly, the formation interval is identified by the SP and GR log responses. High SP usually represent permeable beds, or fresh water, while low SP often represent shale beds, or salt water. If the SP is constant over depth then the formation is impermeable. Additionaly, high GR represent shaly sandstones, or shandy shales, while low GR usually represent sand, or coal, limestone and dolomite. The process of visual evaluation and identification was surely ambiguous but nevertheless capable of resulting enough information. However, the optimum selection of a formation delimitation can best impact the field development and benefit the planning of a drilling program. Thus, the recognition of the stratigraphic boundaries were defined by district lithological and coring analysis reported by the Nalcor Energy Oil & Gas Inc.

6.2 Methodology

Different types of noise result in nonlinear and nonstationarity characteristic behaviour, the convoluted trend and seasonality of the well log data is difficult to extract. We like to examine the effectiveness of a trend in the analysis of the data, so there is no need for any transformations or removal of extremely complected trend models. For the analysis procedure the following variogram models have been used: 1) Exponential, 2) Gaussian, 3) Spherical, 4) Pentaspherical, 5) Circular. The utilized algorithms for statistical analysis and construction of the variograms, as well as the algorithm for calculating the correlations between well logs were developed and run in R and Matlab environment. The algorithms were developed for:

1. Detrending 1D data,
2. Fitting probability distributions to data series,
3. Plotting QQ, Empirical and Theoretical CDFs, and PP graphs,
4. Calculating the experimental variogram and fitting the theoretical variogram model,
5. Performing cross-validation for a given model, and
6. Interpreting several interpolation methods to estimate the query point of the studied data and improve the performance of correlation algorithm .

In order to estimate the empirical semivariogram we used the Cressie-Hawking robust estimator which provides a satisfactory model that improves the variogram estimation of a described geologically continuous process. This developed model deals with outliers and non-normality for distributions particularly heavy in the tails region ([23]).

6.3 Goose (American) Tickle Formation

The Goose (American) Tickle formation is a geological unit dominated by silty argillite with minor sandstones ([72]). The Goose (American) Tickle formation is found in Finnegan well at depth of around 1965m and 2250m. The formation contains more than 60% sandstone and less than 40% shale rocks. The same formation is present in Seamus well in a depth range from about 2225m and 2585m, and contains more than 60% sandstone, less than 40% shale and less than 10% limestone rocks [1].

Finnegan 311mm hole section

To begin with, the statistical moments were calculated and presented in table 6.2. The next step is the elimination of any possible trend in order to remove any long-time scale fluctuations. The chosen logs are the Spontaneous Potential and Gamma Ray logs, as they both contain a describable trend component. The complete expressions of the resulting trend models are shown in table 6.4, while the statistics of the detrended values can be seen in table 6.3. In figure 6.6 the histograms of the original and detrended data-sets are plotted.

In figures 6.7, 6.8, 6.9, 6.10 the fitting of the tested distributions is presented. The values of Spontaneous Potential and Gamma ray determine the total field $\Omega \subset \mathbb{R}$, so the tested distributions were the Gaussian, Cauchy and Gumbel distribution. The data-sets of Array Induction logs determine the total field $\Omega_+ \subset \mathbb{R}^+$, so the tested distributions were the Gaussian, Weibull and Gamma distributions. In Table 6.5 the estimated parameters and validation measures are presented. The formations of Table Point and Aguathuna are presented in Appendix A.2.

Spontaneous Potential

As presented by Figure 6.7 the, the Q-Q plot shows luck-of-fit at the Cauchy and Gumbel distribution tails. The graph of Empirical and theoretical CDFs confirms that the empirical values don't match up well with theoretical distributions. The P-P cross-plot shows that the matching cumulative probabilities from the two cumulative distributions don't agree. The information criteria presented in Table 6.5 agree that the best model is the Gaussian model, followed by Cauchy and Gumbel.

Gamma Ray

As presented by Figure 6.8 the, the Q-Q plot shows luck-of-fit at the Cauchy and Gumbel distribution tails. The graph of Empirical and theoretical CDFs confirms that the empirical values don't match up well with the Cauchy and Gumbel theoretical distributions. The P-P cross-plot shows that the matching cumulative probabilities from the two cumulative distributions of Cauchy and Gumbel don't agree. The information criteria presented in Table 6.5 agree that the best model is the Gaussian model, followed by Cauchy and Gumbel.

Array Induction (10in)

As presented by Figure 6.9 the, the Q-Q plot shows no significant luck-of-fit at the studied distribution tails and a good indication that the dataset comes from a Gaussian distribution. The graph of Empirical and theoretical CDFs confirms that the empirical values don't match up well with the Weibull theoretical distribution. The P-P cross-plot shows that the matching cumulative probabilities from the tested cumulative distributions match up pretty well. The information criteria presented in Table 6.5 agree that the best model is the Gaussian model, followed by Gamma and Weibull. We will continue under the assumption that the dataset comes from a Gaussian distribution.

Array Induction (20in)

As presented by Figure 6.10 the, the Q-Q plot shows no significant luck-of-fit at the studied distribution tails and a good indication that the dataset comes from a Gaussian distribution. The graph of Empirical and theoretical CDFs confirms that the empirical values don't match up well with the Weibull theoretical distribution. The P-P cross-plot shows that the matching cumulative probabilities from the tested cumulative distributions match up pretty well. The information criteria presented in Table 6.5 agree that the best model is the Gamma model, followed by Gaussian and Weibull.

Variogram Analysis

A set of theoretical variograms were constructed for each physical property in order to determine the required nugget, sill and range. The experimental variograms were calculated based on the 5.11 equation. The set maximum correlation distance is split into lag distance bins in order to construct each variogram for each variable. The respective minimum lag distance for the physical property of Spontaneous Potential is equal to

2.5m, Gamma Ray is equal to 4m, Array induction of 10in is equal to 0.5m and Array induction of 20in is equal to 0.2m. The variogram plots are illustrated in Figure 6.11. Four available models were fitted to the studied physical property and the determination modelling parameters for each fit is summarized in table 6.6. The determination coefficients are summarized in Table 6.7. An initial observation is that the best fitted models for the Spontaneous Potential property are the Circular model, followed by Spherical and Pentaspherical. For the Gamma Ray property the best fitted model is the Circular model, followed by Spherical and Pentaspherical. For the Array Induction (10in) property the best fitted model is the Exponential, followed by Penthaspherical and Gaussian, while for the Array Induction (20in) case, the best fitted model is the Gaussian, followed by Circular and Spherical.

The fact that not all properties fit optimally to the same theoretical model is possibly due to different number of points used to each experimental variogram calculation. The sill and range seems to be close for the Spontaneous Potential and Gamma Ray, which were fitted to the same theoretical variogram models.

Logs	Min	Max	Mean	Median	Mode
SP(mV)	-124.31	-11.84	-67.11	-64.39	-93.08
GR(GAPI)	44.83	113.21	78.67	81.12	87.88
A10(Ohmm)	9.77	26.41	17.41	17.43	16.21
A20(Ohmm)	30.8	93.67	56.82	55.93	55.1

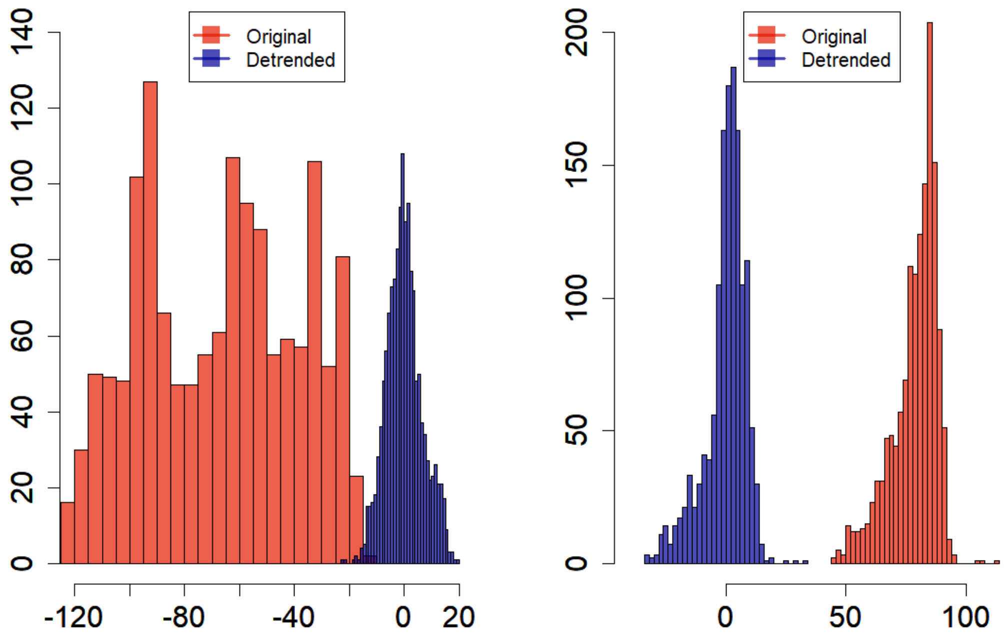
Logs	Variance	SD	Skewness	Kurtosis
SP(mV)	813.43	28.52	-0.03	1.84
GR(GAPI)	90.46	9.51	-1.01	3.90
A10(Ohmm)	5.5	2.34	-0.03	3.04
A20(Ohmm)	82.63	9.09	0.53	3.63

Table 6.2: Data statistics of Finnegan 331mm hole section.

Log	Min	Max	Mean	Median	Mode
SP(mV)	-22.33	19.21	1.54e-14	-0.25	-9.14
GR(GAPI)	-32.98	33.84	4.96e-14	1.56	6.02

Log	Variance	SD	Skewness	Kurtosis
SP(mV)	43.16	6.57	0.18	3.00
GR(GAPI)	73.79	8.60	-1.07	4.65

Table 6.3: Detrended data statistics of Finnegan 331 hole section.



(A) Spontaneous Potential

(B) Gamma Ray

FIGURE 6.6: Histogram of original and detrended data-sets of SP and GR logs of the Goose (American) Tickle formation found in Finnegan 311 hole section. Histograms with binwidth = 30.

Log	Model	Estimated Trend Function
SP	Linear	$-115 + 6.75 \cdot 10^{-2}x + \epsilon_i, \epsilon \sim N(0, 6.57^2)$
GR	Quadratic	$68 + 4.23 \cdot 10^{-2}x - 5.70 \cdot 10^{-5}x^2 + 2.64 \cdot 10^{-8}x^3 + \epsilon_i, \epsilon \sim N(0, 8.6^2)$

TABLE 6.4: Estimated trend models.

	Histograms		
	Distribution	Parameters	Information Criteria
SP	norm	μ=1.546e-14,σ=6.577	AIC=9398.92, BIC=9409.44
	Cauchy	a=-0.399,γ=3.772	AIC=9843.06, BIC=9853.59
	Gumbel	μ=-3.229,b=6.354	AIC=9558.30, BIC=9568.82
GR	norm	μ=4.965e-14,σ=8.587	AIC=10161.95, BIC=10172.47
	Cauchy	a=1.977,γ=3.987	AIC=10216.97, BIC=10227.49
	Gumbel	μ=-4.691,b=10.586	AIC=10824.91, BIC=10835.43
A10	norm	μ=17.408,σ=2.344	AIC=6466.42, BIC=6476.94
	Weibull	a=8.008,λ=18.427	AIC=6543.608, BIC=6554.129
	Gamma	a=53.607,λ= 3.079	AIC= 6489.494, BIC=6500.01
A20	norm	μ=56.817,σ=9.087	AIC=10322.96, BIC=10333.48
	Weibull	a=6.238,λ= 60.757	AIC=10513.14, BIC= 10523.66
	Gamma	a=39.844,λ=0.701	AIC=10272.02, BIC=10282.55

TABLE 6.5: Distributions' estimated parameters and information criteria of the Goose (American) Tickle formation found in Finnegan 311mm hole section. The units of measurement are [mV], [GAPI], [Ωhmm] for SP, GR and A10, A20 respectively.

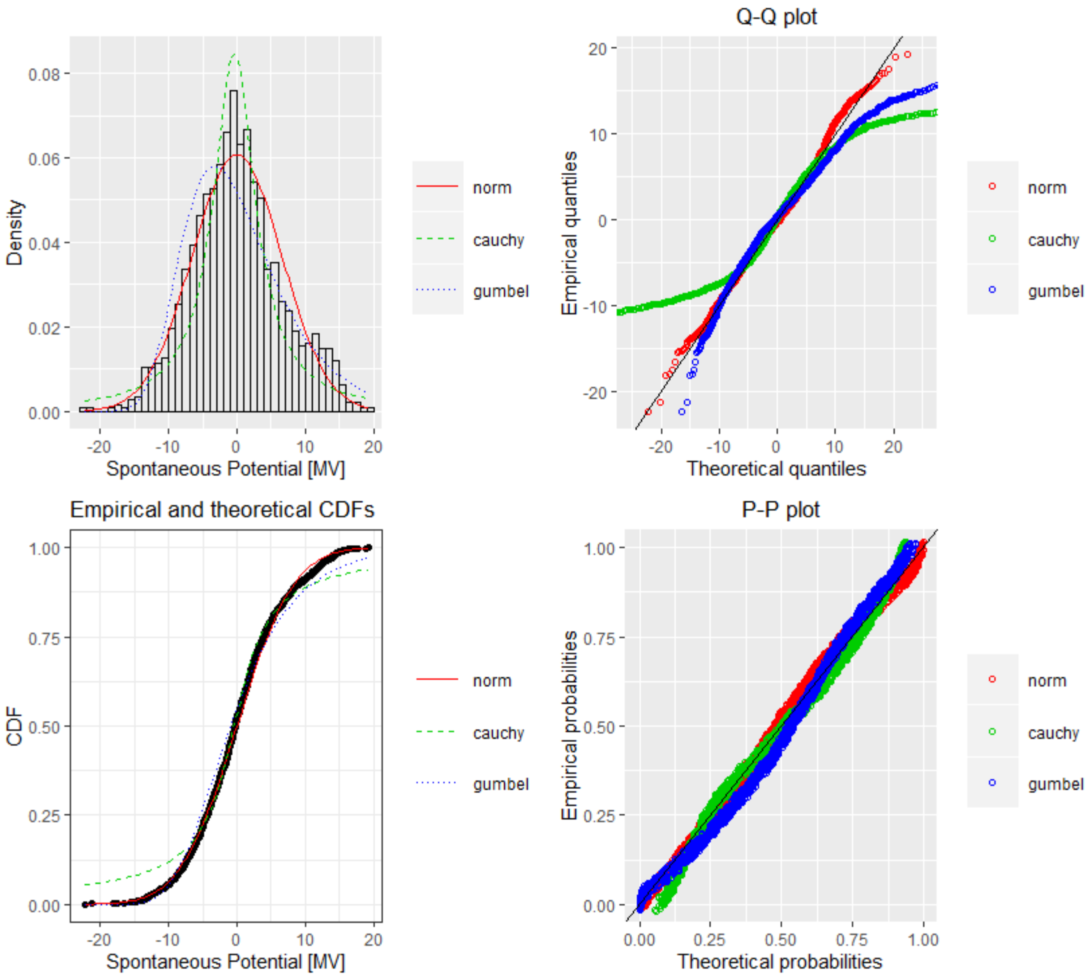


FIGURE 6.7: Fitting of the distributions by maximum likelihood. Featured data-set; SP log of Finnegan 311mm hole section.

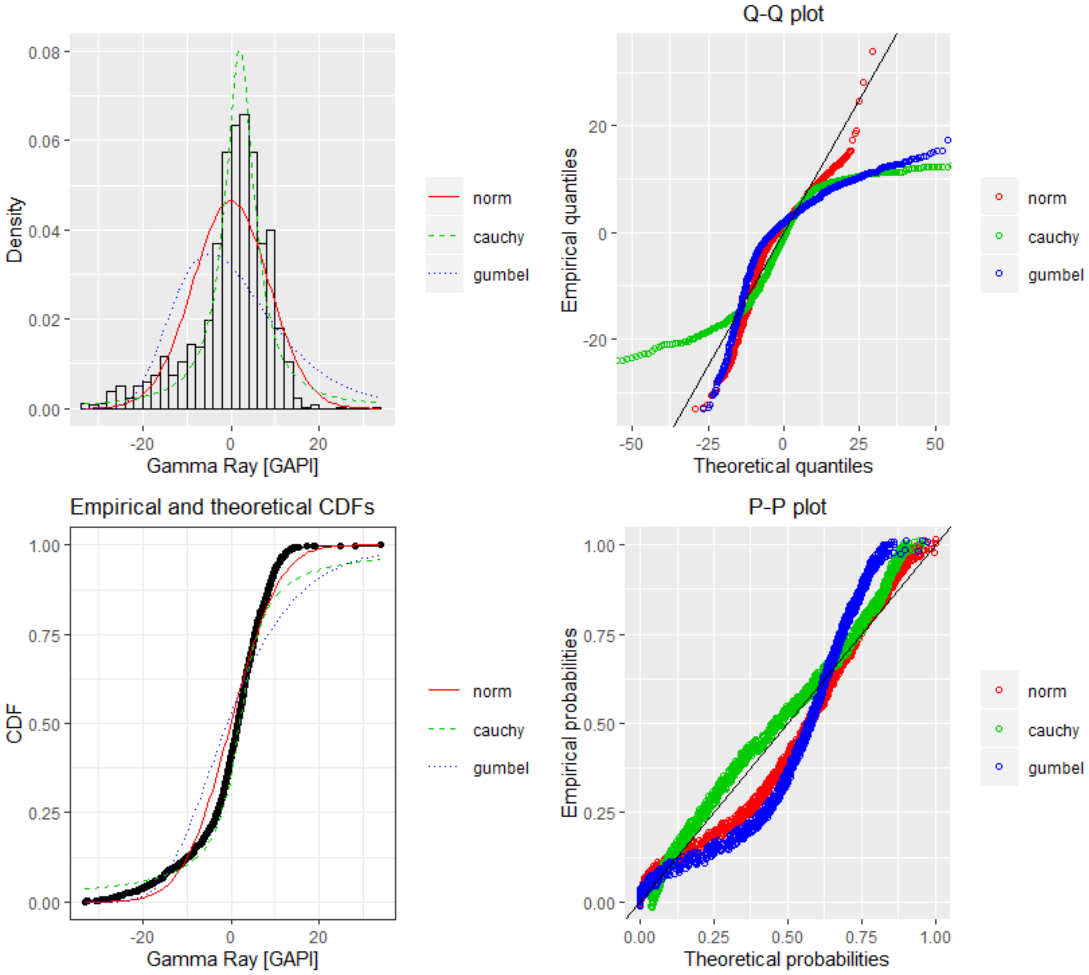


Figure 6.8: Fitting of the distributions by maximum likelihood. Featured data-set; GR log of Finnegan 311mm hole section.

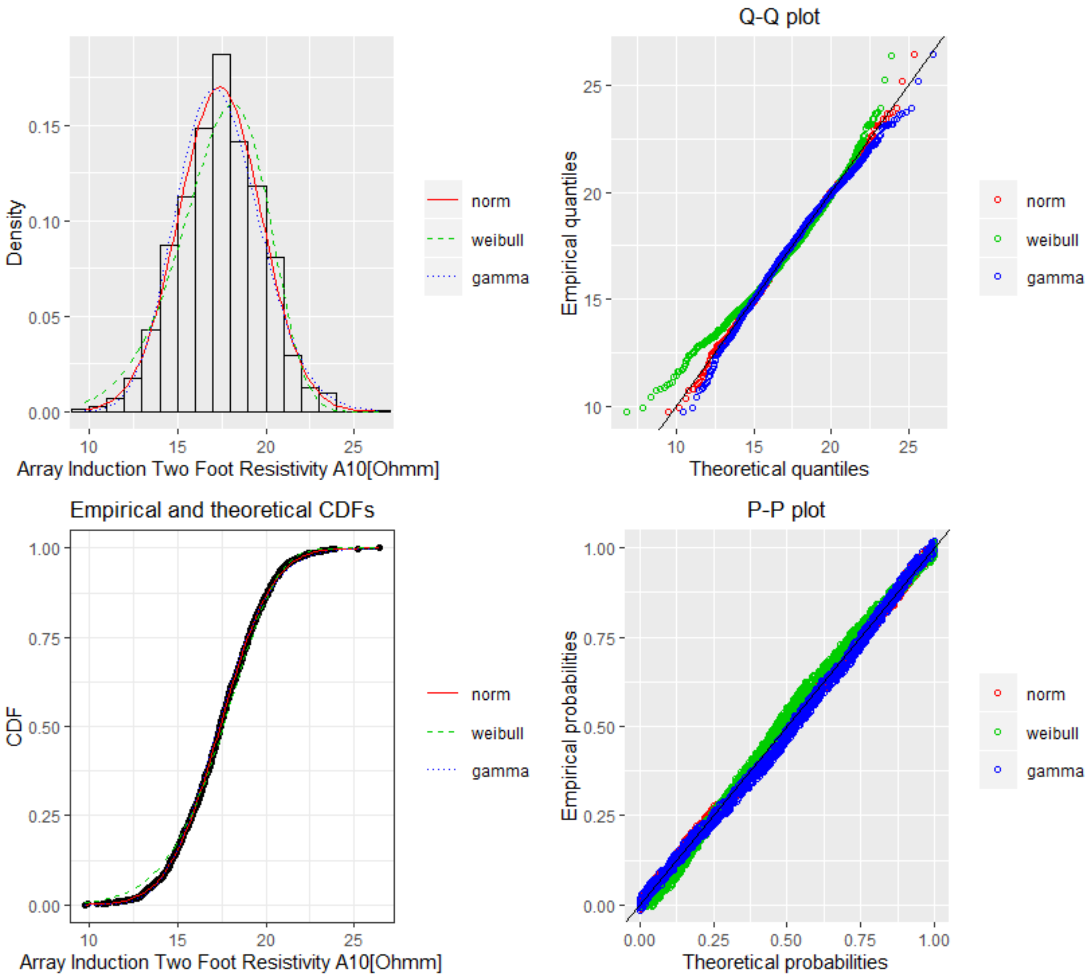


Figure 6.9: Fitting of the distributions by maximum likelihood. Featured data-set; A10 log of Finnegan 311mm hole section.

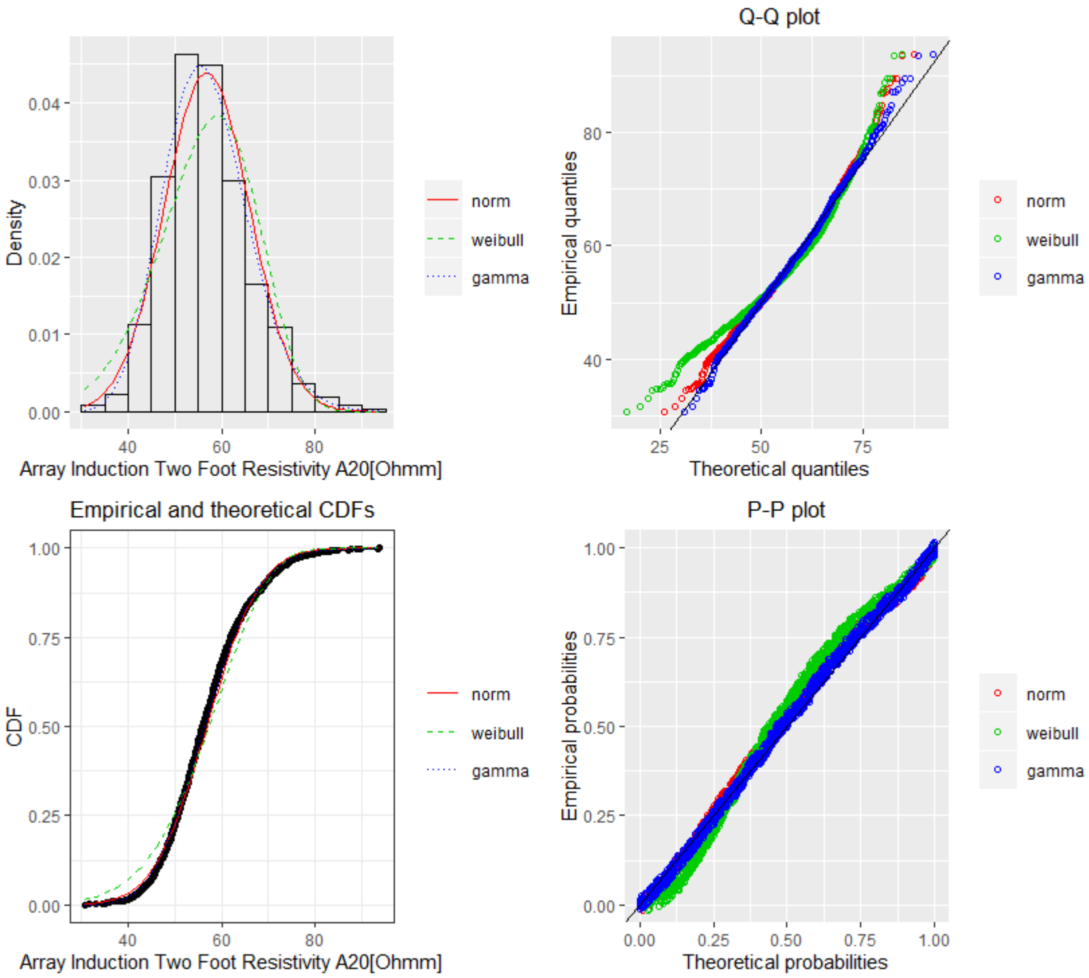
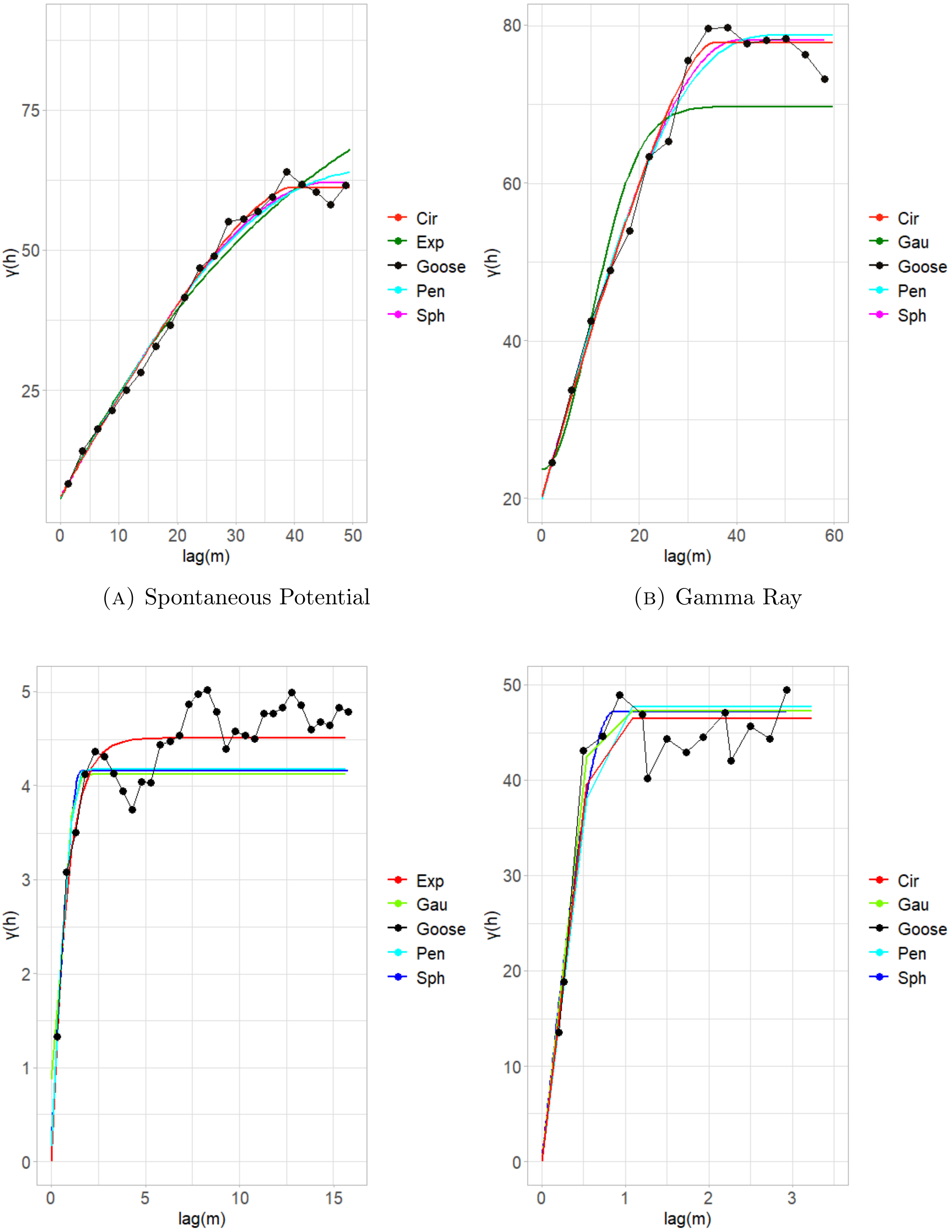


Figure 6.10: Fitting of the distributions by maximum likelihood. Featured data-set; A20 log of Finnegan 311mm hole section.



(A) Spontaneous Potential

(B) Gamma Ray

(C) Induction A10

(D) Induction A20

FIGURE 6.11: Variogram plots. The weights are determined using N_j/h_j^2, where N_j is the number of pairs at certain lag.

	Variograms			
	Model	Sill	Range	Nugget
SP	Cir	55.148	39.170	6.066
	Exp	92.257	43.880	5.625
	Pen	58.584	59.45	5.96
	Sph	56.174	46.010	5.996
GR	Cir	57.55	34.83	20.26
	Gau	46.07	13.83	23.66
	Pen	58.83	49.68	19.96
	Sph	58.03	40.13	20.09
A10	Exp	4.514	0.807	0.000
	Gau	3.258	0.766	0.871
	Pen	4.021	1.885	0.158
	Sph	3.945	1.567	0.221
A20	Cir	46.5	0.725	0.000
	Gau	47.25	0.356	0.000
	Pen	47.69	1.069	0.000
	Sph	47.200	0.858	0.000

Table 6.6: Fitting of the best theoretical model to the experimental variograms of the field.

	Variograms			
	Model	MSE	MAE	RMSE
SP	Cir	3.223	1.369	1.795
	Exp	11.199	2.639	3.346
	Pen	5.726	1.953	2.393
	Sph	4.438	1.668	2.107
GR	Cir	8.348	2.379	2.889
	Gau	40.261	5.522	6.345
	Pen	11.024	2.782	3.320
	Sph	9.278	2.497	3.046
A10	Exp	0.180	0.318	0.425
	Gau	0.296	0.471	0.544
	Pen	0.293	0.459	0.542
	Sph	0.299	0.467	0.547
A20	Cir	27.237	3.672	5.219
	Gau	26.560	3.867	5.154
	Pen	32.164	4.317	5.671
	Sph	29.746	3.963	5.454

Table 6.7: Fitting of the best theoretical model to the experimental variograms of the field.

Seamus 216mm hole section

First, the statistical moments were calculated and presented in table 6.9. The following step is to remove any possible trend in order to eliminate any present variations. The Gamma Ray log is the only one with a present removable trend component. The complete expressions of the resulting trend models are shown in Table 6.11 , while the statistics of the detrended values can be seen in table 6.10. In figure 6.13 the histogram of the original and detrended data-set is plotted.

In figures 6.14, 6.15, 6.16, 6.17 the fitting of the tested distributions is presented. The values of Spontaneous Potential and Gamma ray determine the total field $\Omega \subset \mathbb{R}$, so the tested distributions were the Gaussian, Cauchy and Gumbel distribution. The data-sets of Array Induction logs determine the total field $\Omega_{+} \subset \mathbb{R}^{+}$, so the tested distributions were the Gaussian, Weibull and Gamma distributions. In Table 6.12 the estimated parameters and validation measures are presented. The formations of Table Point and Aguathuna are presented in Appendix A.1.

Spontaneous Potential

As presented by Figure 6.14 the, the Q-Q plot shows luck-of-fit at the Cauchy distribution tails. The graph of Empirical and theoretical CDFs confirms that the empirical values don't match up well with theoretical distributions. The P-P cross-plot shows that the matching cumulative probabilities from the two cumulative distributions don't agree. The information criteria presented in Table 6.12 agree that the best model is the Gaussian model, followed by Gumbel and Cauchy.

Gamma Ray

As presented by Figure 6.15 the, the Q-Q plot shows luck-of-fit at the Cauchy and Gumbel distribution tails. The graph of Empirical and theoretical CDFs confirms that the empirical values don't match up well with the Cauchy and Gumbel theoretical distributions. The P-P cross-plot shows that the matching cumulative probabilities from the two cumulative distributions of Cauchy and Gumbel don't agree. The information criteria presented in Table 6.12 agree that the best model is the Gaussian model, followed by Cauchy and Gumbel.

Array Induction (10in)

As presented by Figure 6.16 the, the Q-Q plot shows luck-of-fit at the studied distribution tails. The graph of Empirical and theoretical CDFs confirms that the empirical values don't match up well with the Weibull theoretical distribution. The P-P cross-plot shows that the matching cumulative probabilities from the tested cumulative distributions match up well. The distribution is higly right skewed (positive skew). The information criteria presented in Table 6.12 agree that the best model is the Gamma model, followed by Weibull and Gaussian.

Array Induction (20in)

As presented by Figure 6.17 the, the Q-Q plot shows luck-of-fit at the studied distribution tails. The graph of Empirical and theoretical CDFs confirms that the empirical values don't match up well with the Weibull theoretical distribution. The P-P cross-plot shows that the matching cumulative probabilities from the tested cumulative distributions match up pretty well. The distribution is higly right skewed (positive skew). The information criteria presented in Table 6.12 agree that the best model is the Gamma model, followed by Weibull and Gaussian.

Variogram Analysis

In figure 6.18 the constructed theoretical variograms of each log are presented. We set the maximum correlation distance bins in way that we can construct each variogram for each physical property log. The respective minimum lag distance for the property of Spontaneous Potential is equal to $2m$, Gamma Ray is equal to $8m$, Array induction of $10in$ is equal to $3m$ and Array induction of $20in$ is equal to $2m$. The modelling parameters of each fitted model are summerized in Table 6.13, while their coefficients are summed up in Table 6.14.First we observed that the best fitted models for the Spontaneous Potential log are the Spherical model, followed by Pentaspherical and Circular. For the Gamma Ray log the best fitted model is the Pentaspherical model, followed by Spherical and Gaussian. For the Array Induction (10in) log the best fitted model is the Gaussian, followed by Penthaspherical and Spherical, while for the Array Induction (20in) case, the best fitted model is the Penthaspherical, followed by Exponential and Spherical.

As previously mentioned, the fact that not all logs fit optimally to the same theoretical model is probably due to different number of points used to each experimental variogram calculation.

6.4 Well Log Correlations

Interpolation Methods Interpolation processes are used to estimate the values of a function between two known points on a line or a curve. The problem of interpolation can be easily described as: Lets consider a range of a function $f(x_0), f(x_1), f(x_2), \ldots, f(x_n)$ that corresponds to $x_0, x_1, x_2, \ldots, x_n$ data points. We need to find a function y_x, that has the same values with the f_x function, at the same $x_0, x_1, x_2, \ldots, x_n$ data points. If we agree that p_x is a known function, then we can "read" the f_x function in the intermediate data points, $x_0, x_1, x_2, \ldots, x_n$, called interpolated points. In this thesis the methods that are going to be applied are the **Linear**, **Nearest Neighbor**, **Cubic** and **Spline** interpolation.

The selected logs of the two wells were correlated with the Pearson and Spearman correlation coefficient; the RMSE measure was as well calculated. The depth measure used in the correlation was the standard true vertical depth. The depth step of Finnegan 216 mm and 311 mm diameter hole section is $0.2m$, while the Seamus 215 mm hole section is $0.1524m$. As the scale changes so does the range of the data sets. In order to correctly correlate two series we need to perform a range standardization. The first step is to insert the data set to the *Matlab* environment. Then, use several interpolation methods to create two data sets with formation alignment and common sampling step. This is done by removing the difference of the initial distances of each set and then set the initial points to zero. The next step is to choose one of the two studied series of a selected formation, with a known and constant depth step. We keep this step constant and we will create a common depth step scale for both series by choosing a maximum cutoff point. Finally we compare the different interpolation methods used in terms of the resulting values of well-to-well log correlations. We will apply the same process for all the studied formations. To illustrate, their graphical representation is presented in figures 6.12, A.21, A.22, A.23, A.24, A.25.

The Gamma Ray log has been used as the lithological indicator for the correlation. Measurements of the gamma ray index are primarily used to correlate stratigraphic sections. Shales and clays found in oil and gas wells are usually responsibly for emitting natural radioactivity as their radioactive isotope content and mineralogy can be tracked down by gamma ray devices. Gamma-ray fluctuations indicate changes in formation mineralogy. Thus, gamma-ray logs taken from different wells within the same region of study can be efficiently used for well to well correlation, since similar formations will result in similar feature measurements.

	Cross Correlation Scores			
	Model	r_P	r_S	RMSE(GAPI)
Goose (American) Tickle	Linear	0.257	0.295	23.613
	NN	0.255	0.294	23.653
	Cubic	0.255	0.294	23.648
	Spline	0.255	0.294	23.656
Table Point	Linear	0.09	0.192	8.083
	NN	0.094	0.199	8.098
	Cubic	0.089	0.19	8.105
	Spline	0.089	0.187	8.110
Aguathuna	Linear	0.261	0.483	18.310
	NN	0.261	0.480	18.322
	Cubic	0.260	0.480	18.371
	Spline	0.257	0.479	18.396
Catoche	Linear	-0.104	-0.142	18.483
	NN	-0.107	-0.140	18.537
	Cubic	-0.103	-0.139	18.556
	Spline	-0.102	-0.138	18.572
Boat Harbour	Linear	-0.070	-0.401	11.415
	NN	-0.066	-0.038	11.460
	Cubic	-0.069	-0.040	11.470
	Spline	-0.102	-0.138	18.572
Watts Bight	Linear	-0.002	0.042	16.025
	NN	0.001	0.043	15.965
	Cubic	-0.003	0.040	16.127
	Spline	-0.001	0.040	16.150

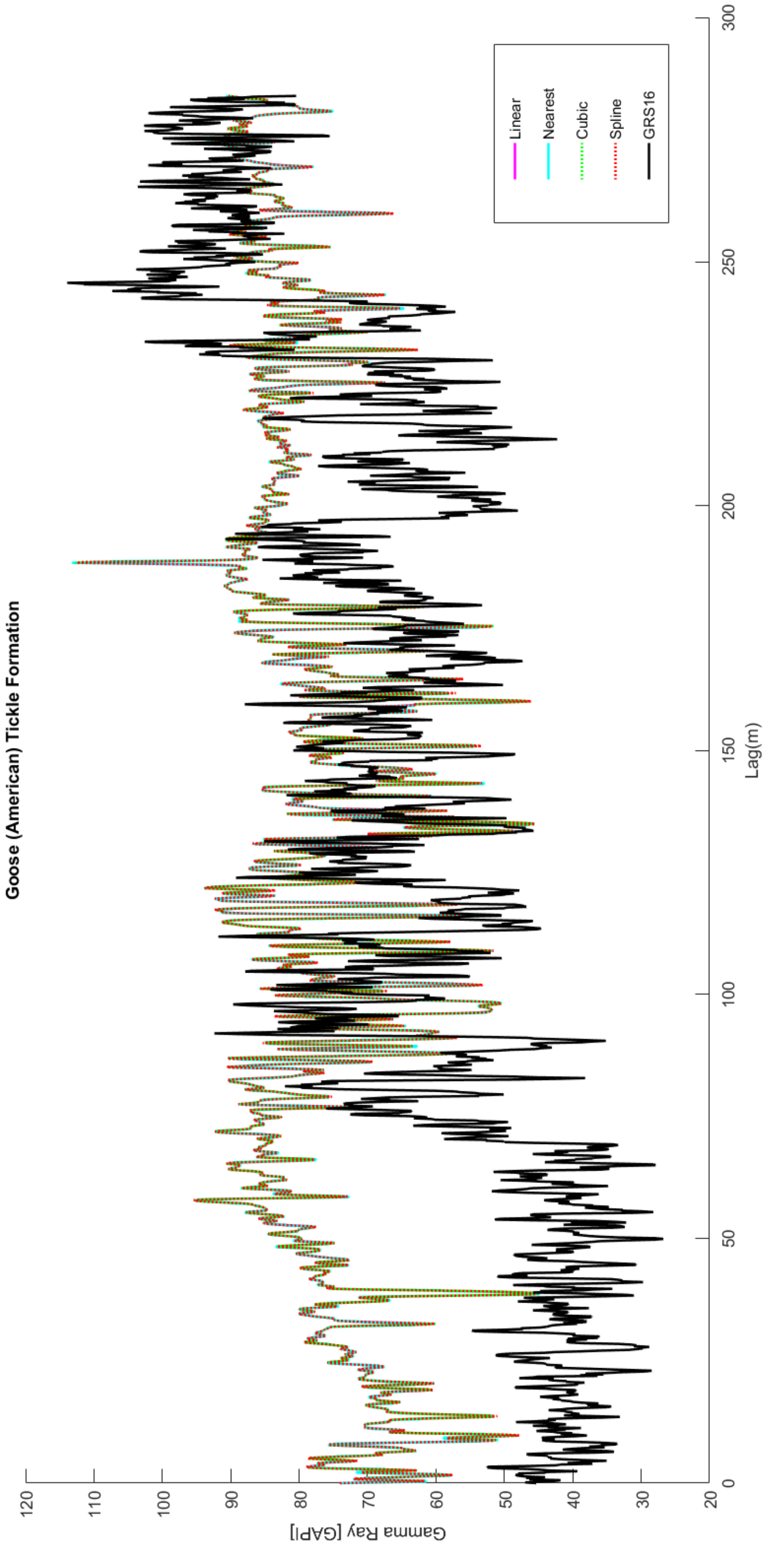
Table 6.8: Leave-One-Out Cross Correlations of the known geological series of Finnegan and Seamus wells.

6.5 Synopsis

Preliminary and exploratory data analysis tools allow the user to examine the data in more quantitative ways. The tools used were Histograms, Normal QQ-plots, Emprirical and theoretical CDF's plots, and P-P plots. Moreover, variogram analysis was used to examine the spatial autocorrelation between the measured respective properties.

Thee exploratory analysis indicates that the majority of the respective properties do not follow the Gaussian distribution. However, after removing a trend function, the residuals are closer to the Gaussian distribution. Results demonstrate that the Spontaneous potential and Gamma radiation indicators can be most often described by Cauchy and Gumbel distributions. In contrast, the Induction indicators can be most often described means of the Gamma and Weibull distributions. The theoretical variograms' formalization was manually achieved by applying WLS, for weights equal to N_j and N_j/h_j^2, accordingly. The weighting scheme of N_j/h_j^2 gives more weight to early lags. On the contrary, the weighting scheme of N_j give more weight to later lags. The results of the variogram analysis indicate that Spontaneous potential and Gamma Radiation indicators are mostly fitted to the same type of theoretical variogram model, with similar sill and range values. The variogram analysis confirmed that high spatial heterogeneity characterizes the entire span of the logging records.

The statistical analysis indicates a weak correlation between the respective properties measured at the two different wells. The association between the data at the neighboring wells is examined by means of statistical dependence measures such as the Pearson's linear correlation coefficient and Spearman's rank correlation coefficient. The cross correlations calculated from the processed data using different interpolation models lead to similar values. The Gamma radiation logs show both positive and negative correlation which are overall higher (in magnitude) than for the other three logs. The values of the positive correlation coefficients range from 0.001 to 0.483, while the values of the negative correlation coefficients range from -0.142 to -0.001. These findings support the notion that the Gamma ray log is influenced by lithological changes.

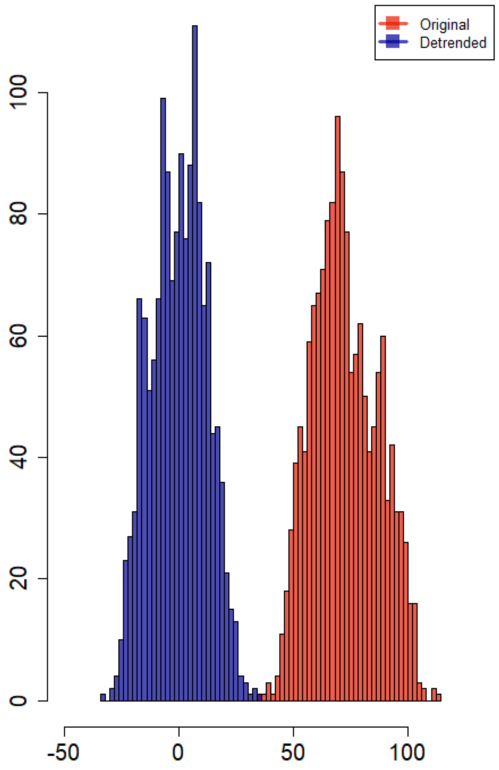

Goose (American) Tickle Formation
Gamma Ray [GAPI]
120
110
100
90
80
70
60
50
40
30
20
0
50
100
150
200
250
300
Lag(m)
Linear
Nearest
Cubic
Spline
GRS16

.png

Logs	Min	Max	Mean	Median	Mode
SP(mV)	110.937	173.937	142.347	141.312	141.312
GR(GAPI)	35.167	113.907	72.472	71.049	62.583
A10(Ohmm)	9.762	1950	94.512	70.810	37.734
A20(Ohmm)	9.001	1950	163.904	117.945	63.713

Logs	Variance	SD	Skewness	Kurtosis
SP(mV)	155.356	12.464	-0.025	2.788
GR(GAPI)	206.604	14.374	0.208	2.351
A10(Ohmm)	20172.2	142.029	9.115	99.550
A20(Ohmm)	26316.1	162.22	3.226	20.405

Table 6.9: Data statistics of Seamus 216mm hole section.



(a) Spontaneous Potential

Figure 6.13: Histogram of original and detrended data-sets of GR logs of the Goose (American) Tickle formation found in Seamus 216 hole section. Histograms with bin-width = 30.

Log	Min	Max	Mean	Median	Mode
GR(GAPI)	-33.230	34.531	1.33e-13	0.642	-4.614

Log	Variance	SD	Skewness	Kurtosis
GR(GAPI)	142.921	11.955	-0.017	2.365

Table 6.10: Detrended data statistics of Seamus 216mm hole section.

Log	Model	Estimated Trend Function
GR	Linear	$58.65 + 1.841 \cdot 10^{-2}x + \epsilon_i, \epsilon \sim N(11.96^2)$

Table 6.11: Estimated trend model.

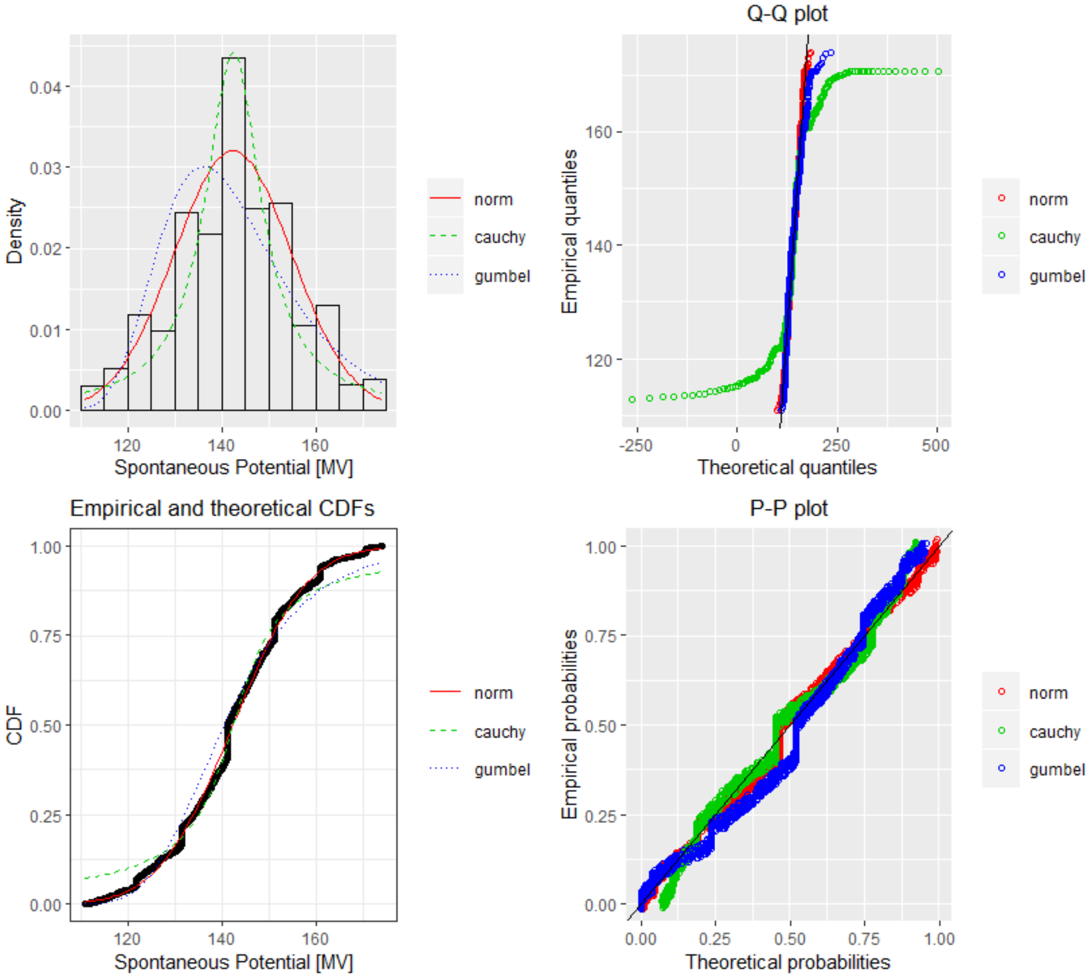


Figure 6.14: Fitting of the distributions by maximum likelihood. Featured data-set; SP log of Seamus 216mm hole section.

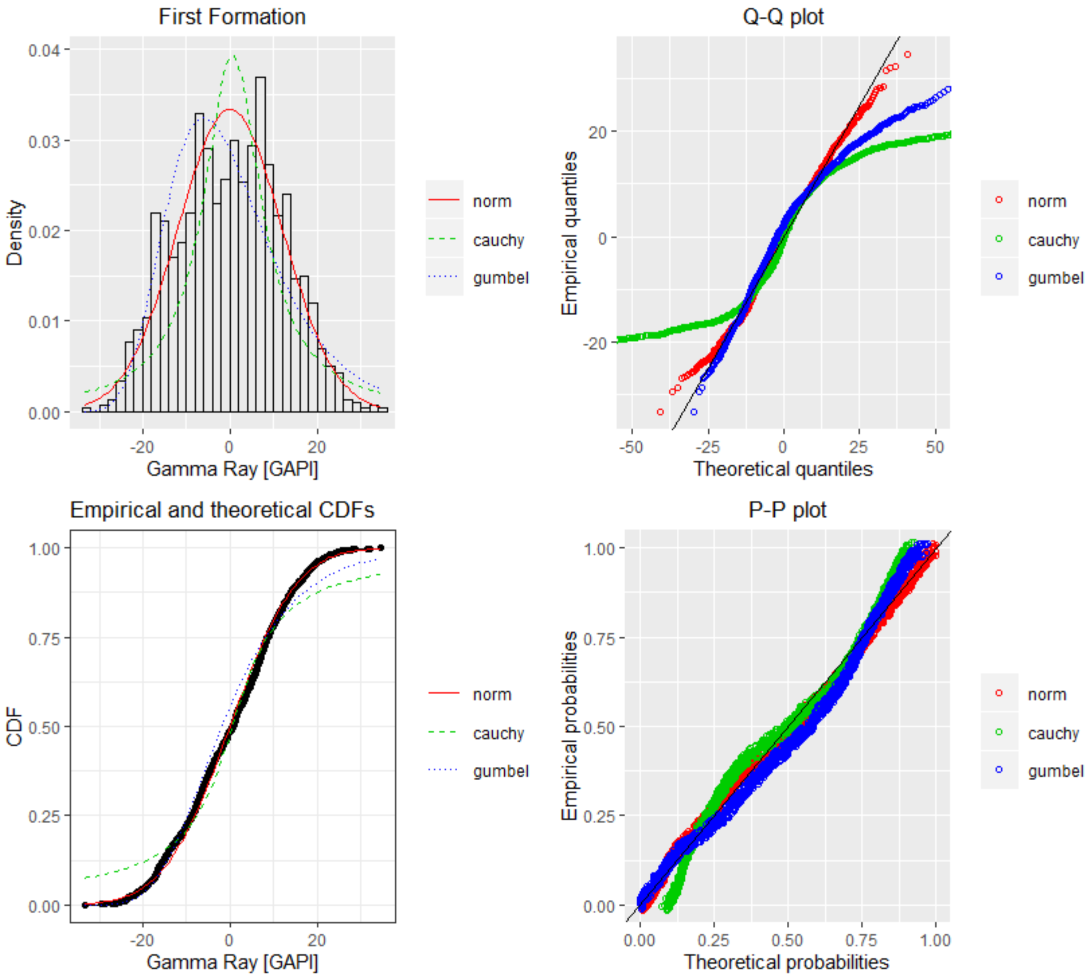


Figure 6.15: Fitting of the distributions by maximum likelihood. Featured data-set; GR log of Seamus 216mm hole section.

	Histograms		
	Distribution	Parameters	Information Criteria
SP	norm	μ=142.347,σ=12.460	AIC=11836.3 BIC=11846.9
	cauchy	a=142.419,γ=7.209	AIC=12340.2, BIC=12350.8
	gumbel	μ=136.085,b=12.219	AIC=12057.2, BIC=12067.8
GR	norm	μ=1.33e-13,σ=11.95	AIC=11711.1, BIC=11721.7
	cauchy	a=0.536,γ=8.080	AIC=12421.3, BIC=12431.9
	gumbel	μ=-5.978,b=11.354	AIC=11880, BIC=11890.6
A10	norm	μ=94.512,σ=141.981	AIC=19140.7, BIC=19151.3
	weibull	a=1.142,λ=100.450	AIC=16596.5, BIC=16607.2
	gamma	a=1.885,λ=0.012	AIC= 16363.8, BIC=16374.5
A20	norm	μ=163.904,σ=162.168	AIC=19539.7, BIC=19550.4
	weibull	a=1.212,λ= 176.322	AIC=18214.7, BIC= 18225.3
	gamma	a=1.638,λ=0.010	AIC=18122.6, BIC=18133.2

Table 6.12: Distributions' estimated parameters and information criteria of the Goose (American) Tickle formation found in Seamus 216mm hole section. The units of measurement are [mV], [GAPI], [Ωhmm] for SP, GR and A10, A20 respectively.

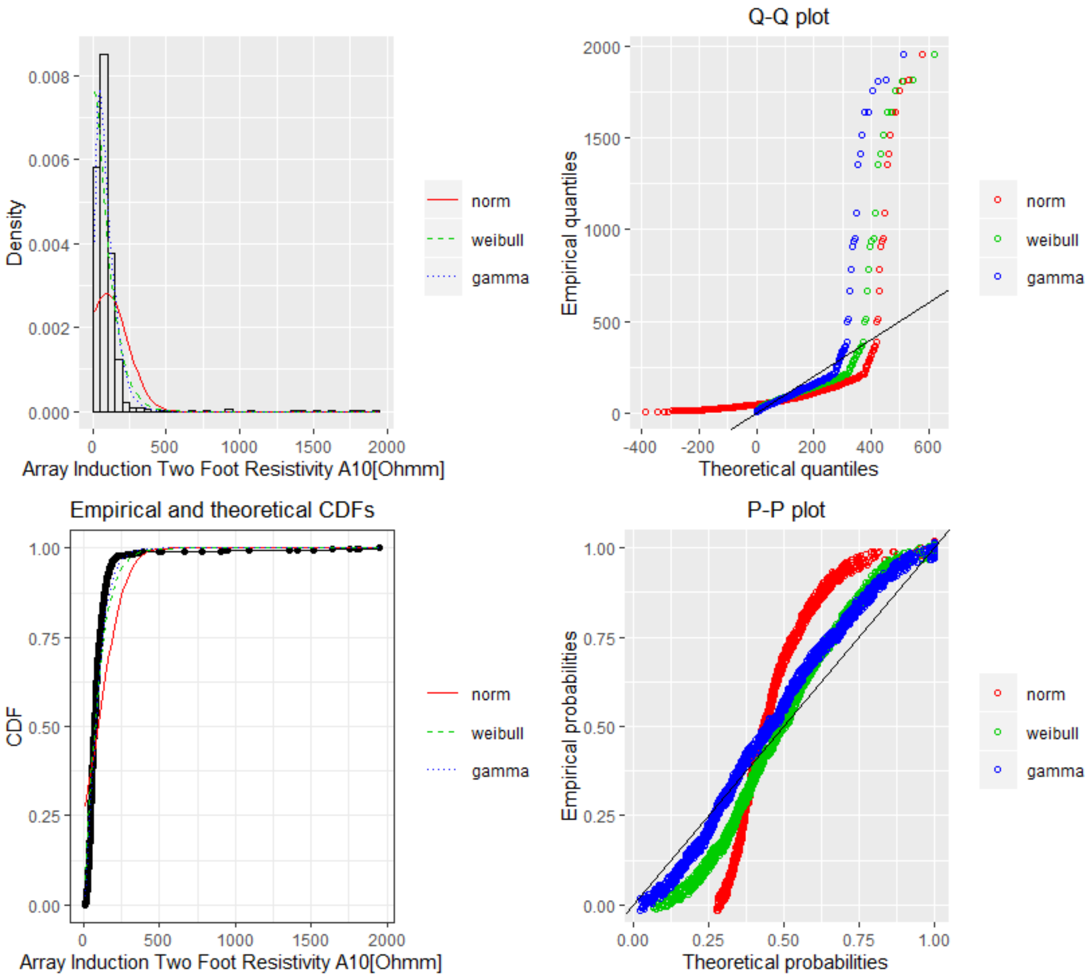


Figure 6.16: Fitting of the distributions by maximum likelihood. Featured data-set; A10 log of Seamus 216mm hole section.

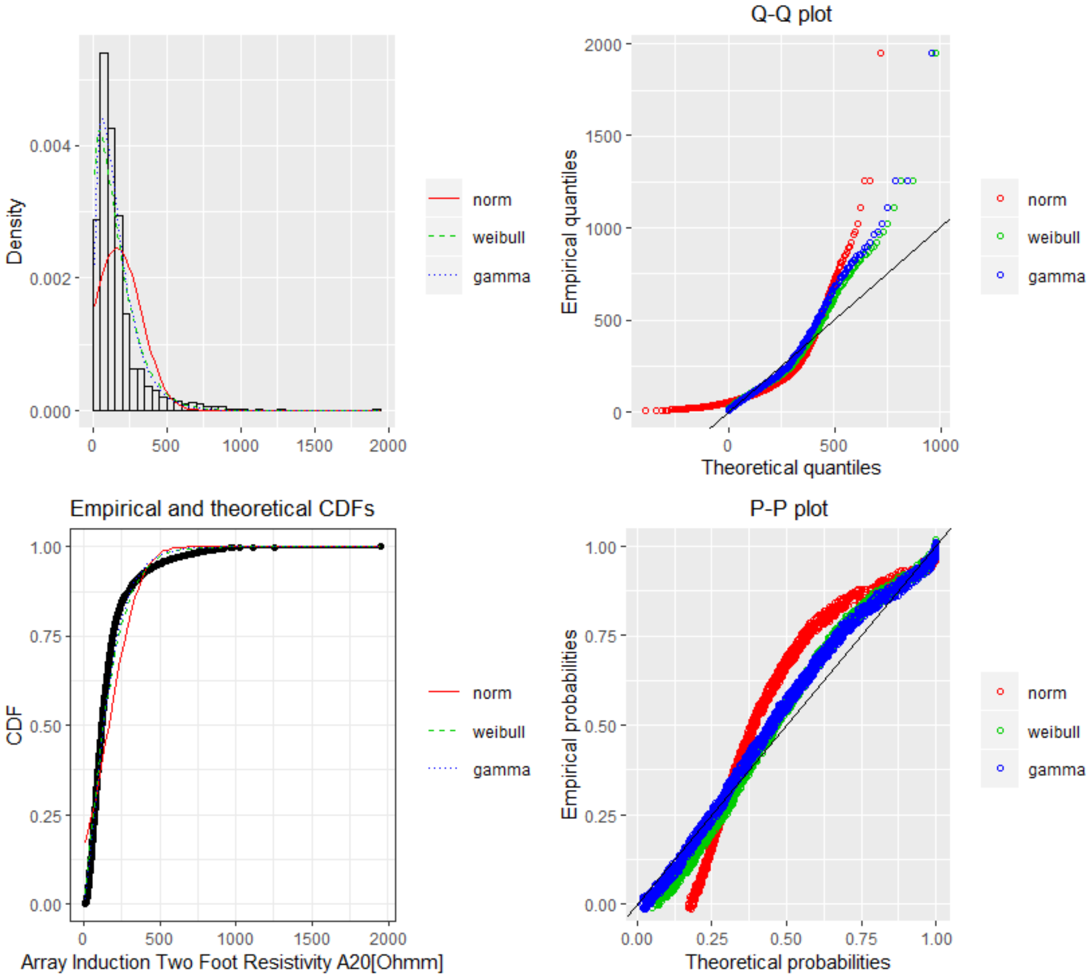
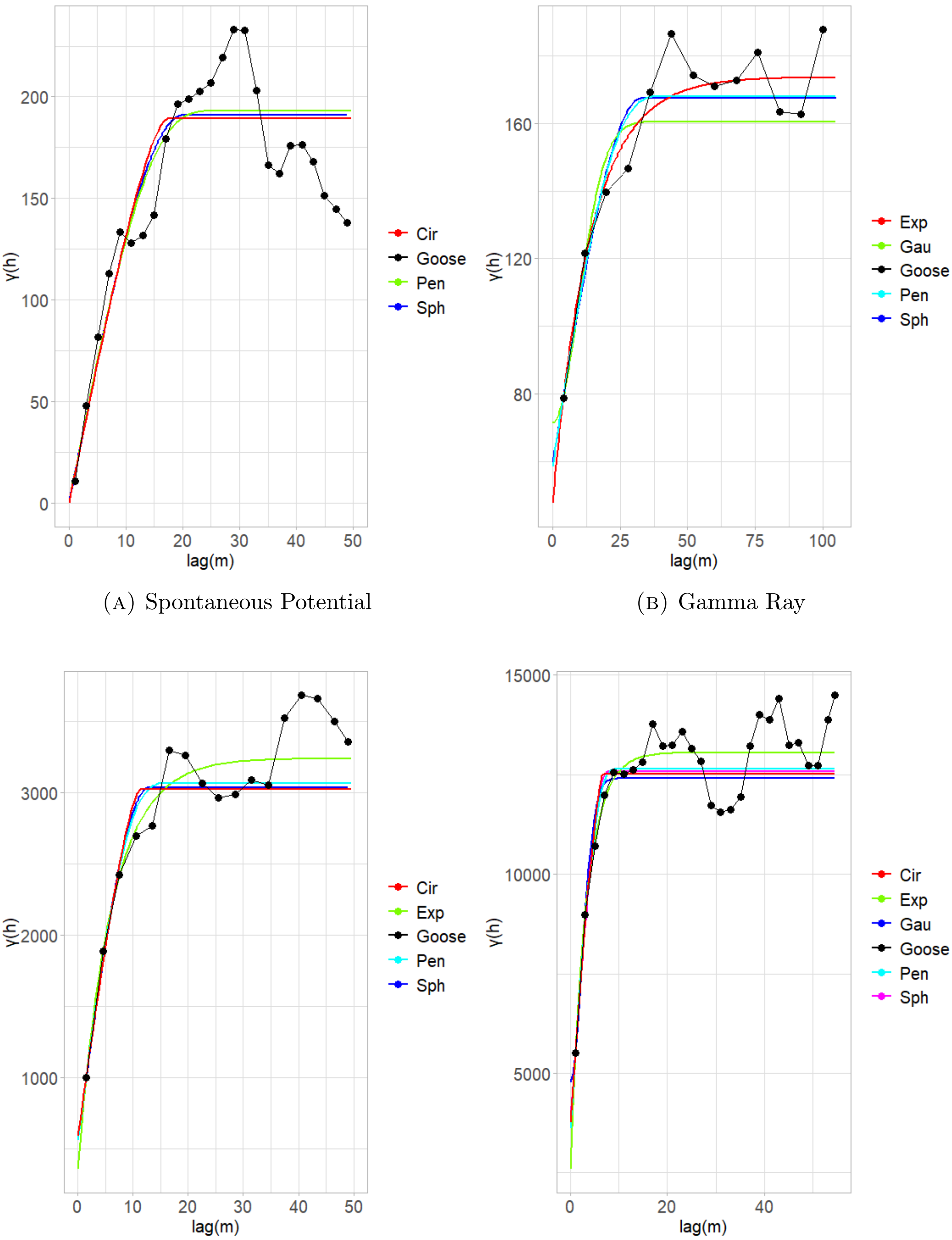


FIGURE 6.17: Fitting of the distributions by maximum likelihood. Featured data-set; A20 log of Seamus 216mm hole section.



(A) Spontaneous Potential

(B) Gamma Ray

(C) Induction A10

(D) Induction A20

FIGURE 6.18: Variogram plots. The weights are determined using N_j, where N_j is the number of pairs at certain lag.

	Variograms			
	Model	Sill	Range	Nugget
SP	Sph	191.089	21.120	0
	Pen	193.064	25.056	0
	Cir	189.217	17.192	0
GR	Sph	107.790	32.895	59.588
	Gau	88.947	13.576	71.532
	Exp	125.756	14.106	47.824
	Pen	109.254	39.894	58.691
A10	Sph	2463.633	12.880	572.131
	Exp	2880.569	6.009	358.367
	Pen	2504.862	15.942	560.626
	Cir	3431.690	11.301	592.550
A20	Sph	8888.640	7.699	3697.260
	Gau	7612.02	3.432	4789
	Exp	10460.82	3.250	2592.67
	Pen	9020.19	9.356	3617.89
	Cir	8755.98	6.664	3757.48

Table 6.13: Fitting of the best theoretical model to the experimental variograms of the field.

	Variograms			
	Model	MSE	MAE	RMSE
SP	Sph	700.453	22.935	26.466
	Pen	707.263	22.936	26.594
	Cir	711.312	23.465	46.670
GR	Sph	164.396	10.737	12.822
	Gau	180.643	10.849	13.440
	Exp	228.621	12.541	15.120
	Pen	164.04	10.769	12.808
A10	Sph	117608	274.766	342.940
	Gau	83937	249.511	289.719
	Pen	107069	259.225	327.213
	Cir	122047	282.321	349.353
A20	Sph	986212	804.021	993.082
	Gau	1060172	853.261	1029.65
	Exp	952405	731.143	975.912
	Pen	952353	784.392	975.886
	Cir	1044511	843.266	1022.01

Table 6.14: Fitting of the best theoretical model to the experimental variograms of the field.

Chapter 7

Interpolation and Imputation Methods

7.1 Missing Data

Missing data cases arise in all types of statistical analysis. In the geophysical literature, the interest rate in evaluation and prediction of a model's performance and accuracy was relatively low until the development and utilization of stimulation models became a necessity in predicting geophysical phenomena ([98], [51]).

In the beginning we need to distinguish the three major missingness patterns. Different imputation methods are requisite for different missing data patterns. Those patterns describe which values are missing and which values are observed as well as denote where those values are located in the dataset ([29]). In this thesis, only one dependent variable has missing data and thus, a *univariate* missing data pattern is formed.

Missing Data Mechanism

Missing Data Mechanisms arouse the interest of data scientists who work with missing data handling tools and methods. Those tools and methods are to a large extend dependent upon the nature of the mechanism impaired in a subset of missing values ([78]).

Let's consider a set $Y = (y_{ij})$ which is supplemented with data and an array of missing data cumulants, $M = (M_{ij})$. The mechanism of the emergence of missing data is characterized by the conditional distribution M, given Y. The above consideration can be mathematically attributed by the expression:

$$f(M|Y, \varphi) \tag{7.1}$$

where φ are the unknown parameters. We then define the missing patterns as Y_{mis}, and the observed patterns as Y_{obs}. In case the values of a missing pattern are not randomly missing, then the analysis will interpret to non-significant results.

Some analysis procedures are used only when specific missing data values are sorted into groups of order. The advantages of identifying the patterns and reasons for missing data are ([93]):

1. Classification of given data in the rows and columns in order to check in which pattern the data is imputed.
2. Finding of the appropriate technical analysis of missing data that will give rise to reliable and accurate results.

In this thesis we consider only one variable with missing data, so we distinguish the *univariate* type of pattern. In figure 7.1 the standard taxonomy of the main types of missing data patterns are displayed. There are three main mechanisms described in the literature. The *Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR)*. For the purposes of the thesis we will analyse the following two mechanisms.

- *Missing Completely at Random (MCAR)*: The missingness of the measurements are not dependent upon neither the observed nor the lost data values of the Y set. We can mathematically describe this statement as follows:

  $$f(M|Y, \varphi) = f(M|\varphi), \forall Y, \varphi$$

  Therefore, the missing variables are unrelated to the measured variables and the missingness rate is completely unsystematic. For example when data is missing for the mud pulse telemetering system for which the signal was lost due to hole sloughing, the presence of mudcake, or the invasion of the formation by drilling mud. Those factors affect the data rate transmission ([35]).

- *Missing at Random (MAR)*: The missingness of the measurements are exclusively related to the observed variables, Y_{obs} and not to the missing patterns. This is described as:

  $$f(M|Y,\varphi) = f(Y_{obs}|\varphi), \forall Y_{mis}, \varphi$$

  For example when data is missing because the mud pulse telemetry has low transmission data rate which is also affected by the input voltage threshold, pulse timing process or the pulse pressure of the fluid drilling site ([35]).

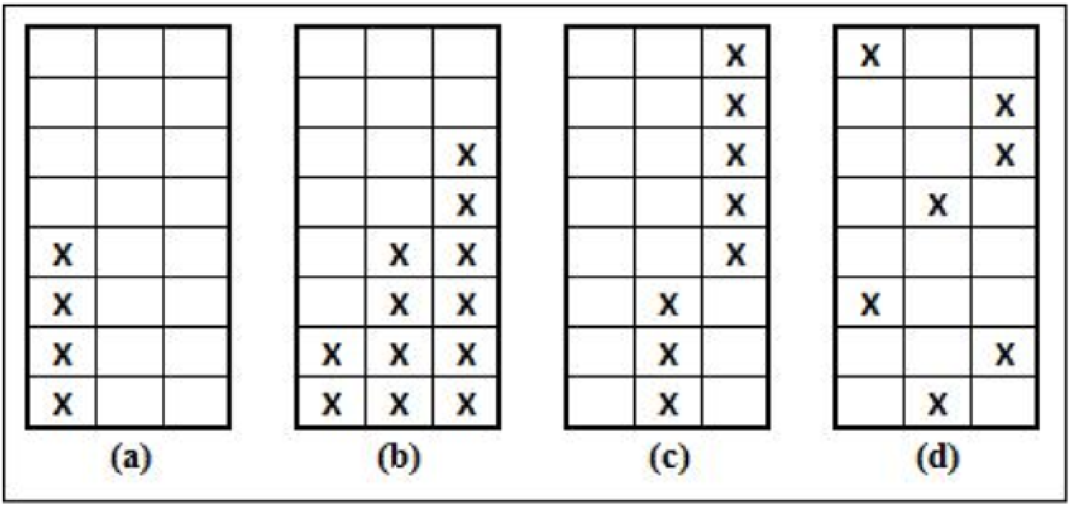


Figure 7.1: Missing Data Patterns. (a) Univariate, (b) Monotone, (c) Connected, (d) Random. The rows correspond to observations, the columns to variables. Annotated by [92].

7.2 Missing Data in a Univariate Sample

The data used for the purposes of this thesis concerns a *equi-spaced* univariate series, meaning that depth increments between successive data observations are equal, $|x_1 - x_2| = |x_2 - x_3| = \cdots = |x_{n-1} - x_n|$.

The simulation algorithm of missing values in a univariate sample data-set that describes the MCAR mechanism was introduced by [51]. In figure 7.2 the algorithm's flowchart used for this thesis is represented. The first step is the analysis of the project's documentation. The next step is to randomly delete 10% of the univariate input data. Now that 10% of the data is lost we check in which pattern the data is imputed and for that reason an univariate *t-test* comparison is used to compare the missing data subgroups. This test checks for statistically significant differences. The null-hypothesis is that the two means are equal and that the test statistics follow a Student-t distribution. The *t-test* statistic is defined by the formula:

$$t = \frac{\bar{y}_{obs} - \bar{y}_{mis}}{\frac{\sigma_1^2}{\sqrt{n_1}} + \frac{\sigma_2^2}{\sqrt{n_2}}} \tag{7.2}$$

where, $\bar{y}_{obs}$, $\bar{y}_{mis}$, σ_1^2, σ_2^2, n_1, n_2 are the mean, the variance and the sample size for the observed and the missing data, respectively. At last, the degrees of freedom ν that are associated with the variability estimate are defined. The ν parameter will eventually specify the t-distribution that is used to calculate the $p - values$ and $t - values$ for the test ([95]). Considering that the MCAR mechanism asserts that both complete and missing data belong to the same population, the null-hypothesis which defines, that the two means and variances are equal, has to be accepted accordingly. If the p-value is less than or equal to 0.05, then the null-hypothesis is accepted and the MCAR is chosen as the main Missing Data Mechanism. If the p-value is greater than 0.05 then the main Missing Data Mechanism is the MAR. The algorithm will run the same commands in order to classify the missing patterns and mechanisms for missing data rates of 0.25, 0.5, 0.8.

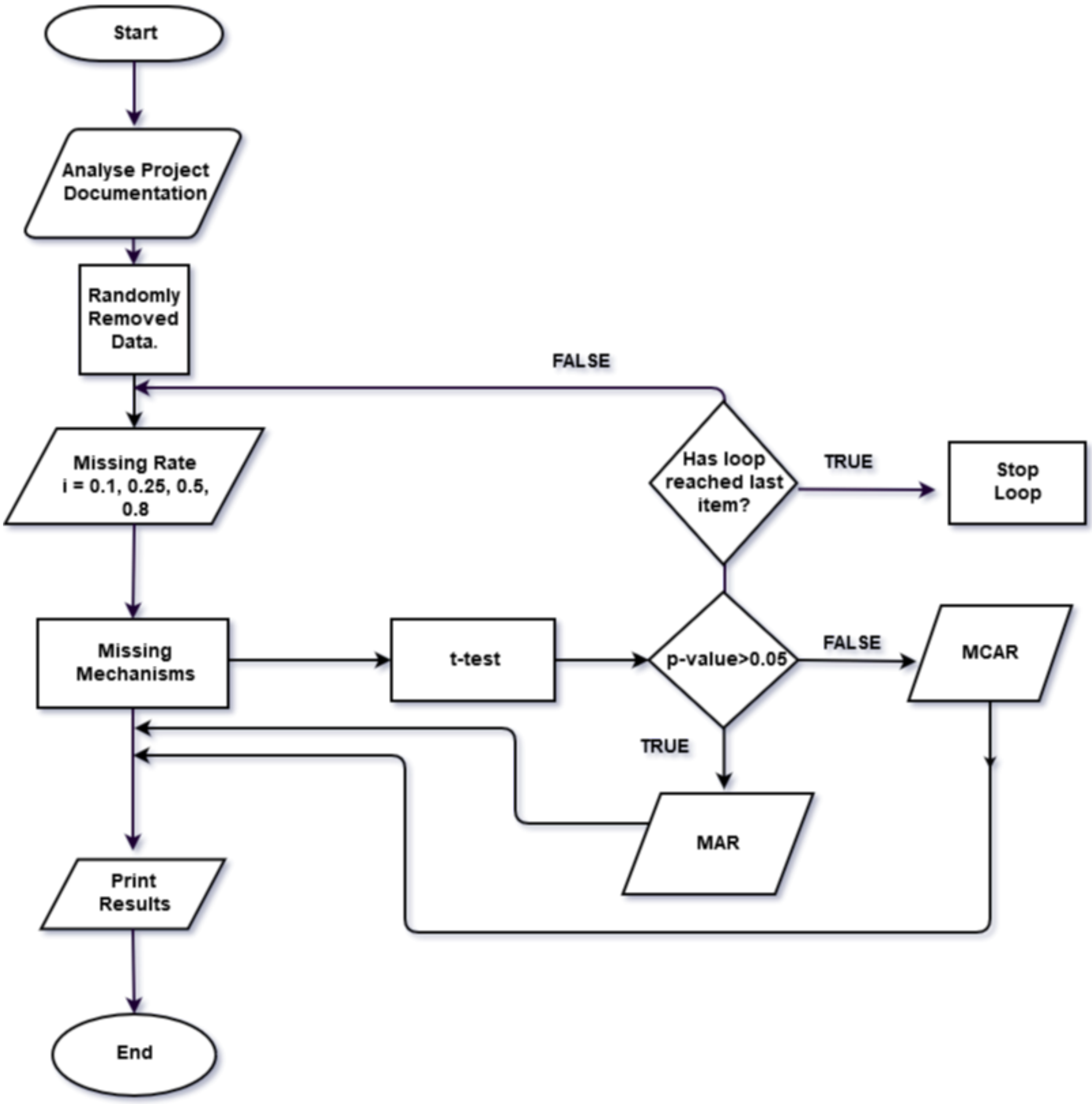


FIGURE 7.2: Algorithm flowchart of created missing data used in the univariate sample of physical properties logging measurements. The algorithm is structured based on [51].

7.3 Imputation Methods

The aim of imputation is to "preserve the characteristics of their distribution and relationships between different variables" as noted by [6].

Consider Y, as a completely observed $n \times p$ matrix and a X as a partially observed $n \times p$ matrix of the complete sample data Y. Imputation techniques are applied to the aforementioned X matrix in order to fully record a matrix Y^* that is the approximation of the previously considered Y matrix. Several methods are reported in the literature to address the process of imputation. In the following lines the main techniques applied in this thesis are described.

7.3.1 Mean Imputation

One of the easiest ways to fill in each missing value is with the sample mean of the corresponding variable of the valid value units. Nonetheless, the major disadvantage of the mean imputation is that it reduces the variability, since all imputed values are equal to mean. That also affects other inferential statistics which are also underestimated, such as the standard deviation and the confident intervals. The method results to bias mean estimates when data are not MCAR. This method should be generally avoided and only be used as a rapid fix when, for example, the handful information is not or hardly related to the studied variable. Let $\hat{y}^*$ the imputed values of the studied observation y. Then the imputed values are estimated by the observed mean by the following formula:

$$\hat{y}^* = \frac{1}{N} \sum_{i \in obs} y_i \tag{7.3}$$

The y_i is defined as the i-th observed value on a set of observed units, while N is the number of the i-th observed values for the studied variable y.

7.4 Kalman Filter

State Space Form

The *State Space Model* was originally developed by electrical engineers to control linear dynamic systems in either continuous or discrete forms. The way a system changes is a function of the current state of the system which can be influenced by external input state variables. Those are defined as the minimum variables that fully describe the studied system. Therefore, the derivatives of a dynamic system are a function of both the current state as well as any external inputs. We can simply describe the state space modelling process as a repackaging of the high order differential equations into a set of first order equations. Thus, we can look at the underlying behavior of the interconnected system as well as how the system is affected by external or even multiple external inputs.

For the purpose of this thesis, we will determine a set of vectors $x_1, x_2, \dots, x_n$ which we will assume to be an unobserved series of unobserved values associated with an observed series of observed values $y_1, y_2, \dots, y_n$. The defined relationship between those two vectors is described by a state space model.

The simplest way to describe a time series state model is by a time series additive form; *additive = trend + seasonality + system noise* or $y_t = \tau_t + s_t + \epsilon_t$ for $t = 1, \dots, n$. A suitable model is then constructed for the trend and seasonal component using a random walk model y_i of size n. Therefore, $y_{t+1} = y_t + \eta_t$, where η_t are i.i.d. random variables of zero mean and variance σ^2_η. Considering that the random walk is a non-stationary process, we conclude that the model is non-stationary as well. Differencing is a technique used to make a model stationary.

A full description of the General Linear Gaussian time series space model is described in the following lines ([30]):

$$\begin{aligned} y_t &= Z_t a_t + \epsilon_t, \epsilon_t \sim N(0, H_t) \\ a_{t+1} &= T_t a_t + R_t \eta_t, \eta_t \sim N(0, H_t) \end{aligned} \tag{7.4}$$

The considered simple classical state model of a random walk plus measurement error exhibits the characteristics of a state model structure. The y_t is the observation equation with $a_1, a_2, \dots, a_n$ unobserved values that form the a_t state equation. The y_t is a $p \times 1$ observation vector and the a_t is a $m \times 1$ state vector. Considering the above, the analysis must be based on the observations y_t. The matrices Z_t, T_t, R_t, H_t and Q_t are assumed to be known.

The analysis of trend, seasonal and error components of the time series will be examined by simple generated state space models.

Trend Component

The model of the trend component is given by the following equations:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_\epsilon) \\ \mu_{t+1} &= \mu_t + v_t + \xi_t, \quad \xi_t \sim N(0, \sigma^2_\xi) \\ v_{t+1} &= v_t + \zeta_t, \quad \zeta_t \sim N(0, \sigma^2_\zeta) \end{aligned} \tag{7.5}$$

where v_t is a slope term generated by a random walk. If the variances of ξ and ζ are both greater than zero then the the trend level and slope will produce a different trend state over time. In the case when the error measurements of ξ and ζ are equal to zero then the slope term remains constant over time while the state equation of a future observation μ_{t+1} is dependent upon the previous observation μ_t and the slope term in that way that the trend becomes linear. Eventually the produced equation will be reduced to the deterministic linear trend and noise model.

Seasonal Component

The model for the seasonal component, when the seasonal pattern is constant over time, say, s, is modelled by the constant γ_j^*, for $j = 1, \ldots, s$, and is given by the form

$$\sum_{j=1}^{s} \gamma_j^* = 0 \tag{7.6}$$

Since, in practise, the seasonality changes over time, we assume an added potential error ω_t in the above equation, considering a $j - th$ number of seasons in the data, for $j = 1, \ldots, s$, and $\gamma_t = \gamma_j^*$, since the observations of the model are constant seasonal components. Thus, the following equations are formed

$$\begin{gathered} \sum_{j=0}^{s-1} \gamma_{t+1-j} + \omega_t = 0 \\ or \\ \gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t \end{gathered} \tag{7.7}$$

where $\omega_t \sim N(0, \sigma_\omega^2)$

ARIMA Models in State Space Form

The components of a state space model are consider the same as in time series y_t, based on [12]. First, the trend and seasonal component needs to eliminated from the series by differencing. Thus, the produced model will have a stationary behavior, meaning that means and covariances will remain invariant over the course of time.

In section 4.4.3 the ARIMA non-negative integers (p, d, q) where defined. The number of differences d is defined by the transformation $\Delta y_t = y_t - y_{t-1}$ and $\Delta^d y_t = \Delta^{d-1}(\Delta t)$ for the first and $d - th$ differences used to eliminate the trend component. At the important special case when there is a seasonal component, and s is the number of seasons in the data, the $\Delta_s y_t = y_t - y_{t-s}$ and $\Delta_s^d y_t = \Delta^{d-1}(\Delta_s t)$ are the first and $s - th$ differences used to eliminate the seasonal component. Finally, when stationarity is achieved the transformed variables are defined as

$$y_t^* = \Delta^d \Delta_s^D y_t \tag{7.8}$$

where $d, D = 0, 1, \ldots$ and now a stationary autoregressive moving average model ARMA(p, q) equation is modelled given the following form:

$$y_t^* = \varphi_1 y_{t-1}^* + \cdots + \varphi_p y_{t-p}^* + \zeta_t + \theta_1 \zeta_{t-1} + \cdots + \theta_q \zeta_{t-p} \tag{7.9}$$

where $\zeta_t \sim N(0, \sigma_\zeta^2)$ is an i.i.d. series of error measurements. The above equation can be rewritten as:

$$y_t^* = \sum_{j=1}^{r} \varphi_j y_{t-j}^* + \zeta_t + \sum_{j=1}^{r-1} \theta_j \zeta_{t-j} \tag{7.10}$$

where $t = 1, \ldots, n$ and $r = max(p, q + 1)$ considering the fact that some coefficients are zero.

Kalman Filter

The Kalman Filter is used to fit an ARIMA model in a time series and is a appropriate form for online real time processing. The Kalman Algorithm Filter initially calculates the distribution of the current state model by taking into consideration the available observation until a certain time, for each time period. Thus, the unobserved state is estimated under the conditions that this estimation is irrelevant to the future observed states. Additionally, it estimates the maximum likelihood of the data in a way that the ARIMA model fits the data optimally. The Kalman Filter can be used to correct ARIMA forecast results by removing measurement errors.

7.5 Interpolation Methods

In section 6.4, a short introduction about what are the interpolation processes was made. In this section, we will use two interpolation methods, named, *Linear* and *Spline* interpolation to predict missing values of a discrete time series.

7.5.1 Linear Interpolation

Linear Interpolation is a method of approximation of the value of function $f(x)$, at a specific point $\hat{x}$ that interprets between to known points x_1, x_2, when $x_1 < \hat{x} < x_2$. We estimate the value of the function $f(\hat{x})$ using a linear line that passes through points $(x_1, f(x_1))$ and (x_2, f_{x2}). Those conditions are satisfied when the linear function is calculated by the formula

$$\begin{aligned} y(\hat{x}) - y(x_1) &= \frac{y(x_2) - y(x_1)}{x_2 - x_2}(\hat{x} - x_1) \\ &\quad\quad\quad\quad\quad\quad\quad\quad \mathit{or} \\ y(\hat{x}) &= \frac{f(x_1(x_2 - \hat{x})) + f(x_2)(\hat{x} - x_1)}{x_2 - x_1} \\ &\quad\quad\quad\quad\quad\quad \hat{x} \in [x_1, x_2] \end{aligned} \tag{7.11}$$

where, $y(x_1) = f(x_1)$ and $y(x_2) = f(x_2)$ with estimated error: $R^f = \frac{f''(\xi)}{2}(\hat{x} - x_2)(\hat{x} - x_2)$, when $\xi \in [x_1, x_2]$.

7.5.2 Spline Interpolation

We need to estimate a function, say $s(x)$, which is defined from a set of point $[x_i, s(x_i)]$, for $i = 0, 1, \ldots, n$, by using low order polynomials pieces on sub-intervals joined together with certain continuity conditions in a domain of the function, $x_0 \leq x \leq x_i$.

A cubic spline $S_{3,i}(x)$ is a piece-wise of third order polynomials. Let's consider a cubic polynomial form: $S_{3,i}(x_i) = a_i + b_i(x - x_1)^2 + d_i(x - x_i)^3$, for $i = 0, 1, \ldots, n - 1$. The four unknown coefficients need to be specified in order to find the cubic splines. Thus,

$$S_{3,i}(x_i) = s(x_i) \tag{7.12}$$

for $i = 0, 1, \ldots, n$. The first $n+1$ conditions are based upon the fact that the S_3 function has to pass through all the points of its domain. Moreover, $n-1$ conditions can possible be produced by the equivalence of the neighboring polynomials at the joint points. Thus,

$$S_{3,i}(x_i) = S_{3,i+1}(x_i) \tag{7.13}$$

for $i = 1, 2, \ldots, n-1$. Additionally the equivalence of the first and second order derivatives of the function, at the same points, can ensure extra $2n-2$ conditions. That is,

$$\begin{aligned} S'_{3,i}(x_i) &= S'_{3,i+1}(x_i), i = 1, 2, \ldots, n-1 \\ S''_{3,i}(x_i) &= S''_{3,i+1}(x_i), i = 1, 2, \ldots, n-1 \end{aligned} \tag{7.14}$$

Therefore, there are a total of $4n-2$ linear constraints on the 4n unknown coefficients and we need two extra constrains. The additional constrains can be specified by the following various ways

- *Natural Cubic Splines.* The imposed conditions are $S''_{3,i+1}(x_i) = 0$ and $S''_{3,1}(x_0) = 0$.
- *Not–a–knot.* The imposed conditions are $S'''_{3,i}(x_i) = S'''_{3,i+1}(x_i)$ and $S'''_{3,1}(x_1) = S'''_{3,2}(x_1)$.
- *Complete cubic spline.* The imposed conditions are $S'_{3,i}(x_i) = f'(x_i)$ and $S'_{3,0}(x_0) = f'(x_0)$.

Chapter 8

Gap Filling

Usually, the gaps of logging records are rather small, especially when compared to the total depth of a well. In this section we take a topic in well log time series analysis where missing data can be estimated by means of interpolation and imputation. This section seeks to address the following concepts:

- Missing data imputation, interpolation and time series analysis algorithms are used to improve missing well log data quality.
- Prediction precision between the original and the imputed time series data is used to quantify the performance of the predictive modeling methods.

Preliminary Data Analysis

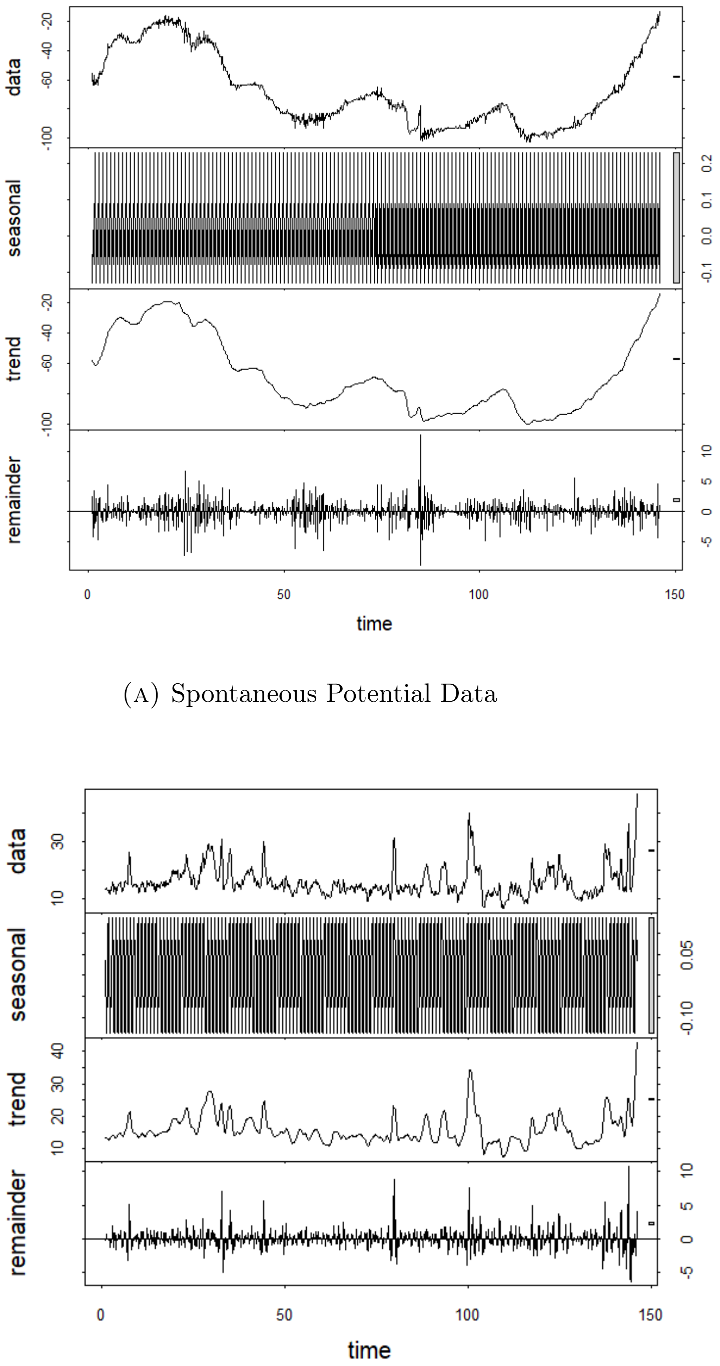
Firstly we will properly convert the scale of depth axis to the scale of time axis. Then we simulate missing values on continuous data sets by performing imputation algorithms of Kalman Smoothing (KS) with a ARIMA model, Spline Interpolation, Linear Interpolation, Simple Moving Average, Linear Weighted Moving Average and Mean Imputation models and then finally compare them to the original selected data sets.

For our example, we select the Table Point Formation data set of the Seamus well. In our case, we need to analyze a formation that is present in both Seamus and Finnegan hydrocarbon and gas wells. We use the available *Spontaneous Potential*, *Gamma Ray* and *Array Induction Two Resistivity* logs to demonstrate our experiment. Missing completely at random (MCAR) and Missing at random (MAR) were used as a generated missing value mechanism ([78]).

8.1 Data Characterization of Table Point Formation

Decomposition

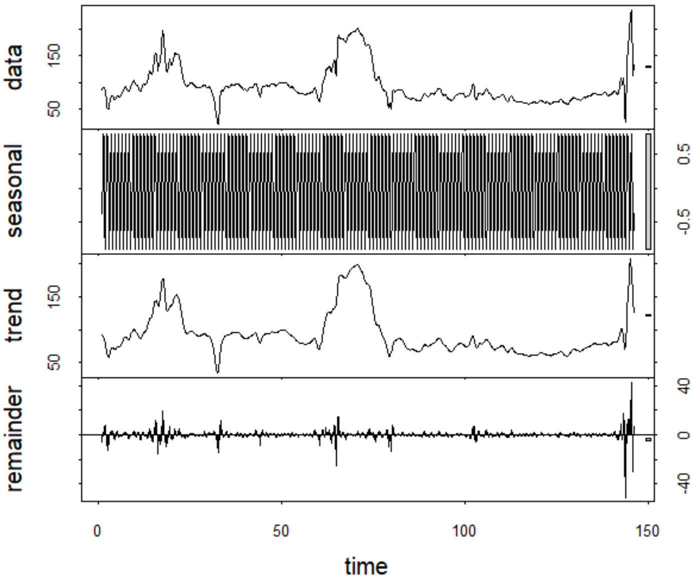
The four different well log data sets of Table Point Formation of the Seamus well, consist of n=871 observations. Before we implement the imputation algorithms, we need to decompose the time series in order to examine their characteristics as refered in section 3.3. The STL (Seasonal and Trend decomposition using Loess) method of decomposition is performed to split the time series into seasonality, trend and remainder component using the *stl* function in *R*. From figures 8.1a to 8.2b we extract the following considerable information. The Spontaneous Potential (Figure 8.1a), Gamma Ray (Figure 8.1b), Array Induction Two Resistivity A10 (Figure 8.2a) and Array Induction Two Resistivity A20 (Figure 8.2b) data-sets show no apparent trend and no regular seasonality and display non-stationary and non-linear characteristics. This is quite common due to well log data complex behavior which is a result of several factors affecting the signal transmission and recording system. Petrophysical properties of the porous media, such as the porosity, permeability and water saturation of the reservoir rock as well as the drilling mud composition, mud weight, mud cake and casing can significantly effect the record and display of the well logging sound waves signals. Those effects must be accounted for to obtain accurate measurements.



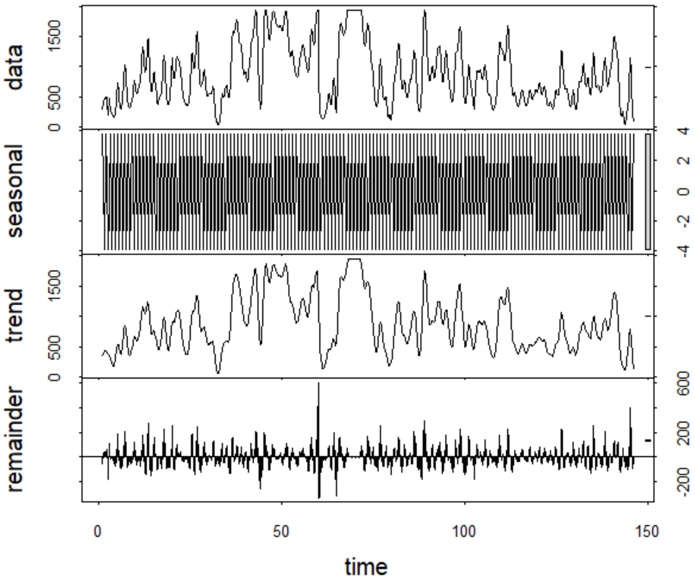
(A) Spontaneous Potential Data

(B) Gamma Ray Data

Figure 8.1: STL Decomposition



(A) Array Induction Two Resistivity A10 Data



(B) Array Induction Two Resistivity A20 Data

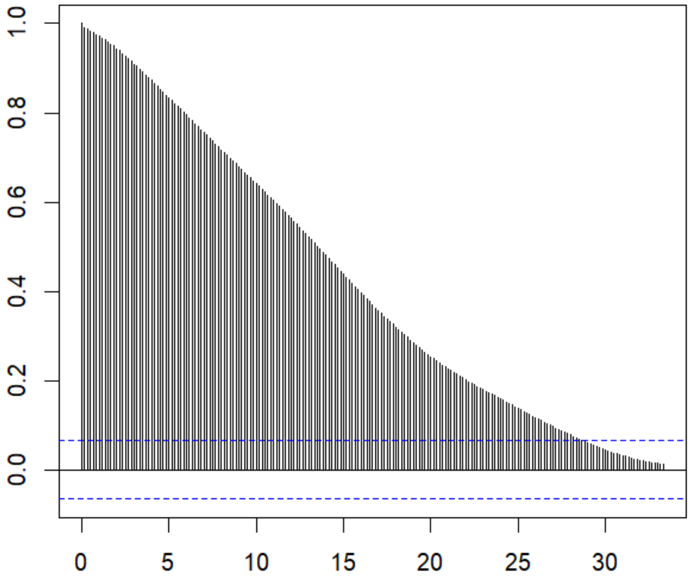
Figure 8.2: STL Decomposition

Autocorrelation Function

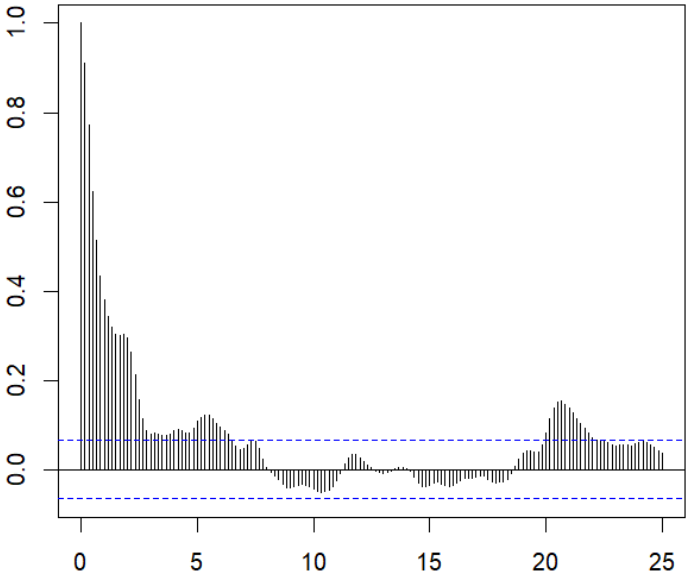
The next step of the analysis is to detect non-randomness in data and at the same time to identify an appropriate forecast and imputation model if our data is a result of a non-random process. Indications of strong correlation across all the lags suggest that the future observations are highly dependent on available past observations, thus the predictions and imputations would be both accurate and precise. In figures 8.3a to 8.4b the lag is returned in units of time. The blue dotted lines indicate bounds for statistical significance. The horizontal lines are at a distance of $\pm 2/\sqrt{n} = \pm 2/\sqrt{871}$. The three first following correlograms demonstrate signs of non-stationary behavior due to very slow decrease of ACF, which means that the mean will change over time. We will compute the KPSS test to accept or reject the null-hypothesis that the series is stationary. In 8.3a the *p-values* $= 0.083$, while in 8.3b, 8.4a and 8.4b the *p-value* is less than 0.01, meaning that in all cases the null-hypothesis is rejected. This behavior is also confirmed by the ADF test. The results of ADF test for the raw data of the given data sets confirm the assumptions of non-stationary. In 8.3a the *p-values* $= 0.99$, in 8.3b the *p-values* $= 0.085$, in 8.4a the *p-values* $= 0.837$ and in 8.4b the *p-values* is less than 0.01. For the three fist cases, the *p-values* of the ADF test is less than the critical value 0.05 and the assumptions about the non-stationarity is confirmed. In the case of 8.4b the *p-values* of the test confirm the null-hypothesis of stationarity. By comparing the two tests we conclude that only in the case of 8.4b the two tests suggest that the time series is stationary.

- In figure 8.3a there is a strong positive correlation decreasing over the course of time.
- In figure 8.3b the autocorrelation function demonstrates a slowly decreasing process and then, at lag 6, it reaches the boundaries of the confidence interval under which the values with either positive or negative change are no longer statistically significant.
- In figure 8.4a the time series dies off positively and slowly. After the lag 10, the autocorrelation function continues to decreases and becomes negative; corresponding to the presence of a trend component.

- In figure 8.4b the time series is trended, since the autocorrelations are large and positive for short lags and then decreasing slowly for large lags.

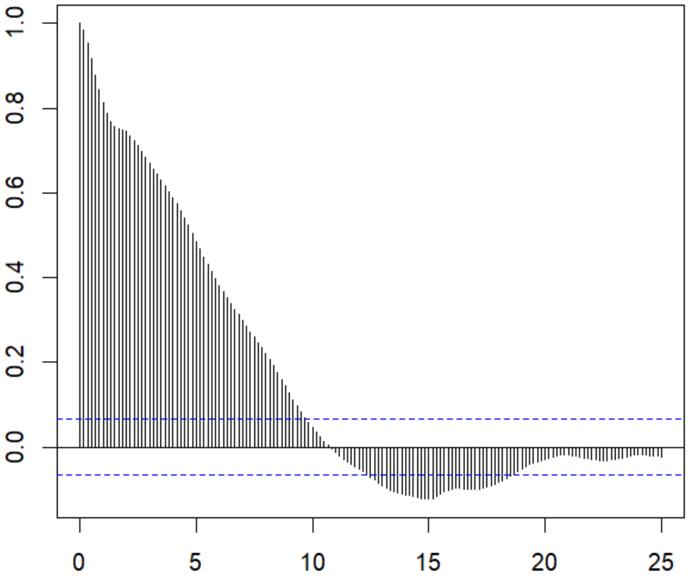


(A) Spontaneous Potential Data

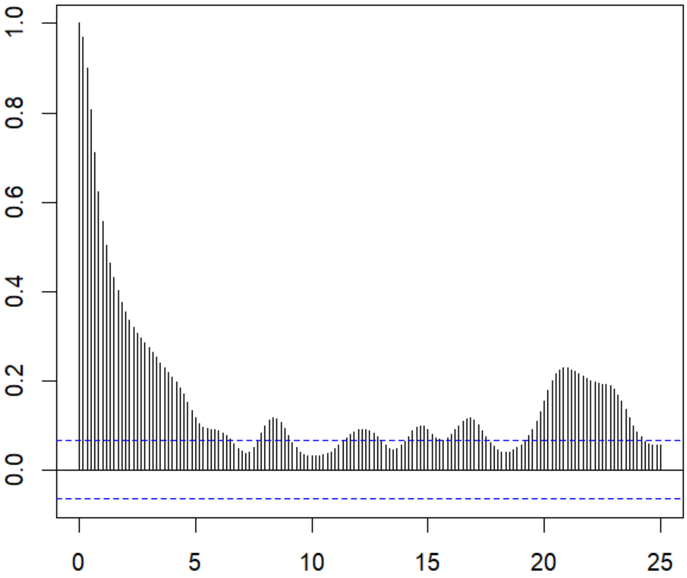


(B) Gamma Ray Data

Figure 8.3: Autocorrelation Function



(A) Array Induction Two Resistivity A10 Data



(B) Array Induction Two Resistivity A20 Data

Figure 8.4: Autocorrelation Function

8.2 Imputation Algorithms

Spontaneous Potential

In error metrics 8.5 and 8.6 two type of errors where calculated, the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), respectively. Results demonstrated that the Linear Weighted Moving Average and Simple Moving Average performance is almost identical. Obviously, the Mean Imputation algorithm exhibits the worst performance due to the presence of a strong trend component. In general, all the rest of the algorithms performed in a similar way, producing more accurate predictions for the observable rate of missingness equal to 0.1. The occurrence of very few high outliers in some cases is the result of . Based on the produced figures we can cite that RMSE and MAPE lead to similar results. The corresponding histograms and scatter diagrams of the original and estimated values of the several missingness factor 0.1, 0.25, 0.5, 0.8 are presented in figures 8.7, 8.8, 8.9 and 8.10, respectively. The model used is the Kalman ARIMA. Generally, the estimated values follow the original observations for a missing rate of 0.1, without, however, exhibiting satisfying proximity of the total distribution. As the missing rate increases to a maximum rate of 0.8, the distribution's' convergence weakens.

Gamma Ray

In error metrics 8.11 and 8.12 the same two type of errors where calculated, the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), respectively. Results show that Linear Interpolation and Linear Weighted Moving Average show the best modeling performance. The Spline Interpolation model display small error values for missing rate equal to 0.1, 0.25 and 0.5, while for rate equal to 0.8 show great variance. On the other hand, the Kalman Arima model produces a few extreme error values including the small rates of missing values. The Mean Imputation algorithm exhibits the worst performance due to the presence of a trend component. Based on the produced figures we can cite that RMSE and MAPE lead to similar results. The corresponding histograms and scatter diagrams of the original and estimated values of the several missingness factor 0.1, 0.25, 0.5, 0.8 are presented in figures 8.13, 8.14, 8.15 and 8.16, respectively. The model used is the Kalman ARIMA. Generally, the estimated values follow the original observations for a missing rate of 0.1, without, however, exhibiting satisfying proximity of the total distribution. As the missing rate increases to a maximum rate of 0.8, the distribution's convergence weakens.

Array Induction Two resistivity A10

In error metrics 8.17 and 8.18 the two type of errors where calculated, the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), respectively. Results show that Linear Interpolation and Linear Weighted Moving Average show the best modeling performance and they are almost identical. The Spline Interpolation model display small error values for missing rate equal to 0.1, 0.25 and 0.5, while for rate equal to 0.8 show great variance. On the other hand, the Kalman Arima model produces a few extreme error values including the small rates of missing values. The Mean Imputation algorithm exhibits the worst performance due to the presence of a trend component. The Simple Moving Average shows good performance. Based on the produced figures we can cite that RMSE and MAPE lead to similar results. The corresponding histograms and scatter diagrams of the original and estimated values of the several missingness factor 0.1, 0.25, 0.5, 0.8 are presented in figures 8.19, 8.20, 8.21 and 8.22, respectively. The model used is the Kalman ARIMA. Generally, the estimated values follow the original observations for a missing rate of 0.1, without, however, exhibiting satisfying proximity of the total distribution. As the missing rate increases to a maximum rate of 0.8, the distribution's convergence weakens.

Array Induction Two resistivity A20

In error metrics 8.23 and 8.24 the two type of errors where calculated, the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), respectively. Results show that Kalman Arima, Linear Interpolation and Linear Weighted Moving Average show the best modeling performance and they are almost identical. The Kalman Arima model produces one extreme error value at missing rate equal to 0.5. The Spline Interpolation model display small error values for missing rate equal to 0.1, 0.25 and 0.5, and then for rate equal to 0.8 show great variance. The Mean Imputation algorithm exhibits the worst performance due to the presence of a trend component. The Simple Moving Average shows good performance. Based on the produced figures we can cite that RMSE and MAPE lead to similar results. The corresponding histograms and scatter diagrams of the original and estimated values of the several missingness factor 0.1, 0.25, 0.5, 0.8 are presented in figures 8.25, 8.26, 8.27 and 8.28, respectively. The model used is the Kalman ARIMA. Generally, the estimated values follow the original observations for a missing rate of 0.1, without, however, exhibiting satisfying proximity of the total distribution. As the missing rate increases to a maximum rate of 0.8, the distribution's convergence weakens.

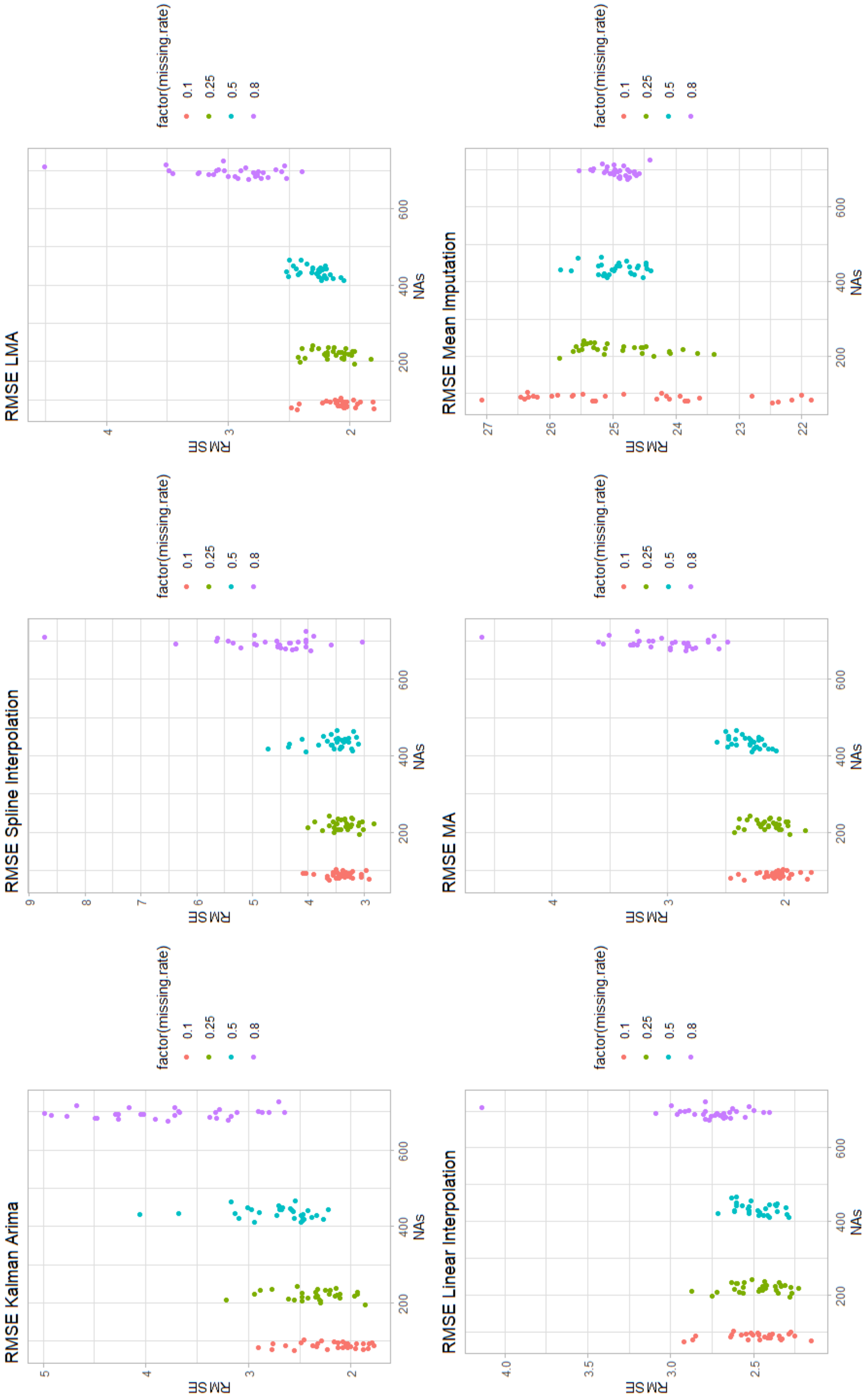


Figure 8.5: RMSE of Spontaneous Potential Data

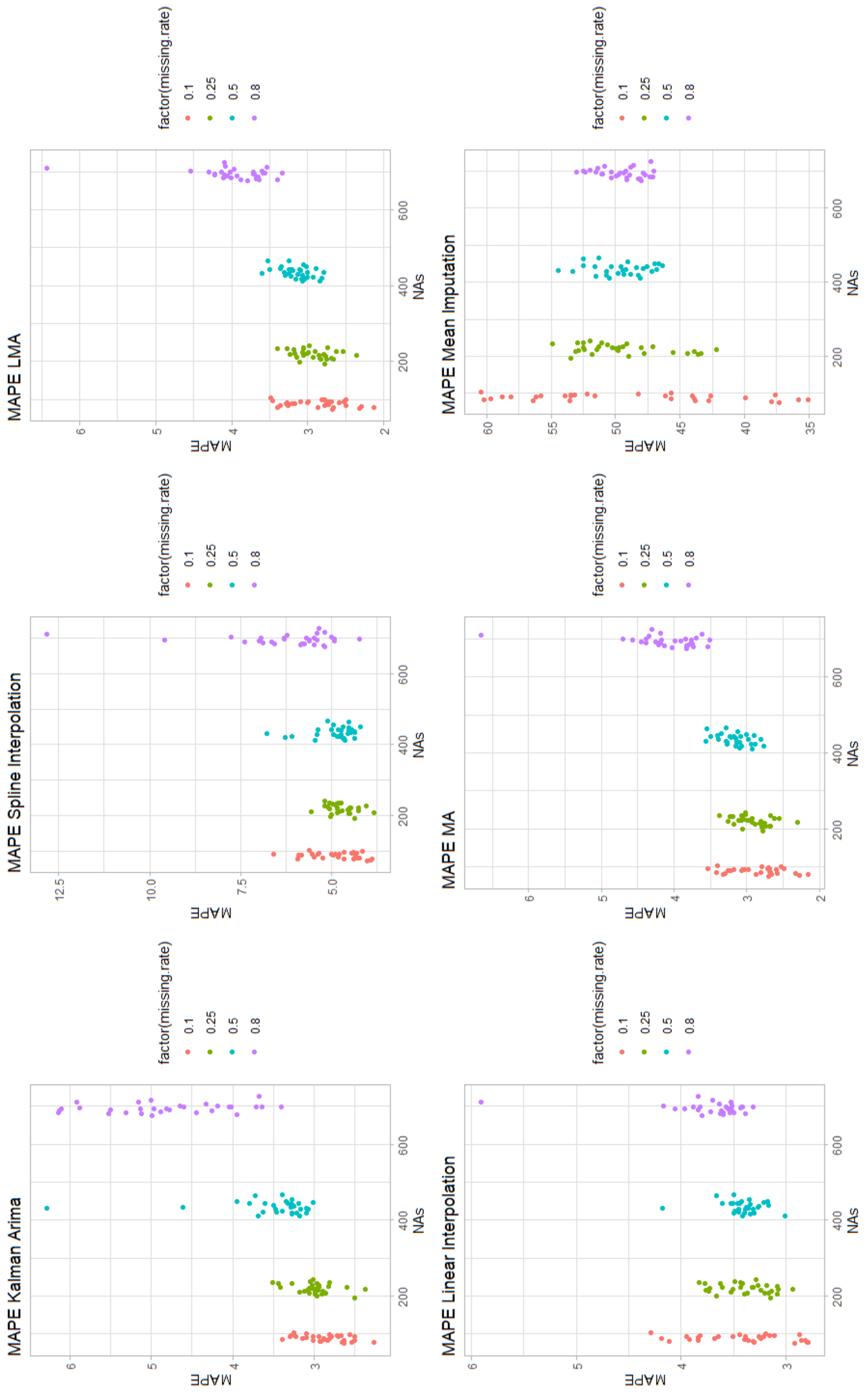
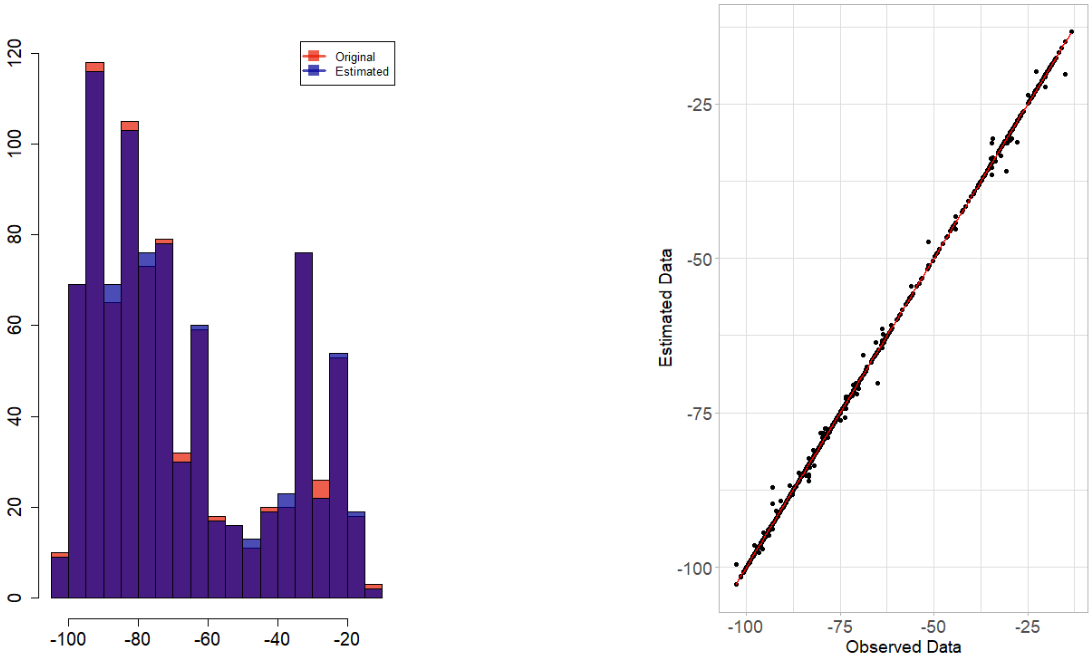
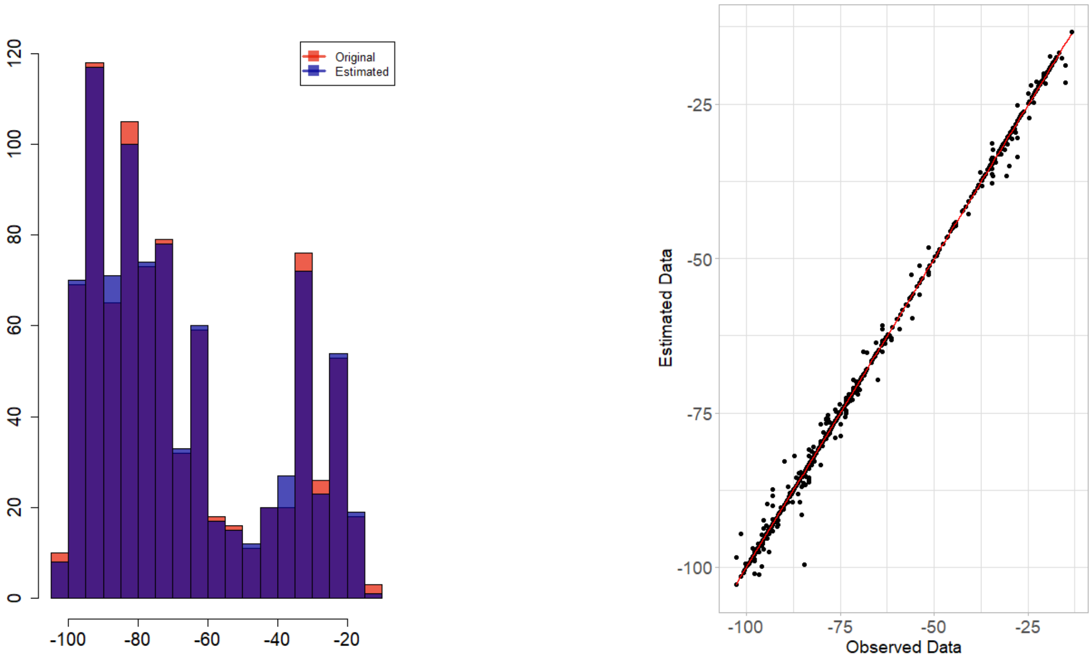


Figure 8.6: MAPE of Spontaneous Potential Data



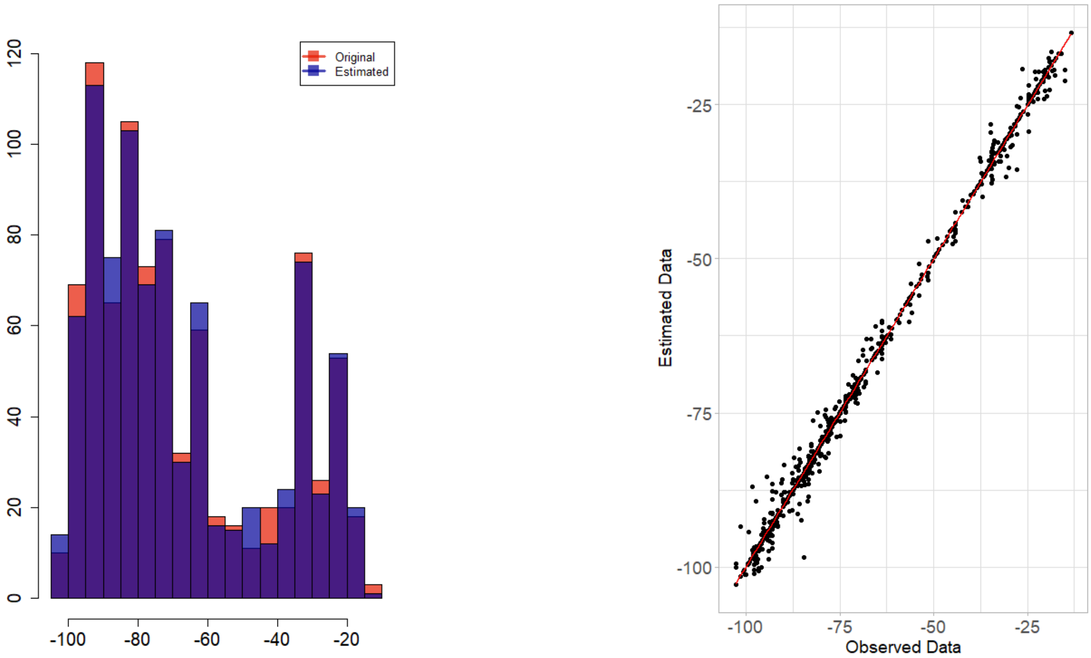
(A) (B)

FIGURE 8.7: Histogram and Scatter plot of the Spontaneous Potential original and estimated values when the missing rate of the data is 0.1. The missing values are imputed by Kalman Arima.



(A) (B)

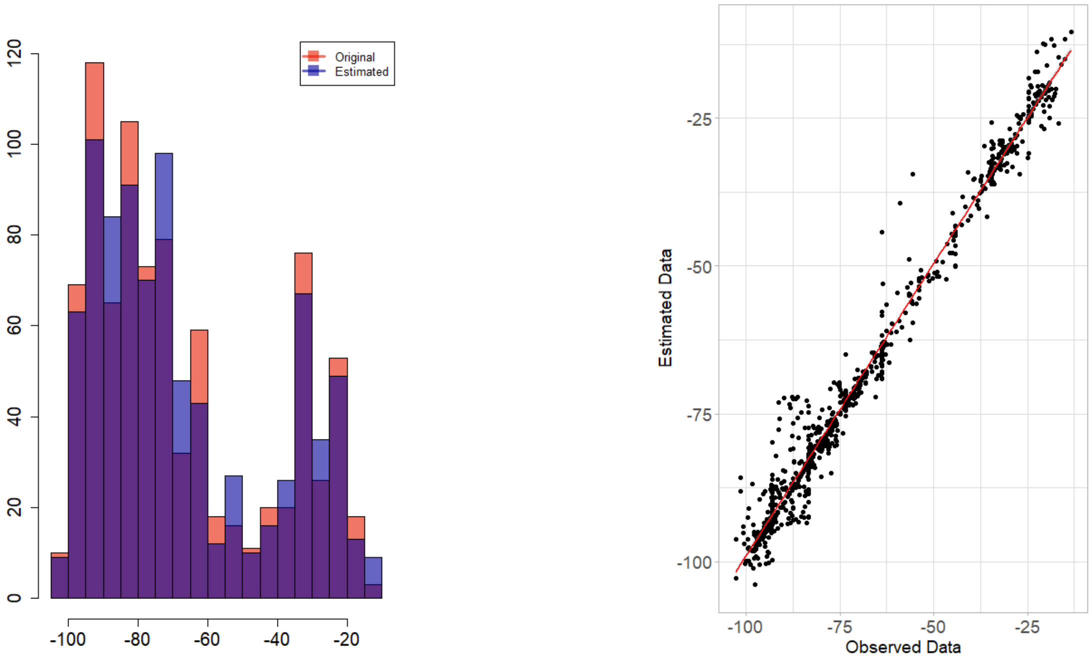
FIGURE 8.8: Histogram and Scatter plot of the Spontaneous Potential original and estimated values when the missing rate of the data is 0.25. The missing values are imputed by Kalman Arima.



(A)

(B)

FIGURE 8.9: Histogram and Scatter plot of the Spontaneous Potential original and estimated values when the missing rate of the data is 0.5. The missing values are imputed by Kalman Arima.



(A)

(B)

FIGURE 8.10: Histogram and Scatter plot of the Spontaneous Potential original and estimated values when the missing rate of the data is 0.8. The missing values are imputed by Kalman Arima.

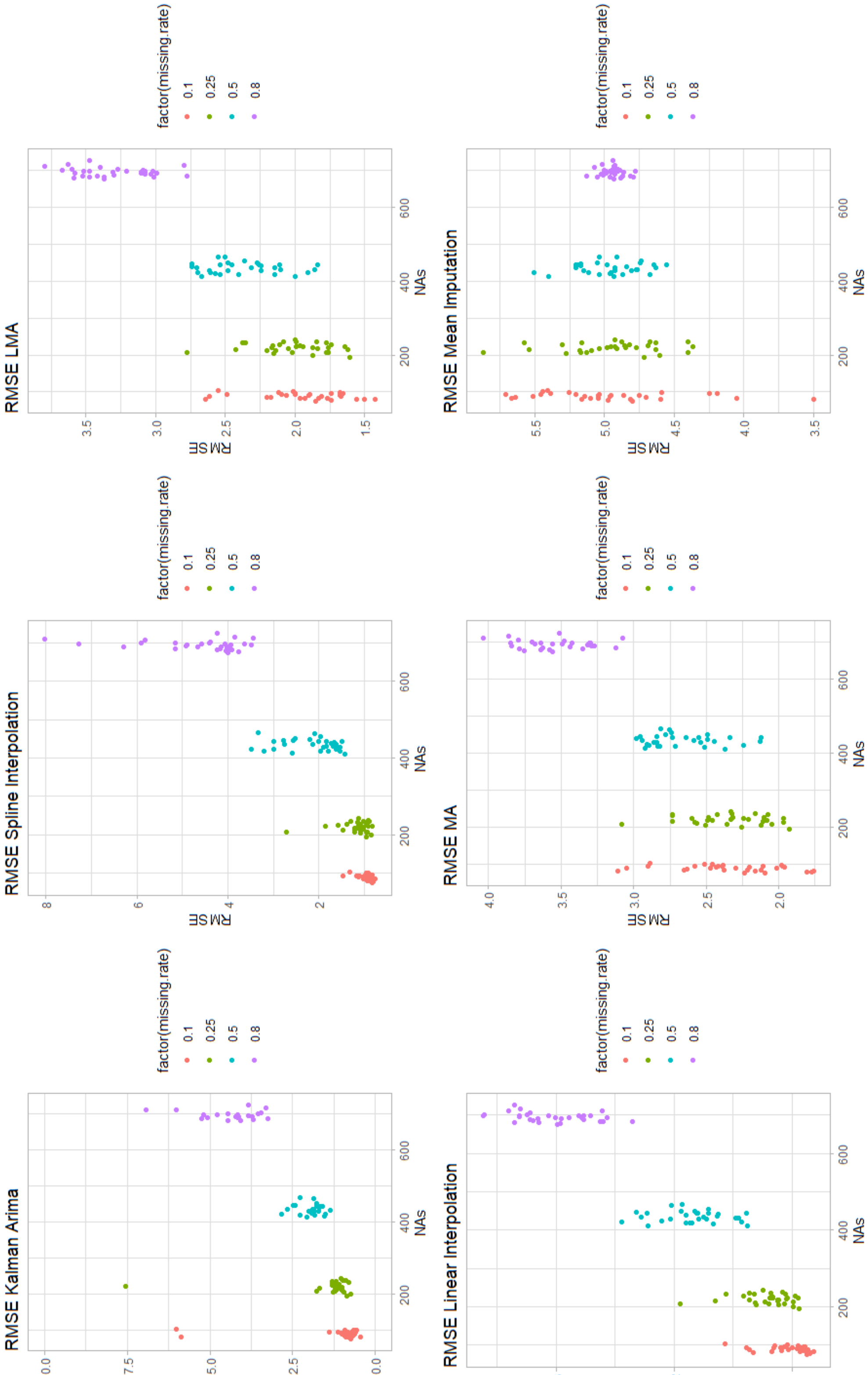


Figure 8.11: RMSE of Gamma Ray Data

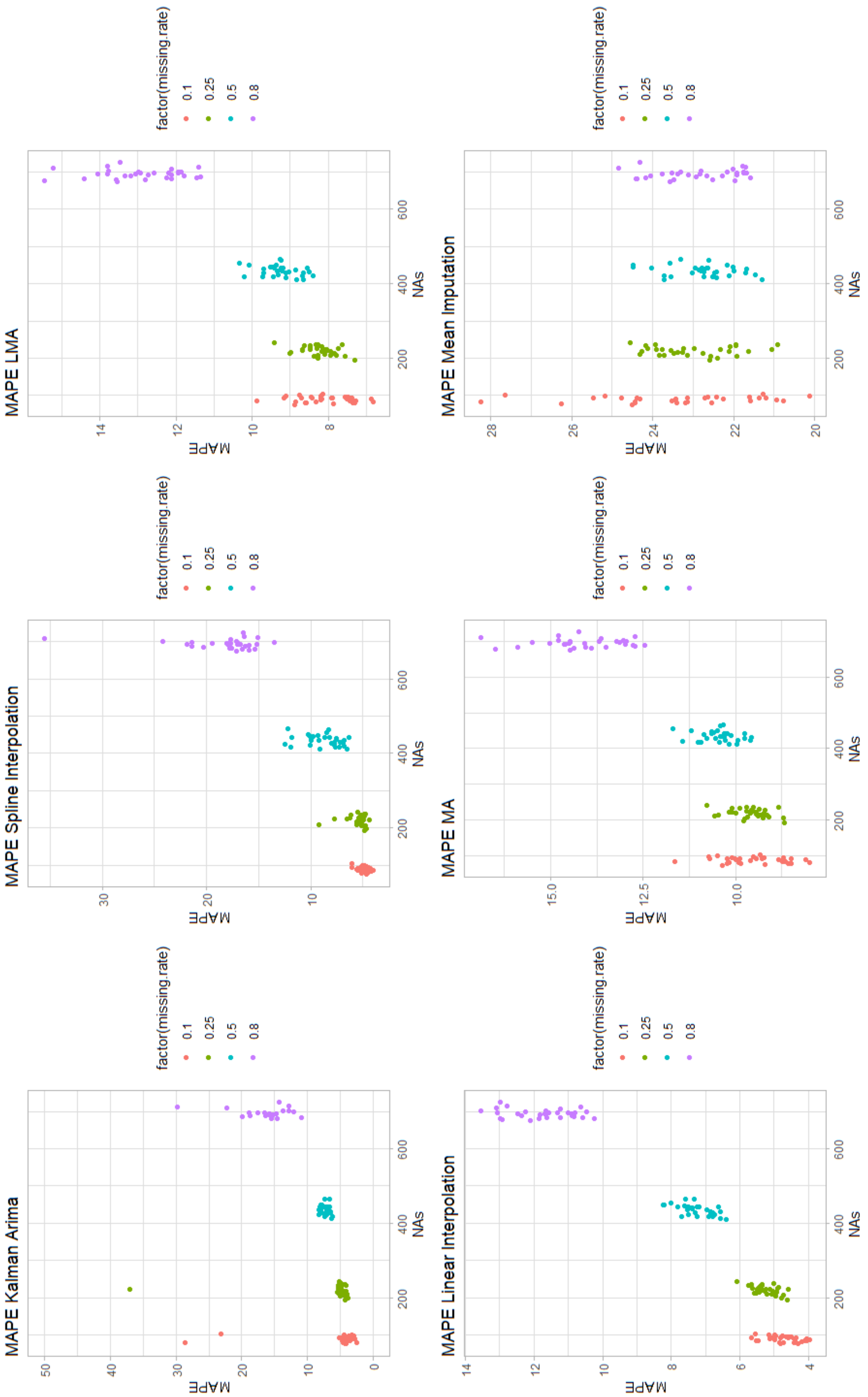
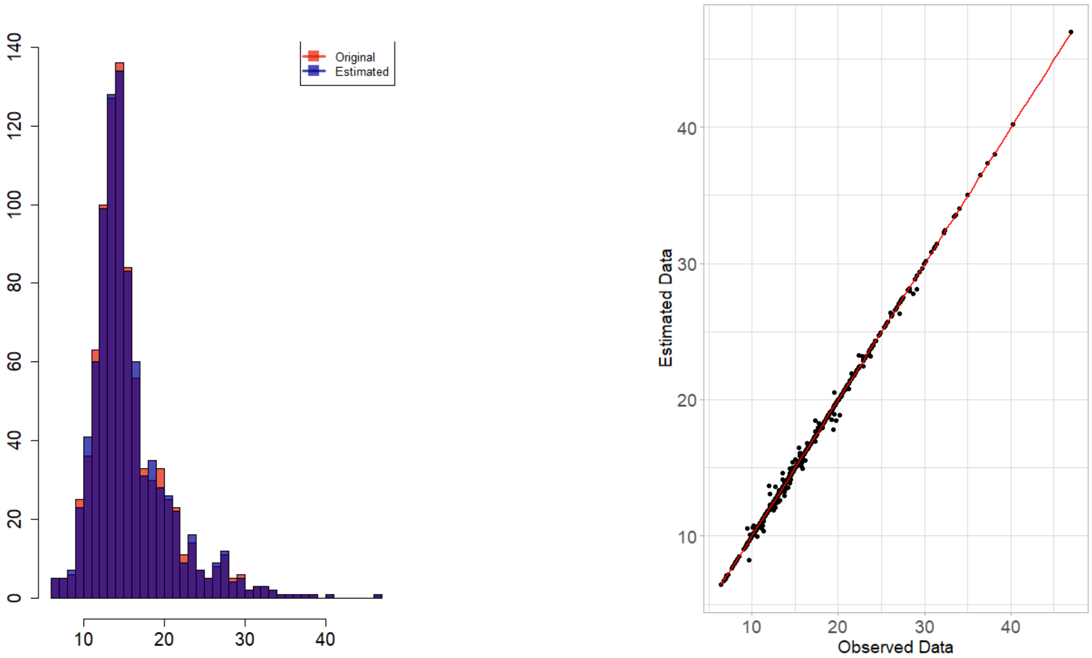


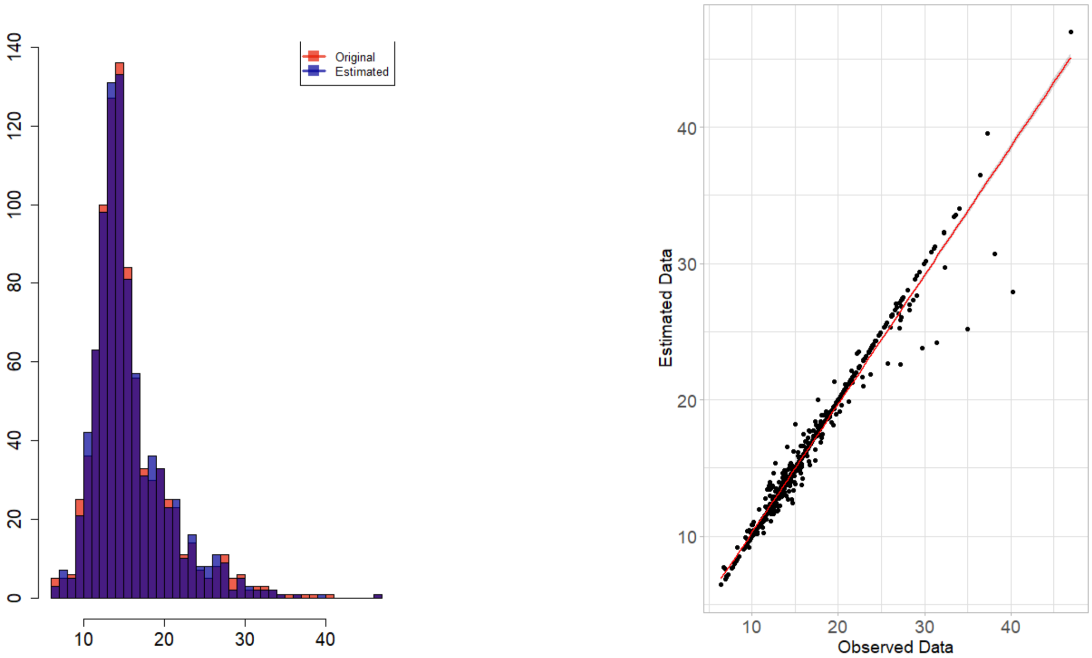
FIGURE 8.12: MAPE of Gamma Ray Data



(A)

(B)

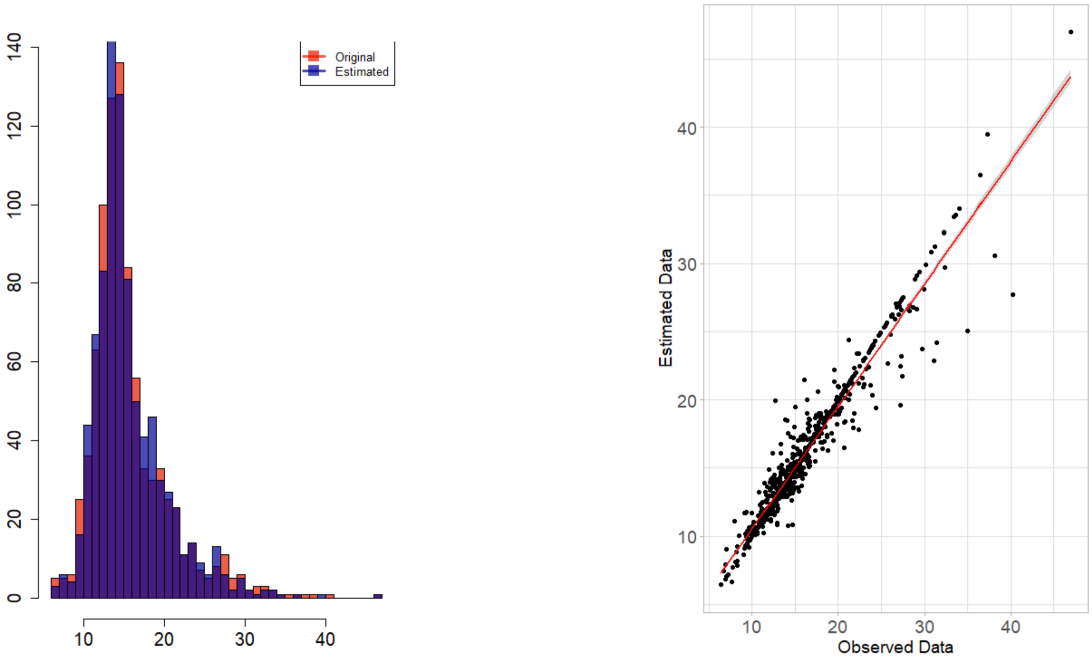
FIGURE 8.13: Histogram and Scatter plot of the Gamma Ray original and estimated values when the missing rate of the data is 0.1. The missing values are imputed by Kalman Arima.



(A)

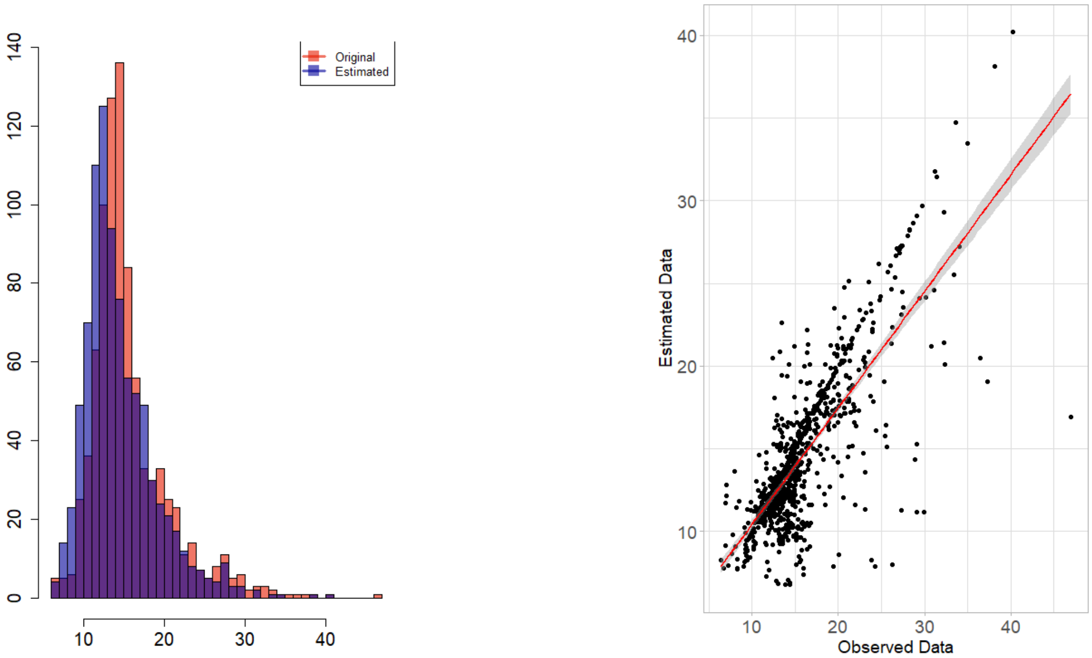
(B)

FIGURE 8.14: Histogram and Scatter plot of the Gamma Ray original and estimated values when the missing rate of the data is 0.25. The missing values are imputed by Kalman Arima.



(A) (B)

Figure 8.15: Histogram and Scatter plot of the Gamma Ray original and estimated values when the missing rate of the data is 0.5. The missing values are imputed by Kalman Arima.



(A) (B)

Figure 8.16: Histogram and Scatter plot of the Gamma Ray original and estimated values when the missing rate of the data is 0.8. The missing values are imputed by Kalman Arima.

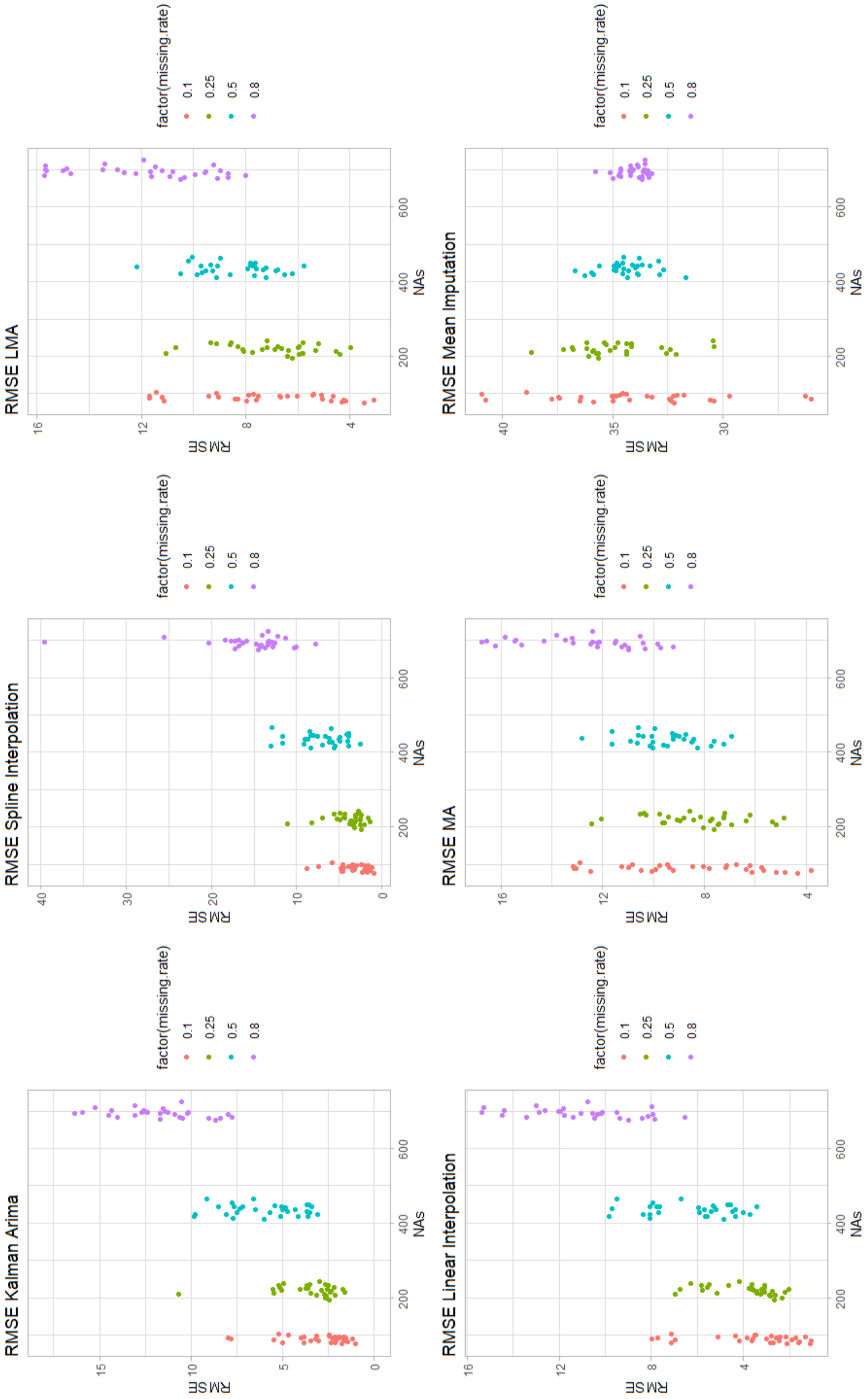


FIGURE 8.17: RMSE of Array Induction Two Resistivity A10 Data

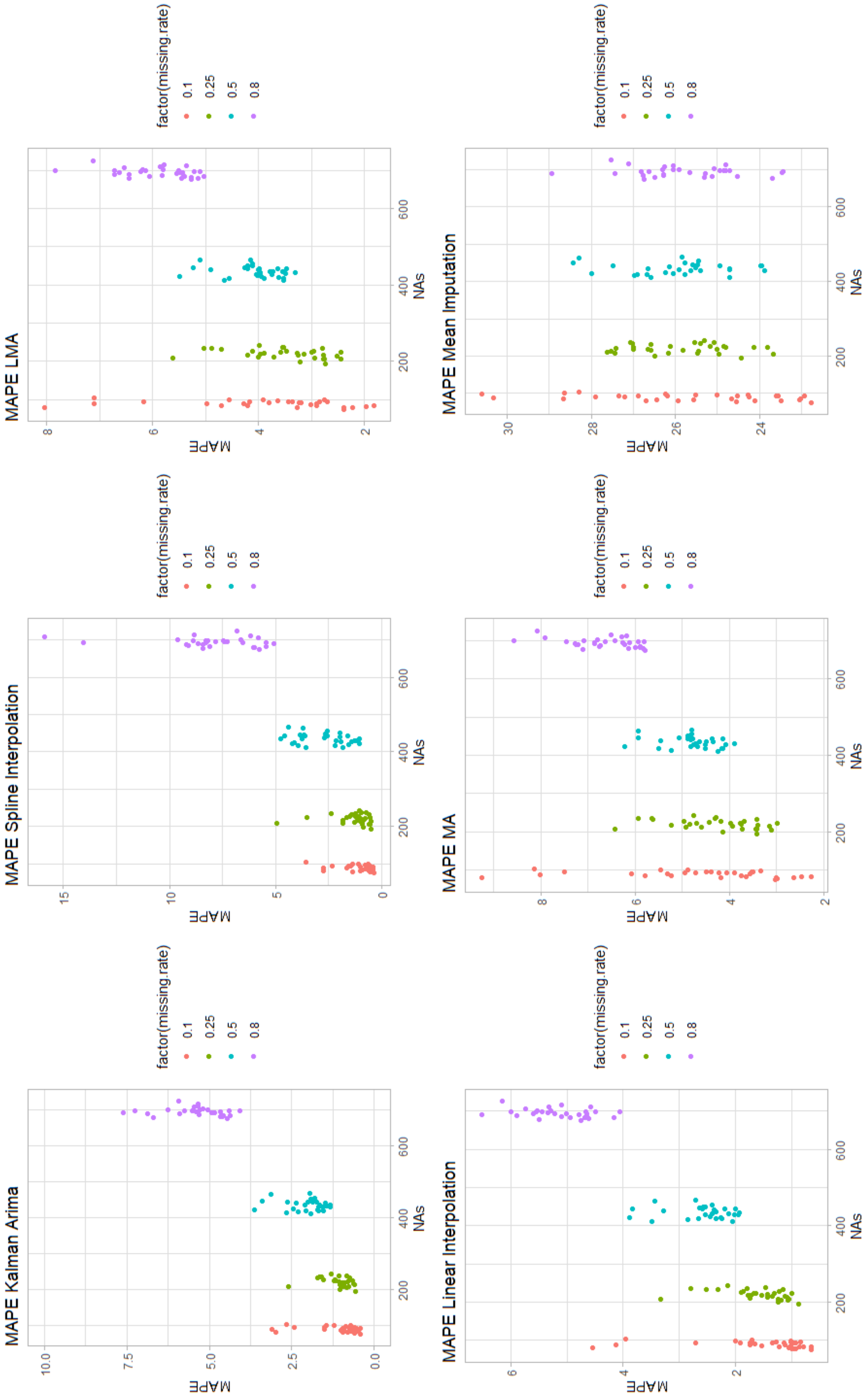
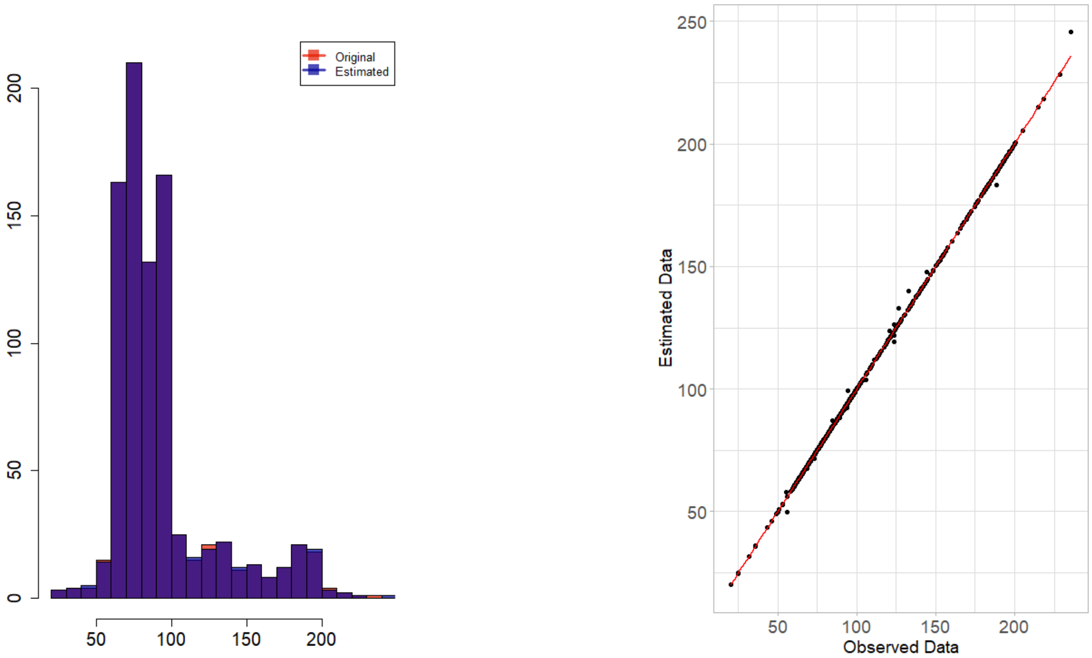


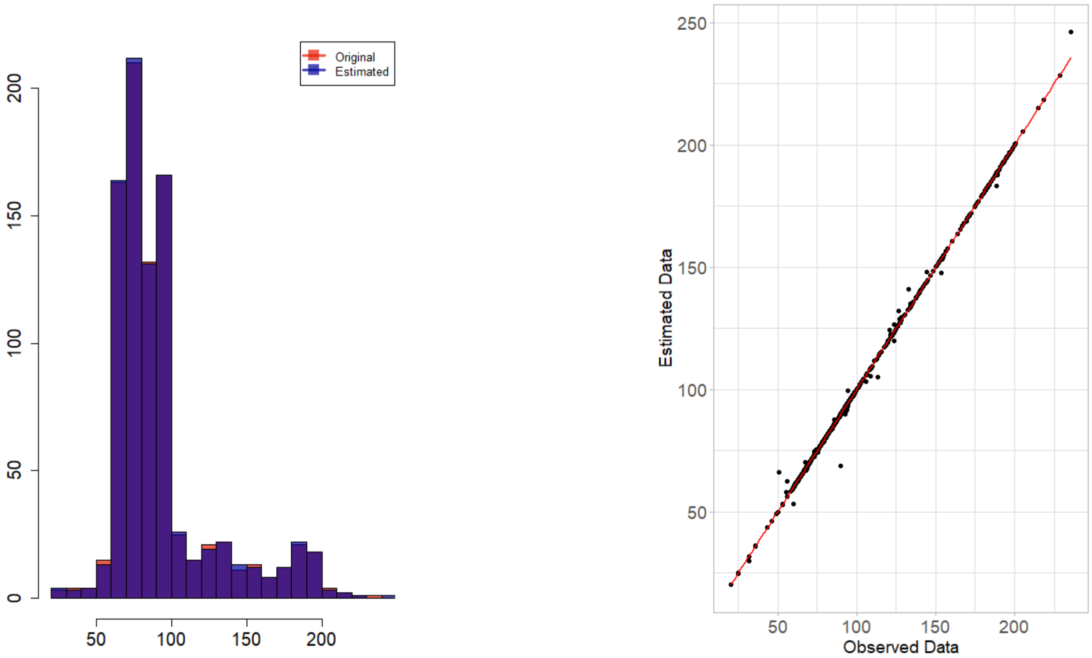
FIGURE 8.18: MAPE of Array Induction Two Resistivity A10 Data



(A)

(B)

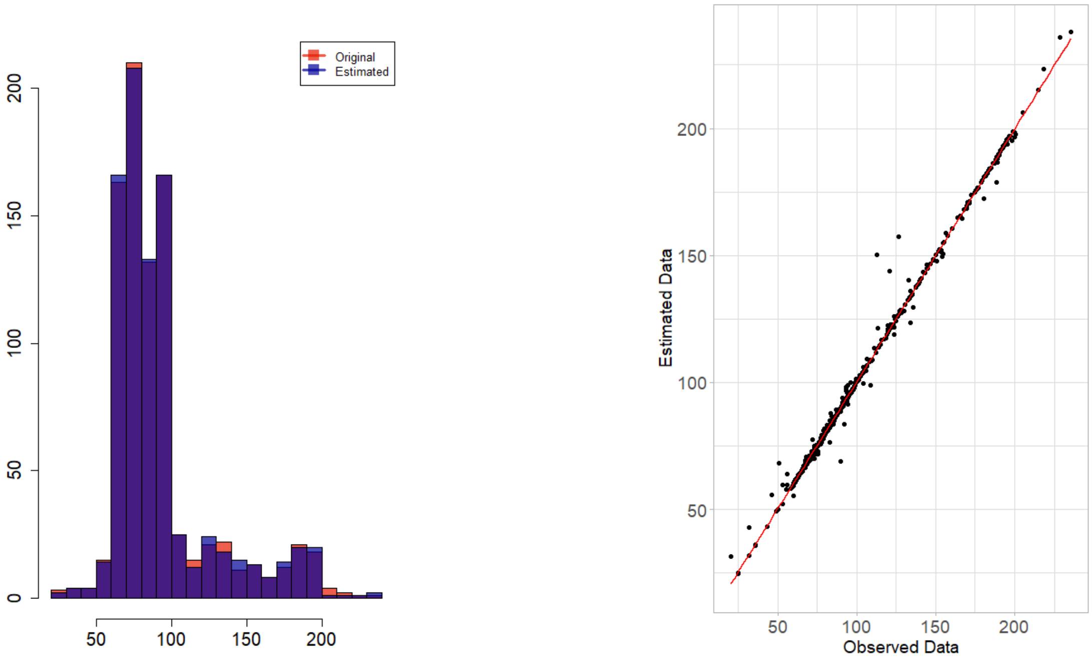
FIGURE 8.19: Histogram and Scatter plot of the Array Induction Two Resistivity A10 original and estimated values when the missing rate of the data is 0.1. The missing values are imputed by Kalman Arima.



(A)

(B)

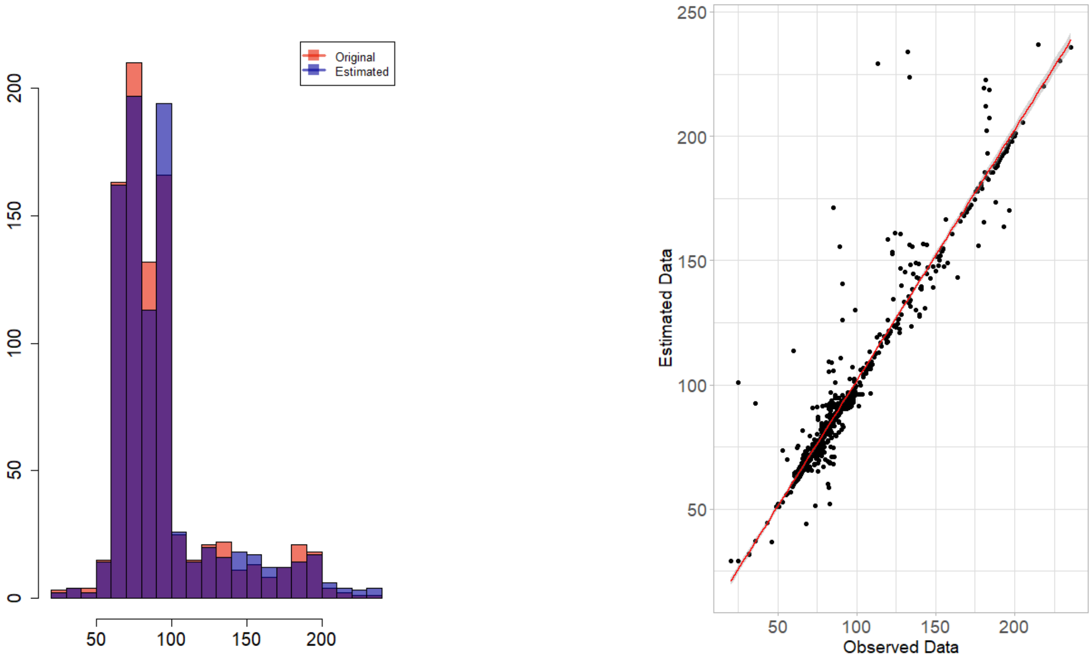
FIGURE 8.20: Histogram and Scatter plot of the Array Induction Two Resistivity 10*in* original and estimated values when the missing rate of the data is 0.25. The missing values are imputed by Kalman Arima.



(A)

(B)

FIGURE 8.21: Histogram and Scatter plot of the Array Induction Two Resistivity 10*in* original and estimated values when the missing rate of the data is 0.5. The missing values are imputed by Kalman Arima.



(A)

(B)

FIGURE 8.22: Histogram and Scatter plot of the Array Induction Two Resistivity 10*in* original and estimated values when the missing rate of the data is 0.8. The missing values are imputed by Kalman Arima.

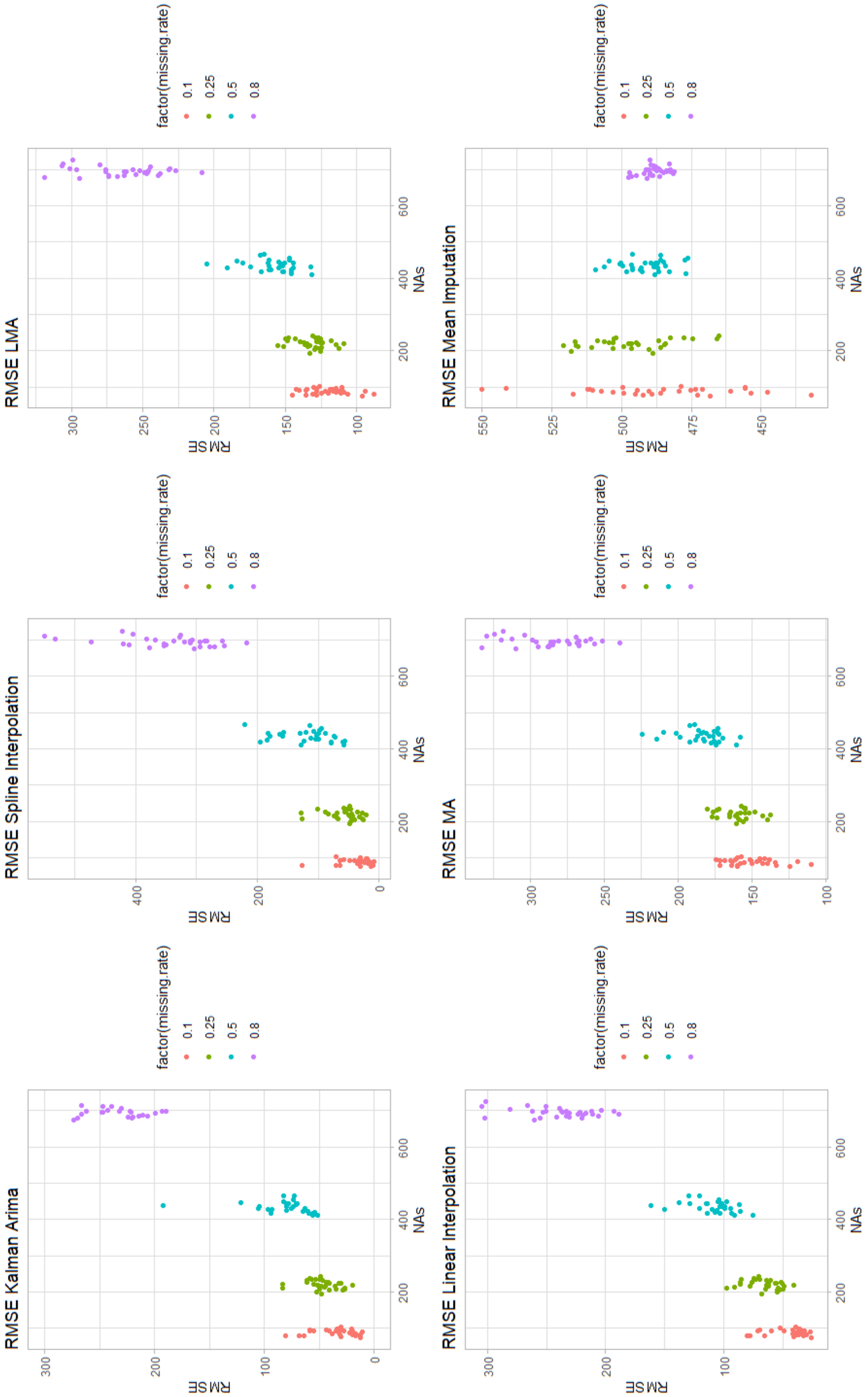


FIGURE 8.23: RMSE of Array Induction Two Resistivity 20*in* Data

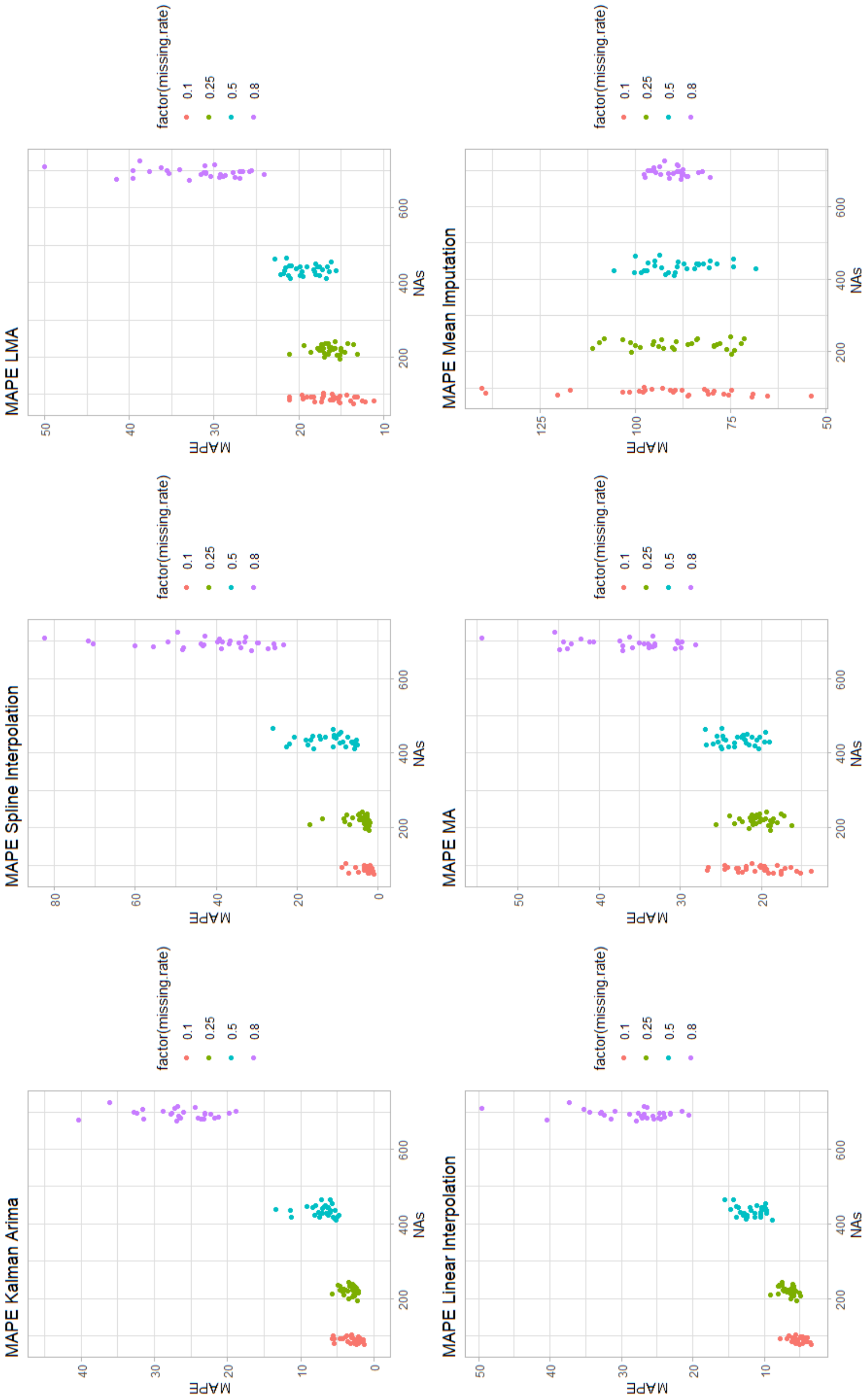
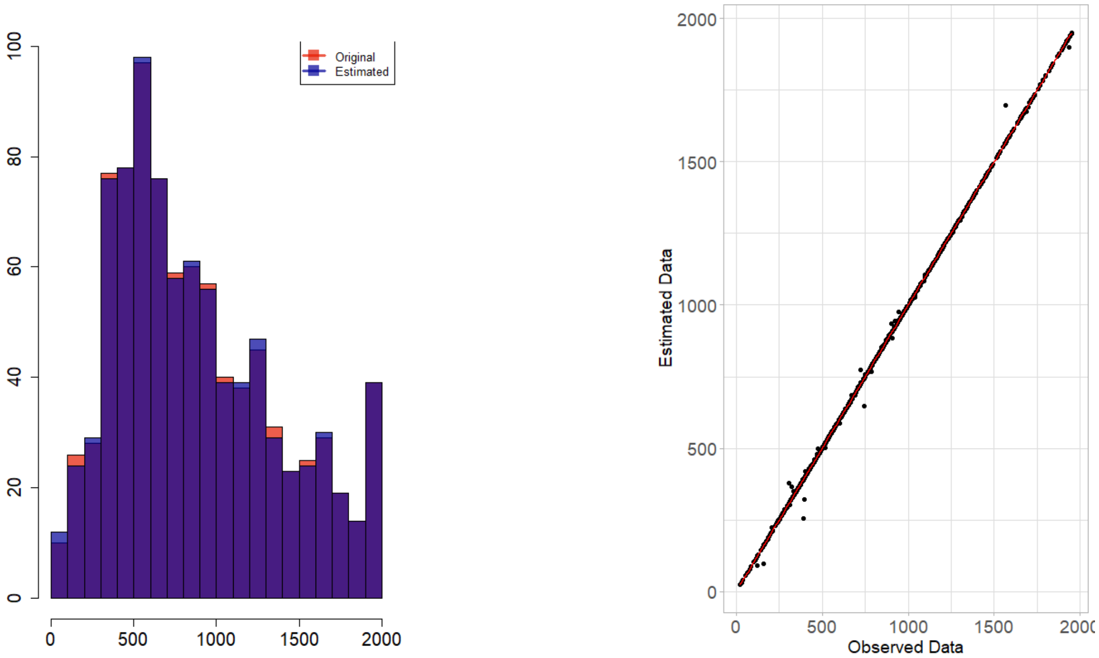


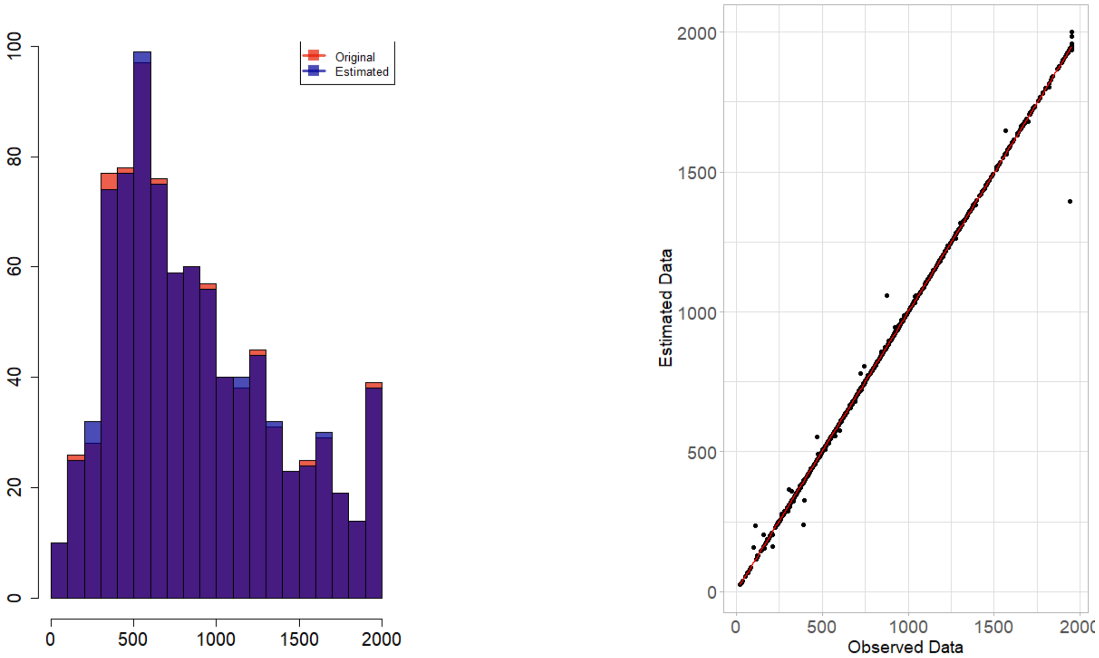
Figure 8.24: MAPE of Array Induction Two Resistivity 20*in* Data



(A)

(B)

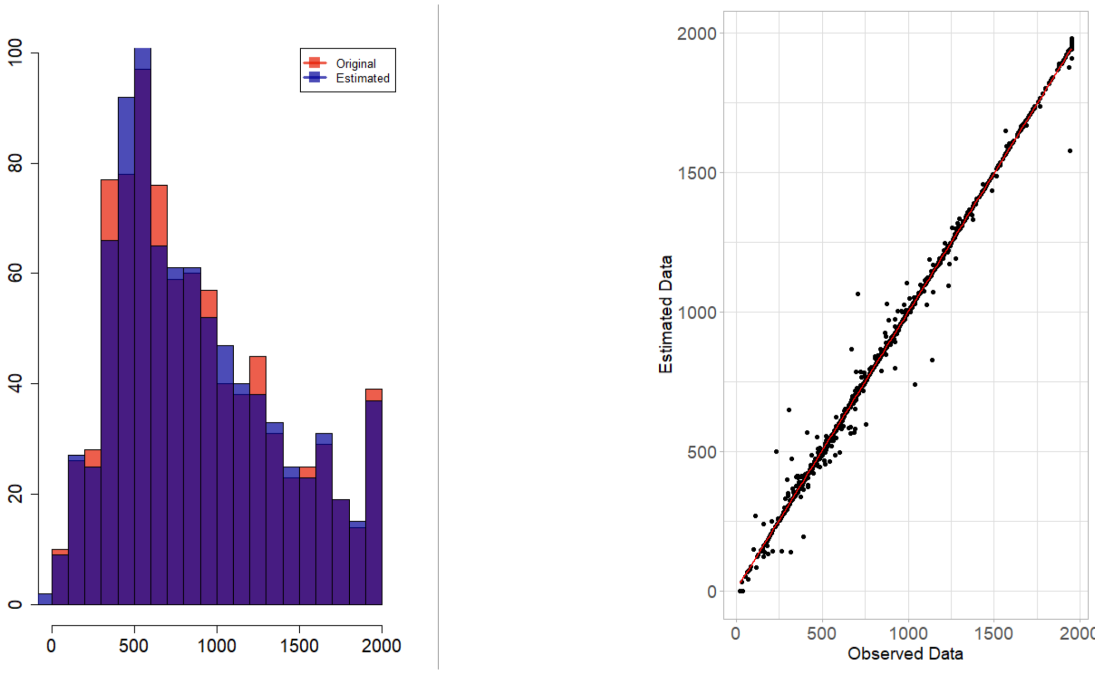
Figure 8.25: Histogram and Scatter plot of the Array Induction Two Resistivity 20*in* original and estimated values when the missing rate of the data is 0.1. The missing values are imputed by Kalman Arima.



(A)

(B)

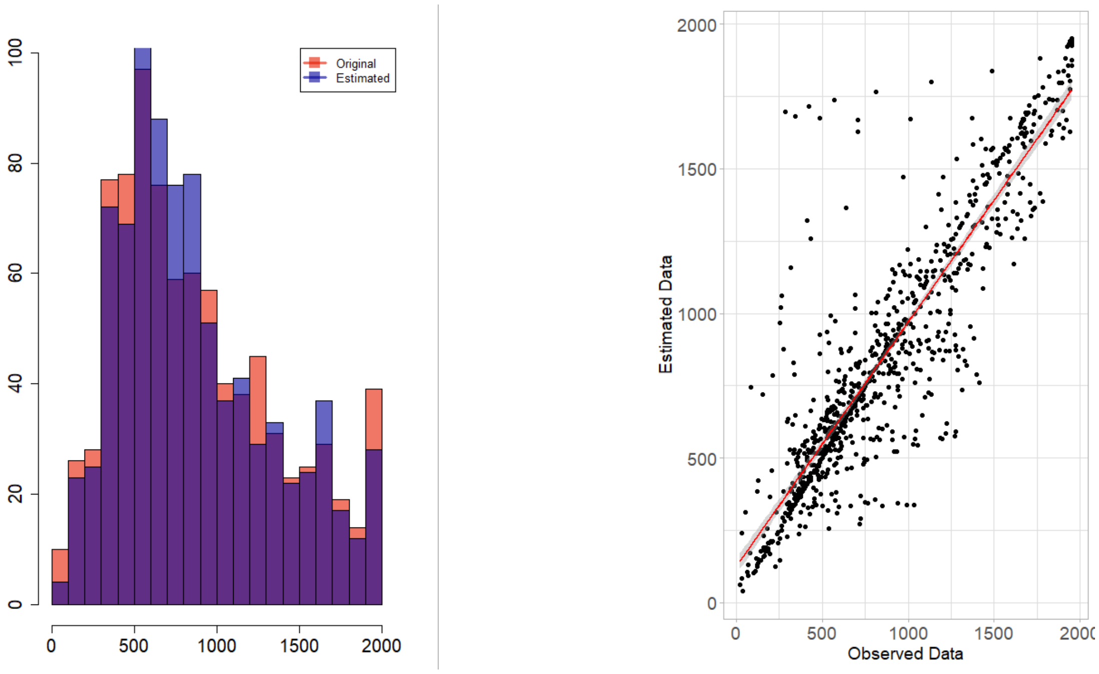
Figure 8.26: Histogram and Scatter plot of the Array Induction Two Resistivity 20*in* original and estimated values when the missing rate of the data is 0.25. The missing values are imputed by Kalman Arima.



(A)

(B)

FIGURE 8.27: Histogram and Scatter plot of the Array Induction Two Resistivity 20in original and estimated values when the missing rate of the data is 0.5. The missing values are imputed by Kalman Arima.



(A)

(B)

FIGURE 8.28: Histogram and Scatter plot of the Array Induction Two Resistivity 20in original and estimated values when the missing rate of the data is 0.8. The missing values are imputed by Kalman Arima.

8.3 Synopsis

One of the goals of this study was to investigate how the analysis of well log data and the resulting models are affected by various amounts of missing data and missing data patterns. Imputation, interpolation and time series algorithms for gap filling in univariate time series (well log data) are compared by means of cross validation. These methods comprise: Kalman ARIMA, mean imputation, linear and spline interpolation, as well as linear weighted and simple moving average method.

The results show that Linear interpolation, Linear weighted Moving Average and in certain cases Kalman Arima, exhibit similar performance, which is superior to the other methods. Histograms and Scatter plots used for the analyses confirm the good performance of the Kalman Arima algorithm. For high rates of missing data, the cross-validation measures tend to deteriorate for all the methods considered. Finally, the Mean-based imputation algorithm produced the largest bias and seems to be most severely affected by the presence of the trend component.

Chapter 9

Conclusions

This thesis seeks to address questions related to the statistical analysis of well log data. For this purpose, we obtained datasets from two hydrocarbon reservoirs that are located in Labrador Island, Western Newfoundland (Canada). The data, which are obtained from two wells (Finnegan and Seamus) that located onshore, contain a significant amount of geophysical information. To simplify the analysis we focused on four logs (corresponding to spontaneous potential, Gamma radiation and two induction logs). The data from these logs span six different formations. Thus, data analysis must face the challenge of handling transitions between different formations.

The thesis has three distinct objectives. The first objective is the estimation of the probability distributions and spatial correlations in data pertaining to the same well log. The second objective is to evaluate potential cross-correlations between logs which are obtained from different wells. The motivation for this task is to investigate if information from one well can be used to fill gaps in the data logs from a neighboring well. Finally, the third objective is to explore methods for the reconstruction of missing well log data using univariate methods (which do not account for cross-correlations between properties in the same well or across different wells).

We report on the conclusions regarding the three main objectives which have been reached by means of the well log data analysis in the two sections below. The first two section comprises conclusions related to the first two objectives, since they both refer to spatial correlations. The second section addresses the goal of missing data reconstruction.

9.1 Spatial Correlations

One of the objectives of this study was to investigate whether geostatistical tools can be used to provide useful information concerning spatial correlations in recorded well logs. Exploratory data analysis was used to summarize the statistical properties of the large data sets from the Seamus and Finnegan wells using graphical tools. We also investigated the fit of several *probability distribution* models to the data. The distribution fitting procedure suggests that regardless of the specific formation that is being considered, Spontaneous potential and Gamma radiation indicators can be best described by means of *Cauchy and Gumbel* distributions. In contrast, the Induction indicators are best described means of the *Gamma and Weibull* probability distributions. In some cases of well logs with skewed histograms, we also investigated the probability distributions of the data logarithms. It was realized that the respective histogram plots of the data logarithms seem to follow more closely the Gaussian distribution than the original values. These observations are useful, since a number of geostatistical methods work best for Gaussian and near-Gaussian data. However, their direct application to data that follow highly skewed distributions and/or fat-tailed (e.g., Gamma, Weibull, Gumbel, Cauchy) is not recommended.

The issue of spatial auto-correlations in logs from a single well was investigated by means of variogram analysis. A thorough analysis of the variogram functions for different logs and within different formations was carried out. This involved the estimation of the empirical (data-based) variogram estimates and their fits with theoretical variogram models using the method of weighted least squares. An overview of the results showed that a single optimal theoretical model for all the properties cannot be established. Furthermore, the results of the variogram analysis indicate that Spontaneous potential and Gamma Radiation indicators are mostly fitted to the same type of theoretical variogram model, with similar sill and range values. The most commonly obtained theoretical model is the Spherical, followed by the Pentaspherical and Gaussian models. The typical values for the range and the sill depend on the formations.

The results of the geostatistical analysis suggest that geostatistical tools can supplement available geophysical methods by providing useful information about regional stratigraphy and the spatial correlation patterns of a given exploration area.

The geostatistical study also involved the calculation of well log cross-correlations between Seamus and Finnegan wells. The respective gamma ray logs for the two wells are displayed in figures 6.12, A.21, A.22, A.23, A.24, A.25. The log data from the two wells were processed by means of interpolation methods to establish a common sampling step in order to calculate cross correlations. Different interpolation models were tested but it was found that they all lead to similar cross-correlation values. The Gamma radiation logs exhibit both positive and negative correlation values which are overall higher (in magnitude) than those of the other three logs that are studied. The values of the positive correlation coefficients range from 0.001 to 0.483, while the values of the negative correlation coefficients range from -0.142 to -0.001. These findings support the notion that the Gamma ray log is influenced by lithological changes according to the explanation provided in Section 6.4.

The analysis of the well log data shows clear signs of non-stationarity. The analysis of data with non-stationary statistics is challenging and remains an open research field. The broad implication of the present study is that methods can only be good as the context within which they are applied. The human factor cannot be eliminated from the process: Experts still need to choose which well log can give meaningful information and which method or set of methods should be applied to extract the information. Alternative and additional suggestions include the calculation of cross-correlation between Spontaneous potential logs and Gamma radiation logs, as well as the calculation of the uncertainty propagation through exemplary algorithms or the estimation of the effect of manually imputed parameters, defined by the user, in the calculation of experimental variograms.

9.2 Missing data reconstruction

The third objective of this study was to explore the performance of different methods that can be used for the reconstruction of missing data in well logs. The results of our analysis confirm that different gap-filling methods may be most suitable for different patterns of missing data.

The algorithms that are used herein relied on the assumption that the handling missing values come from a univariate time series. Importantly, our analysis concluded that identifying the patterns and reasons for the missing data can help to provide reliable and accurate reconstructions.

For the reconstruction of missing values, we used a number of interpolation, imputation and time series methods which included Kalman Smoothing (KS) with an ARIMA model, Spline Interpolation, Linear Interpolation, Simple Moving Average, Linear Weighted Moving Average and Mean Imputation. Based on statistical validation measures and comparison maps, we conclude that Linear Interpolation, Linear Weighted Moving Average and in some cases Kalman Arima, are the methods that exhibit superior and quite similar performance. Histograms and scatter plots confirm the good performance of the Kalman Arima algorithm. Moreover, we can conclude that significant biases occur in the reconstructions if the data sets involve non-modeled spatial trends and when the missing data rate is high (i.e. $> 50\%$)

Imputation and interpolation methods can be easily applied to univariate time series. Future research could investigate the effects of sampling size and number of random-effects (i.e. when performing Multiple Imputation algorithms) and algorithmic improvements. Further studies should focus on exploring different imputation techniques under more comprehensive missing data scenarios (i.e. Complete Case Analysis (CCA), Last Observation Carried Forward (LOCF), Complete Case Missing Value (CCMVPM) restriction, Available Case Missing Value (ACMVPM) restriction, Neighboring Case Missing Value (NCMVPM) restriction, and the selection model (SMPM).

Overall, our results show that well log data analysis can benefit from the application of geostatistical and time series methods. The latter can be effectively applied to one-dimensional spatial data, such as those obtained by well logs. While the current study focused on the modeling of spatial correlations in each well independently of other wells in the area, multivariate time series models could be used to provide jointly analyze data across different wells.

Bibliography

[1] Final well reports. https://www.gov.nl.ca/nr/publications/energy/final-well-report/. Nalcor Enery-Oil and Gas INC.

[2] S. Aldor-Noiman, L. D. Brown, A. Buja, W. Rolke, and R. A. Stine. The power to see: A new graphical test of normality. *The American Statistician*, 67(4):249–260, 2013.

[3] M. Aumüller, E. Bernhardsson, and A. Faithfull. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. In C. Beecks, F. Borutta, P. Kröger, and T. Seidl, editors, *Similarity Search and Applications*, pages 34–49, Cham, 2017. Springer International Publishing.

[4] R. O. Baker, H. W. Yarranton, and J. Jensen. *Practical Reservoir Engineering and Characterization.* Gulf Professional Publishing, 2015.

[5] A. Banerjee, J. J. Dolado, J. W. Galbraith, D. Hendry, et al. Co-integration, error correction, and the econometric analysis of non-stationary data. *OUP Catalogue*, 1993.

[6] C. Barceló. The impact of alternative imputation methods on the measurement of income and wealth: Evidence from the Spanish survey of household finances. *Available at SSRN: https://ssrn.com/abstract=1321827*, 2008.

[7] Z. Bassiouni et al. *Theory, Measurement, and Interpretation of Well Logs*, volume 4. Henry L. Doherty Memorial Fund of AIME, Society of Petroleum Engineers, 1994.

[8] R. M. Bateman. *Openhole Log Analysis and Formation Analysis.* IHRDC Press, Boston, MA, 1985.

[9] G. E. Batista and M. C. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6):519–533, 2003.

[10] A. Bhatt. *Reservoir Properties from Well Logs using neural Networks.* PhD thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2002.

[11] R. S. Bivand, E. J. Pebesma, V. Gomez-Rubio, and E. J. Pebesma. *Applied Spatial Data Analysis with R*, volume 747248717. Springer, New York,NY, 2008.

[12] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time Series Analysis: Forecasting and Control.* John Wiley & Sons, 2015.

[13] W. Boyce, I. Knight, D. Rohr, S. Williams, and E. Measures. The upper St. George Group, western Port au Port Peninsula: lithostratigraphy, biostratigraphy, depositional environments and regional implications. In *Current Research (2000), Newfoundland Department of Mines and Energy, Report 2000-1*, pages 101–125. 2000.

[14] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[15] L. Brillouin. The negentropy principle of information. *Journal of Applied Physics*, 24(9):1152–1163, 1953.

[16] P. J. Brockwell, R. A. Davis, and S. E. Fienberg. *Time Series: Theory and Methods: Theory and Methods.* Springer Science & Business Media, 1991.

[17] M. L. Brown and J. F. Kros. Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103(8):611–621, 2003.

[18] A. Cartellieri, J. Pragt, M. Meister, et al. Advantages and limitations of taking samples while drilling. In *SPWLA 53rd Annual Logging Symposium.* Society of Petrophysicists and Well-Log Analysts, 2012.

[19] R. P. Chapuis and L. Sabourin. Effects of Installation of Piezometers and Wells on Groundwater Characteristics and Measurements. *Canadian Geotechnical Journal*, 26(4):604–613, 1989.

[20] C. Chatfield. *The Analysis of Time Series: an Introduction.* Chapman and Hall/CRC, 2003.

[21] T. C. Coburn, J. M. Yarus, R. L. Chambers, et al. *Stochastic Modeling and Geostatistics: Principles, Methods, and Sase studies, vol. II, AAPG computer Applications in Geology 5*, volume 5. AAPG, 2005.

[22] N. Cressie. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5):563–586, 1985.

[23] N. Cressie and D. M. Hawkins. Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, 12(2):115–125, Apr 1980.

[24] T. Darling. *Well Logging and Formation Evaluation.* Elsevier, 2005.

[25] M. T. Dastorani, A. Moghadamnia, J. Piri, and M. Rico-Ramirez. Application of ann and anfis models for reconstructing missing flow data. *Environmental Monitoring and Assessment*, 166(1-4):421–434, 2010.

[26] M. Deffenbaugh. Geophysical parameter estimation. In D. Havelock, S. Kuwano, and M. Vorländer, editors, *Handbook of Signal Processing in Acoustics*, pages 1593–1626. Springer New York, New York, NY, 2008.

[27] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[28] Y. Ding et al. A Generalized 3D Well Model for Reservoir Simulation. *SPE Journal*, 1(04):437–450, 1996.

[29] J. K. Dixon. Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(10):617–621, 1979.

[30] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods.* Oxford university press, 2012.

[31] J. Edwards, F. Lallier, G. Caumon, and C. Carpentier. Uncertainty management in stratigraphic well correlation and stratigraphic architectures: A training-based method. *Computers & Geosciences*, 111:1–17, 2018.

[32] D. V. Ellis and J. M. Singer. *Well Logging for Earth Scientists*, volume 692. Springer, 2007.

[33] W. Enders. *Applied econometric time series.* John Wiley & Sons, 2008.

[34] J. Fanchi. *Integrated Reservoir Asset Management: Principles and Best Practices.* Gulf Professional Publishing, Elsevier, Amsterdam, Netherlands, 2010.

[35] Z. Fang. *Energy Science and Applied Technology ESAT 2016: Proceedings of the International Conference on Energy Science and Applied Technology (ESAT 2016), Wuhan, China, June 25-26, 2016.* CRC Press, Boca Raton, FL, 2016.

[36] P. Glover. Petrophysics msc course notes. *University of Leeds, UK*, 2015.

[37] J. W. Grzymala-Busse and M. Hu. A comparison of several approaches to missing attribute values in data mining. In *International Conference on Rough Sets and Current Trends in Computing*, pages 378–385. Springer, 2000.

[38] J. Hammond and P. White. The analysis of non-stationary signals using time-frequency methods. *Journal of Sound and Vibration*, 190(3):419–447, 1996.

[39] M. Hazewinkel. *Encyclopaedia of Mathematics: Volume 3 Heaps and Semi-Heaps—Moments, Method of (in Probability Theory)*, volume 3. Springer, 2013.

[40] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice.* OTexts, 2018.

[41] E. Isaaks and R. Srivastava. *An Introduction to Applied Geostatistics.* Oxford University Press, New York, NY, 1989.

[42] J. W. Jennings Jr, F. J. Lucia, et al. Predicting Permeability from Well Logs in Carbonates with a Link to Geology for Interwell Permeability Mapping. In *SPE Annual Technical Conference and Exhibition.* Society of Petroleum Engineers, 2001.

[43] H. K. Kim, Sujung and K. Kurihara. Geostatistical data analysis with outlier detection. *Journal of the Korean Data Analysis Society*, 16(5):2285–2297, 2015.

[44] G. E. King, D. E. King, et al. Environmental risk arising from well-construction failure–differences between barrier and well failure, and estimates of failure frequency across common well types, locations, and well age. *SPE Production & Operations*, 28(04):323–344, 2013.

[45] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145, Montreal, Canada, 1995.

[46] D. Kondrashov and M. Ghil. Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics*, 13(2):151–159, 2006.

[47] D. Koutsoyiannis. *Probability and Statistics for Geophysical Processes.* 01 2008.

[48] D. W. Lee, L. D. Den Boer, C. M. Sayers, and P. J. Hooyman. Methods and Apparatus for Predicting the Hydrocarbon Production of a Well Location, Feb. 3 2009. US Patent 7,486,589.

[49] A. Lichtenstern. *Kriging methods in spatial statistics.* PhD thesis, Technische Universität München, Germany, August 2013. Bachelor thesis.

[50] D. Lineman, J. Mendelson, M. N. Toksoz, et al. Well to well log correlation using knowledge-based systems and dynamic depth warping. In *SPWLA 28th Annual Logging Symposium.* Society of Petrophysicists and Well-Log Analysts, 1987.

[51] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons, New York, NY, 2019.

[52] R. L. Lopes and A. Jorge. Mind the gap: a well log data analysis. *arXiv preprint arXiv:1705.03669*, 2017.

[53] H. Lütkepohl, M. Krätzig, and P. C. Phillips. *Applied Time Series Econometrics.* Cambridge university press, 2004.

[54] W. C. Lyons and G. J. Plisga. *Standard Handbook of Petroleum and Natural Gas Engineering.* Gul Publishing Company, Elsevier, Houston, TX, USA, 2011.

[55] M. Magnani. Techniques for dealing with missing data in knowledge discovery tasks. *Obtido http://magnanim. web. cs. unibo. it/index. html*, 15(01):2007, 2004.

[56] B. Marlin. *Missing data problems in machine learning.* PhD thesis, Department of Computer Science, University of Toronto, 2008.

[57] S. Maus. Variogram analysis of magnetic and gravity data. *Geophysics*, 64(3):776–784, 1999.

[58] M. P. McLaughlin. *A Compendium of Common Probability Distributions.* Michael P. McLaughlin, 2001.

[59] A. Mendoza, C. Torres-Verdin, W. Preeg, et al. Environmental and petrophysical effects on density and neutron porosity logs acquired in highly deviated well. In *SPWLA 47th Annual Logging Symposium*. Society of Petrophysicists and Well-Log Analysts, 2006.

[60] T. C. Mills. *The Foundations of Modern Time Series Analysis*. Palgrave, Hampshire, UK, 2011.

[61] D. C. Montgomery, C. L. Jennings, and M. Kulahci. *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons, Hoboken, NJ, 2015.

[62] S. Moore. Mwd tools improve drilling performance. *Petroleum Engineer International (United States)*, 58(2), 1986.

[63] V. Nikolaevskiy, G. Lopukhov, L. Yizhu, M. Economides, et al. Residual Oil Reservoir Recovery with Seismic Vibrations. *SPE Production & Facilities*, 11(02):89–94, 1996.

[64] M. A. Oliver and R. Webster. *Basic Steps in Geostatistics: the Variogram and Kriging*, volume 106. Springer, 2015.

[65] C. Palagi, K. Aziz, et al. Use of Voronoi Grid in Reservoir Simulation. *SPE Advanced Technology Series*, 2(02):69–77, 1994.

[66] S. J. Pirson. *Geologic Well Log Analysis*. Gulf Publishing Company Houston, Tex., 1970.

[67] B. Podobnik and H. E. Stanley. Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. *Physical review letters*, 100(8):084102, 2008.

[68] Y. Polsky, L. Capuano, J. Finger, M. Huh, S. Knudsen, A. Mansure, D. Raymond, and R. Swanson. Enhanced Geothermal Systems (EGS) Well Construction Technology Evaluation Report. *Sandia National Laboratories, Sandia Report, SAND2008-7866*, pages 1–108, 2008.

[69] D. Price, A. Curtis, and R. Wood. Statistical correlation between geophysical logs and extracted core. *Geophysics*, 73(3):E97–E106, 2008.

[70] M. Priestley and T. S. Rao. A test for non-stationarity of time-series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(1):140–149, 1969.

[71] M. J. Pyrcz and C. V. Deutsch. *Geostatistical reservoir modeling.* Oxford university press, 2014.

[72] L. A. Quinn. *Foreland and trench slope basin sandstones of the Goose Tickle Group and Lower Head Formation, western Newfoundland.* PhD thesis, Memorial University of Newfoundland, 1992.

[73] K. Rantou. Missing data in time series and imputation methods. Master's thesis, Department of Mathematics University of the Aegean, February 2017.

[74] R. M. Rifkin and R. A. Lippert. Notes on regularized least squares. 2007.

[75] J. D. Rogers and J.-W. Chung. Uncertainties Monitoring Groundwater Levels in Exploratory Wells. *Groundwater*, 51(1):2–4, 2013.

[76] S. J. Rogers, J. Fang, C. Karr, and D. Stanley. Determination of Lithology from Well Logs using a Neural Network (1). *AAPG bulletin*, 76(5):731–739, 1992.

[77] M. E. Rossi and C. V. Deutsch. *Mineral Resource Estimation.* Springer Science & Business Media, New York, NY, 2013.

[78] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[79] R. Schoeppel, S. Gilarranz, et al. Use of Well Log Temperatures to Evaluate Regional Geothermal Gradients. *Journal of Petroleum Technology*, 18(06):667–673, 1966.

[80] O. Serra. *Fundamentals of Well-Log Interpretation.* Elsevier Science Pub., New York, NY, 1983.

[81] S. L. Serra O. *Well Logging: Aata Acquisition and Applications.* Serralog Méry Corbon, France, 2004.

[82] A. Settari, Y. Ito, N. Fukushima, and H. Vaziri. Geotechnical Aspects of Recovery Processes in Oil Sands. *Canadian Geotechnical Journal*, 30(1):22–33, 1993.

[83] R. H. Shumway and D. S. Stoffer. Time series analysis and its applications. *Studies In Informatics And Control*, 9(4):13–16, 2000.

[84] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and its Applications: with R Examples*. Springer, 2017.

[85] J. W. Smits, P. T. Wu, Q. Li, and C. B. Liu. Intelligent Diagnosis of Environmental Influence on Well Logs with Model-Based Inversion, Dec. 14 2004. US Patent 6,832,159.

[86] R. Srivastava et al. Reservoir characterization with probability field simulation. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, 1992.

[87] B. Steingrimsson. Geothermal Well Logging: Temperature and Pressure Logs. *Short Course V on Conceptual Modelling of Geothermal Systems*, 2013.

[88] J. B. Surjaatmadja, A. Cheng, and K. A. Rispler. System and Method for FRacturing a Subterranean Well Formation for Improving Hydrocarbon Production, Dec. 16 2003. US Patent 6,662,874.

[89] E. M. Syczewska. Empirical Power of the Kwiatkowski-Phillips-Schmidt-Shin Test. Technical report, Warsaw School of Economics, 2010.

[90] A. E. Taylor and A. S. Judge. *Canadian Geothermal Data Collection-Northern Wells 1975*. Energy, Mines and Resources Canada, Earth Physics Branch, 1976.

[91] R. Tonn. Neural Network Seismic Reservoir Characterization in a Heavy Oil Reservoir. *The Leading Edge*, 21(3):309–312, 2002.

[92] J. A. Torres Munguía. Comparison of imputation methods for handling missing categorical data with univariate pattern. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, 17:101–120, 2014.

[93] S. Van Buuren. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, 2018.

[94] C. B. Vogel. A Seismic Velocity Logging Method. *Geophysics*, 17(3):586–597, 1952.

[95] E. W. Weisstein. Student's t-distribution. *Sigma*, 13:14, 2001.

[96] D. J. Wheeler. *Advanced Topics in Statistical Process Control*, volume 470. SPC Press, Knoxville, TN, 1995.

[97] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis for the analysis of data. *Biometrika*, 55(1):1–17, 1968.

[98] C. J. Willmott, S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'donnell, and C. M. Rowe. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research: Oceans*, 90(C5):8995–9005, 1985.

[99] C. Wittrisch and J.-B. Fay. Measuring Device and Method in a Hydrocarbon Production Well, Apr. 9 1996. US Patent 5,505,259.

[100] W. A. Woodward, H. L. Gray, and A. C. Elliott. *Applied Time Series Analysis with R*. CRC press, 2017.

[101] M. Zukovic and D. T. Hristopulos. An algorithm for spatial data classification and automatic mapping based on "spin" correlations. In *Proceedings of the 22nd European Conference on Modeling and Simlulation*, 2008.

Appendix A

Geostatistical analysis of the selected formations

The Table point formation is a geological unit dominated by dolomitized carbonate conglomerates and calcarenites, while fossils dating back to the Ordovian period are also present. The Aguathuna formation is a geological unit dominated mainly by limestone, dolostone, and shale ([13]).

A.1 Seamus 216mm hole section

The statistical parameters of the the Table Point and Aguathuna formation of the Seamus 216 hole section are presented.

SP (.mV)	Min	Max	Mean	Median	Mode
Table Point	-102.81	-13.37	-67.97	-75.25	-93.06
Aguatha	-75.25	18.68	-30.77	-34.5	14.37

SP (.mV)	Variance	SD	Skewness	Kurtosis
Table Point	618.08	24.86	0.64	2.07
Aguatha	753.19	27.44	0.27	1.84

Table A.1: Spontaneous potential statistical parameters of Seamus 216mm.

GR (GAPI)	Min	Max	Mean	Median	Mode
Table Point	6.43	46.95	15.72	14.52	13.34
Aguatha	8.92	97.44	27.07	22.14	14.18

GR (GAPI)	Variance	SD	Skewness	Kurtosis
Table Point	24.43	4.94	1.74	7.6
Aguatha	254.26	15.95	1.32	4.62

Table A.2: Gamma ray statistical parameters of Seamus 216mm.

A10	Min	Max	Mean	Median	Mode
Table Point	20.27	235.82	92.7	83.08	85.33
Aguatha	65.95	197.78	140.55	143.9	108.65

A10	Variance	SD	Skewness	Kurtosis
Table Point	1162.45	34.09	1.72	5.7
Aguatha	868	29.46	-0.45	2.6

Table A.3: Array Induction 10in statistical parameters of Seamus 216mm.

A20	Min	Max	Mean	Median	Mode
Table Point	24.73	1950	877.8	764.98	1950
Aguatha	83.46	1950	920.15	887.61	1950

A20	Variance	SD	Skewness	Kurtosis
Table Point	239297.4	489.18	0.6	2.43
Aguatha	258693.5	508.62	0.27	2.1

Table A.4: Array Induction 20in statistical parameters of Seamus 216mm.

A.1.1 Table Point Formation

Log	Min	Max	Mean	Median	Mode
SP(mV)	-30.077	22.496	7.219e-14	0.155	-21.828

Log	Variance	SD	Skewness	Kurtosis
SP(mV)	134.33	11.590	-0.111	2.106

TABLE A.5: Detrended data statistics of Table Point formation found in Seamus 216mm hole section.

Log	Model	Estimated Trend Function
SP	Qubic	$-33.9 + 2.61 \cdot 10^{-2}x + -6.34 \cdot 10^{-4}x^2 + 6.96 \cdot 10^{-7}x^3 + \epsilon_i, \epsilon \sim N(0, 11.6^2)$

TABLE A.6: Estimated trend models.

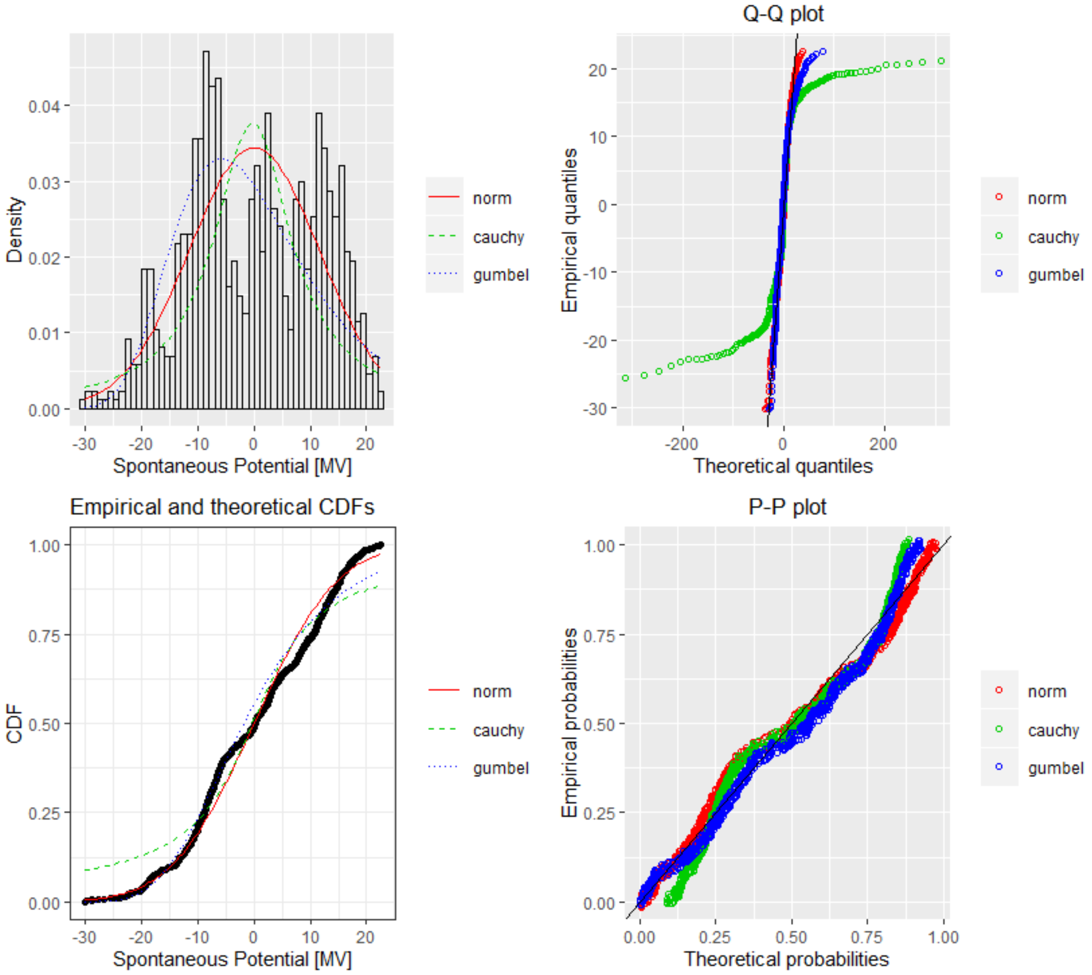


FIGURE A.1: Fitting of the distributions by maximum likelihood. Featured data-set; SP log of Seamus 216mm hole section.

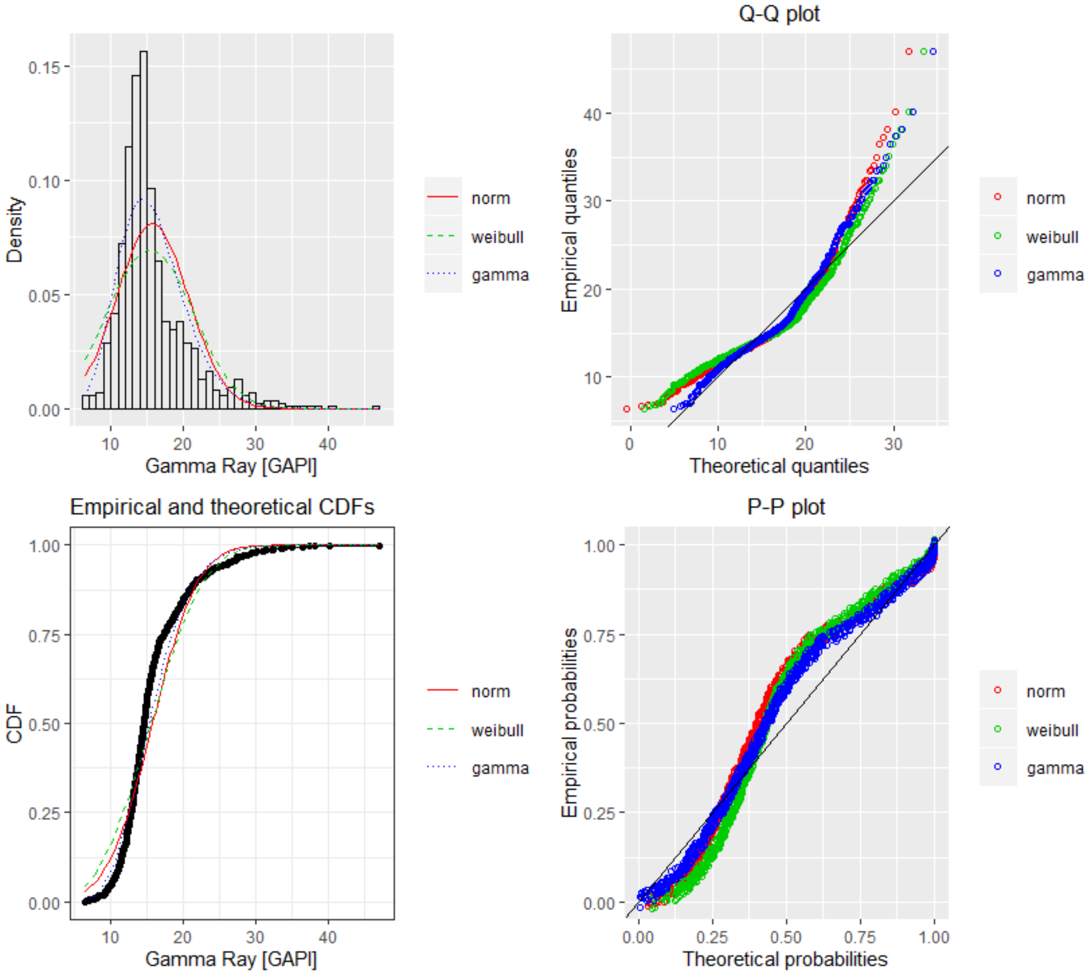


Figure A.2: Fitting of the distributions by maximum likelihood. Featured data-set; GR log of Seamus 216mm hole section.

	Histograms		
	Distribution	Parameters	Information Criteria
SP	norm	μ=7.219e-14,σ=11.58	AIC=6742.95, BIC=6752.49
	Cauchy	a=-0.309,γ=8.465	AIC= 7219.65, BIC=7229.19
	Gumbel	μ=-5.826,b=11.160	AIC=6857.59, BIC=6867.13
GR	norm	μ=15.718,σ=4.940	AIC=5258.35, BIC= 5267.89
	Weibull	a=3.102,λ=17.473	AIC=8622.55, BIC=8632.09
	gamma	a=12.167,λ= 0.774	AIC= 5049.42, BIC= 5058.95
A10	norm	μ=92.701,σ=34.075	AIC=5258.35, BIC= 5267.889
	Weibull	a=2.746,λ=103.987	AIC=8593.23, BIC=8602.76
	gamma	a=9.318,λ=0.101	AIC= 8358.49, BIC=8368.03
A20	norm	μ=877.803,σ=488.899	AIC=13262.5, BIC=13272.1
	Weibull	a= 1.878,λ=990.129	AIC=13132.3 , BIC= 13141.8
	gamma	a=2.856,λ=0.003	AIC=13146.4, BIC= 13156

Table A.7: Distributions' estimated parameters and information criteria of the Table Point formation found in Seamus 216mm hole section. The units of measurement are [mV], [GAPI], [Ωhmm] for SP, GR and A10, A20 respectively.

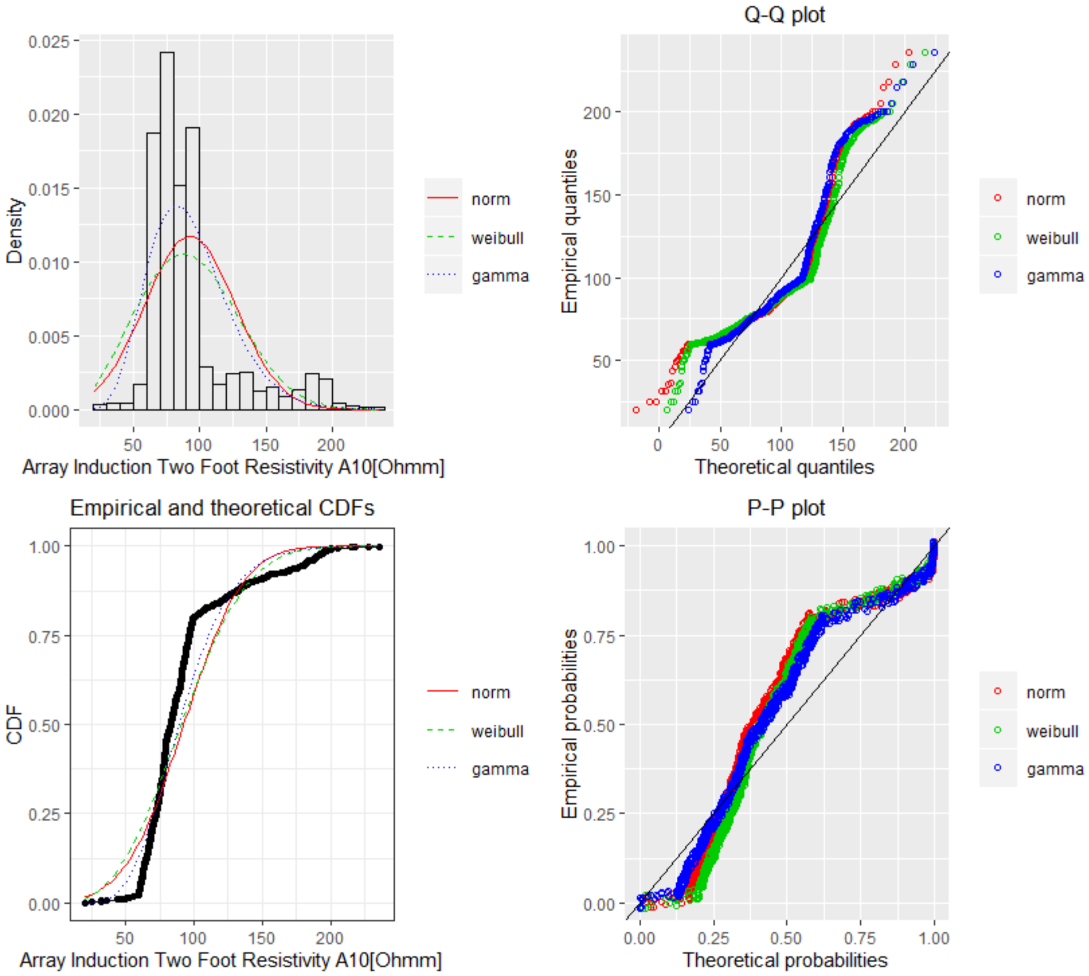


Figure A.3: Fitting of the distributions by maximum likelihood. Featured data-set; A10 log of Seamus 216mm hole section.

	Variograms			
	Model	Sill	Range	Nugget
SP	Gau	229.263	10.078	3.410
GR	Cir	13.893	3.095	1.303
	Gau	11.992	1.744	3.231
	Pen	10.738	8.043	4.933
	Sph	14.216	3.447	0.981
A10	Cir	713.728	16.272	0.000
	Pen	717.241	22.522	0.000
	Sph	714.986	18.505	0.000
A20	Exp	213056	2.171	0.000
	Gau	170931.4	2.143	32864.3
	Pen	181428.5	7.005	27292.6
	Sph	177039.4	5.915	31402.4

Table A.8: Fitting of the best theoretical model to the experimental variograms of the field.

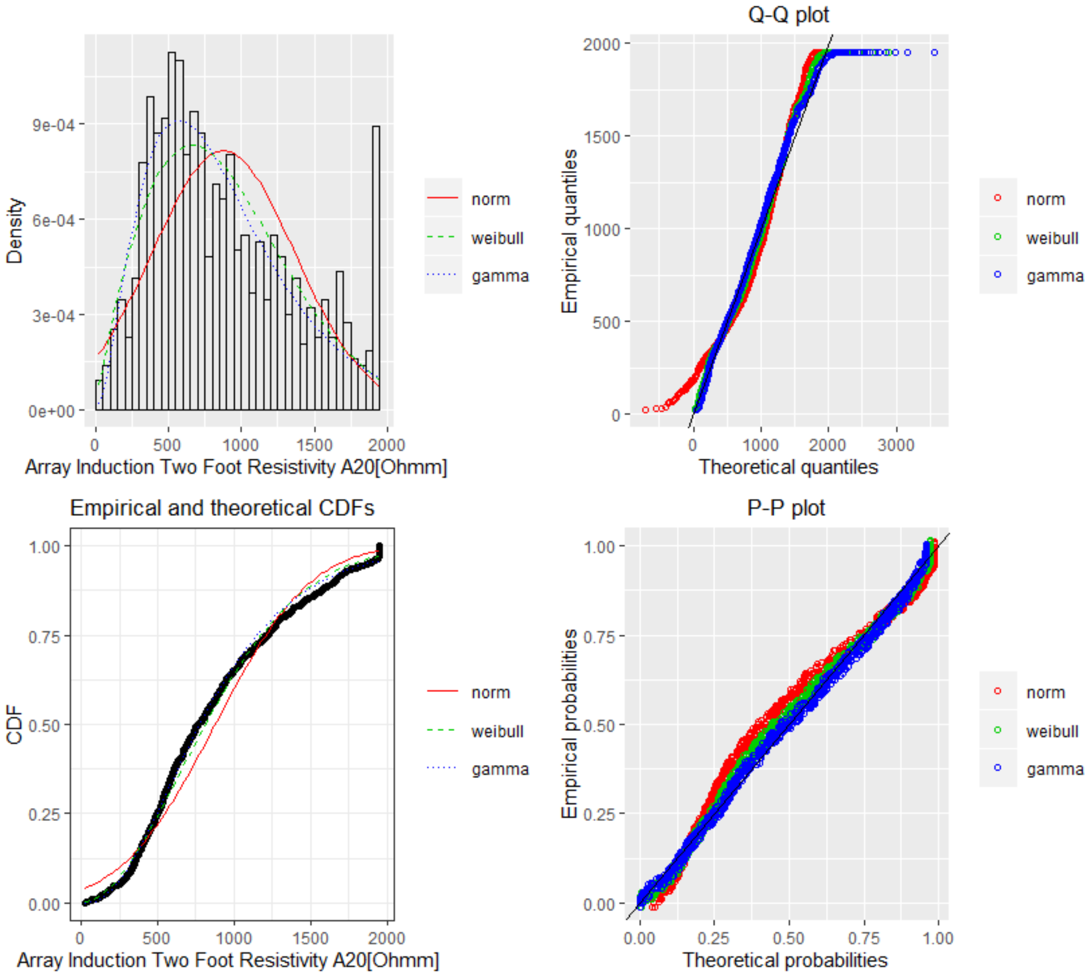
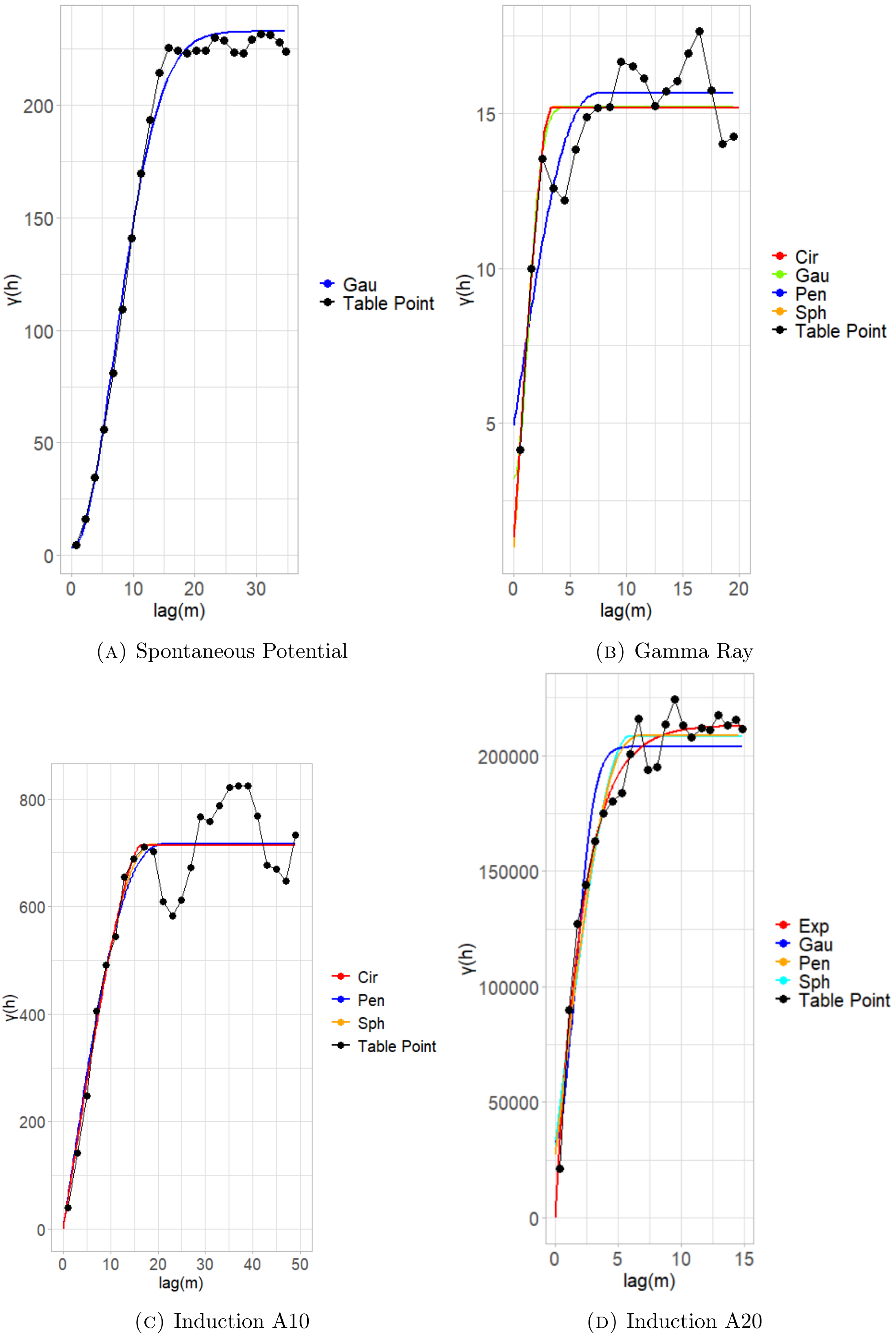


Figure A.4: Fitting of the distributions by maximum likelihood. Featured data-set; A20 log of Seamus 216mm hole section.

	Variograms			
	Model	MSE	MAE	RMSE
SP	Gau	82.054	7.493	9.058
GR	Cir	2.706	1.329	1.645
	Gau	2.285	1.212	1.511
	Pen	1.850	1.045	1.360
	Sph	2.712	1.326	1.647
A10	Cir	4440.83	56.005	66.639
	Pen	4394.45	56.105	66.291
	Sph	4395	56.042	66.295
A20	Exp	$1386\cdot10^5$	8455.11	11774.3
	Gau	$25524\cdot10^5$	13309.9	15976.3
	Pen	$17091\cdot10^5$	10084.6	13073.2
	Sph	$18224\cdot10^5$	10704.3	13500

Table A.9: Fitting of the best theoretical model to the experimental variograms of the field.



(A) Spontaneous Potential

(B) Gamma Ray

(C) Induction A10

(D) Induction A20

FIGURE A.5: Variogram plots. The weights are determined using N_j, where N_j is the number of pairs at certain lag. For the calculation of the Spontaneous potential the weights are determined using N_j/h_i^2.

A.1.2 Aguathuna Formation

Log	Min	Max	Mean	Median	Mode
A10(Ohmm)	-83.268	45.952	1.233e-13	4.303	1.847

Log	Variance	SD	Skewness	Kurtosis
A10(Ohnmm)	485.109	22.025	-0.781	3.715

Table A.10: Detrended data statistics of Aguathuna formation found in Seamus 216 hole section.

Log	Model	Estimated Trend Function
A10	Linear	$106.61 - 0.19x + \epsilon_i, \epsilon \sim N(0, 22.1^2)$

Table A.11: Estimated trend models.

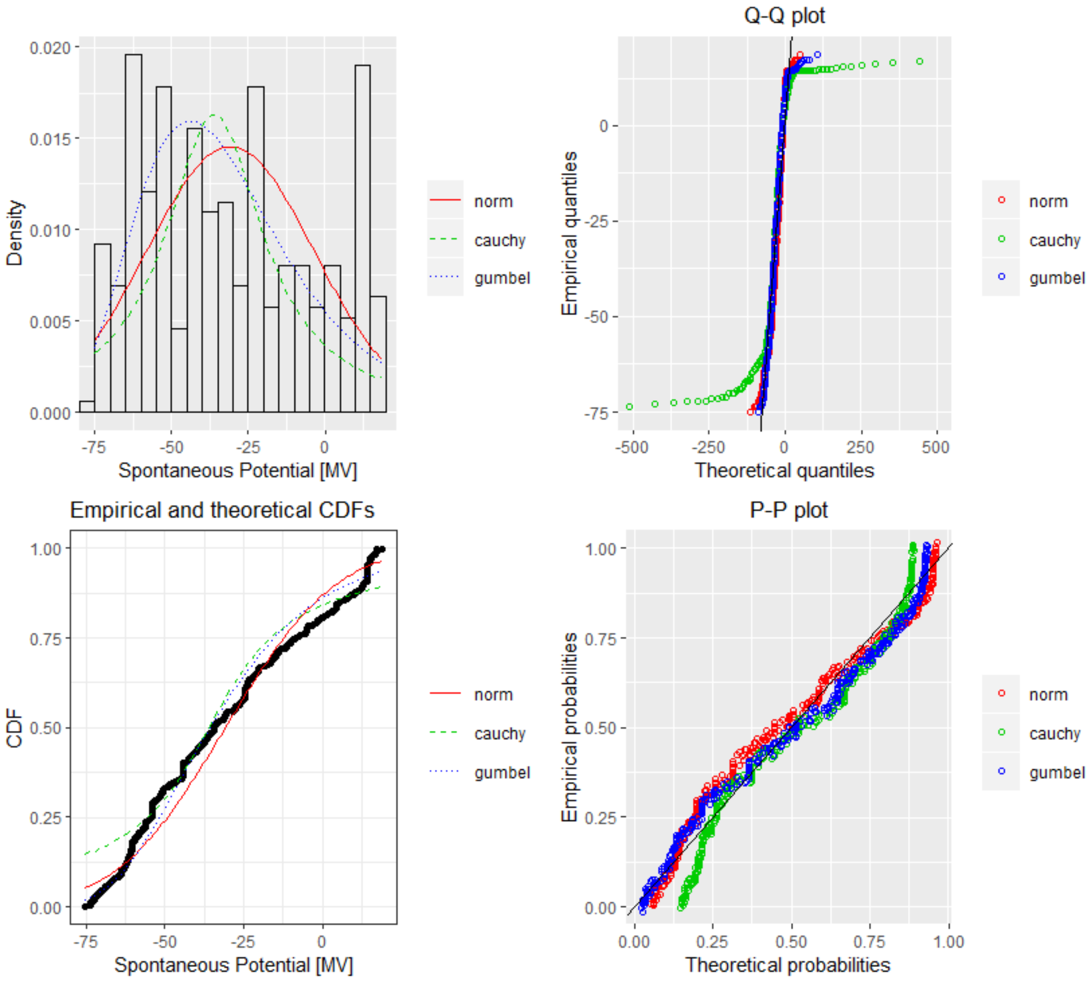


Figure A.6: Fitting of the distributions by maximum likelihood. Featured data-set; SP log of Seamus 216mm hole section.

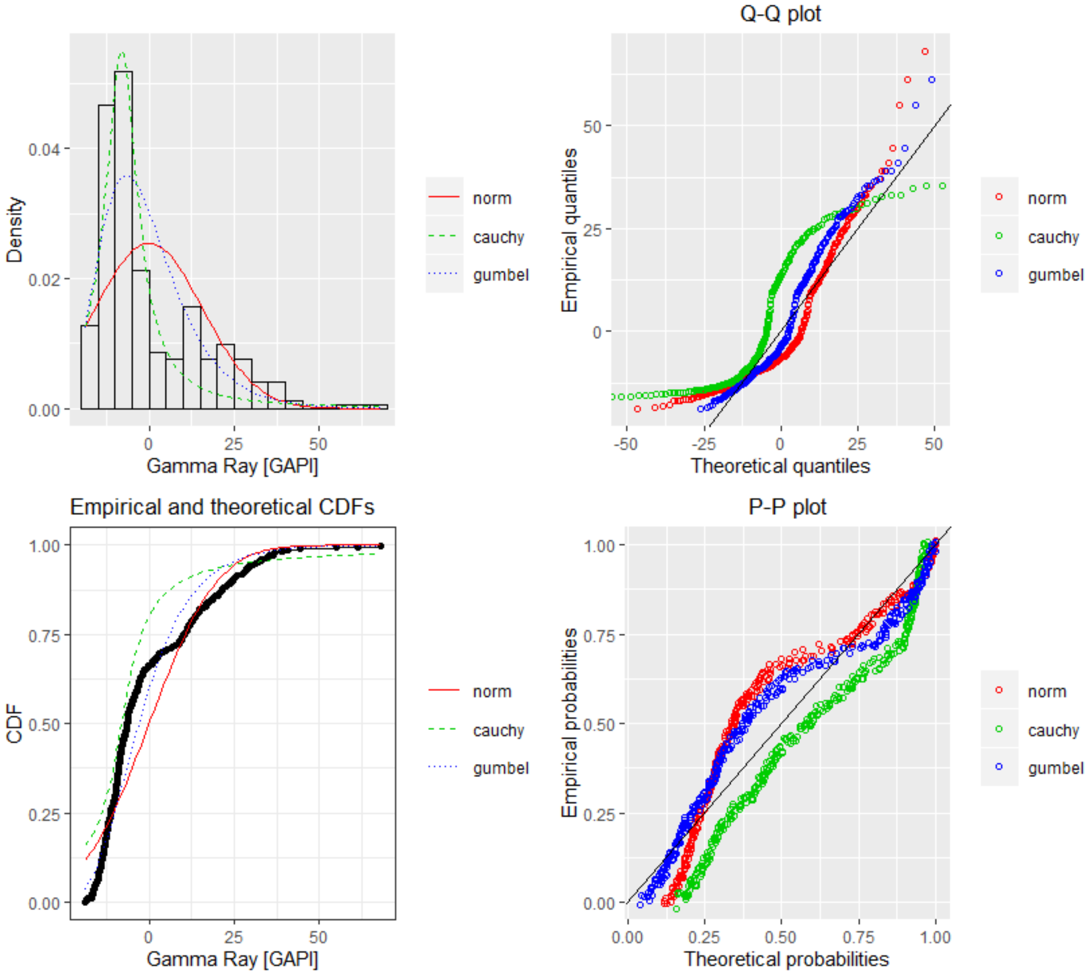


Figure A.7: Fitting of the distributions by maximum likelihood. Featured data-set; GR log of Seamus 216mm hole section.

	Histograms		
	Distribution	Parameters	Information Criteria
SP	norm	μ=-30.771,σ=27.405	AIC=3286.38, BIC=3294.08
	Cauchy	a=-36.001,γ=19.518	AIC=3476.17, BIC=3483.87
	Gumbel	μ=-44.006,b=23.065	AIC=3274.43, BIC=3282.13
GR	norm	$\mu = -4.895e{-}15$,σ=15.68	AIC=2898.93, BIC= 2906.63
	Cauchy	a=-8.045,λ=5.786	AIC=2886.79, BIC=2894.49
	Gumbel	a=-6.710,λ= 10.262	AIC= 2767.84, BIC= 2775.54
A10	norm	$\mu = 1.23e{-}13$,σ=21.99	AIC=3133.72, BIC= 3141.42
	Cauchy	a=5.192,λ=11.841	AIC=3222.16, BIC=3229.86
	Gumbel	$a = -11.710$,λ=25.847	AIC= 3269.59, BIC=3277.29
A20	norm	μ=920.154,σ=570.886	AIC=5312.54, BIC=5320.24
	Weibull	a=1.863,λ=1034.914	AIC=5278.68, BIC= 5286.38
	Gamma	a=2.568, λ=0.003	AIC=5298.38, BIC= 5306.08

Table A.12: Distributions' estimated parameters and information criteria of the Table Point formation found in Seamus 216 hole section. The units of measurement are [mV], [GAPI], [Ωhmm] for SP, GR and A10, A20 respectively.

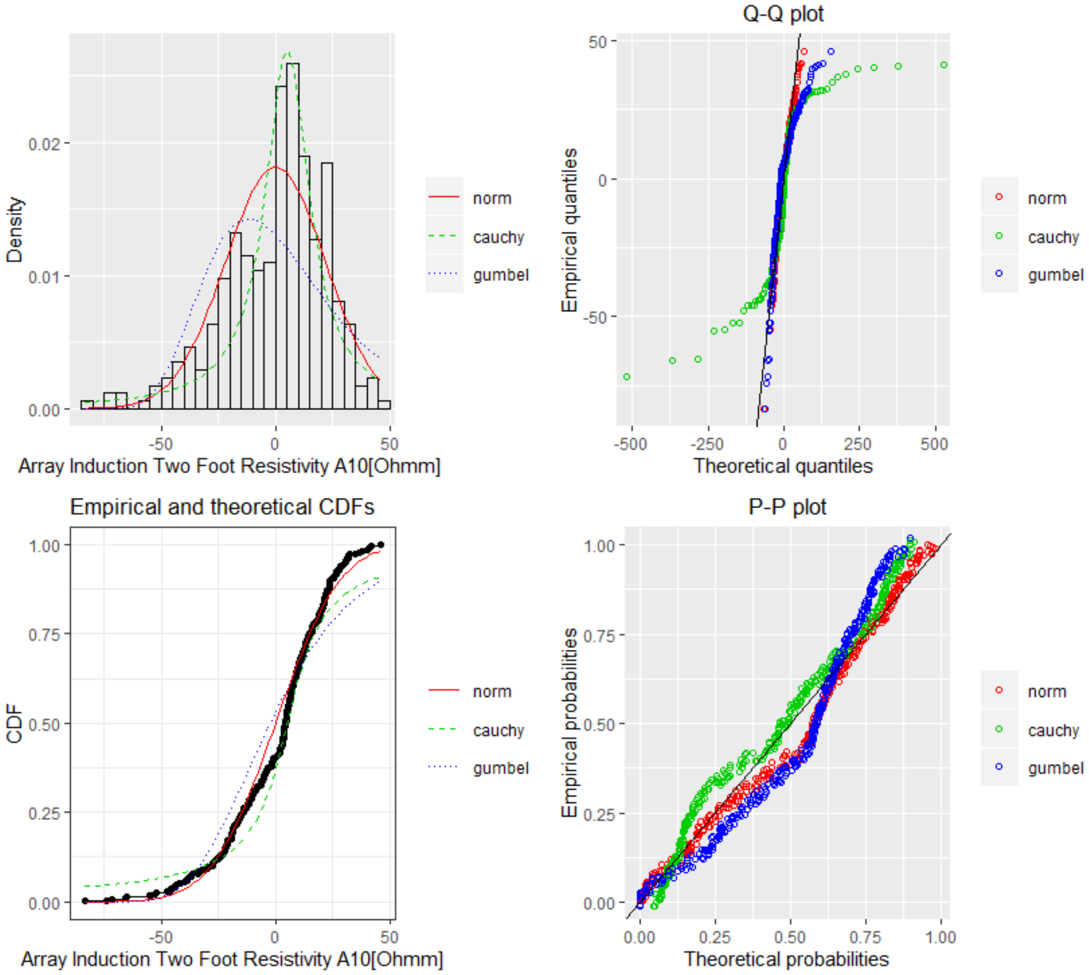


Figure A.8: Fitting of the distributions by maximum likelihood. Featured data-set; A10 log of Seamus 216mm hole section.

	Variograms			
	Model	Sill	Range	Nugget
SP	Gau	2035.890	7.629	0.000
A10	Gau	403.185	4.816	0.000
	Sph	390.383	10.151	0.000
	Pen	408.810	13.288	0.000
	Cir	388.817	8.944	0.000
A20	Sph	269635.7	6.979	36447.3
	Gau	234617.4	3.38148	71416.5
	Exp	310531	2.642	0.000
	Pen	272134.8	8.463	34517.7
	Cir	264979.3	6.240	40977.7

Table A.13: Fitting of the best theoretical model to the experimental variograms of the field.

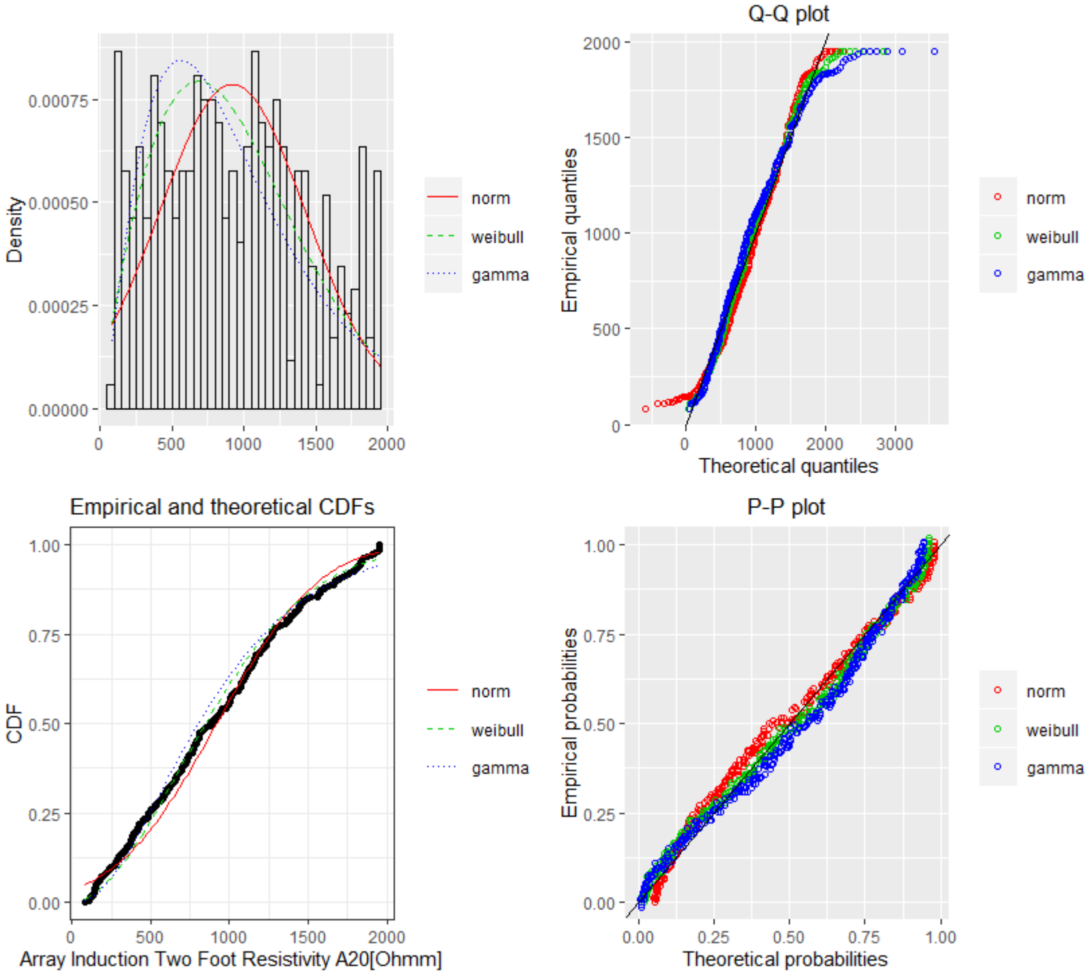
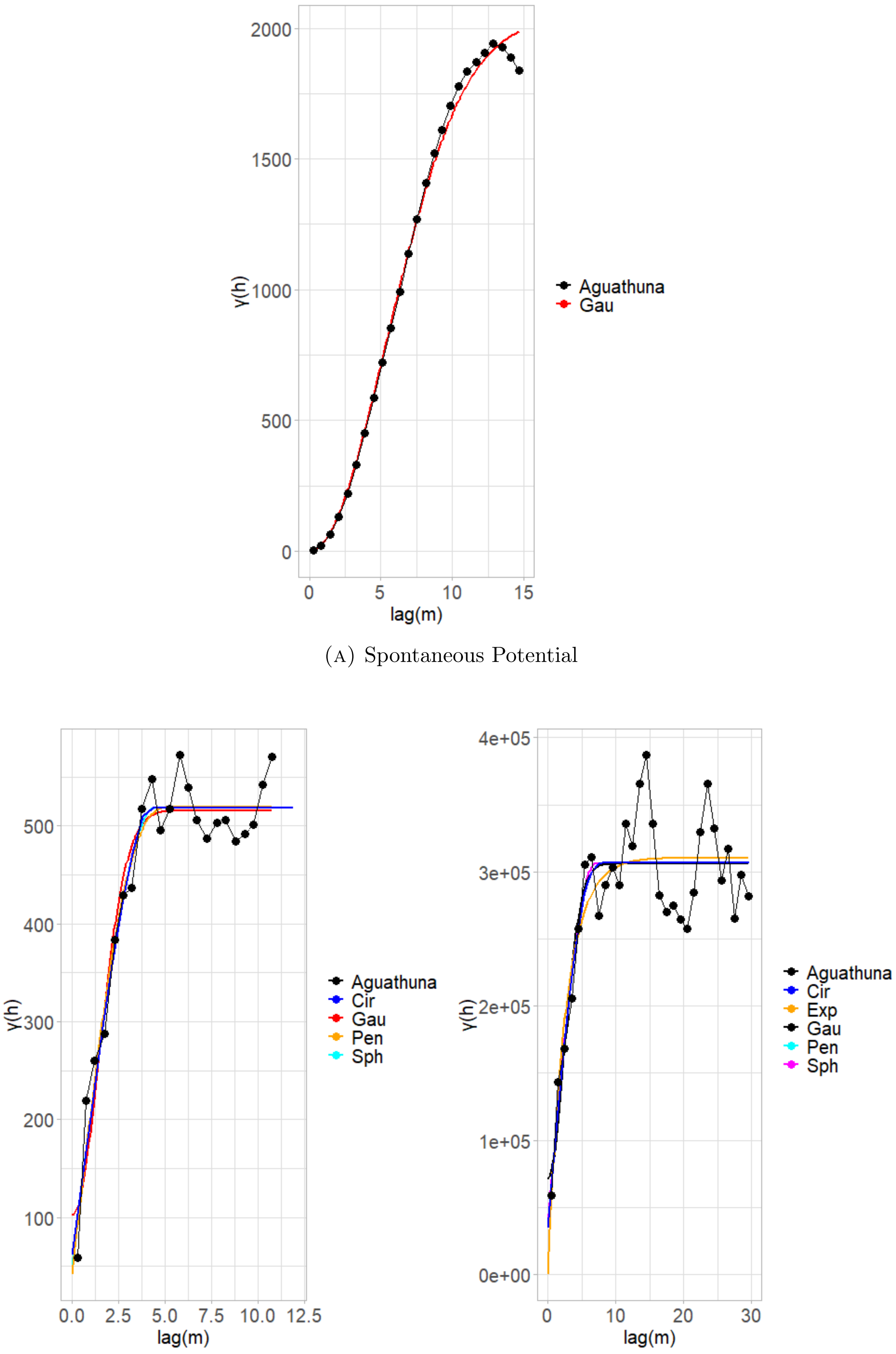


Figure A.9: Fitting of the distributions by maximum likelihood. Featured data-set; A20 log of Seamus 216mm hole section.

	Variograms			
	Model	MSE	MAE	RMSE
SP	Gau	2314.63	38.7431	48.111
A10	Gau	634.203	18.409	25.183
	Sph	1098.68	28.921	33.146
	Pen	1292.77	30.844	35.955
	Cir	900.201	25.840	30.003
A20	Sph	10965×10^5	27872.2	33112.8
	Gau	10938×10^5	27604.8	33072.3
	Exp	12133×10^5	29224.7	34832.4
	Pen	110415×10^5	28174.4	33228.7
	Cir	10854×10^5	27639.9	32945.5

Table A.14: Fitting of the best theoretical model to the experimental variograms of the field.



(A) Spontaneous Potential

(B) Induction A10

(C) Induction A20

FIGURE A.10: Variogram plots. The weights are determined using N_j, where N_j is the number of pairs at certain lag.

A.2 Finnegan 216mm hole section

The statistical parameters of the the Table Point and Aguathuna formation of the Finnegan 216 hole section are presented.

SP (.mV)	Min	Max	Mean	Median	Mode
Table Point	206.11	316.73	254.49	243.96	245.17
Aguathuna	195.36	257.63	224.02	220.66	212.17

SP (.mV)	Variance	SD	Skewness	Kurtosis
Table Point	697.33	26.40	0.46	1.92
Aguathuna	204.38	14.29	0.41	2.06

Table A.15: Spontaneous potential statistical parameters of Finnegan 216mm.

GR (GAPI)	Min	Max	Mean	Median	Mode
Table Point	3.65	34.78	11.64	11.82	12.42
Aguathuna	5.26	70.38	21.63	17	12.3

GR (GAPI)	Variance	SD	Skewness	Kurtosis
Table Point	29.54	5.44	0.85	4.13
Aguathuna	143.21	11.97	1.51	5.36

Table A.16: Gamma ray statistical parameters of Finnegan 216mm.

A10 (Ohmm)	Min	Max	Mean	Median	Mode
Table Point	189.76	2927.43	1239.76	1239.32	1232.15
Aguathuna	58.56	9280.18	1290.24	1288.16	1157.22

A10 (Ohmm)	Variance	SD	Skewness	Kurtosis
Table Point	33911.99	184.15	1.9	27.62
Aguathuna	664848.5	815.38	6.04	54.82

Table A.17: Array induction 10in statistical parameters of Finnegan 216mm.

A20 (Ohmm)	Min	Max	Mean	Median	Mode
Table Point	245.35	1818.34	1231.72	1238.03	1232.15
Aguathuna	33.38	3567.02	1206.96	1276.09	1218.03

A20 (Ohmm)	Variance	SD	Skewness	Kurtosis
Table Point	13612.58	116.67	-2.22	24.79
Aguathuna	173256.6	416.24	0.41	10.13

Table A.18: Array induction 20in statistical parameters of Finnegan 216mm.

A.2.1 Table Point Formation

Log	Min	Max	Mean	Median	Mode
SP(.mV)	-48.818	31.691	2.673e-13	4.513	-1.636
GR(.GAPI)	-6.928	21.143	-1.247e-14	-0.909	3.745
A20(Ohmm)	-988.255	584.777	4.568e-13	6.186	17.826

Log	Variance	SD	Skewness	Kurtosis
A10(Ohnmm)	270.041	16.433	-0.795	3.113
GR(.GAPI)	18.696	4.324	1.285	5.325
A20(Ohmm)	13603	116.632	-2.257	25.008

Table A.19: Detrended data statistics of Table Point found in Finnegan 216 hole section.

Log	Model	Estimated Trend Function
SP	Linear	$290.32 - 0.09x + \epsilon_i, \epsilon \sim N(0, 16.4^2)$
GR	Linear	$17.35 - 0.016x + \epsilon_i, \epsilon \sim N(0, 4.33^2)$
A20	Linear	$1237.09 - 0.015x + \epsilon_i, \epsilon \sim N(0, 117^2)$

Table A.20: Estimated trend models.

	Histograms		
	Distribution	Parameters	Information Criteria
SP	norm	μ=2.672e-13,σ=16.42	AIC=6127.86, BIC=6137.04
	Cauchy	a=5.232,γ=8.332	AIC= 6291.04, BIC= 6300.22
	Gumbel	μ=-8.779,b=18.466	AIC=6379.97, BIC=6389.15
GR	norm	$\mu = -1.247e{-}14$,σ=4.321	AIC=4189.26, BIC= 4198.43
	Cauchy	a=-1.205,λ=2.293	AIC=4306.18, BIC=4315.35
	Gumbel	a=-1.903,λ= 3.175	AIC= 4002.98, BIC= 4012.16
A10	norm	μ = 1239.764,σ=184.025	AIC=9636.58, BIC= 9645.76
	Cauchy	a=1239.361,λ=39.761	AIC=8841.62, BIC=8850.8
	Gumbel	a = 1150.319,λ=255.331	AIC=10011.5 , BIC=10020.7
A20	norm	μ=4.965$e{-}14$,σ=8.587	AIC=10161.9 , BIC=10172.5
	Cauchy	a=1.967,λ=3.977	AIC=10217, BIC= 10227.5
	Gumbel	a=-4.691, λ=10.576	AIC=10824.9, BIC= 10835.4

Table A.21: Distributions' estimated parameters and information criteria of the Table Point formation found in Seamus 216 hole section. The units of measurement are [mV], [GAPI], [Ωhmm] for SP, GR and A10, A20 respectively.

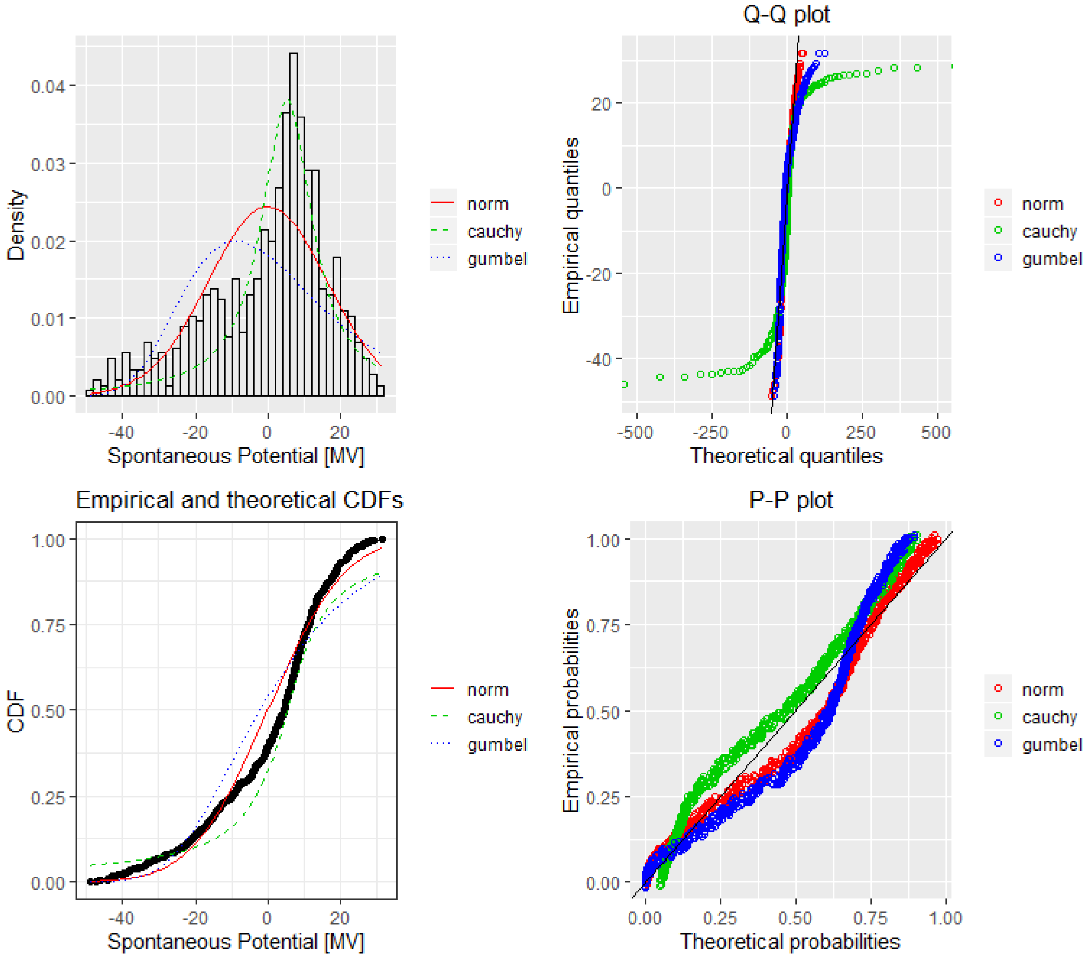


Figure A.11: Fitting of the distributions by maximum likelihood. Featured data-set; SP log of Finnegan 216mm hole section.

	Variograms			
	Model	Sill	Range	Nugget
SP	Cir	2035.890	7.629	0.000
	Gau	2035.890	7.629	0.000
	Pen	2035.890	7.629	0.000
	Sph	2035.890	7.629	0.000
GR	Exp	403.185	4.816	0.000
	Gau	390.383	10.151	0.000
	Pen	408.810	13.288	0.000
	Sph	388.817	8.944	0.000
A10	Exp	269635.7	6.979	36447.3
	Gau	234617.4	3.38148	71416.5

Table A.22: Fitting of the best theoretical model to the experimental variograms of the field.

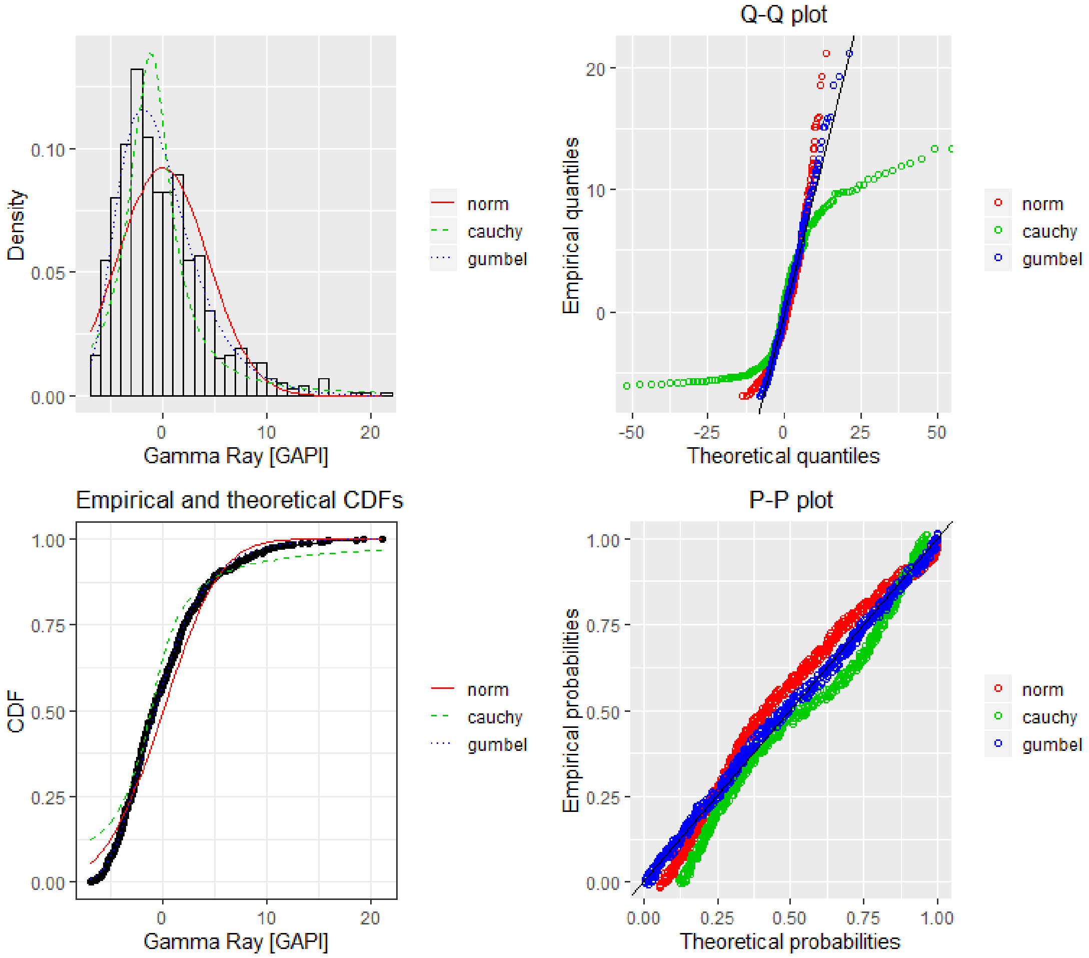


Figure A.12: Fitting of the distributions by maximum likelihood. Featured data-set; GR log of Finnegan 216mm hole section.

	Variograms			
	Model	MSE	MAE	RMSE
SP	Cir	21.871	3.786	4.676
	Gau	51.565	5.575	7.181
	Pen	22.166	3.849	4.708
	Sph	22.070	3.825	4.698
GR	Exp	0.933	0.771	0.966
	Gau	0.993	0.759	0.996
	Pen	0.710	0.700	0.842
	Sph	0.673	0.695	0.820
A10	Exp	1784640	956.02	1335.9
	Gau	1148419	852.498	1071.64

Table A.23: Fitting of the best theoretical model to the experimental variograms of the field.

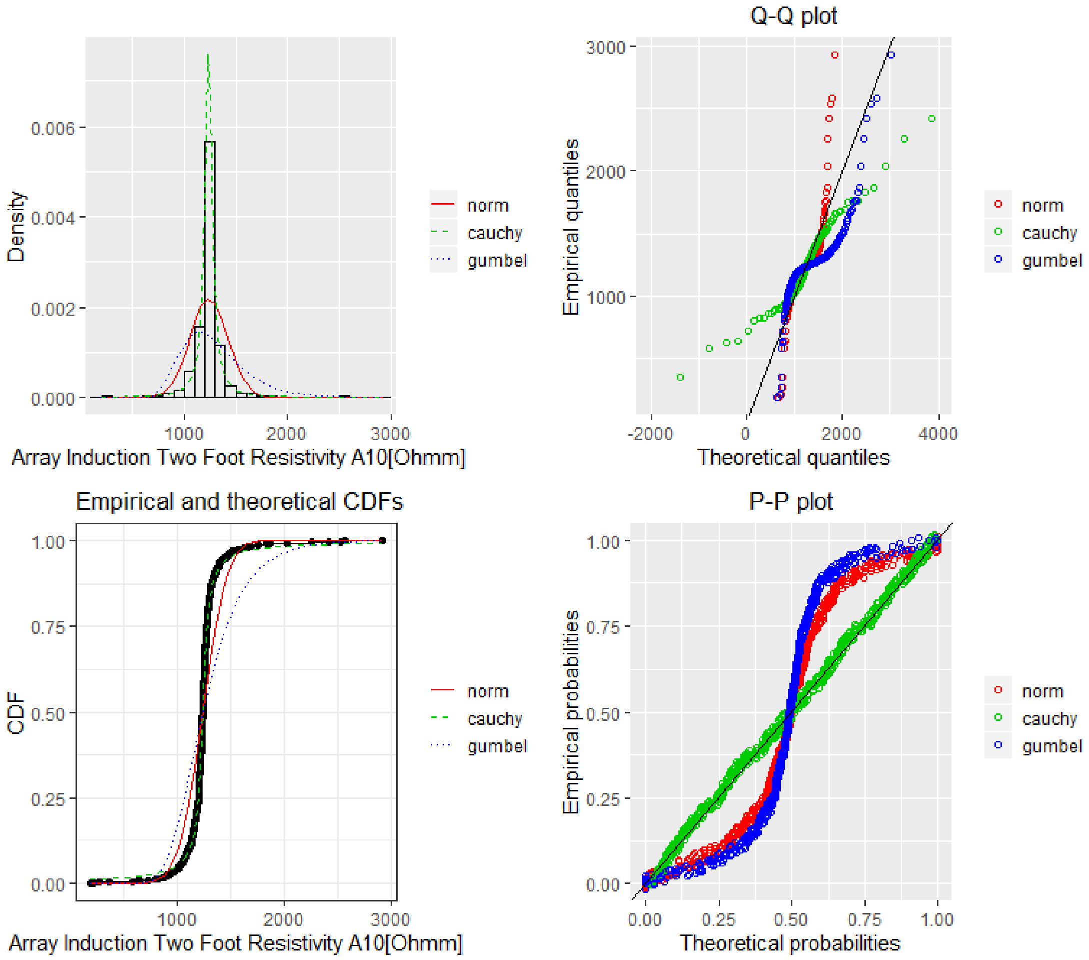


Figure A.13: Fitting of the distributions by maximum likelihood. Featured data-set; A10 log of Finnegan 216mm hole section.

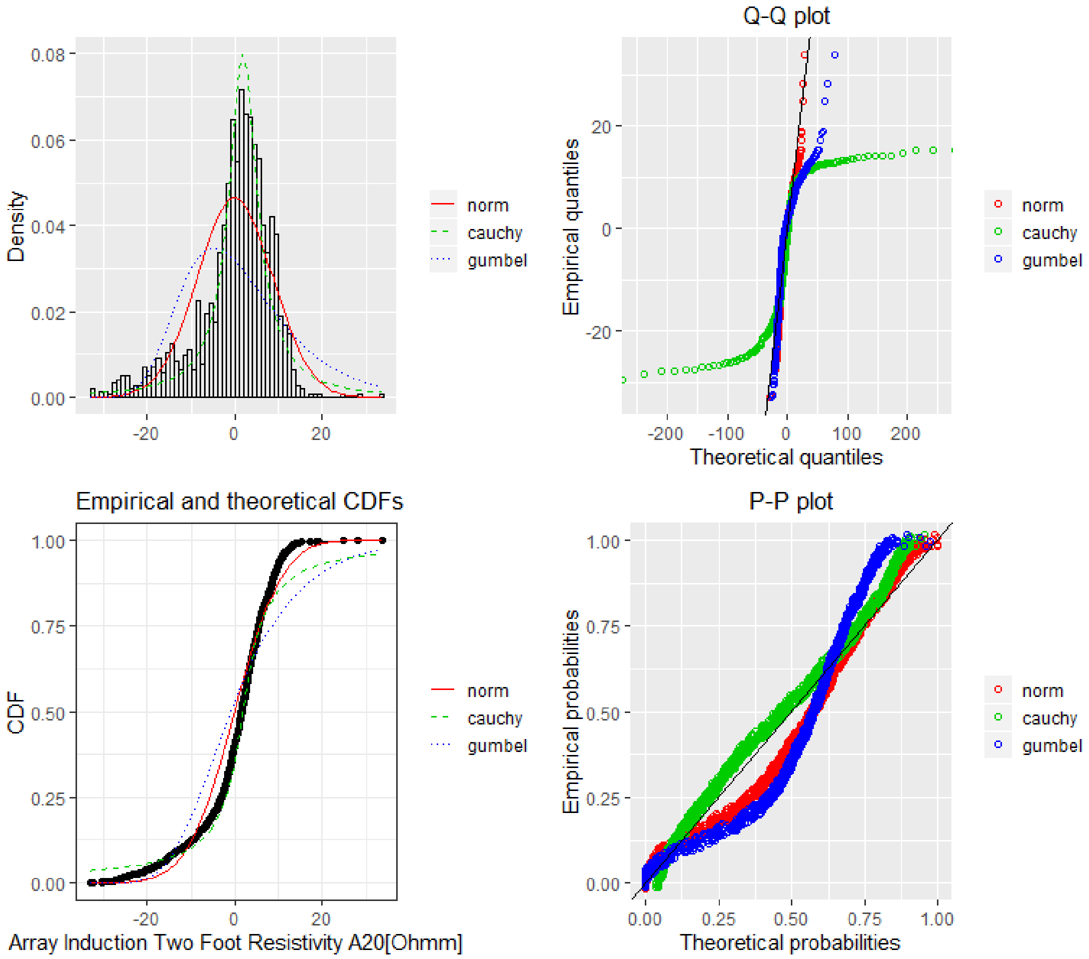
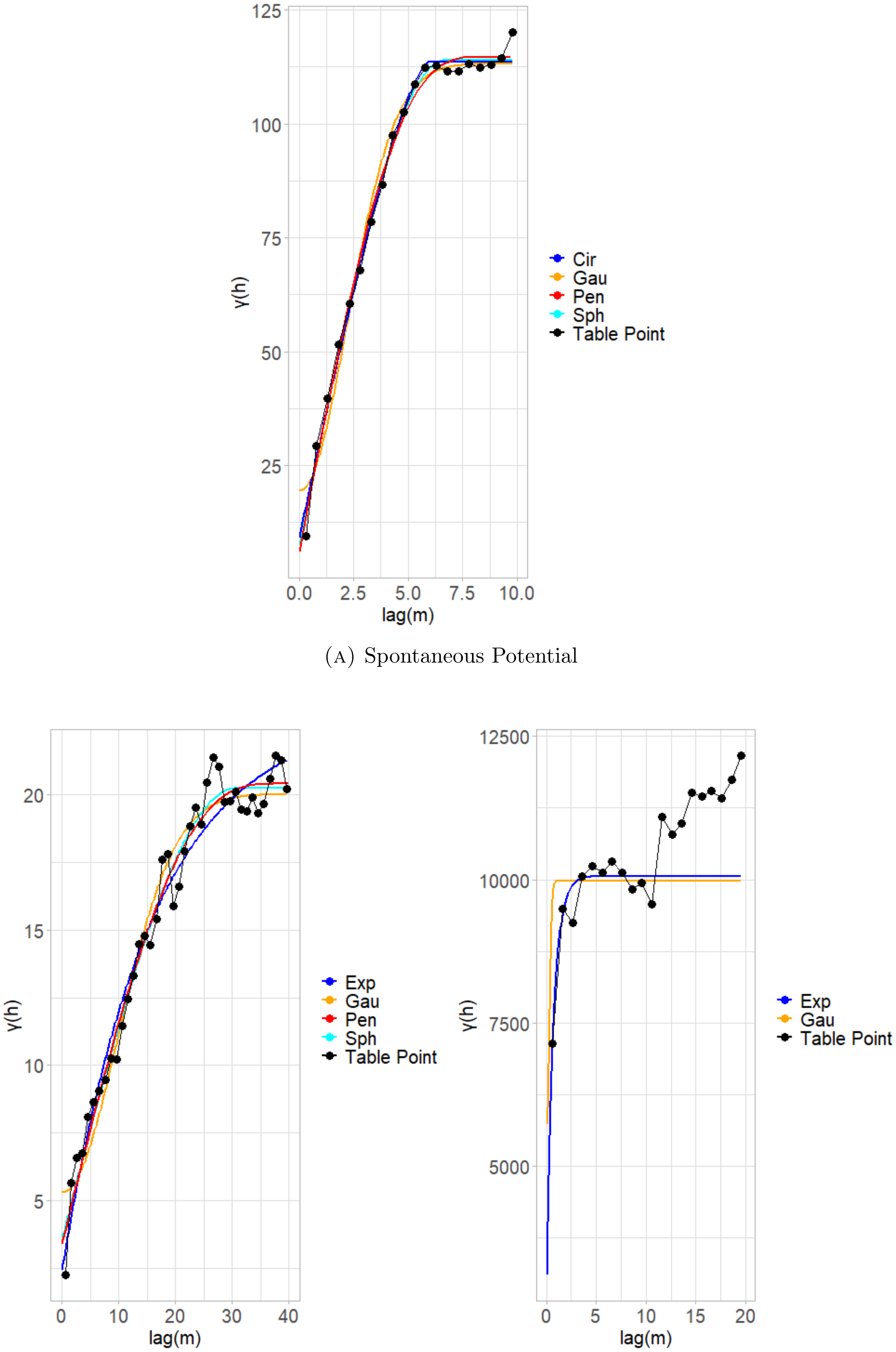


Figure A.14: Fitting of the distributions by maximum likelihood. Featured data-set; A20 log of Finnegan 216mm hole section.



(A) Spontaneous Potential

(B) Gamma Ray

(C) Induction A10

FIGURE A.15: Variogram plots. The weights are determined using N_j, where N_j is the number of pairs at certain lag.

A.2.2 Aguathuna Formation

Log	Min	Max	Mean	Median	Mode
SP(.mV)	-29.441	28.190	-1.473e-15	-1.159	28.189
GR(.GAPI)	-20.638	41.896	1.553e-14	-1.453	-3.914
A10(Ohmm)	-1270.69	7955.62	7.846e-14	-65.771	-193.574
A20(Ohmm)	-1150.17	2321.79	2.274e-13	22.513	-79.281

Log	Variance	SD	Skewness	Kurtosis
A10(Ohnmm)	114.036	10.679	0.200	2.600
GR(.GAPI)	107.700	10.378	1.217	5.166
A10(Ohmm)	649360	805.829	6.141	56.023
A20(Ohmm)	162932	403.649	0.505	10.477

Table A.24: Detrended data statistics of Aguathuna formation found in Finnegan 216 hole section.

Log	Model	Estimated Trend Function
SP	Linear	$202 + 0.271x - 5.81 \cdot 10^{-4}x^2 + \epsilon_i, \epsilon \sim N(0, 10.7^2)$
GR	Linear	$11.193 - 0.082x + \epsilon_i, \epsilon \sim N(0, 10.4^2)$
A10	Linear	$1486.817 - 1.560x + \epsilon_i, \epsilon \sim N(0, 807^2)$
A20	Linear	$1377.926 - 1.354x + \epsilon_i, \epsilon \sim N(0, 404^2)$

Table A.25: Estimated trend models.

	Histograms		
	Distribution	Parameters	Information Criteria
SP	norm	μ= -1.472e-15,σ=10.657	AIC=1904.17, BIC= 1911.22
	Cauchy	a=-1.493,γ=6.990	AIC= 2015.16, BIC=2022.21
	Gumbel	μ=-5.212,b= 9.976	AIC=1922.86, BIC=1929.91
GR	norm	$\mu = 1.553e{-}14$,σ=10.357	AIC=1889.82, BIC= 1896.87
	Cauchy	a=-1.986,λ=5.168	AIC=1909.45, BIC=1916.5
	Gumbel	a=-4.604,λ= 8.095	AIC= 1841.2, BIC=1848.25
A10	norm	$\mu = 7.846e{-}14$,σ=804.222	AIC=4074.62, BIC= 4081.68
	Cauchy	a=39.125,λ=113.621	AIC=3566.45, BIC=3573.5
A20	norm	μ=2.273$e{-}13$,σ=402.844	AIC=3727.58 , BIC=3734.63
	Cauchy	a=39.125,λ=113.621	AIC=3566.45, BIC= 3573.5

Table A.26: Distributions' estimated parameters and information criteria of the Table Point formation found in Seamus 216 hole section. The units of measurement are [mV], [GAPI], [Ωhmm] for SP, GR and A10, A20 respectively.

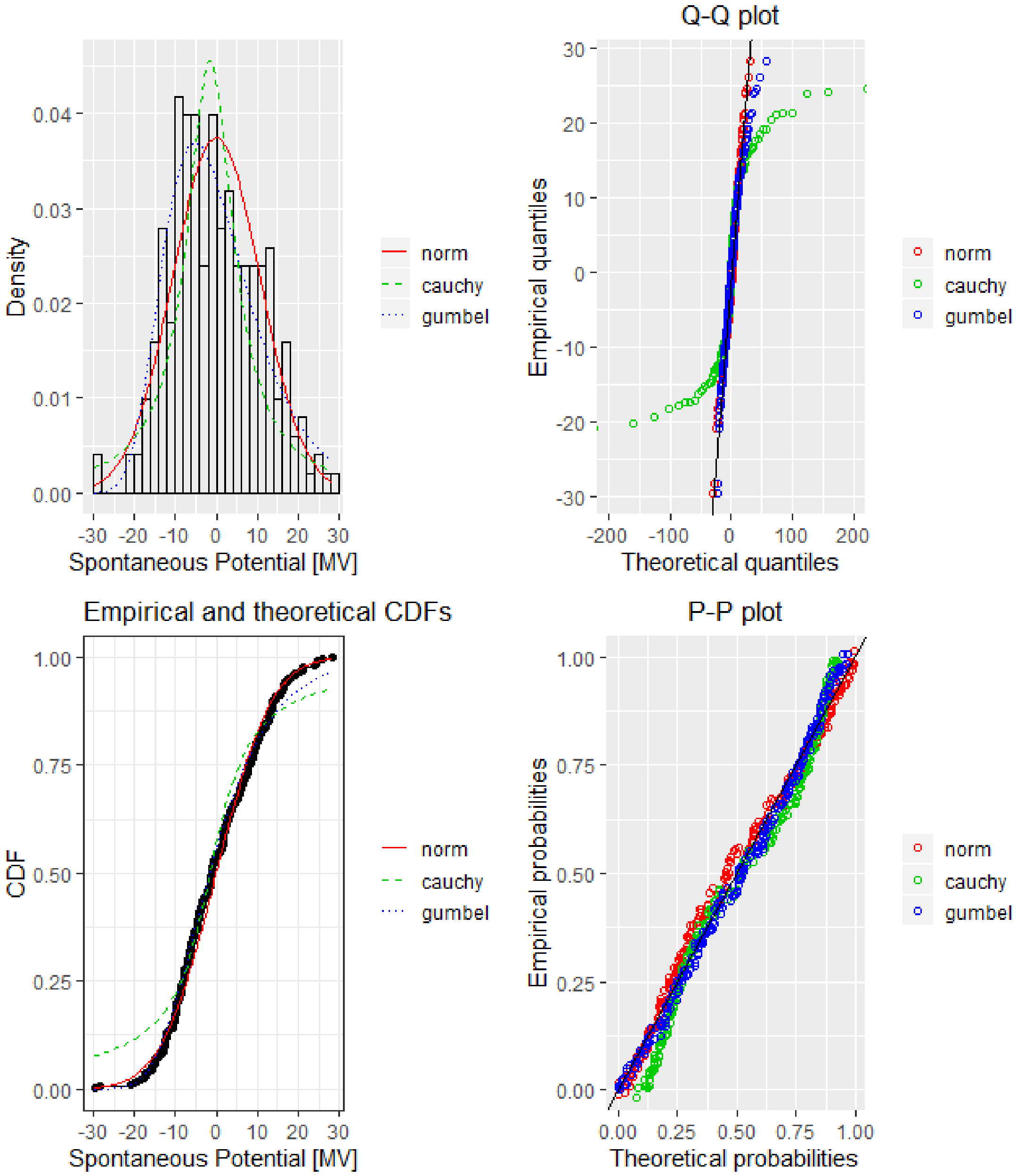


Figure A.16: Fitting of the distributions by maximum likelihood. Featured data-set; SP log of Finnegan 216mm hole section.

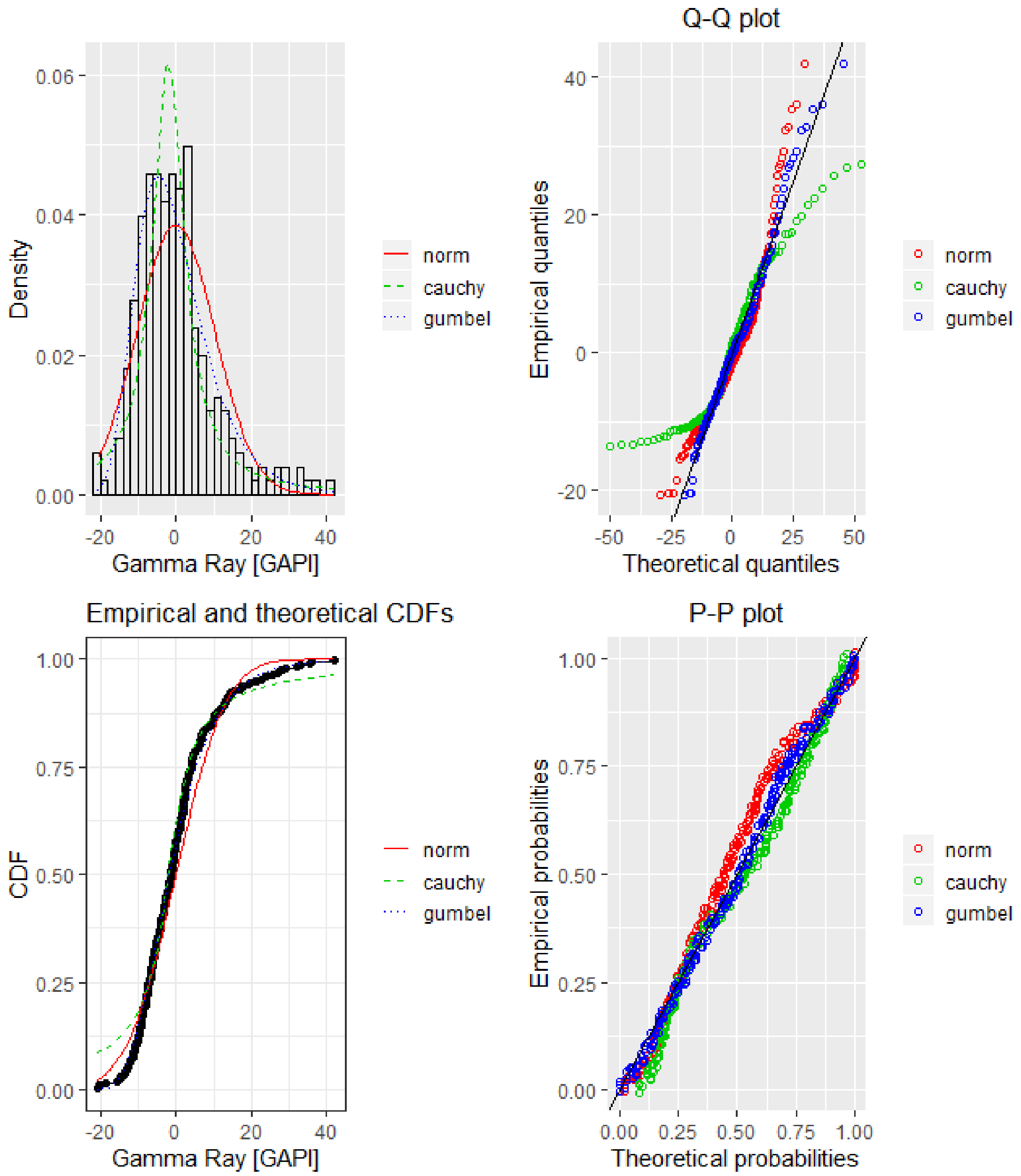


FIGURE A.17: Fitting of the distributions by maximum likelihood. Featured data-set; GR log of Finnegan 216mm hole section.

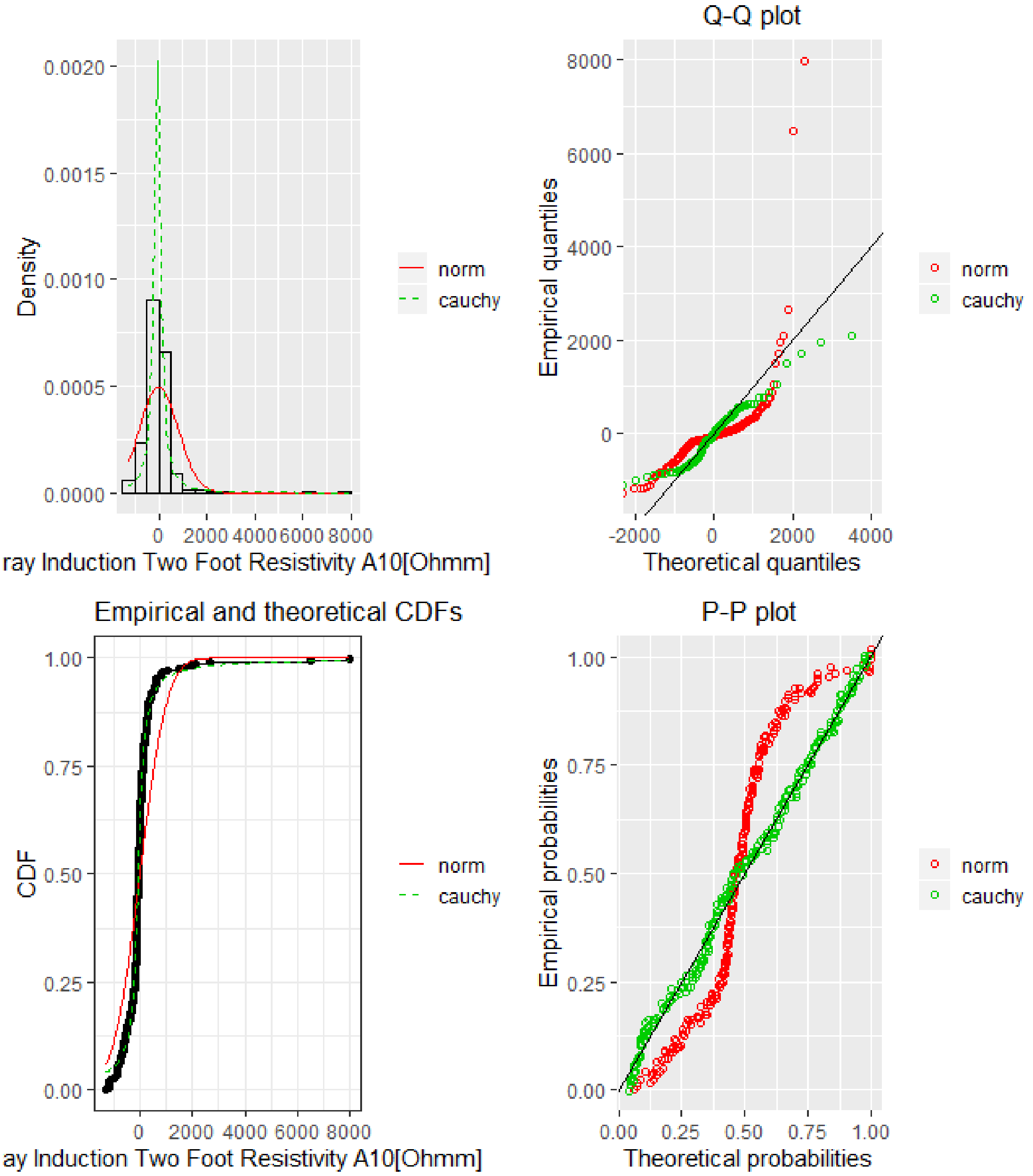


Figure A.18: Fitting of the distributions by maximum likelihood. Featured data-set; A10 log of Finnegan 216mm hole section.

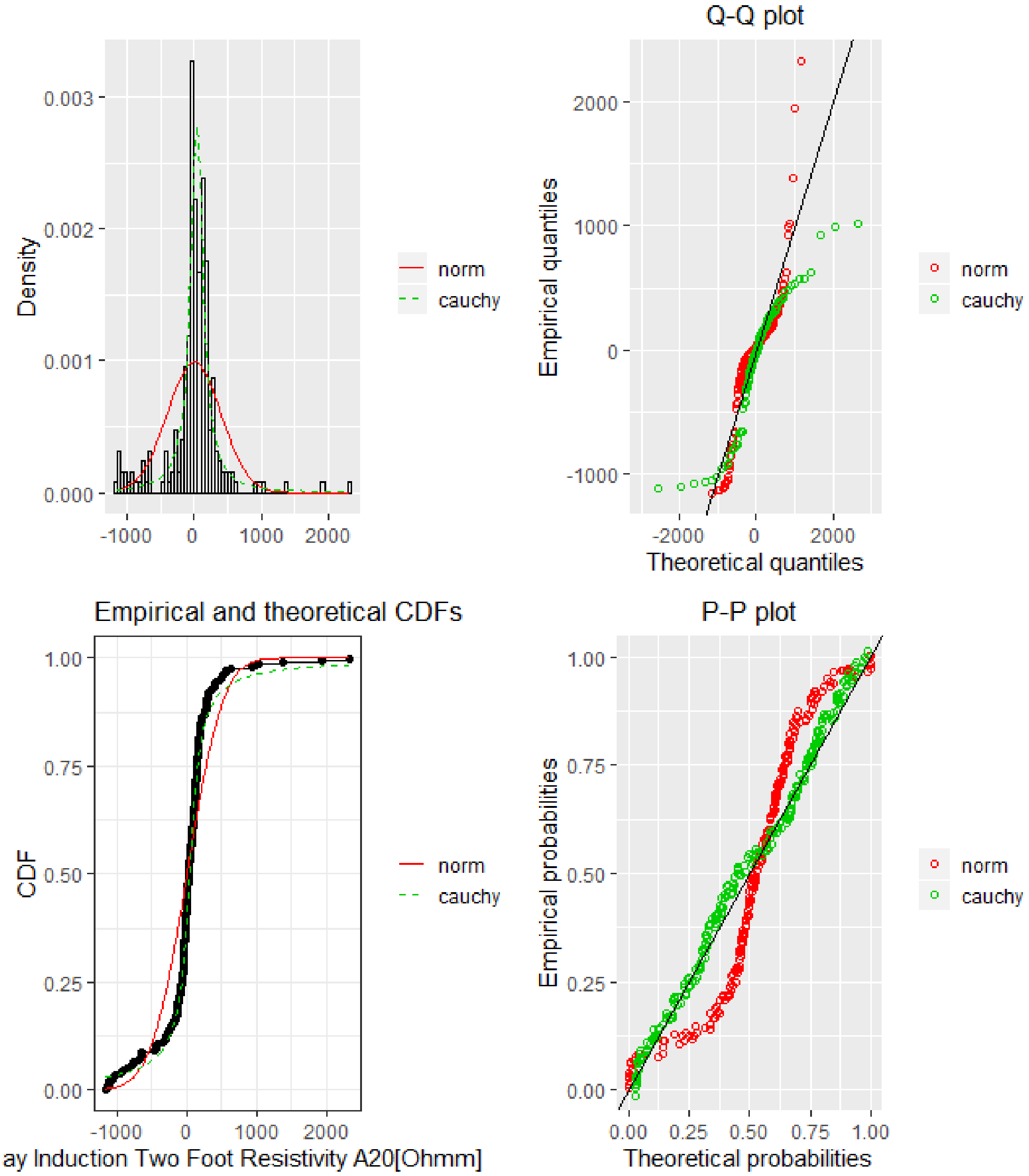
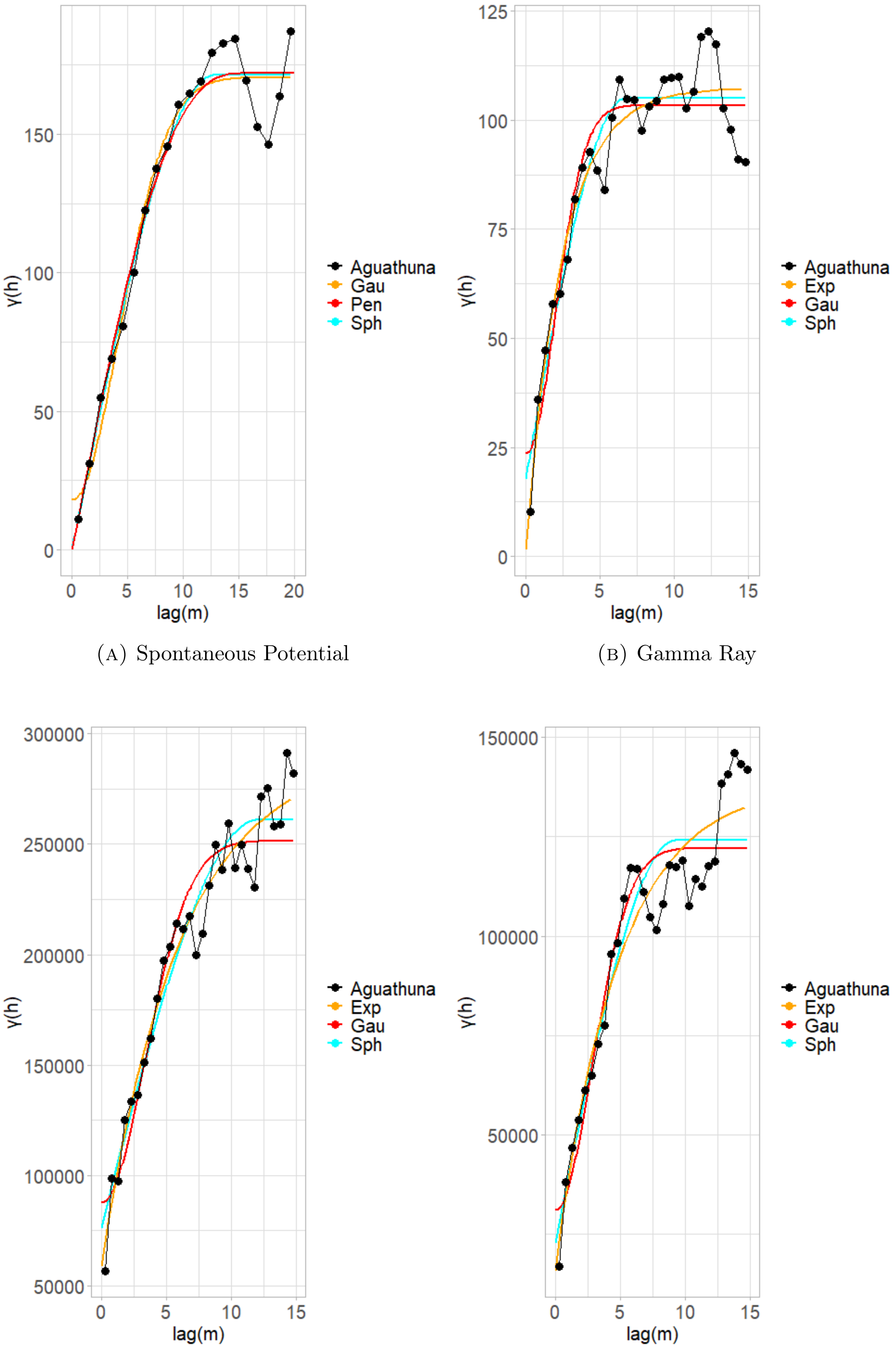


FIGURE A.19: Fitting of the distributions by maximum likelihood. Featured data-set; A20 log of Finnegan 216mm hole section.



(A) Spontaneous Potential

(B) Gamma Ray

(C) Induction A10

(D) Induction A20

FIGURE A.20: Variogram plots. The weights are determined using N_j, where N_j is the number of pairs at certain lag.

	Variograms			
	Model	Sill	Range	Nugget
SP	Gau	156.957	6.263	19.116
	Pen	181.122	16.780	0.000
	Sph	178.355	13.704	0.527
GR	Exp	259.38	10.858	0.000
	Gau	158.032	6.341	19.529
	Sph	180.957	14.201	1.651
A10	Exp	105.904	2.428	1.500
	Gau	79.836	2.781	23.569
	Sph	87.332	6.720	17.702
A20	Exp	232382.0	6.037	58595.3
	Gau	163960.4	4.791	87586.9
	Sph	184620.1	12.042	76319.8

Table A.27: Fitting of the best theoretical model to the experimental variograms of the field.

	Variograms			
	Model	MSE	MAE	RMSE
SP	Gau	52.395	5.294	7.238
	Pen	33.896	4.782	5.822
	Sph	178.355	13.704	0.527
GR	Sph	$45809 \cdot 10^6$	205523	214031
	Gau	$45815 \cdot 10^6$	205540	214045
	Exp	$45817 \cdot 10^6$	205543	214050
A10	Exp	$87474 \cdot 10^3$	8042.73	9352.79
	Gau	$12790 \cdot 10^4$	9350.49	11309.4
	Sph	$21954 \cdot 10^4$	12556.6	14817
A20	Exp	$15356 \cdot 10^4$	10089.4	12392
	Gau	$30253 \cdot 10^4$	14065.4	17393.5
	Sph	$21955 \cdot 10^4$	12555.8	14817.5

Table A.28: Fitting of the best theoretical model to the experimental variograms of the field.

A.3 Correlation Graphs

All correlations computation presented in appendices, are based on Pearson's rank correlation and have been verified through Spearman's rank correlation. The graphical outcomes of the correlation table 6.12 are compiled in figures A.21, A.22, A.23, A.24, A.25. Results show no strong sign of correlation between the values of Gamma Ray log in Seamus and Finnegan well.

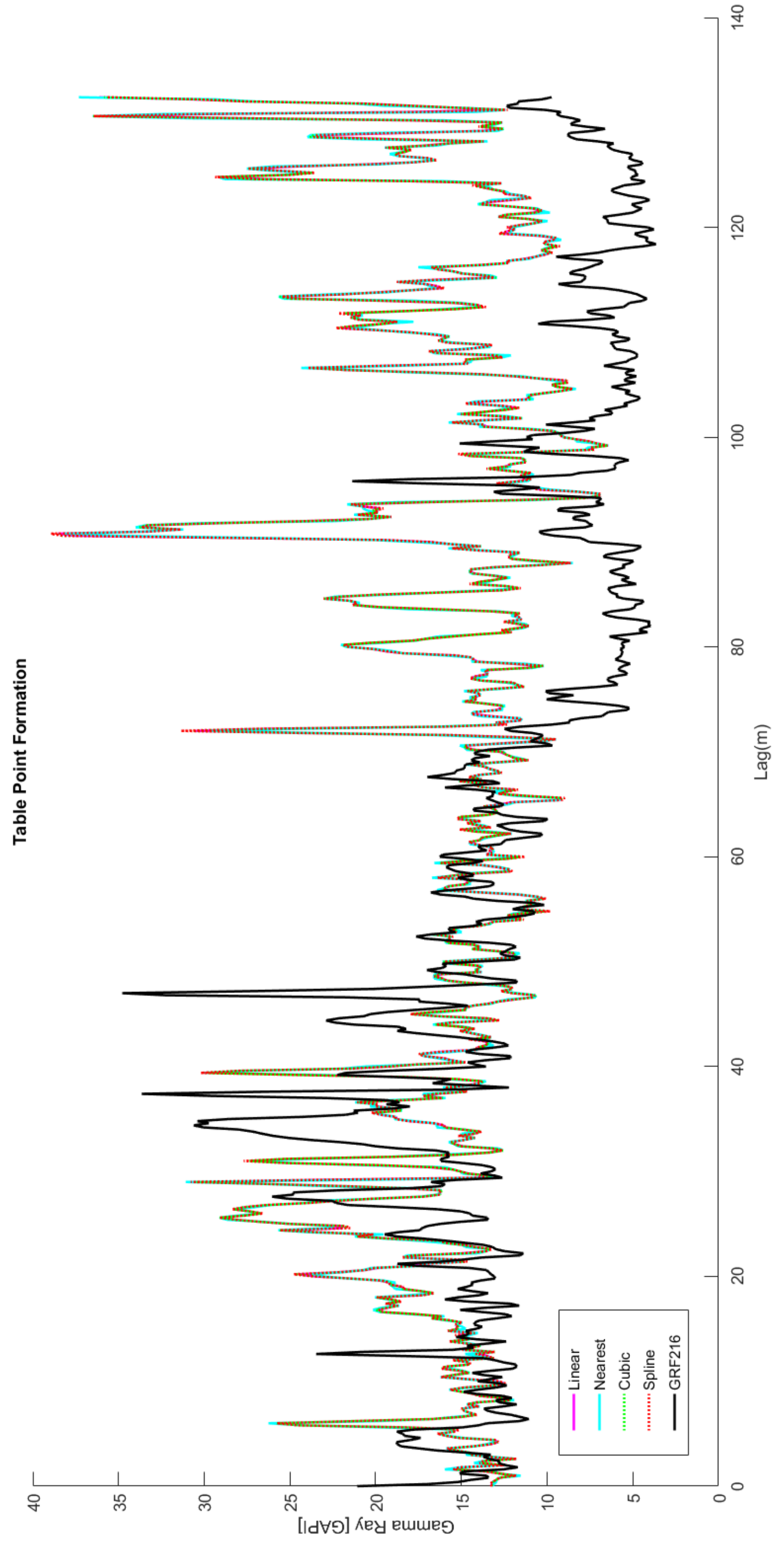
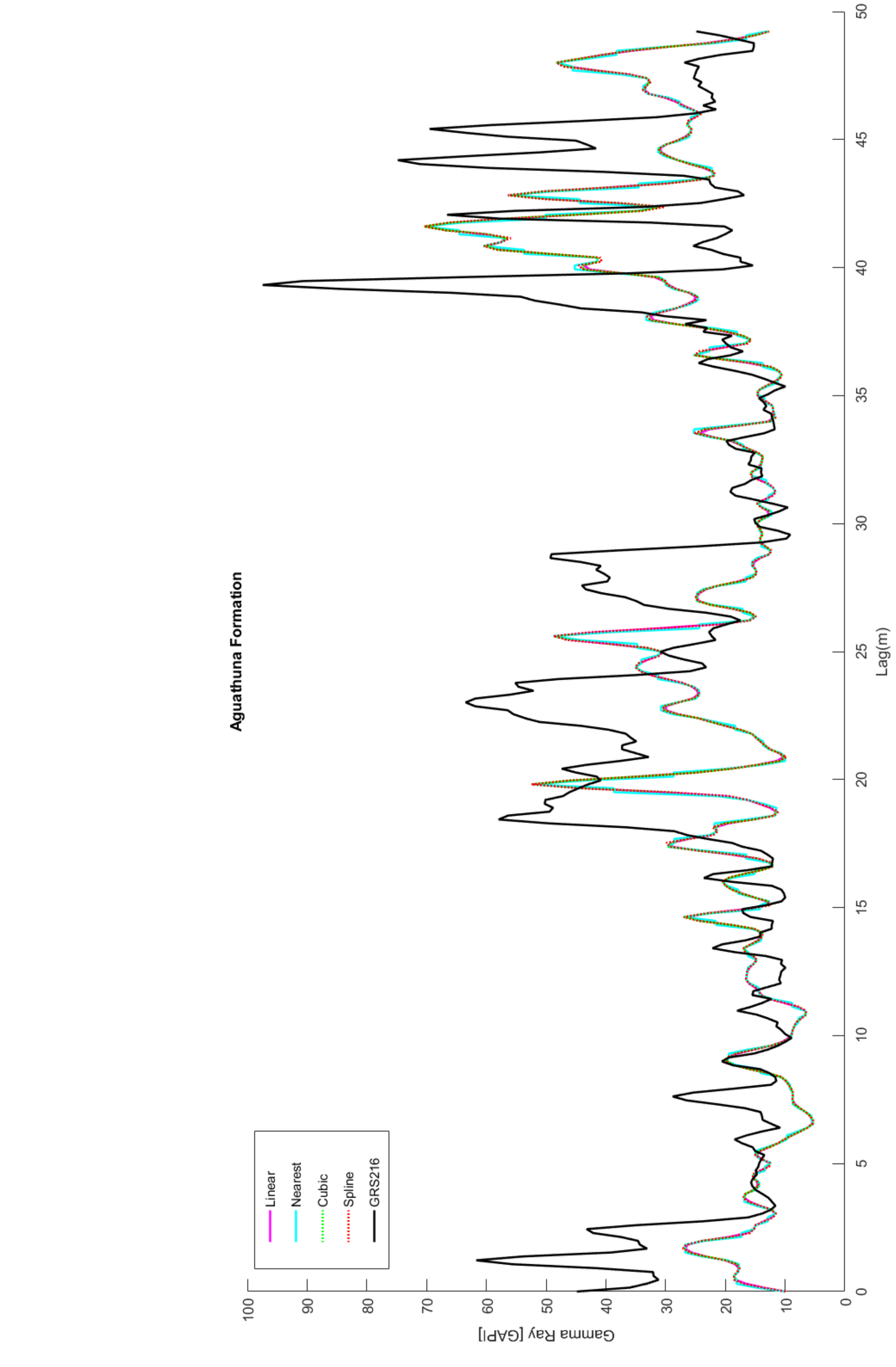
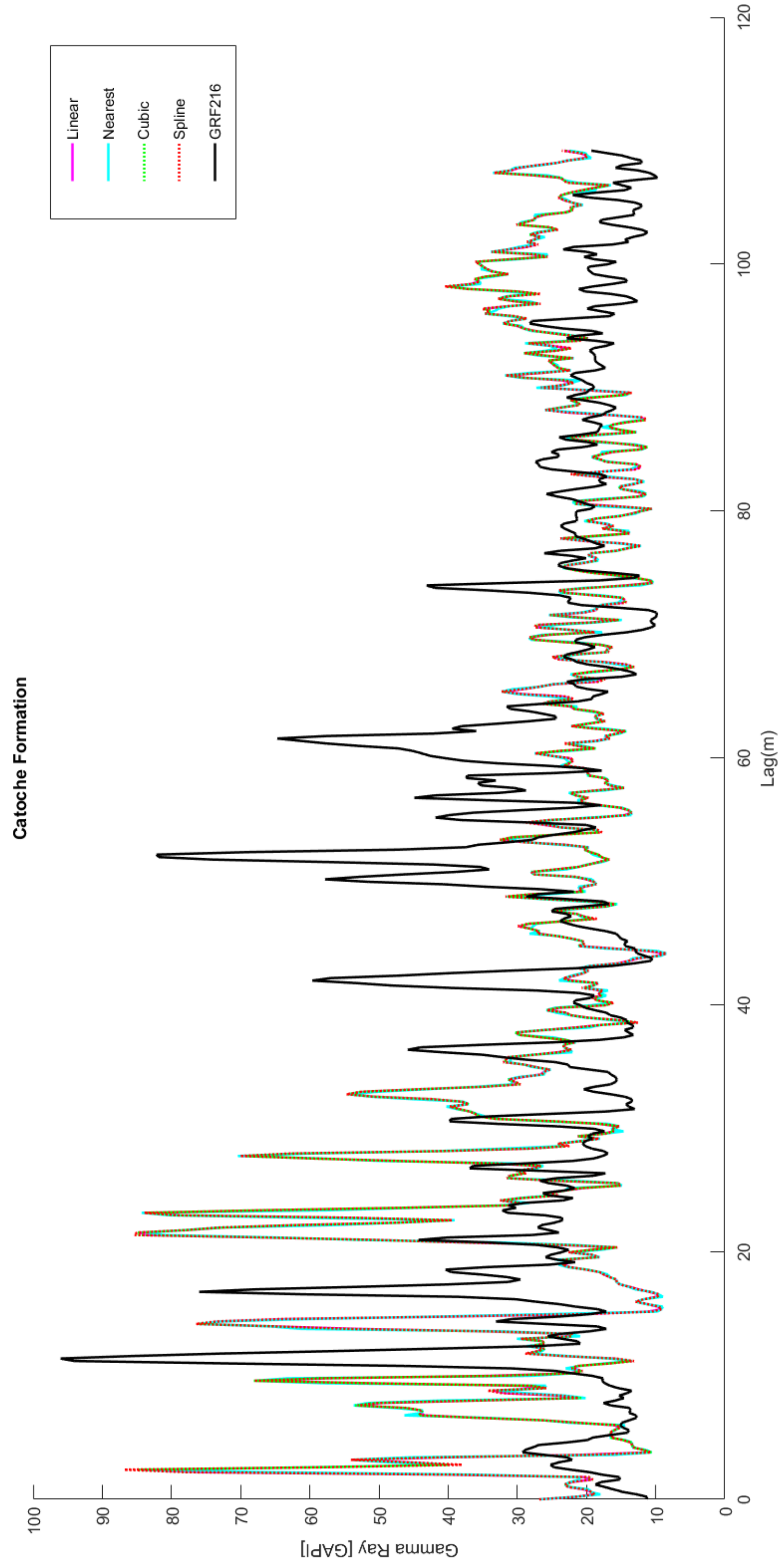


Table Point Formation
Gamma Ray [GAPI]
0
5
10
15
20
25
30
35
40
Lag(m)
0
20
40
60
80
100
120
140
Linear
Nearest
Cubic
Spline
GRF216

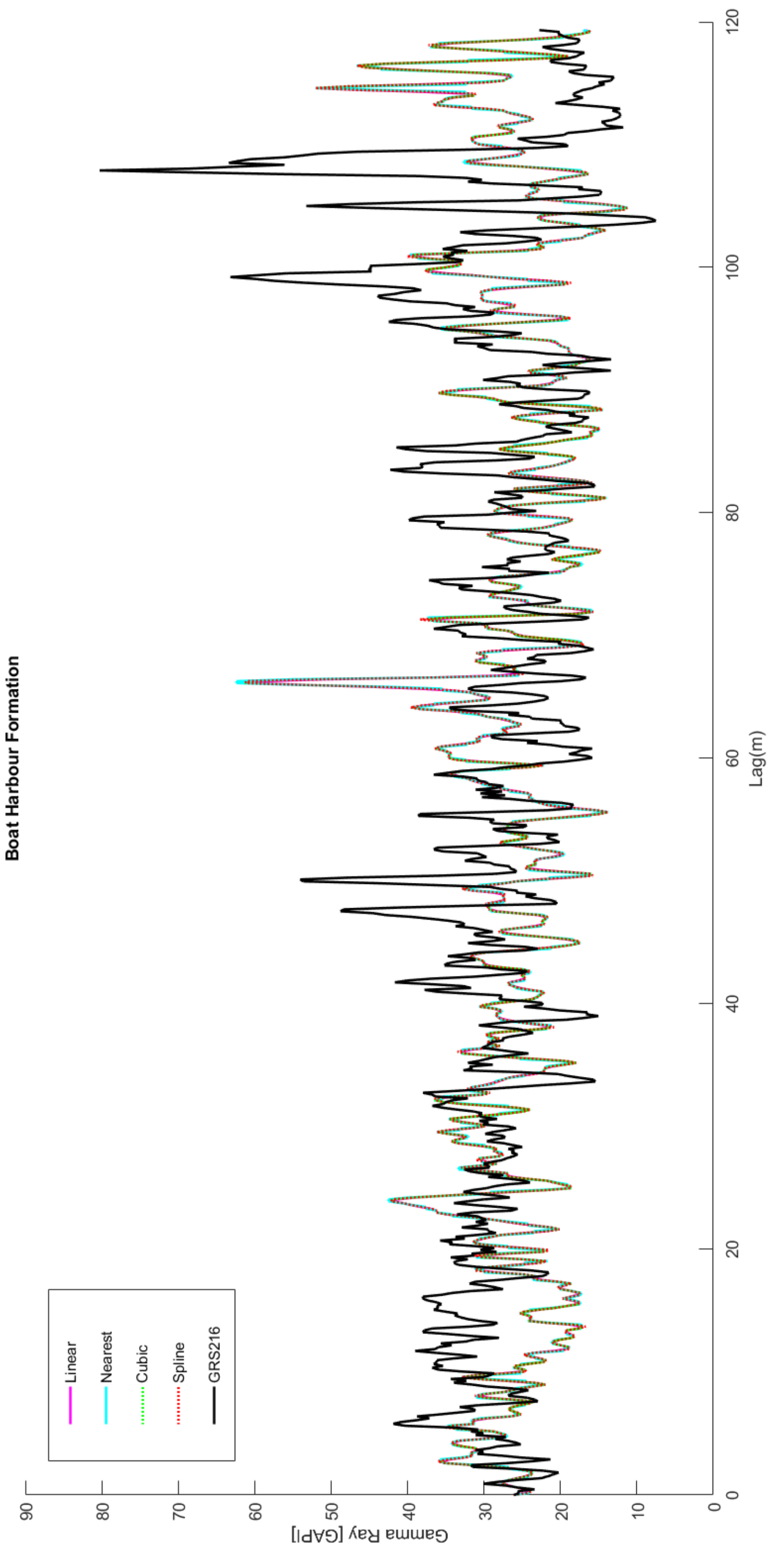
.png


Aguathuna Formation
Linear
Nearest
Cubic
Spline
GRS216
Gamma Ray [GAPI]
100
90
80
70
60
50
40
30
20
10
0
Lag(m)
0
5
10
15
20
25
30
35
40
45
50

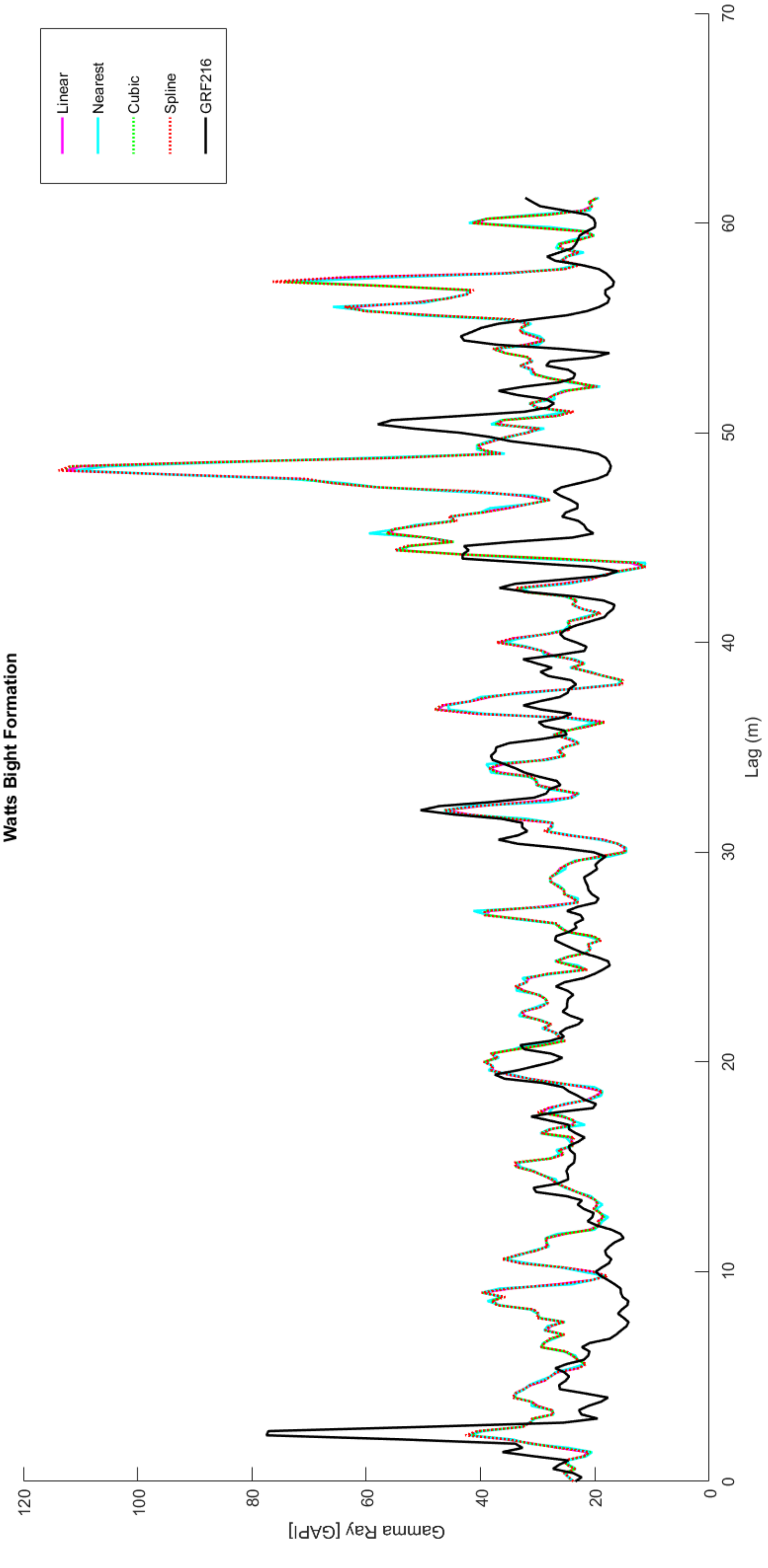
.png


Catoche Formation
Gamma Ray [GAPI]
100
90
80
70
60
50
40
30
20
10
0
Lag(m)
0
20
40
60
80
100
120
Linear
Nearest
Cubic
Spline
GRF216

.png


Boat Harbour Formation
Linear
Nearest
Cubic
Spline
GRS216
Gamma Ray [GAPI]
0
10
20
30
40
50
60
70
80
90
Lag(m)
0
20
40
60
80
100
120

.png


Watts Bight Formation
Gamma Ray [GAPI]
0
20
40
60
80
100
120
Lag (m)
0
10
20
30
40
50
60
70
Linear
Nearest
Cubic
Spline
GRF216

.png

Appendix B

Data Structures and Algorithms

All algorithms for statistical and spatial analysis as well as the algorithms for the estimation of missing values in data, were developed and run in *R* and *Matlab* environment.

B.1 Correlations

The preliminary, exploratory and variogram analysis was developed and run in *R* environment. The calculation of the well-log correlations was developed and run in *Matlab* environment.

```
#R packages
library(gstat)
library(automap)
library(ggplot2)
library(MASS)
library(fitdistrplus)
library(gridExtra)
library(actuar)
library(extraDistr)
library(imputeTS)
library(Amelia)
library(forecast)
library(readr)
library(caret)

#Set graph size
x11(width=8, height=9, pointsize=15)
par(mfrow=c(1,1), mar=c(3,3,3,3))

#Import data
setwd("c:\\users\\Anastasia\\Desktop\\NalcorEnergy"')
rm(list=ls(all.names=TRUE))
graphics.off()
F311<-read_delim("Finnegan311.csv", ";", escape_double = FALSE,
trim_ws = TRUE)
F216<-read_delim("Finnegan216.csv", ";", escape_double = FALSE,
trim_ws = TRUE)
S216<-read_delim("Seamus311.csv", ";", escape_double = FALSE,
trim_ws = TRUE)

#Select formations
#Formations of the Finnegan 311 hole section
#Goose (American) Tickle
Goose311_1=matrix(F311$DEPT[7087:8509])
Goose311_2=matrix(F311$SP[7087:8509])
Goose311_3=matrix(F311$GR[7087:8509])
Goose311_4=matrix(F311$M2R1[7087:8509])
Goose311_5=matrix(F311$M2R2[7087:8509])

GooseDataF<-cbind(Goose311_1,Goose311_2,Goose311_3,Goose311_4,Goose311_5)

#Formations of the Finnegan 216 hole section
#Table Point
TableF_1<-F216$DEPT[180:905]
TableF_2<-F216$SP[180:905]
TableF_3<-F216$GR[180:905]
TableF_4<-F216$M2R1[180:905]
TableF_5<-F216$M2R2[180:905]

TableDataF<-cbind(TableF_1,TableF_2,TableF_3,TableF_4,TableF_5)
```

```
#Aguathuna
AguathF_1<-F216$DEPT[906:1155]
AguathF_2<-F216$SP[906:1155]
AguathF_3<-F216$GR[906:1155]
AguathF_4<-F216$M2R1[906:1155]
AguathF_5<-F216$M2R2[906:1155]

AguathDataF<-cbind(AguathF_1,AguathF_2,AguathF_3,AguathF_4,AguathF_5)

#Formations of the Seamus 216 hole section
#Goose (American) Tickle
Goose216_1<-S216$DEPT[500:2000]
Goose216_2<-S216$SP[500:2000]
Goose216_3<-S216$GR[500:2000]
Goose216_4<-S216$M2R1[500:2000]
Goose216_5<-S216$M2R2[500:2000]

GooseDataS<-cbind(Goose216_1,Goose216_2,Goose216_3,Goose216_4,Goose216_5)

#Table Point
TableS_1<-S216$DEPT[2700:3570]
TableS_2<-S216$SP[2700:3570]
TableS_3<-S216$GR[2700:3570]
TableS_4<-S216$M2R1[2700:3570]
TableS_5<-S216$M2R2[2700:3570]

TableDataS<-cbind(TableS_1,TableS_2,TableS_3,TableS_4,TableS_5)

#Aguathuna
AguathS_1<-S216$DEPT[3571:3917]
AguathS_2<-S216$SP[3571:3917]
AguathS_3<-S216$GR[3571:3917]
AguathS_4<-S216$M2R1[3571:3917]
AguathS_5<-S216$M2R2[3571:3917]

AguathDataS<-cbind(AguathS_1,AguathS_2,AguathS_3,AguathS_4,AguathS_5)

###############################################################
```

```
#CALCULATION OF MOMENTS
set.seed(1)
options(digits = 6)
Moments.gen<-function(x){
min=min(x)
max=max(x)
mean=mean(x)
med=median(x)
mod=getmode(x)
var=var(x)
sd=sqrt(var(x))
sk=skewness(x)
kur=kurtosis(x)
print(min)
print(max)
print(mean)
print(med)
print(mod)
print(var)
print(sd)
print(sk)
print(kur)
}

x<-c(2:5)
    for (j in x){
        k<-(GooseDataF[,j])
        print(Moments.gen(k))
        print("Next Formation")
        l<-(TableDataF[,j])
        print(Moments.gen(l))
        print("Next Formation")
        m<-(AguathDataF[,j])
        print(Moments.gen(m))
        print("Next Formation")
        n<-(GooseDataS[,j])
        print(Moments.gen(n))
        print("Next Formation")
        q<-(TableDataS[,j])
        print(Moments.gen(q))
        print("Next Formation")
        p<-(AguathDataS[,j])
        print(Moments.gen(p))
    }
```

```
################################################################
#DETRENDED DATA
#Transform the data into time series
#Formations of the Finnegan 311 hole section
#Goose (American) Tickle
#Spontaneous Potential
AAPL1 <-ts(Goose311_2,start(Goose311_1,0.2))
reg1 <- lm(AAP1L~time(AAPL1))
detrended<-as.numeric(AAPL1-predict.lm(reg1))
summary(req1)

#Gamma Ray
AAPL2<-ts(Goose311_3,start(Goose311_1,0.2))
reg2 <- lm(AAPL2~time(AAPL2)+ I(time(AAPL2)^2) + I(time(AAPL2)^3))
detrended1<-as.numeric(AAPL2-predict.lm(reg2))
summary(req2)

#Formations of the Finnegan 216 hole section
#Table Point
#Spontaneous Potential
AAPL3 <-ts(TableF_2,start(TableF_1,0.2))
reg3 <- lm(AAPL3~time(AAPL3))
detrended3<-as.numeric(AAPL3-predict.lm(reg3))
summary(req3)

#Gamma Ray
AAPL4 <-ts(TableF_3,start(TableF_1,0.2))
reg4 <- lm(AAPL4~time(AAPL4))
detrended4<-as.numeric(AAPL4-predict.lm(reg4))
summary(req4)
```

```
#Array Induction 20
AAPL9 <-ts(TableF_5,start(TableF_1,0.2))
reg9 <- lm(AAPL9~time(AAPL9))
detrended9<-as.numeric(AAPL9-predict.lm(reg9))
summary(req9)

#Aguathuna
#Spontaneous Potential
AAPL12 <-ts(AguathF_2,start(AguathF_1,0.2))
reg12 <- lm(AAPL12~time(AAPL12)+ I(time(AAPL12)^2))
detrended12<-as.numeric(AAPL12-predict.lm(reg12))
summary(req12)

#Gamma Ray
AAPL5 <-ts(AguathF_2,start(AguathF_1,0.2))
reg5 <- lm(AAPL5~time(AAPL5))
detrended5<-as.numeric(AAPL5-predict.lm(reg5))
summary(req5)

#Array Induction 10
AAPL10 <-ts(AguathF_4,start(AguathF_1,0.2))
reg10 <- lm(AAPL10~time(AAPL10))
detrended10<-as.numeric(AAPL10-predict.lm(reg10))
summary(req10)

#Array Induction 20
AAPL11 <-ts(AguathF_5,start(AguathF_1,0.2))
reg11 <- lm(AAPL11~time(AAPL11))
detrended11<-as.numeric(AAPL11-predict.lm(reg11))
summary(req11)

#Formations of the Seamus 216 hole section
#Goose (American) Tickle
#Gamma Ray
AAPL6 <-ts(Goose216_3,start(Goose216_1,0.1524))
reg6 <- lm(AAPL6~time(AAPL6))
detrended6<-as.numeric(AAPL6-predict.lm(reg6))
summary(req6)
```

```
#Table Point
AAPL7<-ts(TableS_3,start(TableS_1,0.1524))
reg7<-lm(AAPL7~time(AAPL7)+ I(time(AAPL7)^2) + I(time(AAPL7)^3))
detrended7<-as.numeric(AAPL7-predict.lm(reg7))
summary(req7)

#Aguathuna
AAPL8<-ts(AguathS_4,start(AguathS_1,0.1524))
reg8<-lm(AAPL8~time(AAPL8))
detrended8<-as.numeric(AAPL8-predict.lm(reg8))
summary(req8)
```

```
################################################################
#PLOT DISTRIBUTIONS

list.of.data.sets <- list(
  y1<-as.vector(detrended),
  y2<-as.vector(detrended1),
  y3<-as.vector(Goose311_4),
  y4<-as.vector(Goose311_5),
  y5<-as.vector(detrended3),
  y6<-as.vector(detrended4),
  y7<-as.vector(TableF_4),
  y8<-as.vector(TableF_5),
  y9<-as.vector(AguathF_2),
  y10<-as.vector(detrended5),
  y11<-as.vector(AguathF_4),
  y12<-as.vector(AguathF_5),
  y13<-as.vector(Goose216_2),
  y14<-as.vector(detrended6),
  y15<-as.vector(Goose216_4),
  y16<-as.vector(Goose216_5),
  y17<-as.vector(detrended7),
  y18<-as.vector(TableS_3),
  y19<-as.vector(TableS_4),
  y20<-as.vector(TableS_5),
  y21<-as.vector(detrended8),
  y22<-as.vector(AguathS_2),
  y23<-as.vector(AguathS_3),
  y24<-as.vector(AguathS_5)
  )


my.distr.function1<-function(neg.data){
  fg <- fitdist(y1,"norm")
  fm  <- fitdist(y1,"cauchy")
  fk <- fitdist(y1,"gumbel", start=list(a=100, b=100))
  f_list <- list(fg,fm,fk)
  plot.legend <- sapply(c(1:length(f_list)),
  function(x) f_list[[x]]$distname)
  f1 <- denscomp(f_list, legendtext = plot.legend,
  xlegend = "right",
  plotstyle = "ggplot",breaks=30)
  f2 <- qqcomp(f_list, legendtext = plot.legend,
  xlegend = "right",
  plotstyle = "ggplot")
  f3 <- cdfcomp(f_list, legendtext = plot.legend,
  xlegend = "right",
  plotstyle = "ggplot")
  f4 <- ppcomp(f_list, legendtext = plot.legend,
  xlegend = "right",
  plotstyle = "ggplot")
  grid.arrange(f1,f2,f3,f4)
  summary(fg,fm,fk)}

my.distr.function2<-function(pos.data){
  fg <- fitdist(y1,"norm")
  fl <- fitdist(y1,"weibull")
  fmm <-fitdist(y1,"gamma")
  f_list <- list(fg,fl,fmm)
  plot.legend <- sapply(c(1:length(f_list)),
  function(x) f_list[[x]]$distname)
  f1 <- denscomp(f_list, legendtext = plot.legend,
  xlegend = "right",
  plotstyle = "ggplot",breaks=30)
  f2 <- qqcomp(f_list, legendtext = plot.legend,
  xlegend = "right",
  plotstyle = "ggplot")
  f3 <- cdfcomp(f_list, legendtext = plot.legend,
  xlegend = "right",
  plotstyle = "ggplot")
  f4 <- ppcomp(f_list, legendtext = plot.legend,
  xlegend = "right",
  plotstyle = "ggplot")
  grid.arrange(f1,f2,f3,f4)
  summary(fg,fl,fmm)
}

for(i in 1:length(list.of.data.sets)){
 if(list.of.data.sets[[i]]>0) {
 my.distr.function2(pos.data=list.of.data.sets[[i]])
    } else if {list.of.data.sets[[i]]<0} {
 my.distr.function1(neg.data=list.of.data.sets[[i]]) }}

results.of.all.data.sets1 <- lapply(list.of.data.sets,
FUN=c(my.distr.function1)

results.of.all.data.sets2 <- lapply(list.of.data.sets,
FUN=c(my.distr.function2)
```

```
###############################################################
#EMPIRICAL VARIOGRAMS

##Finnegan 311###
#Goose(American) Tickle
#Spontaneous potential
Test = data.frame(DEPTH=Goose311_1, SP=detrended)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Goose",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=50,cressie=TRUE,width=2.5)
```

```
#Gamma ray
Test = data.frame(DEPTH=Goose311_1, SP=detrended1)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Goose",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=60,cressie=TRUE,width=4)

#Array Induction 10
Test = data.frame(DEPTH=Goose311_1, SP=Goose311_4)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Goose",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=16,cressie=TRUE,width=0.5)

#Array Induction 20
Test = data.frame(DEPTH=Goose311_1, SP=Goose311_5)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Goose",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=3,cressie=TRUE,width=0.2)

##Seamus 216##
#Goose (American) Tickle
#Spontaneous potential
Test = data.frame(DEPTH=Goose216_1, SP=Goose216_2)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Goose",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=50,cressie=TRUE,width=2)

#Gamma Ray
Test = data.frame(DEPTH=Goose216_1, SP=detrended6)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Goose",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=150,cressie=TRUE,width=8)
```

```
#Array Induction 10
Test = data.frame(DEPTH=Goose216_1, SP=Goose216_4)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Goose",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=50,cressie=TRUE,width=3)

#Array Induction 20
Test = data.frame(DEPTH=Goose216_1], SP=Goose216_5)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Goose",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=55,cressie=TRUE,width=2)


#Table Point

#Spontaneous Potential
Test = data.frame(DEPTH=TableS_1, SP=TableS_2)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Table Point",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=35,cressie=TRUE,width=1.5)


#Gamma Ray
Test = data.frame(DEPTH=TableS_1, SP=detrended7)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Table Point",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=20,cressie=TRUE,width=1)


#Array Induction 10
Test = data.frame(DEPTH=TableS_1, SP=TableS_4)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
```

```
k1 <- gstat(id="Table Point",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=50,cressie=TRUE,width=2)



#Array Induction 20
Test = data.frame(DEPTH=TableS_1, SP=TableS_5)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Table Point",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=15,cressie=TRUE,width=0.7)

#Aguathuna

#Spontaneous Potential
Test = data.frame(DEPTH=AguathS_1, SP=AguathS_2)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Aguathuna",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=15,cressie=TRUE,width=0.6)

#Array Induction 10
Test = data.frame(DEPTH=AguathS_1, SP=detrended8)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Aguathuna",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=11,cressie=TRUE,width=0.5)

#Array Induction 20
Test = data.frame(DEPTH=AguathS_1, SP=AguathS_5)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Aguathuna",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=30,cressie=TRUE,width=1)

##Finnegan 216
#Table Point
#Spontaneous Potential
```

```
Test = data.frame(DEPTH=TableF_1, SP=detrended3)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Table Point",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=10,cressie=TRUE,width=0.5)

#Gamma Ray
Test = data.frame(DEPTH=TableF_1, SP=detrended4)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Table Point",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=40,cressie=TRUE,width=1)

#Array Induction 20
Test = data.frame(DEPTH=TableF_1, SP=detrended9)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Table Point",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=20,cressie=TRUE,width=1)

#Aguathuna
#Spontaneous Potential
Test = data.frame(DEPTH=AguathF_1, SP=AguathF_2)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Aguathuna",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=16,cressie=TRUE,width=0.7)

#Gamma Ray
Test = data.frame(DEPTH=AguathF_1, SP=detrended5)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Aguathuna",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=15,cressie=TRUE,width=0.5)


#Array Induction 10
Test = data.frame(DEPTH=AguathF_1, SP=detrended10)
```

```
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Aguathuna",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=15,cressie=TRUE,width=0.5)

###A20
Test = data.frame(DEPTH=AguathF_1, SP=detrended11)
Test$y = matrix(1L, nrow = length(Test$DEPTH), ncol = 1)*50
coordinates(Test) = ~ DEPTH+y
k1 <- gstat(id="Aguathuna",formula = SP~1, data=Test)
vk1 = variogram(k1,cutoff=15,cressie=TRUE,width=0.5)
```

```
###############################################################
#THEORETICAL VARIOGRAMS

#Finnegan 311 hole section
#Goose (American) Tickle
#Spontaneous Potential
tested_par1=fit.variogram(vk1,model=vgm(55,"Sph",42,5 ))
tested_par2=fit.variogram(vk1,model=vgm(55,"Exp",42,5 ))
tested_par3=fit.variogram(vk1,model=vgm(55,"Pen",42,5 ))
tested_par5=fit.variogram(vk1,model=vgm(55,"Cir",42,5 ))

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk2$dist)),id="Exp"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk2$dist)),id="Pen"),
               cbind(variogramLine(tested_par5,maxdist =
               max(vk2$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+ geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size =15))+ scale_y_continuou
```

```
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FE2712",
"#008000","#100C08","#00FFFF","#FF00FF","#0000FF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle =element_text(size =
15),plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15))
+ xlim(0,50) + xlab("lag(m)")

#Gamma Ray

tested_par1=fit.variogram(vk1,model=vgm(45,"Sph",75,
nugget =25 ))
tested_par2=fit.variogram(vk1,model=vgm(45,"Gau",75,
nugget =25 ))
tested_par3=fit.variogram(vk1,model=vgm(45,"Pen",75,
nugget =25 ))
tested_par4=fit.variogram(vk1,model=vgm(45,"Cir",75,
nugget =25 ))

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
              max(vk1$dist)),id="Sph"),
              cbind(variogramLine(tested_par2,maxdist =
              max(vk2$dist)),id="Gau"),
              cbind(variogramLine(tested_par3,maxdist =
              max(vk2$dist)),id="Pen"),
              cbind(variogramLine(tested_par4,maxdist =
              max(vk2$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3) +
geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FE2712",
"#000000","#100C08","#00FFFF","#FF00FF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,60) +
xlab("lag(m)")


#Array Induction 10

tested_par1=fit.variogram(vk1,model=vgm(5,"Sph",2,5 ))
tested_par2=fit.variogram(vk1,model=vgm(5,"Gau",2,5 ))
tested_par3=fit.variogram(vk1,model=vgm(5,"Exp",2,5 ))
tested_par4=fit.variogram(vk1,model=vgm(5,"Pen",2,5 ))
tested_par6=fit.variogram(vk1,model=vgm(5,"Cir",2,5 ))


vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist=
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk2$dist)),id="Gau"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk2$dist)),id="Exp"),
               cbind(variogramLine(tested_par4,maxdist =
               max(vk2$dist)),id="Pen"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#7CFC00","#000000","#00FFFF","#0000FF","#FF00FF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme(legend.text=element_text(size=15)) + xlim(0,16) +
xlab("lag(m)")


#Array Induction 20

tested_par1=fit.variogram(vk1,model=vgm(50,"Sph",1,0))
tested_par2=fit.variogram(vk1,model=vgm(50,"Gau",1,0))
tested_par3=fit.variogram(vk1,model=vgm(50,"Pen",1,0))
tested_par5=fit.variogram(vk1,model=vgm(50,"Cir",1,0))

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk2$dist)),id="Gau"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk2$dist)),id="Pen"),
               cbind(variogramLine(tested_par5,maxdist =
               max(vk2$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#7CFC00","#000000","#00FFFF","#0000FF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,3.5) +
xlab("lag(m)")

#Seamus 216
```

```
#Goose (American) Tickle
#Spontaneous Potential

tested_par1=fit.variogram(vk1,model=vgm(200,"Sph",30,10))
tested_par2=fit.variogram(vk1,model=vgm(200,"Pen",30,10))
tested_par3=fit.variogram(vk1,model=vgm(200,"Cir",30,10))

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk2$dist)),id="Pen"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk2$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#000000","#7CFC00","blue")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,50) +
xlab("lag(m)")

#Gamma Ray

tested_par1=fit.variogram(vk1,model=vgm(200,"Sph",40,0))
tested_par2=fit.variogram(vk1,model=vgm(200,"Gau",40,0))
tested_par3=fit.variogram(vk1,model=vgm(200,"Exp",40,0))
tested_par4=fit.variogram(vk1,model=vgm(200,"Pen",40,0))
```

```
vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
              max(vk1$dist)),id="Sph"),
              cbind(variogramLine(tested_par2,maxdist =
              max(vk2$dist)),id="Gau"),
              cbind(variogramLine(tested_par3,maxdist =
              max(vk2$dist)),id="Exp"),
              cbind(variogramLine(tested_par4,maxdist =
              max(vk2$dist)),id="Pen"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#7CFC00","#000000","#00FFFF","#0000FF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme(legend.text=element_text(size=15)) + xlim(0,105) +
xlab("lag(m)")

#Array Induction 10

tested_par1=fit.variogram(vk1,model=vgm(2000,"Sph",10,0))
tested_par2=fit.variogram(vk1,model=vgm(2000,"Exp",10,0))
tested_par3=fit.variogram(vk1,model=vgm(2000,"Pen",10,0))
tested_par5=fit.variogram(vk1,model=vgm(2000,"Cir",10,0))

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
              max(vk1$dist)),id="Sph"),
              cbind(variogramLine(tested_par2,maxdist =
              max(vk2$dist)),id="Exp"),
              cbind(variogramLine(tested_par3,maxdist =
              max(vk2$dist)),id="Pen"),
                cbind(variogramLine(tested_par5,maxdist =
                max(vk2$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#7CFC00","#000000","#00FFFF","#0000FF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,50) +
xlab("lag(m)")

#Array Induction 20

tested_par1=fit.variogram(vk1,model=vgm(1500,"Sph",20,0))
tested_par2=fit.variogram(vk1,model=vgm(1500,"Gau",20,0))
tested_par3=fit.variogram(vk1,model=vgm(1500,"Exp",20,0))
tested_par4=fit.variogram(vk1,model=vgm(1500,"Pen",20,0))
tested_par6=fit.variogram(vk1,model=vgm(1500,"Cir",20,0))

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
                max(vk1$dist)),id="Sph"),
                cbind(variogramLine(tested_par2,maxdist =
                max(vk2$dist)),id="Gau"),
                cbind(variogramLine(tested_par3,maxdist =
                max(vk2$dist)),id="Exp"),
                cbind(variogramLine(tested_par4,maxdist =
                max(vk2$dist)),id="Pen"),
                cbind(variogramLine(tested_par6,maxdist =
                max(vk2$dist)),id="Cir"))
```

```
ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#7CFC00","#0000FF","#000000","#00FFFF","#FF00FF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,55) +
xlab("lag(m)")



#Table Point
#Spontaneous Potential

tested_par2=fit.variogram(vk1,model=vgm(300,"Gau",10,20),
fit.method = 7)

vgLine1<-rbind(cbind(variogramLine(tested_par2,maxdist =
               max(vk2$dist)),id="Gau"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#0000FF","#000000")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
```

```
+ theme( legend.text=element_text(size=15)) + xlim(0,35) +
xlab("lag(m)")

#Gamma Ray

tested_par1=fit.variogram(vk1,model=vgm(200,"Sph",2,0),
fit.method = 1)
tested_par4=fit.variogram(vk1,model=vgm(200,"Pen",2,0),
fit.method = 1)
tested_par5=fit.variogram(vk1,model=vgm(200,"Gau",2,0),
fit.method = 1)
tested_par6=fit.variogram(vk1,model=vgm(200,"Cir",2,0),
fit.method = 1)

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par4,maxdist =
               max(vk1$dist)),id="Pen"),
               cbind(variogramLine(tested_par5,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par6,maxdist =
               max(vk2$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#7CFC00","#0000FF","#FFA500","#000000")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,20) +
xlab("lag(m)")
```

```
#Array Induction 10

tested_par1=fit.variogram(vk1,model=vgm(1000,"Sph",20,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(1000,"Pen",20,0),
fit.method = 1)
tested_par5=fit.variogram(vk1,model=vgm(1000,"Cir",20,0),
fit.method = 1)

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk1$dist)),id="Pen"),
               cbind(variogramLine(tested_par5,maxdist =
               max(vk1$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#0000FF","#FFA500","#000000")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,50) +
xlab("lag(m)")

#Array Induction 20

tested_par1=fit.variogram(vk1,model=vgm(200000,"Sph",3,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(200000,"Gau",3,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(200000,"Exp",3,0),
fit.method = 1)
tested_par4=fit.variogram(vk1,model=vgm(200000,"Pen",3,0),
fit.method = 1)


vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk1$dist)),id="Exp"),
               cbind(variogramLine(tested_par4,maxdist =
               max(vk2$dist)),id="Pen"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#0000FF","#FFA500","#00FFFF","#000000")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,15) +
xlab("lag(m)")


#Aguathuna
#Spontaneous Potential

tested_par1=fit.variogram(vk1,model=vgm(1000,"Gau",15,20),
fit.method = 1)
```

```
vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Gau"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#000000",
"#FF0000")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic")) + theme(
legend.text=element_text(size=15)) + xlim(0,15) +
xlab("lag(m)")


#Array Induction 10

tested_par1=fit.variogram(vk1,model=vgm(550,"Gau",1,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(550,"Sph",1,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(550,"Pen",1,0),
fit.method = 1)
tested_par4=fit.variogram(vk1,model=vgm(550,"Cir",1,0),
fit.method = 1)

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk1$dist)),id="Pen"),
               cbind(variogramLine(tested_par4,maxdist =
               max(vk2$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#0000FF","#000000","#FFA500","#00FFFF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic")) + theme(
legend.text=element_text(size=15)) + xlim(0,12) +
xlab("lag(m)")

#Array Induction 20

tested_par1=fit.variogram(vk1,model=vgm(300000,"Sph",2,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(300000,"Gau",2,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(300000,"Exp",2,0),
fit.method = 1)
tested_par4=fit.variogram(vk1,model=vgm(300000,"Pen",2,0),
fit.method = 1)
tested_par5=fit.variogram(vk1,model=vgm(300000,"Cir",2,0),
fit.method = 1)

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk1$dist)),id="Exp"),
               cbind(variogramLine(tested_par4,maxdist =
               max(vk2$dist)),id="Pen"),
               cbind(variogramLine(tested_par4,maxdist =
               max(vk2$dist)),id="Cir"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c("#FF0000",
"#0000FF","#FFA500","#000000","#00FFFF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic")) + theme(
legend.text=element_text(size=15)) + xlim(0,30) +
xlab("lag(m)")

#Finnegan 216
#Table Point
#Spontaneous Potential

tested_par1=fit.variogram(vk1,model=vgm(150,"Gau",5,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(150,"Sph",5,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(150,"Cir",5,0),
fit.method = 1)
tested_par4=fit.variogram(vk1,model=vgm(150,"Pen",5,0),
fit.method = 1)

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk1$dist)),id="Sph"),
                cbind(variogramLine(tested_par3,maxdist =
                max(vk1$dist)),id="Cir"),
                cbind(variogramLine(tested_par4,maxdist =
                max(vk2$dist)),id="Pen"))


ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c(
"#0000FF","#FFA500","#FF0000","#00FFFF","#000000")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15)
, plot.caption = element_text(size = 12,face="italic")) +
theme( legend.text=element_text(size=15)) + xlim(0,10) +
xlab("lag(m)")

#Gamma Ray
tested_par1=fit.variogram(vk1,model=vgm(25,"Gau",20,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(25,"Sph",20,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(25,"Exp",20,0),
fit.method = 1)
tested_par4=fit.variogram(vk1,model=vgm(25,"Pen",20,0),
fit.method = 1)


vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
                max(vk1$dist)),id="Gau"),
                cbind(variogramLine(tested_par2,maxdist =
                max(vk1$dist)),id="Sph"),
                cbind(variogramLine(tested_par3,maxdist =
                max(vk1$dist)),id="Exp"),
                cbind(variogramLine(tested_par4,maxdist =
                max(vk2$dist)),id="Pen"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c(
"#0000FF","#FFA500","#FF0000","#00FFFF","#000000")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,40) +
xlab("lag(m)")

#Array Induction 10

tested_par1=fit.variogram(vk1,model=vgm(10000,"Gau",1,7000))
tested_par2=fit.variogram(vk1,model=vgm(10000,"Exp",1,7000))

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
                max(vk1$dist)),id="Gau"),
                cbind(variogramLine(tested_par2,maxdist =
                max(vk1$dist)),id="Exp"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c( "#000000","#FFA500"))
+ theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme(legend.text=element_text(size=15)) + xlim(0,20) +
xlab("lag(m)")



#Aguathuna
#Spontaneous Potential

tested_par1=fit.variogram(vk1,model=vgm(50,"Gau",5,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(50,"Sph",5,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(50,"Pen",5,0),
fit.method = 1)

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk2$dist)),id="Pen"))

ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c(
"#000000","#FFA500","#FF0000","#00FFFF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
```

```
+ theme(legend.text=element_text(size=15)) + xlim(0,15) +
xlab("lag(m)")

#Gamma Ray

tested_par1=fit.variogram(vk1,model=vgm(100,"Sph",5,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(100,"Gau",5,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(100,"Exp",5,0),
fit.method = 1)



vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk2$dist)),id="Exp"))



ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c(
"#000000","#FFA500","#FF0000","#00FFFF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic")) +
theme( legend.text=element_text(size=15)) + xlim(0,15) +
xlab("lag(m)")

#Array Induction 10
```

```
tested_par1=fit.variogram(vk1,model=vgm(300000,"Sph",5,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(300000,"Gau",5,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(300000,"Exp",5,0),
fit.method = 1)

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk2$dist)),id="Exp"))



ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c(
"#000000","#FFA500","#FF0000","#00FFFF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme(legend.text=element_text(size=15)) + xlim(0,15) +
xlab("lag(m)")



#Array Induction 10

tested_par1=fit.variogram(vk1,model=vgm(150000,"Sph",5,0),
fit.method = 1)
tested_par2=fit.variogram(vk1,model=vgm(150000,"Gau",5,0),
fit.method = 1)
tested_par3=fit.variogram(vk1,model=vgm(150000,"Exp",5,0),
fit.method = 1)

vgLine1<-rbind(cbind(variogramLine(tested_par1,maxdist =
               max(vk1$dist)),id="Sph"),
               cbind(variogramLine(tested_par2,maxdist =
               max(vk1$dist)),id="Gau"),
               cbind(variogramLine(tested_par3,maxdist =
               max(vk2$dist)),id="Exp"))



ggplot(vk1, aes(x=dist,y=gamma,colour=id)) +
geom_line(data=vgLine1,size=0.8) + geom_point(size=3)
+geom_line()  + theme_light()+ theme(axis.title.x =
element_text(size=15),axis.title.y =
element_text(size=15))+theme(axis.text.x =
element_text(size=15),axis.text.y = element_text(size = 15)) +
scale_y_continuous(name="$\gamma(h)$") +
scale_x_continuous(name="Distance (h)") +
scale_color_manual(values = c(
"#000000","#FFA500","#FF0000","#00FFFF")) +
theme(legend.title=element_blank())+theme(plot.title =
element_text(size=15), plot.subtitle = element_text(size = 15),
plot.caption = element_text(size = 12,face="italic"))
+ theme( legend.text=element_text(size=15)) + xlim(0,15) +
xlab("lag(m)")




#Calculate Validation Scores
#Experimental variogram
exp.var<-vk1$gamma
#Estimated variogram
est.var<-variogramLine(tested_par1,maxdist =
                       max(vk1$dist),n=nrow(vk1))
```

```
my.val.scores <- function(exp.var,est.var){
d = exp.var[[i]]-est.var[[i]]$gamma)
mse = mean((d)^2)
mae = mean(abs(d))
rmse = sqrt(mse)
}

#Print the following code for each exp.var and est.var
print<-my.val.scores(exp.var,est.var)

##############################################################
#MATLAB
##############################################################
#CORRELATIONS

clc; clear variable; close all;
load(’Finnegan311.dat’)
load(’Finnegan216.dat’)
load(’Seamus216.dat’)

% Table Point

DEF=DEPT(180:905); %Depth of Finegan
GRF=GR(180:905);
DES=DEPT1(2700:3571); %Depth of Seamus
GRS=GR1(2700:3571);
save TEMP DES GRF GRS DEF
cle
load TEMP
DEF=DEF-min(DEF);
DES=DES-min(DES);
M=max(DES);
find(DEF<=M+0.5 & DEF>=M-0.5)
Cutoff=663;
StepDEF=DEF(2)-DEF(1);
DEKOINO=(0:StepDEF:DEF(Cutoff))’;
plot(DEKOINO, GRF(1:Cutoff))
GRS2=interp1(DES,GRS,DEKOINO); %Linear
```

```
GRS3=interp1(DES,GRS,DEKOINO,'nearest');
GRS4=interp1(DES,GRS,DEKOINO,'cubic');
GRS5=interp1(DES,GRS,DEKOINO,'spline');
%corrcoef(GRS2,GRS3)
% corrcoef(GRS4,GRS3)
% corrcoef(GRS5,GRS3)
figure
hold on
plot(DEKOINO,GRS2,'m','LineWidth',1.5)
plot(DEKOINO,GRS3,'c','LineWidth',1.5)
plot(DEKOINO,GRS4,'g:','LineWidth',1.5)
plot(DEKOINO,GRS5,'r:','LineWidth',1.5)
plot(DEKOINO,GRF(1:Cutoff),'k','LineWidth',1.5)
legend('Linear','Nearest','Cubic','Spline','GRF')
RS=corr(GRS3,GRF(1:Cutoff),'type','Spearman');
[RP,P]=corr(GRS3,GRF(1:Cutoff),'type','Pearson');
rmse=sqrt(mean((GRS3,GRF(1:Cutoff)).^2))

% Aguathuna

DEF=DEPT(906:1155); %Depth of Finnegan
GRF=GR(906:1155);
DES=DEPT1(3572:3917); %Depth of Seamus
GRS=GR1(3572:3917);
save TEMP DES GRF GRS DEF
cle
load TEMP
DEF=DEF-min(DEF);
DES=DES-min(DES);
M=max(DEF);
find(DES<=M+0.5 & DES>=M-0.5)
Cutoff=325;
StepDEF=DES(2)-DES(1);
DEKOINO=(0:StepDEF:DES(Cutoff-1))';
plot(DEKOINO, GRS(1:Cutoff-1))
GRS2=interp1(DEF,GRF,DEKOINO); %Linear
GRS3=interp1(DEF,GRF,DEKOINO,'nearest');
GRS4=interp1(DEF,GRF,DEKOINO,'cubic');
```

```
GRS5=interp1(DEF,GRF,DEKOINO,'spline');
% corrcoef(GRS2,GRS3)
% corrcoef(GRS4,GRS3)
% corrcoef(GRS5,GRS3)
figure
hold on
plot(DEKOINO,GRS2,'m','LineWidth',1.5)
plot(DEKOINO,GRS3,'c','LineWidth',1.5)
plot(DEKOINO,GRS4,'g:','LineWidth',1.5)
plot(DEKOINO,GRS5,'r:','LineWidth',1.5)
plot(DEKOINO, GRS(1:Cutoff-1),'k','LineWidth',1.5)
legend('Linear','Nearest','Cubic','Spline','GRF')
RS=corr(GRS3,GRF(1:Cutoff),'type','Spearman');
[RP,P]=corr(GRS3,GRF(1:Cutoff),'type','Pearson');
rmse=sqrt(mean((GRS3,GRF(1:Cutoff)).^2))

%Catoche

DEF=DEPT(1156:1780); %Depth of Finegan
GRF=GR(1156:1780);
DES=DEPT1(3918:4639); %Depth of Seamus
GRS=GR1(3918:4639);
save TEMP DES GRF GRS DEF
cle
load TEMP
DEF=DEF-min(DEF);
DES=DES-min(DES);
M=max(DES);
find(DEF<=M+0.5 & DEF>=M-0.5)
Cutoff=548;
StepDEF=DEF(2)-DEF(1);
DEKOINO=(0:StepDEF:DEF(Cutoff))';
plot(DEKOINO, GRF(1:Cutoff-1))
GRS2=interp1(DES,GRS,DEKOINO); %Linear
GRS3=interp1(DES,GRS,DEKOINO,'nearest');
GRS4=interp1(DES,GRS,DEKOINO,'cubic');
GRS5=interp1(DES,GRS,DEKOINO,'spline');
%corrcoef(GRS2,GRS3)
```

```
%corrcoef(GRS4,GRS3)
%corrcoef(GRS5,GRS3)
figure
hold on
plot(DEKOINO,GRS2,'m','LineWidth',1.5)
plot(DEKOINO,GRS3,'c','LineWidth',1.5)
plot(DEKOINO,GRS4,'g:','LineWidth',1.5)
plot(DEKOINO,GRS5,'r:','LineWidth',1.5)
plot(DEKOINO,GRF(1:Cutoff-1),'k','LineWidth',1.5)
legend('Linear','Nearest','Cubic','Spline','GRF')
RS=corr(GRS3,GRF(1:Cutoff-1),'type','Spearman');
[RP,P]=corr(GRS3,GRF(1:Cutoff-1),'type','Pearson');
rmse=sqrt(mean((GRS3,GRF(1:Cutoff-1)).^2))

%Boat Harbour

DEF=DEPT(1781:2380); %Depth of Finegan
GRF=GR(1781:2380);
DES=DEPT1(4640:5459); %Depth of Seamus
GRS=GR1(4640:5459);
save TEMP DES GRF GRS DEF
cle
load TEMP
DEF=DEF-min(DEF);
DES=DES-min(DES);
M=max(DEF);
find(DES<=M+0.5 & DES>=M-0.5)
Cutoff=784;
StepDEF=DES(2)-DES(1);
DEKOINO=(0:StepDEF:DES(Cutoff))';
plot(DEKOINO, GRF(1:Cutoff-1))
GRS2=interp1(DEF,GRF,DEKOINO); %Linear
GRS3=interp1(DEF,GRF,DEKOINO,'nearest');
GRS4=interp1(DEF,GRF,DEKOINO,'cubic');
GRS5=interp1(DEF,GRF,DEKOINO,'spline');
%corrcoef(GRS2,GRS3)
%corrcoef(GRS4,GRS3)
%corrcoef(GRS5,GRS3)
```

```
figure
hold on
plot(DEKOINO,GRS2,'m','LineWidth',1.5)
plot(DEKOINO,GRS3,'c','LineWidth',1.5)
plot(DEKOINO,GRS4,'g:','LineWidth',1.5)
plot(DEKOINO,GRS5,'r:','LineWidth',1.5)
plot(DEKOINO,GRS(1:Cutoff),'k','LineWidth',1.5)
legend('Linear','Nearest','Cubic','Spline','GRF')
RS=corr(GRS3,GRF(1:Cutoff),'type','Spearman');
[RP,P]=corr(GRS3,GRF(1:Cutoff),'type','Pearson');
rmse=sqrt(mean((GRS3,GRF(1:Cutoff-1)).^2))

% Watts Bight

DEF=DEPT(2381:2730); %Depth of Finegan
GRF=GR(2381:2730);
DES=DEPT1(5460:5866); %Depth of Seamus
GRS=GR1(5460:5866);
save TEMP DES GRF GRS DEF
cle
load TEMP
DEF=DEF-min(DEF);
DES=DES-min(DES);
M=max(DES);
find(DEF<=M+0.5 & DEF>=M-0.5)
Cutoff=308;
StepDEF=DEF(2)-DEF(1);
DEKOINO=(0:StepDEF:DEF(Cutoff))';
plot(DEKOINO, GRF(1:Cutoff-1))
GRS
GRS2=interp1(DES,GRS,DEKOINO); %Linear
GRS3=interp1(DES,GRS,DEKOINO,'nearest');
GRS4=interp1(DES,GRS,DEKOINO,'cubic');
GRS5=interp1(DES,GRS,DEKOINO,'spline');
% corrcoef(GRS2,GRS3)
% corrcoef(GRS4,GRS3)
% corrcoef(GRS5,GRS3)
figure
```

```
hold on
plot(DEKOINO,GRS2,'m','LineWidth',1.5)
plot(DEKOINO,GRS3,'c','LineWidth',1.5)
plot(DEKOINO,GRS4,'g:','LineWidth',1.5)
plot(DEKOINO,GRS5,'r:','LineWidth',1.5)
plot(DEKOINO,GRF(1:Cutoff-1),'k','LineWidth',1.5)
legend('Linear','Nearest','Cubic','Spline','GRF')
RS=corr(GRS3,GRF(1:Cutoff-1),'type','Spearman');
[RP,P]=corr(GRS3,GRF(1:Cutoff-1),'type','Pearson');
rmse=sqrt(mean((GRS3-GRF(1:Cutoff)).^2))

% Goose (American) Tickle

DEF=DEPT2(7087:8509); %Depth of Finegan
GRF=GR2(7087:8509);
DES=DEPT1(13:2200); %Depth of Seamus
GRS=GR1(13:2200);
save TEMP DES GRF GRS DEF
cle
load TEMP
DEF=DEF-min(DEF);
DES=DES-min(DES);
M=max(DEF);
find(DES<=M+0.5 & DES>=M-0.5)
Cutoff=1864;
StepDEF=DES(2)-DES(1);
DEKOINO=(0:StepDEF:DES(Cutoff))';
GRF
plot(DEKOINO, GRS(1:Cutoff))
GRS
GRS2=interp1(DEF,GRF,DEKOINO); %Linear
GRS3=interp1(DEF,GRF,DEKOINO,'nearest');
GRS4=interp1(DEF,GRF,DEKOINO,'cubic');
GRS5=interp1(DEF,GRF,DEKOINO,'spline');
% corrcoef(GRS2,GRS3)
% corrcoef(GRS4,GRS3)
% corrcoef(GRS5,GRS3)
figure
```

```
hold on
plot(DEKOINO,GRS2,’m’,’LineWidth’,1.5)
plot(DEKOINO,GRS3,’c’,’LineWidth’,1.5)
plot(DEKOINO,GRS4,’g:’,’LineWidth’,1.5)
plot(DEKOINO,GRS5,’r:’,’LineWidth’,1.5)
plot(DEKOINO,GRS(1:Cutoff),’k’,’LineWidth’,1.5)
legend(’Linear’,’Nearest’,’Cubic’,’Spline’,’GRF’)
RS=corr(GRS3,GRF(1:Cutoff-1),’type’,’Spearman’);
[RP,P]=corr(GRS3,GRF(1:Cutoff-1),’type’,’Pearson’);
rmse=sqrt(mean((GRS3-GRF(1:Cutoff)).^2))
```

B.2 Missing data

The missing values generator and MCAR test is achieved by using the [73]’s code, as well as the *Amelia*, *imputeTS* and *forecast* packages in *R* .

```
#Table Point
#Spontaneous Potential

  Timeseries1 <- ts(TableS_2,start(TableS_1,0.1524),frequency = 6)
  plot(stl(Timeseries1,s.window = c("periodic")),
  main="Seamus 216 SP Data Decomposition")
  acf(Timeseries1,main="")
  complete.ts <-Timeseries1
  seeds <- 30
  n <- length(complete.ts)
  miss.rate <- c(0.1,0.25,0.5,0.8)
  incomplete.ts <- array(,dim=c(n,seeds,length(miss.rate)))
  NAs <- array(,dim=c(seeds,length(miss.rate)))
  Impute1<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute2<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute3<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute4<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute5<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute6<- array(,dim=c(n,seeds,length(miss.rate)))
```

```
for (c in 1:length(miss.rate)){
  for (i in 1:seeds){
    set.seed(i)
    incomplete.ts[,i,c] <-ts(miss.gen(complete.ts,miss.rate[c]))
    b=0
    for (a in 1:length(complete.ts)){if(is.na(incomplete.ts[a,i,c]))
    {b=b+1}
    NAs[i,c] <-b}
    incomp.ts <- as.numeric(incomplete.ts[,i,c])
    Impute1[,i,c] <- na.kalman(incomp.ts,model="auto.arima")
    Impute2[,i,c] <- na.interpolation(incomp.ts)
    Impute3[,i,c] <- na.interpolation(incomp.ts,option="spline")
    Impute4[,i,c] <- na.ma(incomp.ts,weighting="simple")
    Impute5[,i,c] <- na.ma(incomp.ts,weighting="linear")
    Impute6[,i,c] <- na.mean(incomp.ts)
  }}
  inc <- as.data.frame(incomplete.ts[,5,])
  names(inc) <- c("0.1","0.25","0.5","0.8")
  missmap(inc,rank.order = TRUE)


#Gamma Ray

  Timeseries2 <- ts(TableS_3,start(TableS_1,0.1524),frequency = 6)
  plot(stl(Timeseries2,s.window = c("periodic")),
  main="Seamus 216 GR Data Decomposition")
  acf(Timeseries2,main="")
  complete.ts <-Timeseries2
  seeds <- 30
  n <- length(complete.ts)
  miss.rate <- c(0.1,0.25,0.5,0.8)
  incomplete.ts <- array(,dim=c(n,seeds,length(miss.rate)))
  NAs <- array(,dim=c(seeds,length(miss.rate)))
  Impute1<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute2<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute3<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute4<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute5<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute6<- array(,dim=c(n,seeds,length(miss.rate)))

  for (c in 1:length(miss.rate)){
    for (i in 1:seeds){
      set.seed(i)
      incomplete.ts[,i,c] <-ts(miss.gen(complete.ts,miss.rate[c]))
      b=0
      for (a in 1:length(complete.ts)){if(is.na(incomplete.ts[a,i,c]))
      {b=b+1}
      NAs[i,c] <-b }
      incomp.ts <- as.numeric(incomplete.ts[,i,c])
      Impute1[,i,c] <- na.kalman(incomp.ts,model="auto.arima")
      Impute2[,i,c] <- na.interpolation(incomp.ts)
      Impute3[,i,c] <- na.interpolation(incomp.ts,option="spline")
      Impute4[,i,c] <- na.ma(incomp.ts,weighting="simple")
      Impute5[,i,c] <- na.ma(incomp.ts,weighting="linear")
      Impute6[,i,c] <- na.mean(incomp.ts)
    }}
  inc <- as.data.frame(incomplete.ts[,5,])
  names(inc) <- c("0.1","0.25","0.5","0.8")
  missmap(inc)


#Array Induction 10
  Timeseries3 <- ts(TableS_3,start(TableS_1,0.1524),frequency = 6)
  plot(stl(Timeseries3,s.window = c("periodic")),
  main="Seamus 216 AT10 Data Decomposition")
  acf(Timeseries3,main="")
  complete.ts <-Timeseries3
  seeds <- 30
  n <- length(complete.ts)
  miss.rate <- c(0.1,0.25,0.5,0.8)
  incomplete.ts <- array(,dim=c(n,seeds,length(miss.rate)))
  NAs <- array(,dim=c(seeds,length(miss.rate)))
  Impute1<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute2<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute3<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute4<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute5<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute6<- array(,dim=c(n,seeds,length(miss.rate)))

  for (c in 1:length(miss.rate)){
    for (i in 1:seeds){
      set.seed(i)
      incomplete.ts[,i,c] <-ts(miss.gen(complete.ts,miss.rate[c]))
      b=0
      for (a in 1:length(complete.ts)){if(is.na(incomplete.ts[a,i,c]))
      {b=b+1}
      NAs[i,c] <-b }
      incomp.ts <- as.numeric(incomplete.ts[,i,c])
      Impute1[,i,c] <- na.kalman(incomp.ts,model="auto.arima")
      Impute2[,i,c] <- na.interpolation(incomp.ts)
      Impute3[,i,c] <- na.interpolation(incomp.ts,option="spline")
      Impute4[,i,c] <- na.ma(incomp.ts,weighting="simple")
      Impute5[,i,c] <- na.ma(incomp.ts,weighting="linear")
      Impute6[,i,c] <- na.mean(incomp.ts)
    }}

  inc <- as.data.frame(incomplete.ts[,5,])
  names(inc) <- c("0.1","0.25","0.5","0.8")
  missmap(inc)

#Array Induction 20

  Timeseries4 <- ts(TableS_5,start(TableS_1,0.1524),frequency = 6)
  plot(stl(Timeseries4,s.window = c("periodic")),
  main="Seamus 216  AT20 Data Decomposition")
  acf(Timeseries4,main="")
  complete.ts <-Timeseries4
  seeds <- 30
  n <- length(complete.ts)
  miss.rate <- c(0.1,0.25,0.5,0.8)
  incomplete.ts <- array(,dim=c(n,seeds,length(miss.rate)))
  NAs <- array(,dim=c(seeds,length(miss.rate)))
  Impute1<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute2<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute3<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute4<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute5<- array(,dim=c(n,seeds,length(miss.rate)))
  Impute6<- array(,dim=c(n,seeds,length(miss.rate)))

  for (c in 1:length(miss.rate)){
    for (i in 1:seeds){
      set.seed(i)
      incomplete.ts[,i,c] <-ts(miss.gen(complete.ts,miss.rate[c]))
      b=0
      for (a in 1:length(complete.ts)){if(is.na(incomplete.ts[a,i,c]))
      {b=b+1}
      NAs[i,c] <-b }
      incomp.ts <- as.numeric(incomplete.ts[,i,c])
      Impute1[,i,c] <- na.kalman(incomp.ts,model="auto.arima")
      Impute2[,i,c] <- na.interpolation(incomp.ts)
      Impute3[,i,c] <- na.interpolation(incomp.ts,option="spline")
      Impute4[,i,c] <- na.ma(incomp.ts,weighting="simple")
      Impute5[,i,c] <- na.ma(incomp.ts,weighting="linear")
      Impute6[,i,c] <- na.mean(incomp.ts)
    }}

  inc <- as.data.frame(incomplete.ts[,5,])
  names(inc) <- c("0.1","0.25","0.5","0.8")
  missmap(inc)

#Calculated errors of simulated data

MRSE1<- array(,dim=c(seeds,length(miss.rate)))
MRSE2<- array(,dim=c(seeds,length(miss.rate)))
MRSE3<- array(,dim=c(seeds,length(miss.rate)))
MRSE4<- array(,dim=c(seeds,length(miss.rate)))
MRSE5<- array(,dim=c(seeds,length(miss.rate)))
MRSE6<- array(,dim=c(seeds,length(miss.rate)))


MAPE1<- array(,dim=c(seeds,length(miss.rate)))
MAPE2<- array(,dim=c(seeds,length(miss.rate)))
```

```
MAPE3<- array(,dim=c(seeds,length(miss.rate)))
MAPE4<- array(,dim=c(seeds,length(miss.rate)))
MAPE5<- array(,dim=c(seeds,length(miss.rate)))
MAPE6<- array(,dim=c(seeds,length(miss.rate)))

for(l in 1:length(miss.rate)){
  for(f in 1:seeds){
    MRSE1[f,l]<-sqrt((sum((Impute1[,f,l]-complete.ts[,f,l])^2)
    )/NAs[f,l])
    MRSE2[f,l]<-sqrt((sum((Impute2[,f,l]-complete.ts[,f,l])^2)
    )/NAs[f,l])
    MRSE3[f,l]<-sqrt((sum((Impute3[,f,l]-complete.ts[,f,l])^2)
    )/NAs[f,l])
    MRSE4[f,l]<-sqrt((sum((Impute4[,f,l]-complete.ts[,f,l])^2)
    )/NAs[f,l])
    MRSE5[f,l]<-sqrt((sum((Impute5[,f,l]-complete.ts[,f,l])^2)
    )/NAs[f,l])
    MRSE6[f,l]<-sqrt((sum((Impute6[,f,l]-complete.ts[,f,l])^2)
    )/NAs[f,l])

    MAPE1[f,l]<-(100/NAs[f,l])*(sum(abs((Impute1[,f,l]
    -complete.ts[,f,l])/complete.ts)))
    MAPE2[f,l]<-(100/NAs[f,l])*(sum(abs((Impute2[,f,l]
    -complete.ts[,f,l])/complete.ts)))
    MAPE3[f,l]<-(100/NAs[f,l])*(sum(abs((Impute3[,f,l]
    -complete.ts[,f,l])/complete.ts)))
    MAPE4[f,l]<-(100/NAs[f,l])*(sum(abs((Impute4[,f,l]
    -complete.ts[,f,l])/complete.ts)))
    MAPE5[f,l]<-(100/NAs[f,l])*(sum(abs((Impute5[,f,l]
    -complete.ts[,f,l])/complete.ts)))
    MAPE6[f,l]<-(100/NAs[f,l])*(sum(abs((Impute6[,f,l]
    -complete.ts[,f,l])/complete.ts)))
}}


#Calculated errors of original data

MRSE1<- array(,dim=c(seeds,length(miss.rate)))
```

```
MRSE2<- array(,dim=c(seeds,length(miss.rate)))
MRSE3<- array(,dim=c(seeds,length(miss.rate)))
MRSE4<- array(,dim=c(seeds,length(miss.rate)))
MRSE5<- array(,dim=c(seeds,length(miss.rate)))
MRSE6<- array(,dim=c(seeds,length(miss.rate)))

MAPE1<- array(,dim=c(seeds,length(miss.rate)))
MAPE2<- array(,dim=c(seeds,length(miss.rate)))
MAPE3<- array(,dim=c(seeds,length(miss.rate)))
MAPE4<- array(,dim=c(seeds,length(miss.rate)))
MAPE5<- array(,dim=c(seeds,length(miss.rate)))
MAPE6<- array(,dim=c(seeds,length(miss.rate)))


for(l in 1:length(miss.rate)){
  for(f in 1:seeds){
    MRSE1[f,l]<-sqrt((sum((Impute1[,f,l]-complete.ts)^2))/
    NAs[f,l])
    MRSE2[f,l]<-sqrt((sum((Impute2[,f,l]-complete.ts)^2))/
    NAs[f,l])
    MRSE3[f,l]<-sqrt((sum((Impute3[,f,l]-complete.ts)^2))/
    NAs[f,l])
    MRSE4[f,l]<-sqrt((sum((Impute4[,f,l]-complete.ts)^2))/
    NAs[f,l])
    MRSE5[f,l]<-sqrt((sum((Impute5[,f,l]-complete.ts)^2))/
    NAs[f,l])
    MRSE6[f,l]<-sqrt((sum((Impute6[,f,l]-complete.ts)^2))/
    NAs[f,l])

    MAPE1[f,l]<-(100/NAs[f,l])*(sum(abs((Impute1[,f,l]
    -complete.ts)/complete.ts)))
    MAPE2[f,l]<-(100/NAs[f,l])*(sum(abs((Impute2[,f,l]
    -complete.ts)/complete.ts)))
    MAPE3[f,l]<-(100/NAs[f,l])*(sum(abs((Impute3[,f,l]
    -complete.ts)/complete.ts)))
    MAPE4[f,l]<-(100/NAs[f,l])*(sum(abs((Impute4[,f,l]
    -complete.ts)/complete.ts)))
    MAPE5[f,l]<-(100/NAs[f,l])*(sum(abs((Impute5[,f,l]
    -complete.ts)/complete.ts)))
    MAPE6[f,l]<-(100/NAs[f,l])*(sum(abs((Impute6[,f,l]
    -complete.ts)/complete.ts)))
  }}

missing.rate <- c()
missing.rate[1:30] <- 0.1
missing.rate[31:60] <- 0.25
missing.rate[61:90] <- 0.5
missing.rate[91:120] <- 0.8



#Visualization of data

z <- cbind(as.vector(NAs),as.vector(MRSE1),missing.rate)
r <- cbind(as.vector(NAs),as.vector(MRSE2),missing.rate)
p <- cbind(as.vector(NAs),as.vector(MRSE3),missing.rate)
q <- cbind(as.vector(NAs),as.vector(MRSE4),missing.rate)
w <- cbind(as.vector(NAs),as.vector(MRSE5),missing.rate)
x <- cbind(as.vector(NAs),as.vector(MRSE6),missing.rate)
data1 <- as.data.frame(z)
data2 <- as.data.frame(r)
data3 <- as.data.frame(p)
data4 <- as.data.frame(q)
data5 <- as.data.frame(w)
data6 <- as.data.frame(x)
names(data1) <-c("NAs","RMSE","missing.rate")
names(data2) <-c("NAs","RMSE","missing.rate")
names(data3) <-c("NAs","RMSE","missing.rate")
names(data4) <-c("NAs","RMSE","missing.rate")
names(data5) <-c("NAs","RMSE","missing.rate")
names(data6) <-c("NAs","RMSE","missing.rate")


plot1 <- ggplot(data=data1, aes(x=NAs, y=RMSE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("RMSE Kalman Arima") + theme_light() + ylim(0,18)
```

```
plot2 <- ggplot(data=data2, aes(x=NAs, y=RMSE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("RMSE Linear Interpolation") + theme_light()
plot3 <- ggplot(data=data3, aes(x=NAs, y=RMSE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("RMSE Spline Interpolation") + theme_light()
plot4 <- ggplot(data=data4, aes(x=NAs, y=RMSE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("RMSE MA") +theme_light()
plot5 <- ggplot(data=data5, aes(x=NAs, y=RMSE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("RMSE LMA") +theme_light()
plot6 <- ggplot(data=data6, aes(x=NAs, y=RMSE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("RMSE Mean Imputation") +theme_light()
multiplot(plot1,plot2, plot3, plot4, plot5, plot6, cols=3)

z1 <- cbind(as.vector(NAs),as.vector(MAPE1),missing.rate)
r1 <- cbind(as.vector(NAs),as.vector(MAPE2),missing.rate)
p1 <- cbind(as.vector(NAs),as.vector(MAPE3),missing.rate)
q1 <- cbind(as.vector(NAs),as.vector(MAPE4),missing.rate)
w1 <- cbind(as.vector(NAs),as.vector(MAPE5),missing.rate)
x1 <- cbind(as.vector(NAs),as.vector(MAPE6),missing.rate)

data11 <- as.data.frame(z1)
data22 <- as.data.frame(r1)
data33 <- as.data.frame(p1)
data44 <- as.data.frame(q1)
data55 <- as.data.frame(w1)
data66 <- as.data.frame(x1)

names(data11) <-c("NAs","MAPE","missing.rate")
names(data22) <-c("NAs","MAPE","missing.rate")
names(data33) <-c("NAs","MAPE","missing.rate")
names(data44) <-c("NAs","MAPE","missing.rate")
names(data55) <-c("NAs","MAPE","missing.rate")
names(data66) <-c("NAs","MAPE","missing.rate")
```

```
plot11 <- ggplot(data=data11, aes(x=NAs, y=MAPE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("MAPE Kalman Arima") + theme_light()
plot21 <- ggplot(data=data22, aes(x=NAs, y=MAPE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("MAPE Linear Interpolation") + theme_light()
plot31 <- ggplot(data=data33, aes(x=NAs, y=MAPE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("MAPE Spline Interpolation") + theme_light()
plot41 <- ggplot(data=data44, aes(x=NAs, y=MAPE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("MAPE MA")+ theme_light()
plot51 <- ggplot(data=data55, aes(x=NAs, y=MAPE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("MAPE LMA")+ theme_light()
plot61 <- ggplot(data=data66, aes(x=NAs, y=MAPE,
colour=factor( missing.rate)))+geom_point()+
ggtitle("MAPE Mean Imputation") + theme_light()
multiplot(plot11,plot21, plot31, plot41, plot51, plot61, cols=3)

################################################################
#HISTOGRAMS AND SCATTER PLOTS

complete.ts <-c(Timeseries1,Timeseries2,
Timeseries3,Timeseries4)
seedd<-30
miss.rate1 <- c(0.1,0.25,0.50,0.80)
incomplete.ts <- array(,dim=c(n,seeds,length(miss.rate1)))
NAss <- array(,dim=c(seeds,length(miss.rate1)))
Impute11<- array(,dim=c(n,seeds,length(miss.rate1)))

Original.data <- Timeseries
Estimated.data <- Impute1[,i,c]

make.scatters.function<-function(Original.data,Estimated.data)
{
data<-data.frame(x=Original,y=Simulated)
ggplot(data,aes(x,y)) + geom_point() + theme_light() +
xlab("Observed Data") + ylab("Estimated Data")+
geom_smooth(method="lm",col="red") + theme(axis.title.y =
element_text(size=14)) + theme(axis.title.x =
element_text(size=14)) + theme(axis.text.x =
element_text(size=14)) + theme(axis.text.y =
element_text(size=14))}

make.hist.function<-function((Original.data,Estimated.data))
{
p1<-hist(Original,breaks=30, col=alpha(rgb(0.9,0.1,0),0.7),
xlab="", ylab="", main="" ,cex.lab=1.5, cex.axis=1.5)
#Second distribution with add=T to plot on top
p2<-hist(Simulated,breaks=30, col=alpha(rgb(0,0,0.6),0.7),
add=T,cex.lab=1.5, cex.axis=1.5)
#Add legend
legend("topright", legend=c("Original","Estimated"),
col=c(alpha(rgb(0.9,0.1,0),0.7), alpha(rgb(0,0,0.6),0.7)),
pt.cex=2, pch=15,lwd=3)
}


for (c in 1:length(miss.rate1)){
  for (i in 1:seeds){
    set.seed(i)
    incomplete.ts[,i,c]<-ts(miss.gen(complete.ts,miss.rate1[c])
    b=0
    for (a in 1:length(complete.ts)){if(is.na(incomplete.ts[a,
    i,c])){b=b+1}
      NAs[i,c] <-b }
    incomp.ts <- as.numeric(incomplete.ts[,i,c])
    Impute11[,i,c] <-na.kalman(incomp.ts,model="auto.arima")
    make.scatters.function(Original.data,Estimated.data)
    make.hist.function(Original.data,Estimated.data)
    }}
```