

Statistical Methods for the Analysis of Corrosion Data for Integrity Assessments

A Thesis Submitted for the Degree of Doctor of Philosophy

by

Hwei-Yang Tan

College of Engineering, Design and Physical Sciences

Brunel University London

March 2017

Dedication

To my parents and family

Abstract

In the oil and gas industry, statistical methods have been used for corrosion analysis for various asset systems such as pipelines, storage tanks, and so on. However, few industrial standards and guidelines provide comprehensive stepwise procedures for the usage of statistical approaches for corrosion analysis. For example, the UK HSE (2002) report “Guidelines for the use of statistics for analysis of sample inspection of corrosion” demonstrates how statistical methods can be used to evaluate corrosion samples, but the methods explained in the document are very basic and do not consider risk factors such as pressure, temperature, design, external factors and other factors for the analyses. Furthermore, often the industrial practice that uses linear approximation on localised corrosion such as pitting is considered inappropriate as pitting growth is not uniform.

The aim of this research is to develop an approach that models the stochastic behaviour of localised corrosion and demonstrate how the influencing factors can be linked to the corrosion analyses, for predicting the remaining useful life of components in oil and gas plants.

This research addresses a challenge in industry practice. Non-destructive testing (NDT) and inspection techniques have improved in recent years making more and more data available to asset operators. However, this means that these data need to be processed to extract meaningful information. Increasing computer power has enabled the use of statistics for such data processing. Statistical software such as R and OpenBUGS is available to users to explore new and pragmatic statistical methods (e.g. regression models and stochastic models) and fully use the available data in the field.

In this thesis, we carry out extreme value analysis to determine maximum defect depth of an offshore conductor pipe and simulate the defect depth using geometric Brownian motion in Chapter 2. In Chapter 3, we introduce a Weibull density regression that is based on a gamma transformation proportional hazards model to analyse the corrosion data of piping deadlegs. The density regression model takes multiple influencing factors into account; this model can be used to extrapolate the corrosion

density of inaccessible deadlegs with data available from other piping systems. In Chapter 4, we demonstrate how the corrosion prediction models in Chapters 2 and 3 could be used to predict the remaining useful life of these components. Chapter 1 sets the background to the techniques used, and Chapter 5 presents concluding remarks based on the application of the techniques.

Acknowledgements

I want to show my sincere gratitude to everyone who helped me through the PhD study.

Firstly, I would like to thank National Structural Integrity Research Centre (NSIRC) for sponsoring my research and giving me such an invaluable opportunity to explore the hidden treasures of data and helping me prepare to become an integrity engineer and data analyst.

My deepest gratitude goes to my academic supervisor, Professor Keming Yu, for his continuous guidance, various help, encouragement and great patience. Professor Yu inspired me with the research ideas, in particular on the mathematical models in my research.

I am also very grateful to my industrial supervisor from TWI Ltd, Dr Ujjwal Bharadwaj, for his mentorship on the application of mathematical approaches on the real-life industrial cases. Dr Ujjwal is well-experienced in the structural reliability assessment in energy industries; he shared his experiences and much useful information with me that allow me to have a clear picture of how the structural integrity assessment is conducted in a plant. I feel fortunate to have him as one of my supervisors.

I would like to thank also my second academic supervisor, Dr Bin Wang for sharing his knowledge in structural integrity and fracture mechanics with me. Dr Wang has also helped me a lot especially in administrative issues during my research.

I want to take this opportunity to thank some of my colleagues from TWI Ltd. They kindly supported me in various ways and extended my knowledge in integrity engineering. Special thanks to Professor Tat-Hean Gan, Dr Shervin Maleki, Dr Elvin Eren, Dr Payam Jamshidi, Antony Jopen, XiaoFei Cui, and Nik Noor.

Lastly, I would like to thank my parents and my wife, for their numberless love, comprehensive support, and constant encouragement. They are giving me strength and faith to finish my PhD and continue my life journey.

Contents

DEDICATION.....	II
ABSTRACT	III
ACKNOWLEDGEMENTS	V
CONTENTS.....	VI
LIST OF FIGURES.....	X
LIST OF TABLES.....	XII
CHAPTER 1. GENERAL INTRODUCTION.....	1
1.1 Background	1
1.2 Aim.....	1
1.2.1 Analysing inspection data from offshore conductor pipes using extreme value analysis techniques.....	2
1.2.2 Application of Weibull density regression analyses on piping deadlegs.....	2
1.2.3 The use of corrosion data analyses for predicting remaining life for integrity assessments	3
1.3 Thesis organisation	4
CHAPTER 2. APPLICATION OF EXTREME VALUE THEORY FOR THE ANALYSIS OF THE MAXIMUM DEFECT DEPTH	5
Abstract.....	5
2.1 Introduction	5
2.1.1 Motivation and purposes	8
2.1.2 Previous works on extreme value analysis for sample corrosion	9

2.1.3	Offshore conductor pipe data.....	10
2.2	Generalised Extreme Value distribution.....	12
2.2.1	The Block Maxima method.....	13
2.2.2	Parameter estimation	14
2.3	Generalised Pareto distribution	15
2.3.1	The Peak-Over-Threshold (POT) method.....	16
2.3.2	Parameter estimation	19
2.3.3	Extremes of locally-dependent defect depth.....	20
2.3.3.1	Modelling block maxima.....	20
2.3.3.2	Modelling threshold exceedances.....	20
2.3.3.3	Density-based Spatial Clustering of Applications with Noise (DBSCAN) cluster analysis	21
2.3.3.3.1	Selection of DBSCAN parameters	23
2.3.3.4	Euclidean distance	24
2.3.3.5	Clustering tendency test.....	25
2.4	Geometric Brownian motion-based corrosion simulation	26
2.5	Application.....	28
2.5.1	Conductor A	28
2.5.2	Block Maxima method.....	30
2.5.3	Peak-over-threshold method.....	36
2.5.4	Defect depth simulation and prediction.....	43
2.6	Summary.....	48
2.7	Appendix.....	49
2.7.1	Appendix A: Stationary data.....	49
2.7.2	Appendix B: Standard Brownian motion	51
2.7.3	Appendix C: Kolmogorov–Smirnov test	51
2.8	References.....	53
CHAPTER 3. APPLICATION OF WEIBULL DENSITY REGRESSION ANALYSIS ON PIPING DEADLEGS.....		57

Abstract	57
3.1 Introduction	58
3.1.1 Motivation and purposes	58
3.1.2 Previous works on deadlegs corrosion	59
3.1.3 Previous works on regression analyses for sample corrosion	60
3.1.4 Piping deadlegs corrosion data.....	61
3.2 Weibull density regression	62
3.2.1 Weibull distribution	62
3.2.2 Gamma transformation proportional hazards (Gt-PH) model	63
3.3 Bayesian hierarchical regression analysis	66
3.3.1 Markov Chain Monte Carlo (MCMC) methods.....	68
3.3.1.1 Gibbs sampling.....	69
3.3.2 Hierarchical stages.....	70
3.3.2.1 First stage.....	70
3.3.2.2 Second stage	71
3.3.2.3 Third stage	71
3.4 Application	72
3.4.1 Weibull density regression	72
3.4.2 Bayesian hierarchical regression analysis	79
3.4.3 Convergence checking.....	82
3.4.4 Mann-Whitney-Wilcoxon test	87
3.5 Summary	88
3.6 References	89
CHAPTER 4. THE USE OF CORROSION DATA ANALYSES FOR PREDICTING REMAINING LIFE FOR INTEGRITY ASSESSMENTS	94
Abstract	94
4.1 Introduction	94
4.1.1 Motivation and purposes	95
4.1.2 Previous works on remaining useful life estimation with corroded assets.....	96

4.2	Remaining useful life assessment for inaccessible piping deadleg	97
4.2.1	First order reliability method.....	98
4.2.2	Case study	100
4.3	Remaining useful life assessment for offshore conductor pipe.....	105
4.4	Summary	107
4.5	References	108
CHAPTER 5.	GENERAL CONCLUSION.....	112

List of Figures

Figure Number	Caption	Page Number
Figure 2.1	Wellhead conductor and casing system.	7
Figure 2.2	Partitioned block maxima for sample defect depth in mm.	14
Figure 2.3	Defect depth distribution, the values that exceed the threshold u are of interest.	18
Figure 2.4	Exceedances over threshold u .	19
Figure 2.5	Centre-based density in DBSCAN clustering.	22
Figure 2.6	Core, border, and noise points in DBSCAN clustering.	22
Figure 2.7	Sample data for k-dist plot.	24
Figure 2.8	K-dist plot for sample data in Figure 2.7.	24
Figure 2.9	Euclidean distance.	25
Figure 2.10	Conductor A consists of 4 sections, where section 2 is in sea splash zone and severe external corrosion is identified.	29
Figure 2.11	Defects of conductor A2 at mean sea level.	30
Figure 2.12	Block maxima that are divided equally into squares.	31
Figure 2.13	Block maxima divided by matrix columns.	31
Figure 2.14	Block maxima divided by matrix rows.	32
Figure 2.15	K-S test results for the GEVD for different block sizes.	32
Figure 2.16	Block maxima fitted with GEVD.	34
Figure 2.17	Mean residual life plot for defect depth data.	37
Figure 2.18	Parameter estimates against threshold for defect depth data.	38
Figure 2.19	Hopkins statistic against a range of threshold values.	38
Figure 2.20	Scale parameters against a range of thresholds for de-clustered data.	39
Figure 2.21	Shape parameter against a range of thresholds for de-clustered data.	40
Figure 2.22	K-dist plot for defect depth data above threshold.	41

Figure 2.23	Defect depth data before de-clustering.	41
Figure 2.24	Defect depth data after de-clustering.	42
Figure 2.25	The algorithm to find diffusion parameter for geometric Brownian motion and simulation of block maxima.	45
Figure 2.26	GEVD for actual defect depth block maxima from conductor A2.	46
Figure 2.27	GEVD for simulated maximum defect depth block maxima.	46
Figure 2.28	Simulated future maximum defect depth using the geometric Brownian motion.	47
Figure 2.29	An example of block maxima that contains a large number of non-extrema.	48
Figure 2.30	An example of peak-over-threshold extrema.	49
Figure 3.1	Corrosion rate from the field vs. model.	75
Figure 3.2	Effect of factors on deadleg corrosion.	76
Figure 3.3	Model corrosion rates at different Weibull quantiles for vertical deadlegs.	77
Figure 3.4	Model corrosion rates at different Weibull quantiles for horizontal deadlegs.	78
Figure 3.5	Model corrosion rates at different Weibull quantiles for mix (vertical + horizontal) deadlegs.	79
Figure 3.6	Trace plots for 200000 iterations of the model.	84
Figure 3.7	Autocorrelation plots for 200000 iterations of the model.	85
Figure 3.8	Gelman-Rubin diagnostic plots for 200000 iterations of the model.	86
Figure 4.1	Probability integration in FORM.	100
Figure 4.2	Life distribution of piping deadleg.	103
Figure 4.3	Remaining useful life and probability of failure of piping deadleg.	105
Figure 4.4	Allowable defect depth of conductor pipe.	107

List of Tables

Table Number	Caption	Page Number
Table 2.1	Estimated parameters for GEVD fitting to block maxima..	33
Table 2.2	Estimated parameters for GPD fitting to defect data over threshold.	40
Table 2.3	Estimated parameters for GEVD fitting to defect block maxima and simulated block maxima using geometric Brownian motion.	47
Table 3.1	Covariates that are used in the Weibull density regression.	73
Table 3.2	Coefficients estimated in density regression.	74
Table 3.3	Estimated coefficients in Bayesian hierarchical modelling.	82
Table 3.4	Potential scale reduction factor (PSRF) of Gelman-Rubin convergence diagnostic.	8788
Table 3.5	Mann-Whitney-Wilcoxon test for frequentist Weibull median and the median of Weibull estimated using the Bayesian approach.	88
Table 4.1	Example data of piping deadleg for a demonstration of the probability of failure estimation using FORM.	101
Table 4.2	KS and AD tests for $X_{\text{nom}}^{\text{deadleg}}$	102
Table 4.3	KS and AD tests for $X_{\text{min}}^{\text{deadleg}}$	103

Chapter 1. General Introduction

1.1 Background

Corrosion is one of the main factors that causes failures in a variety of industrial assets. Data on general and localised corrosion under commonly encountered conditions such as steel exposed to the atmosphere, marine environments, soils, and internal fluids can be used for predicting the life of the structure subject to corrosion. With developments in monitoring and inspection technology, such data are increasingly available. However, often there is a lot of data that requires processing to extract meaningful information that can be used in integrity assessments. Also, at times, due to cost and technical factors such as the inaccessibility of certain areas for inspection, hundred percent coverage of the asset for inspection and monitoring may not be possible. This requires the use of techniques to extrapolate information from data that are available. Although there have been developments, aided by better computing power, in the use of statistical methods to address such issues, challenges remain in determining the best way forward in such situations.

1.2 Aim

This study aims to support analyses of data in order to determine the damage caused by corrosion by:

- Analysing inspection data from offshore conductor pipes using extreme value analysis techniques.
- Analysing inspection data from piping deadlegs in oil and gas wellhead towers using Weibull density regression based on gamma transformation proportional hazard models.
- Demonstrating how the analyses of corrosion data using the techniques shown can be employed for predicting remaining life for integrity assessments.

Although the techniques mentioned above have been applied to specific types of assets where data have been available for the purpose of this research, the techniques will find wider application.

1.2.1 Analysing inspection data from offshore conductor pipes using extreme value analysis techniques

In the oil and gas industries, offshore conductor pipe is a large diameter pipe that is set in the seabed to provide structural foundation for the offshore wellhead. A conductor prevents the sea water from going into the inner layer of the structure that consists of several casings of drill pipe. As it is installed offshore, it is subjected to marine environmental forces such as current and waves that can cause the corrosion and fatigue damage to the conductor. The number of ageing offshore well conductor assets which are being operated beyond their original design life is increasing worldwide. These ageing conductors require periodic maintenance in order to continue operating. The increased frequency of maintenance works that incur high cost such as inspection and repair have now become a major challenge to the oil and gas operators. This problem is further accentuated by the incomplete well construction records, unknown operational conditions and inadequate past maintenance records. Thus there is an urgent need for an approach to prioritise the inspections or repairs of conductors in order to reduce the cost of maintenance. This can be done by analysing the thickness measurements taken from the inspection to make inference of maximum defect depth, growth of corrosion and remaining useful life of these assets. Research done here shows the use of extreme value theory on inspection data to find the maximum defect of conductors, and stochastic process to predict the growth of corrosion.

1.2.2 Application of Weibull density regression analyses on piping deadlegs

Deadleg are segments in the piping systems that are continuously exposed to the process stream but with relatively different flow velocity where the process is stagnant.

These inactive pipes are normally connected to active pipes that carry the main stream. Deadlegs are prone to internal corrosion as a result of water separation from the primary product due to the low flow velocity. Non-destructive testing techniques are the commonly used inspection methods to evaluate the properties of a material and health of a pipe, however, these stagnant pipe segments are usually found at the dead end of piping systems, so they are difficult for inspectors to access to for inspection. The inaccessibility of some of the deadlegs poses some challenges to the plant integrity engineers as the healths of these segments are unknown, and their thickness measurements cannot be obtained for integrity assessment. Failure of these stagnant pipes can result in the failure of the piping systems that can cost millions of pounds in repair and business interruption costs. In extreme cases, such failures could also result in fatalities. For these reasons, a mathematical predictive method for degradation is needed for those inaccessible deadlegs, to allow plant operator to know whether the deadlegs are safe to be left unremoved, or with an optimised cost to eliminate those deadlegs by prioritising each deadleg removal; removing deadlegs requires shutting down the piping system and causes some business costs.

1.2.3 The use of corrosion data analyses for predicting remaining life for integrity assessments

The main interest of plant inspectors in analysing the inspection data of an asset is to know the remaining useful life (RUL) of the asset. The remaining useful life of an asset is defined as the length from the present time to the end of the useful life. Determining RUL is a challenge in many situations as it is directly related to the degradation which has many uncertainties and is not easily measured. The idea of the RUL has been broadly applied as a part of operational research, reliability, and statistics literature with important applications in many fields such as material science, biostatistics and econometrics. RUL estimation is one of the key factors in oil and gas asset integrity and maintenance management. The RUL estimation helps the integrity engineers to make decision on the next inspection and the time to repair or replace the components of an asset. By using the RUL estimation, plant operators can

maximise the operation time of an asset and prioritise the inspection and maintenance of the assets to minimise the maintenance and potential business interruption cost due to failure of these assets. In this research, we show how the analyses of raw inspection and monitoring data feed into integrity assessments, prediction of remaining useful life of conductors and deadleg piping.

1.3 Thesis organisation

This thesis consists of five chapters. Chapter 1 introduces the background and the aim of the research. Chapter 2 develops methods to estimate the maximum defect depth and defect growth of an offshore conductor pipe. Chapter 3 uses a regression model to extrapolate the corrosion density of piping deadlegs in various designs and operating conditions. Chapter 4 studies the remaining useful life of piping and offshore conductor pipe by using a prognosis model that utilises the corrosion rate and defect growth model determined in Chapters 2 and 3. Chapter 5 is the general conclusion that presents concluding remarks based on the application of these techniques.

Chapter 2. Application of Extreme Value Theory for the Analysis of the Maximum Defect Depth

Abstract

The offshore conductor pipe is a large diameter pipe that is set into the ground or seabed. It provides a stable structural foundation for a subsea oil well. The increase in ageing offshore well conductor pipes is beginning to be a major challenge faced by operators worldwide. The structural integrity of these offshore conductor assets is required to be maintained in order to prevent the structural collapse of the well and extend the life further to allow for safe and planned retirement. Two different extreme value approaches are used to analyse the inspection data of the conductor pipes, namely a block maxima (BM) method and a peak-over-threshold method (POT). By using these methods, maximum defect depths are estimated for inspected areas of the conductors and then information is extrapolated for uninspected areas. Geometrical Brownian motion is used to predict the future growth of the defect depth.

Keywords: conductor; extreme value; block maxima; peak-over-threshold; DBSCAN clustering; geometric Brownian motion.

2.1 Introduction

One of the main factors that cause failures in a variety of industrial assets is corrosion. Various types of corrosion have been defined and categorised as uniform, galvanic, crevice, pitting, intergranular, de-alloying, erosion-corrosion, and environmentally-assisted cracking. Other forms of corrosion such as microbiological-induced corrosion, filiform corrosion, and liquid-metal embrittlement are typically covered as sub-categories of these types [1]. Data on general and localised corrosion under commonly encountered conditions such as steel exposed to the atmosphere, marine environments, soils and internal fluids can be used for corrosion rate estimation and

predict the life of the structure subject to corrosion. Data on localised corrosion, however, are limited and corrosion rate cannot be used to predict the lifetime of assets. For systems that undergo localised corrosion, the concept of corrosion probability has been introduced for lifetime prediction [2-3]. With developments in monitoring and inspection technology, such data is increasingly available. However, often there is a lot of data that requires processing to extract meaningful information that can be used in integrity assessments. In addition, at times, due to cost and technical factors such as the inaccessibility of certain areas for inspection, hundred percent coverage of the asset for inspection and monitoring is not possible. This requires the use of techniques to extrapolate information from data that is available. Although there have been developments, aided by better computing power, in the use of statistical methods to address such issues, challenges remain in determining the best way forward in such situations.

The offshore wellhead conductor pipe is a large diameter pipe that is set into the seabed to provide an oil well with a structural foundation. These well conductors are ageing and have been in service for more than 30 years; the ageing conductors have begun to be one of the challenges worldwide that cannot be ignored by the operators. The lack of well construction records, unknown operational conditions and inadequate maintenance further aggravate the problem. Offshore conductors are subjected to the impact of seawater spray, especially the pipe section which is situated at mean sea level or commonly known as sea splash zone, causing the area of pipe in this zone to become uttermost corroded. The loss of cement in the internal annular space between the conductor and surface casing also allows the exposure of surface casing to sea spray and eventual wall loss due to corrosion. Figure 2.1 shows a simple example of wellhead conductor and casings system. A certain level of remaining wall thickness of conductor has to be maintained for it to be strong enough to support the entire wellhead and casings system.

Extreme value analyses have been applied to offshore conductor pipes inspection data. We are given inspection data that result from the Magnetic Eddy Current (MEC) method performed on 84 conductors, and these data are analysed using statistical methods. Two different extreme value approaches are presented in the research,

which are Block Maxima (BM) method using generalised extreme value distribution (GEVD), and Peak-over-Threshold (POT) method together with generalised-Pareto distribution (GPD). These methods are used to estimate the maximum defect depth distribution and extrapolate the defect depth from inspected area to the uninspected area of conductors. For BM method, inspection data is partitioned into the equally-sized blocks, and maximum defect depth is taken from each block and then fitted with GEVD. The maximum defect depth is calculated using return period concept which is by using the quantile function of GEVD. For POT method, inspection data is firstly partitioned into clusters maximum exceedance within each cluster is taken to fit with the GPD. Maximum defect depth is estimated using return period method and the uninspected area is extrapolated. Corrosion growth of these maxima and exceedances are estimated using geometrical Brownian motion simulation, with the fact that localised corrosion is in stochastic behaviour hence the linear corrosion rate is not appropriate to predict the future wall losses of these conductors.

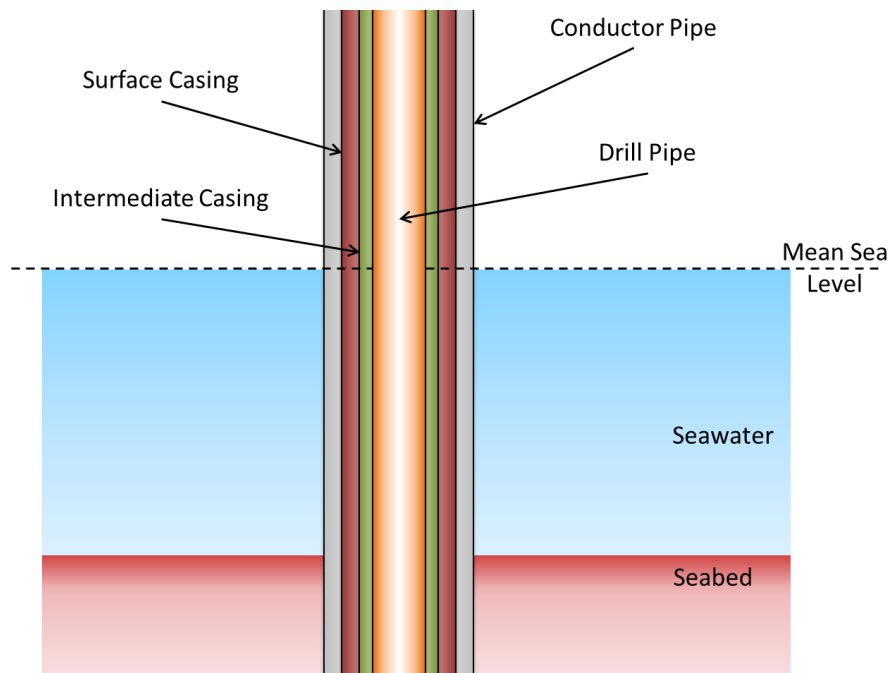


Figure 2.1. Wellhead conductor and casings system

2.1.1 Motivation and purposes

The offshore conductor pipes in oil and gas wellhead tower provide a stable structural foundation for the oil well, and their structural integrities have to be maintained for the plant to continue operating. These conductors are ageing, and corrosion is one of the main damage mechanisms especially the sections of the conductor pipes which are situated at the mean sea splash zone are found to be severely corroded. The lack of design documents and operational information makes the problems further aggravated. A non-destructive technique using Magnetic Eddy Current (MEC) was conducted to measure the remaining wall thicknesses of these conductors. These data have to be analysed to make useful inferences for the operators to understand the fitness of conductors for service, and estimate the future corrosion level for short and long-term actions whether to repair and replace the conductors.

This research shows how extreme value theory can be used to analyse such inspection data. In corrosion engineering field, the inspection data is usually analysed by using simple fitting of corrosion samples with commonly used statistical distributions such as normal and lognormal distributions, and maximum defect depth is estimated by using the upper bound in the confidence interval. However, setting the confidence interval is arbitrary, and only few industrial standards provide comprehensive guidelines for the usage of statistical approaches for corrosion analysis. Extreme value theory allows the integrity assessor to select the extrema from the corrosion samples in a more systematic way. Particularly for localised corrosion analysis, these extrema are what integrity assessor concerned about.

Once the distribution of maximum defect depths is obtained, future growths of these defects are simulated using geometric Brownian motion simulation. The simulation provides a predictive growth model of defects and allows oil and gas operators to estimate the remaining useful life and fitness for service of the conductors.

2.1.2 Previous works on extreme value analysis for sample corrosion

Kowaka et al. [2] use extreme value statistical methods to investigate corrosion phenomena. Shibata reviewed the application of extreme value theory to statistics and emphasised that concept of probability is important for corrosion failure evaluation and prediction by using the concept of the return period, where the return period is defined by dividing the larger surface of the component by sample corrosion which is used for extreme value analysis. Shibata [3] gave an example of the application of EVA on calculations of the most likely maximum pit depth in an oil tank. Scarf and Laycock [4] demonstrated the use of different extreme value statistical methods on pitting corrosion data from industry and laboratory experiments. Block minima with generalised extreme value distribution and peak-over-threshold method using generalised-Pareto distribution were demonstrated.

Vajo et al. [5] applied extreme value theory to the crevice corrosion front that advances under elastomeric seals to aluminium surfaces. They discussed that extreme value analysis could be applied to seals of small test samples to estimate the most probable maximum corroded depths of larger structures and used empirical power law function to model the maximum depth of corrosion under seals versus test time. Glegola [6] presented the block maxima and peak-over-threshold inference methods applied to extreme defect depths. Glegola conducted the statistical test for multiple sets of different block maxima sizes, and the best quality of the fit was used to fit with GEVD. Hierarchical clustering analysis was used to de-cluster the local dependence of the underlying observations and results showed that data de-clustering improves the consistency of the results given by both extreme value and generalised-Pareto distributions.

Rivas et al. [7] presented both block maxima and peak-over-threshold methods on analyses of pitting corrosion data from laboratory-simulated buried line pipe steel. They proved that the threshold approach is more robust to maximum depth reduction, which results from the dependency between pit depths provided the necessary data be available. Stone [8] discussed that extreme value analyses applied to the wall loss distribution that is other than exponential are not that representative. Schneider [9]

illustrated the extreme value approach on data collected from a pipeline system in an oil platform using an ultrasonic corrosion mapping technique; the 2D auto-correlation function is used to determine the size of block minima to ensure the minimum cluster among the blocks.

Melchers [10] suggested that the Fréchet distribution is more appropriate and plausible than Gumbel to represent the maximum pit depth for long-term marine pitting corrosion of steel, considering the modern pitting theory that pit populations consist of stable pitting and meta-stable pitting, as the longer-term exposures sulphate-reducing bacterial activity is the main corrosive agent in a marine environment. Valor et al. [11] improved the existing Markov chains stochastic model for assessing extreme pitting corrosion in the light of new experimental evidence. Pit initiation was modelled using a non-homogenous Poisson process. The distribution of pit nucleation times was simulated using a Weibull process. On the other hand, the time evolution of pit depth was modelled using a non-homogenous Markov process. Valor et al. have shown that the model is able to provide a better physical understanding of the pitting process. Benstock and Cegla [12] introduced a framework for building extreme value models of inspections data to choose a proper set of minima for substantially correlated exponential and Gaussian surfaces. Naoya et al. [13] used a combination of extreme value analysis and Bayesian inference methods to predict the maximum depths of corrosion on oil storage tank, and more accurate prediction is obtained by using the proposed method.

2.1.3 Offshore conductor pipe data

A conductor pipe is a large diameter pipe that is set into the seabed to provide the stable structural foundation for an oil well. It is typically set on oil wells before any drilling operations. An offshore conductor pipe is set in the seabed and is a key structural foundation for the subsea wellhead.

We are provided with 84 conductors' inspection data. The name of the company that provided the data and the location of the conductors will not be disclosed to protect

the sensitive information. These ageing conductors have been in service for more than 30 years, and structural integrity assessments are needed to ensure that they are still fit for service. One of the methods is to measure the remaining wall thickness of the conductor. Offshore structures are often exposed to sea spray where the members of the structures are located near the mean sea level; they are vulnerable in the sea splash zone, and higher external corrosion is expected in this area. There are also reported observations on the loss of cement in the internal annular space between the conductor and surface casing, leaving the surface casing exposed to seawater spray and eventual wall loss due to corrosion.

Magnetic Eddy Current (MEC) inspection was performed on each of these conductors. MEC can detect the corrosion on either side of the wall inspected, and is fast in corrosion screening. This method is practical to inspect the conductor pipes from the external surface, while they are still in service and at operating temperatures. The eddy current sensor covers a circumferential width of 200 mm. The conductors were marked circumferentially into 14 tracks which mean it gives approximately 10 mm spacing from one measurement to the next measurement circumferentially. The MEC scans were taken over one vertical track at a time until 360 degree coverage of the general conductor pipe sections was achieved. All accessible areas of the conductor pipes were targeted for inspection with the exception of specific dead zones, which could not be inspected due to the design of the scanner. This gives in total more than 1 million thickness measurements for each conductor.

Localised pitting corrosion was found on the external surface at the splash zone. The wall loss on external surface at the splash zone is usually found to be between 20% and 40% of nominal thickness, and the wall loss which falls into this range is considered severe. Extreme value theory is used to find the current maximum wall loss for the inspected and uninspected area of the conductors, then the remaining useful life is predicted using the geometric Brownian motion simulation. Predicted remaining useful life allows the integrity engineers to make decision on the maintenance of these conductors.

2.2 Generalised Extreme Value distribution

Given independent and identically distributed (iid) random variables X_1, X_2, \dots, X_n with a cumulative distribution function F , the distribution of the maximum $M_n = \max\{X_1, X_2, \dots, X_n\}$ can be derived exactly for all values of n [15]:

$$\begin{aligned} Pr\{M_n \leq z\} &= Pr\{X_1 \leq z, X_2 \leq z, \dots, X_n \leq z\} \\ &= Pr\{X_1 \leq z\} \times Pr\{X_2 \leq z\} \times \dots \times Pr\{X_n \leq z\} \\ &= \{F(z)\}^n. \end{aligned} \quad (2.1)$$

However, this is not immediately helpful in practice since the distribution function F is unknown and inferences on the maximum imply extrapolation into the upper tail of the distribution. We accept the fact that F is unknown and by the Extreme Value Theorem, the limiting distribution of properly standardised maxima extreme values is a generalised extreme value distribution (GEVD). The GEVD includes three classes of extreme value distributions as special cases, and these are called the Gumbel, Fréchet, and Weibull distributions respectively.

The three classes of extreme value distributions are embedded in the maximum generalised extreme value distribution with cumulative distribution function (CDF)

$$G(z) = \exp\left\{-\left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad (2.2)$$

where $1 + \xi(z - \mu)/\sigma > 0$, $-\infty < \mu < \infty$, $\sigma > 0$, and $-\infty < \xi < \infty$. It has three parameters: a location parameter μ ; a scale parameter σ ; a shape parameter ξ . The GEVD with $\xi = 0$ is interpreted as the limit of (2.2) as $\xi \rightarrow 0$, leading to the Gumbel family with distribution function

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z - \mu}{\sigma}\right)\right\}\right], \quad -\infty < z < \infty. \quad (2.3)$$

Estimates of extreme quantiles of the maximum distribution are obtained by solving $G(z_p) = 1 - p$ for z_p giving:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi} [1 - \{-\log(1-p)\}^{-\xi}], & \xi \neq 0 \\ \mu - \sigma \log\{-\log(1-p)\}, & \xi = 0. \end{cases} \quad (2.4)$$

If the parent distribution F has a maximum limiting distribution G , then F is said to be in the maxima domain of attraction of G . Castillo et al. summarise the maxima and minima domain of attraction of parametric limiting distributions in [16].

2.2.1 The Block Maxima method

To estimate the distribution of maxima using the extreme value distributions, we need to obtain data from a maximum process. The block maxima method described by Coles [15] provides an alternative method for estimating the distribution of maxima by grouping the data into blocks of equal or approximately equal size and taking as data the maximum in each block. For n iid observations X_1, X_2, \dots, X_n , let m denote the number of blocks, so there are $B = n/m$ observations in each block. Let $X_{max_i} = \max\{X_{i1}, \dots, X_{iB}\}$, $i = 1, \dots, m$ be the maximum value in block i . Then the block maxima $X_{max_1}, \dots, X_{max_m}$ are independent observations that will follow, approximately, a maximum extreme value distribution.

In corrosion applications, inspection data (defect depths) are usually populated in a $m \times n$ matrix. The matrix will be partitioned into several smaller matrices (blocks), a maximum value is taken from every block and these maximum values are then fitted with the GEVD. Figure 2.2 shows as an example a 16×16 corrosion matrix that is divided into 64 blocks.

0.2	0.2	0.2	0.2	0.2	0.2	3.7	3.7	5.5	5.5	5.5	5.5	0.2	0.2
0.2	2.9	2.9	2.9	0.2	0.2	5.1	5.1	5.1	2.9	2.9	2.9	0.2	0.2
0.2	2.9	2.9	2.9	3.7	3.7	3.7	3.7	5.5	5.5	5.5	0.2	0.2	0.2
0.2	2.9	2.9	3.3	3.3	4.6	4.6	4.6	5.9	5.9	5.9	0.2	0.2	0.2
0.2	2.9	2.9	3.7	3.7	5.1	5.1	5.1	6.8	6.8	6.8	0.2	0.2	0.2
0.2	2.9	2.9	3.7	3.7	5.1	5.1	5.1	6.8	6.8	6.8	0.2	0.2	0.2
0.2	0.2	0.2	4.2	4.2	5.5	5.5	5.5	6.8	6.8	6.8	0.2	2.9	2.9
0.2	0.2	0.2	4.6	4.6	5.1	5.1	5.1	9.5	9.5	9.5	0.2	5.1	5.1
0.2	0.2	0.2	5.1	5.1	5.1	5.1	4.6	4.6	6.8	6.8	6.8	0.2	5.5
0.2	0.2	0.2	5.5	5.5	5.5	3.3	3.3	6.4	6.4	6.4	0.2	6.4	6.4
0.2	0.2	0.2	5.9	5.9	5.9	0.2	0.2	5.5	5.5	5.5	0.2	6.8	6.8
0.2	0.2	0.2	5.9	5.9	5.9	0.2	0.2	5.1	5.1	5.1	0.2	7.3	7.3
0.2	0.2	0.2	5.9	5.9	5.9	0.2	0.2	4.2	4.2	4.2	3.3	3.3	6.8
0.2	0.2	0.2	5.9	5.9	5.9	3.7	3.7	3.7	3.3	3.3	2.9	2.9	6.4
0.2	0.2	0.2	5.5	5.5	5.5	4.6	4.6	4.6	2.9	2.9	0.2	0.2	5.9
0.2	0.2	0.2	4.6	4.6	5.1	5.1	5.1	5.1	4.2	4.2	0.2	0.2	5.1

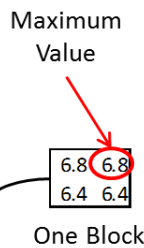
Maximum Value

 One Block

Figure 2.2. Partitioned block maxima for sample defect depth in mm.

2.2.2 Parameter estimation

One of the methods which is commonly used to estimate the unknown parameters of GEVD is the maximum likelihood estimation (MLE) method. Under the assumption that Z_1, \dots, Z_m are independent variables having the GEVD, the log-likelihood of the unknown parameters when $\xi \neq 0$ is:

$$l(\mu, \sigma, \xi) = -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^m \log \left[1 + \xi \left(\frac{Z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^m \left[1 + \xi \left(\frac{Z_i - \mu}{\sigma}\right)\right]^{-1/\xi}, \quad (2.5)$$

provided that $1 + \xi \left(\frac{Z_i - \mu}{\sigma}\right) > 0$, for $i = 1, \dots, m$.

The case $\xi = 0$ requires separate treatment using the Gumbel limit of the GEVD. The log-likelihood function is given by:

$$l(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^m \left(\frac{Z_i - \mu}{\sigma}\right) - \sum_{i=1}^m \exp\left\{-\left(\frac{Z_i - \mu}{\sigma}\right)\right\}. \quad (2.6)$$

For any given dataset the maximisation is performed using standard numerical optimisation algorithms.

2.3 Generalised Pareto distribution

Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables, having marginal distribution function F . It is natural to regard as extreme events those of the X_i that exceed some high threshold u . Denoting an arbitrary term in the X_i sequence by X , the conditional probability that describes the stochastic behaviour of extreme events is given by:

$$Pr\{X > u + y \mid X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, y > 0. \quad (2.7)$$

Since we accept that parent distribution F is unknown, approximations that are broadly applicable for high values of the threshold are sought.

Let M_n denotes the distribution of block maxima where $M_n = \max\{X_1, X_2, \dots, X_n\}$. For large enough n , the GEVD gives

$$Pr\{M_n \leq z\} = \exp\left\{-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

for some $\mu, \sigma > 0$ and ξ .

It is shown by Coles [15] that under appropriate conditions, for large enough values of u , the distribution function $X - u$ conditioned on $X > u$ is approximately within the Generalised Pareto family of distributions. A random variable Y is said to have the Generalised-Pareto distribution (GPD) with shape parameter ξ and scale parameter $\tilde{\sigma}$, if its cumulative distribution function is given by:

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \quad (2.8)$$

where $y > 0$, $\xi \neq 0$, $(1 + \xi y/\tilde{\sigma}) > 0$, and

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \quad (2.9)$$

The GPD with parameter $\xi = 0$ is interpreted by taking the limit $\xi \rightarrow 0$ in (2.8), leading to

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), y > 0 \quad (2.10)$$

corresponding to an exponential distribution with parameter $1/\tilde{\sigma}$.

The return-level associated return period $1/p$, for probability $0 < p \leq 1$, is given by:

$$y_p = \begin{cases} \frac{\tilde{\sigma}}{\xi} [p^{-\xi} - 1], & \xi \neq 0 \\ -\tilde{\sigma} \log(p), & \xi = 0. \end{cases} \quad (2.11)$$

The level y_p is expected to be exceeded on average once every $N = 1/p$ exceedances. It corresponds to some extreme excess and for the wall loss should be rewritten as $u + y_p$.

2.3.1 The Peak-Over-Threshold (POT) method

The Peak-over-Threshold (POT) method uses a more natural way of determining whether an observation is extreme. The data consist of a sequence of iid measurements x_1, \dots, x_n , and extreme events are identified by defining a high threshold u , for which the exceedances are $\{x_i: x_i > u\}$. Denoting these exceedances by $x_{(1)}, \dots, x_{(k)}$, and defining threshold excesses by $y_j = x_{(j)} - u$, for $j = 1, \dots, k$, y_j may be regarded as independent realisations of a random variable whose distribution can be approximated by a member of the generalised Pareto family. If the GPD is a valid model for excesses over the threshold u_0 , then it is valid for excesses over all

thresholds $u > u_0$ [15]. Denoting by σ_{u_0} the GPD scale parameter for excesses over threshold u_0 , the expected value of the excesses, conditional on being greater than the threshold, is

$$E(X - u | X > u) = \frac{\sigma_{u_0} + \xi u}{1 - \xi}. \quad (2.12)$$

Thus, for all $u > u_0$, $E(X - u | X > u)$ is a linear function of u . Furthermore, $E(X - u | X > u)$ is simply the mean of the excesses of the threshold u , for which the sample mean of the threshold excesses of u provided an empirical estimate. This leads to the mean residual life plot, a graphical procedure for identifying a suitably high threshold for modelling extremes via the GPD. The mean residual life plotting procedure is as follows:

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{max} \right\} \quad (2.13)$$

where $x_{(1)}, \dots, x_{(n_u)}$ consist of the n_u observations that exceed u , and x_{max} is the largest of the X_i [15].

Mean residual life plots are commonly used as aids for threshold value selection.

Another method based on the estimation of the model at a range of thresholds is also widely used to decide the proper threshold for the peak-over-threshold model. If a GPD is a reasonable model for excesses of a threshold u_0 , then excesses of a higher threshold u should also follow a GPD. The shape parameters of the two distributions are identical. However, denoting by $\tilde{\sigma}_u$ the values of the generalised Pareto scale parameter for a threshold of $u > u_0$, it follows from (2.9) that

$$\tilde{\sigma}_u = \tilde{\sigma}_{u_0} - \xi u_0 + \xi u. \quad (2.14)$$

Let us reparametrize $\tilde{\sigma}$ according to:

$$\sigma^* = \tilde{\sigma}_u - \xi u \quad (2.15)$$

then σ^* should be approximately constant for all $u > u_0$. This makes the threshold selection easier. We plot the estimated σ^* and ξ as a function of threshold u and look for such a threshold value u_0 for which both estimated parameters are approximately constant whenever $u > u_0$. By the delta method, the confidence bounds to the plot can be added.

Figure 2.3 and Figure 2.4 show the values above threshold u , where these exceedances are of interest that we want to model.

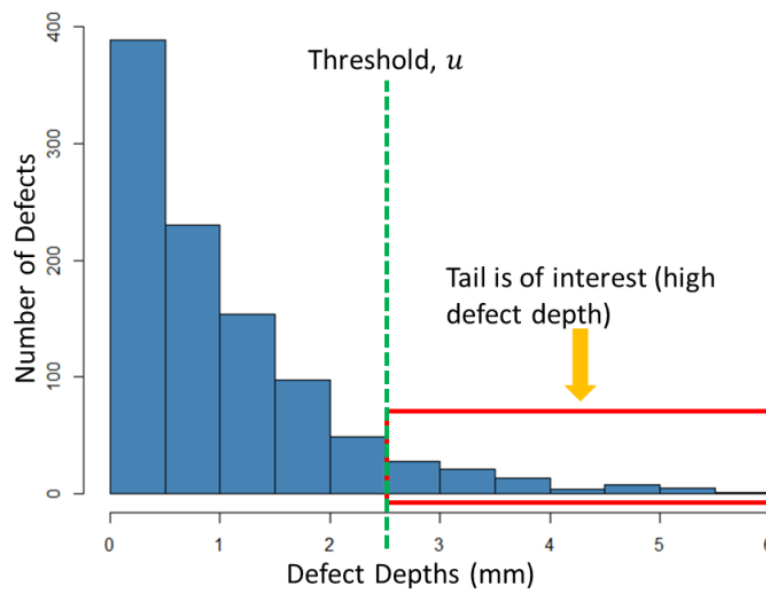


Figure 2.3. Defect depth distribution, the readings that exceed the threshold u are of interest.

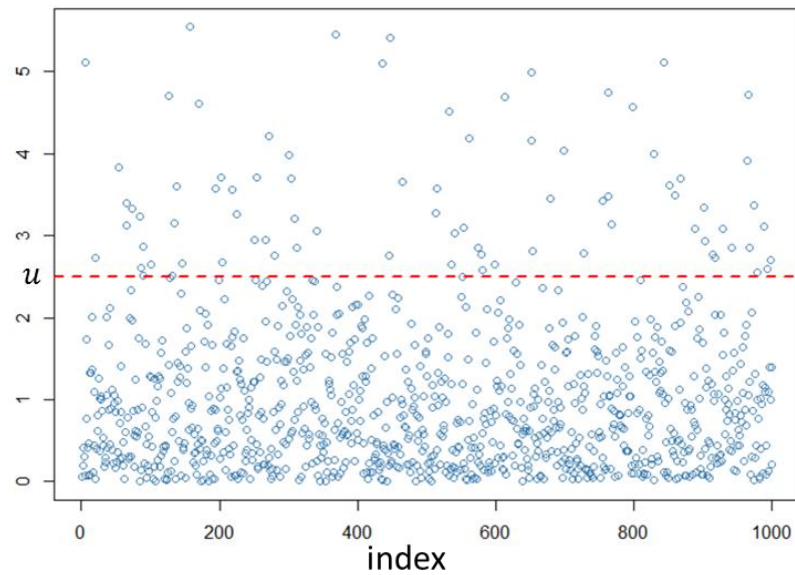


Figure 2.4. Exceedances over threshold u

2.3.2 Parameter estimation

Once the threshold has been determined, the parameters of the GPD can be estimated by using maximum likelihood estimation. Suppose that the values y_1, \dots, y_k are the k excesses of a threshold u . For $\xi \neq 0$ the log-likelihood is derived from (2.8) as

$$l(\sigma, \xi) = -k \log \sigma - (1 + 1/\xi) \sum_{i=1}^k \log(1 + \xi y_i / \sigma), \quad (2.16)$$

provided $(1 + \xi y_i / \sigma) > 0$ for $i = 1, \dots, k$; otherwise, $l(\sigma, \xi) = -\infty$. In the case $\xi = 0$ the log-likelihood is obtained from (2.10) as

$$l(\sigma) = -k \log \sigma - \frac{1}{\sigma} \sum_{i=1}^k y_i. \quad (2.17)$$

Numerical optimisation techniques are required to obtain the maximisation of the log-likelihood function.

2.3.3 Extremes of locally-dependent defect depth

The assumption of independence of underlying observations is usually unrealistic for the types of data to which extreme value models are commonly applied. For corrosion data, it is very likely the growth of one defect can influence the growth of the neighbouring defects. These defects are said to be locally dependent. It is shown by Leadbetter et Al. [40] and Coles [15] that for the stationary and locally dependent observations, the statistical behaviour of extreme events can be modelled by extreme value distributions (Appendix A). The dependence condition requires that the extent of long-range dependence at extreme levels be limited. It implies that as long as the extreme events are far enough apart, they can be considered as approximately independent. In the application to corrosion, this means that if the distance of two extreme depth defects is long enough, they can be considered as nearly independent. The assumption is reasonable on corrosion and motivates the usage of extreme value distributions to model corrosion data.

2.3.3.1 Modelling block maxima

According to Coles [15], modelling block maxima of stationary and locally dependent data does not differ from modelling block maxima of stationary but independent data. This is because if the long-range dependence at extreme levels is weak, observations can be considered as approximately independent for block maxima. The corrosion data can be assumed to be characterised by a limited extent of long-range dependence. Hence, the GEV method introduced in 2.2 can still be applied to the corrosion data.

2.3.3.2 Modelling threshold exceedances

The generalised Pareto distribution can still be used to model the statistical behaviour of excesses over the threshold of stationary data that satisfy the assumptions about the long-range dependence. However, since neighbouring exceedances may be dependent on each other, we cannot use maximum likelihood estimation for this data

directly. Data de-clustering is the widely used method to solve this dependency problem by eliminating dependent observations such that the remaining exceedances are nearly independent. The following is the framework to model excesses over the threshold of stationary data with the generalised Pareto distribution:

- Identify clusters of exceedances;
- Identify the maximum excess within each cluster;
- Fit the generalised Pareto distribution to cluster maxima.

The key issue of data declustering is cluster identification. Many algorithms can be used to find a given number of clusters in data. Wrongly specifying this number can cause the algorithms to group data into artificial clusters. This can lead to the creation of a different data structure than the actual one. Therefore the estimation of the number of clusters is of great importance.

2.3.3.3 Density-based Spatial Clustering of Applications with Noise (DBSCAN) cluster analysis

Many methods can be used to identify clusters in data such as K-means, hierarchical clustering, fuzzy clustering, canopy clustering, autocorrelation function, and other methods. Some of these methods are used in corrosion applications to remove or reduce the temporal dependence of localised defects. Schneider [9] computed a two-dimensional autocorrelation function to select a block size ensuring the minima from each block were independent. Glegola [6] used an agglomerative hierarchical clustering method to identify clusters of exceedances and the maximum value within each cluster is selected before fitting with generalised-Pareto distribution for inference. Saleh et al. [17] applied K-means clustering and expectation-maximization methods to classify pulsed eddy current inspection data and automatically determine corrosion distribution of multiple layers in aluminium structures. Ayako [18] performed model-based clustering analysis via Gaussian mixture models to categorise soil corrosion and prioritise areas to be inspected regarding underground conditions. Bi et al. [19] used a K-means clustering algorithm to classify pitting corrosion characteristics of low

carbon steel generated by acoustic emission signals. There are other examples of the application of clustering methods on corrosion data analyses [20-23].

We proposed density-based spatial clustering of application with noise (DBSCAN) for cluster analyses of corrosion data. We selected this clustering method because of its ability in discovering clusters with an arbitrary shape such as linear, concave, oval, and so on. Furthermore, compared to some of the clustering methods, predetermination of the number of clusters is not required in DBSCAN algorithms. Also, it has been proven in its ability to process enormous databases [24]. DBSCAN is robust to outliers as it has a notion of noise. Although DBSCAN has trouble with high-dimensional data because density is harder to define for such data, the corrosion data are two-dimensional, so it can still be used for such data without problems.

DBSCAN is a density-based clustering algorithm that produces a partitional clustering, in which the number of clusters is automatically determined by the algorithm [25]. DBSCAN is based on a centre-based approach where the density is estimated at a particular point in the data set by counting the number of points within a specified radius, Eps, of that point. This includes the point itself. The technique is graphically illustrated in Figure 2.5. The number of points within a radius of Eps of point A is 6, including A itself.

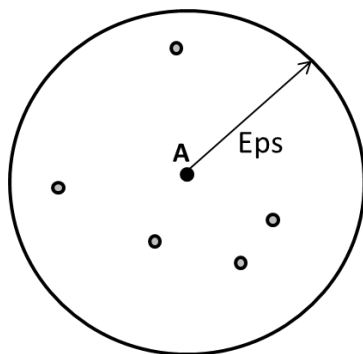


Figure 2.5. Centre-based density in DBSCAN clustering.

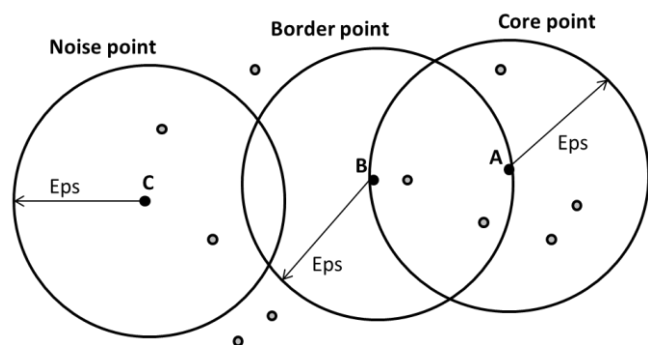


Figure 2.6. Core, border, and noise points in DBSCAN clustering.

The centre-based approach to density allows classification of points:

- **Core points:** These points are in the interior of a density-based cluster. A point is a core point if the number of points within a given neighbourhood around the point which is determined by the distance function and a user-specified parameter of distance, *Eps*, exceeds a certain threshold, *MinPts*. *MinPts* is also a user-specified parameter. In Figure 2.6, point *A* is a core point, for the indicated radius (*Eps*) if $MinPts \leq 6$
- **Border points:** A border point falls inside the area of a core point, yet it is not a core point. In Figure 2.6, point *B* is a border point. A border point can fall within the neighbourhoods of multiple core points.
- **Noise points:** A noise point is any point that is neither a core point nor a border point. In Figure 2.6, point *C* is a noise point.

The DBSCAN algorithm has the following procedures [25]:

1. Identify all points as core, border, or noise points.
2. Get rid of noise points.
3. Put a border between all core points that are inside *Eps* of each other.
4. Create a separate cluster for each group of connected core points.
5. For each border point, assign them to one of the clusters of their associated core points.

We used Euclidean distance (2.3.3.4) for the distance calculation for the cluster analysis.

2.3.3.3.1 Selection of DBSCAN parameters

DBSCAN algorithm requires the parameters *Eps* and *MinPts* to be determined. The basic approach is to look at the behaviour of the distance from a point to its *k*th nearest neighbour, which we will call the *k*-dist. For points that belong to some cluster, the value of *k*-dist will be small if *k* is not larger than the cluster size. Note that there will be some variation, depending on the density of the cluster and the random distribution of points but, on average, the range of variation will not be huge if the cluster densities are not radically different. However, for points that are not in a cluster, such

as noise points, the k -dist will be relatively large. Therefore, if k -dist is computed for all the data points for some k and we sort them in increasing order, a sharp change is expected in the value of k -dist that corresponds to a suitable value of Eps when these sorted values are plotted. If we select this distance as the Eps parameter and take the value of k as the $MinPts$ parameter, then the points for which k -dist is less than Eps will be labelled as core points, and other points will be labelled as either noise or border points.

Figure 2.7 shows a sample data set, while the k -dist graph for the data is given in Figure 2.8. The estimation of Eps that is resolved along these lines relies on upon k , yet does not change significantly as k changes. If the value of k is too small, then even a small number of closely spaced points that are noise or outliers will be incorrectly labelled as clusters. If the value of k is too large, then small clusters are likely to be labelled as noise. The value $k = 4$ is commonly used for DBSCAN algorithm, which is a reasonable value for most two-dimensional data sets.

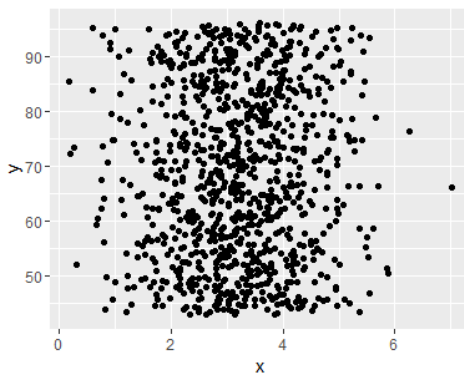


Figure 2.7. Sample data for k -dist plot.

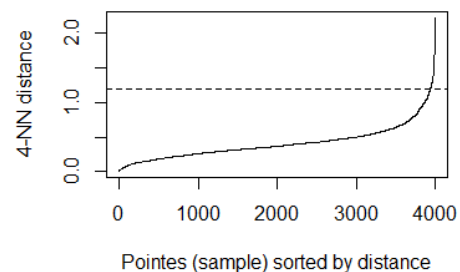


Figure 2.8. K -dist plot for sample data in Figure 2.7.

2.3.3.4 Euclidean distance

The most commonly used distance between two points is Euclidean distance. Most of the time when people talk about distance, they refer to the Euclidean distance. It examines the square root of squared differences between coordinates of a pair

objects. For example, Figure 2.9 shows that the distance between two points in the plane with coordinates (x, y) and (a, b) is given by:

$$\text{dist}[(x, y), (a, b)] = \sqrt{(x - a)^2 + (y - b)^2} \quad (2.18)$$

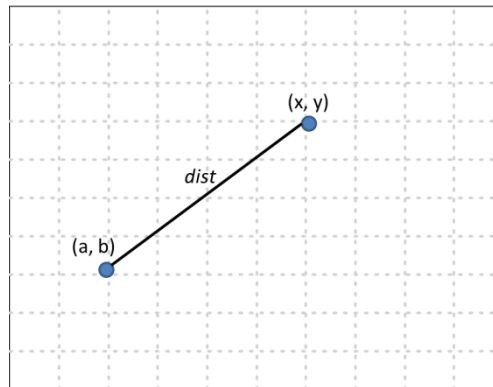


Figure 2.9. Euclidean distance

2.3.3.5 Clustering tendency test

One way to determine whether a dataset has clusters is to try to cluster it and then check it. However, on checking, clusters are always found in a given dataset by almost all clustering algorithms. To get over this issue, we can try to assess whether a dataset has clusters without clustering, by measuring the clustering tendency of the dataset in Euclidean space.

We used the Hopkins statistic to evaluate the clustering tendency. For this approach, we generate p points that are randomly distributed across the data space and also sample p actual data points [26]. For both sets of points, we find the distance to the nearest neighbour in the original dataset. Let the u_i be the nearest neighbour distances of the artificially generated points, while the w_i are the nearest neighbour distances of the sample of points from the original data set. The Hopkins statistic H is then defined by:

$$H = \frac{\sum_{i=1}^p w_i}{\sum_{i=1}^p u_i + \sum_{i=1}^p w_i}. \quad (2.19)$$

If the sample data points have roughly the same nearest neighbour distances, then H will be near 0.5. Data are highly clustered when the H value is close to 0 and data are regularly distributed in the data space when the H value is close to 1.

2.4 Geometric Brownian motion-based corrosion simulation

Localised corrosion has been recognised to have stochastic behaviour [27-30]. Many types of research have discussed the use of stochastic processes to model the deterioration of assets, including corrosion. The generation of new corrosion defects is modelled by a Poisson process in [31]. Zhang et al. [32-33] compared the gamma process-based, inverse Gaussian process and geometric Brownian motion process-based models to characterize the growth of corrosion depth on underground pipelines based on multiple inline inspection data (ILI); uncertainties are considered from different sources by formulating the model in the hierarchical Bayesian framework including the error and random scattering with the inspection tools. Zhang et al. illustrated that the models could predict the growth of corrosion defects reasonably well. Caleyó et al. [34] proposed a new Markov chain model for pitting corrosion that avoids a reduction of the number of pitting states for simplicity. The model has been validated using both synthetic and experimental pitting corrosion data. Describing it as a non-homogeneous Poisson process, one can determine the evolution of the total number of pits. The stochastic behaviour is described by the gamma process because the mean and variance have a well-recognised evolution. This process is also very useful because of the capacity to couple successfully the stochastic attribute of the pitting initiation and the pitting corrosion growth [35]. Guida & Pulcini [36] proposed a state-dependent inverse gamma process for modelling pitting corrosion growth in nonlinear trend. The model is mathematically more tractable as it does not require discretization of time and state.

The model presented here is based on the geometric Brownian motion simulation as discussed by Zhang et al. [33] to predict the growth rate of the defect depth of pitting corrosion.

In Zhang's paper the instantaneous degradation rate at time t , $r(t)$, is given by:

$$r(t) = r_0 \cdot \exp[\beta t + \sigma W(t)] \quad (2.20a)$$

i.e.

$$\log(r(t)) = \log(r_0) + \beta t + \sigma W(t) \quad (2.20b)$$

where r_0 denotes the initial degradation rate, and β and σ denote the drift and diffusion parameters, respectively and $W(t)$ is a standard Brownian process (Appendix B). Implicit in Eq. (2.20b) is that $\log(r(t_2)) - \log(r(t_1)) = \beta \Delta t + \sigma W(\Delta t)$ with $\Delta t = t_2 - t_1$ and $t_2 > t_1$. Δt is the time interval and $W(\Delta t)$ characterises the noise in Δt . It follows that the logarithm of the instantaneous degradation rate at the present time t_2 , $\log(r(t_2))$ can be related to that at the previous time t_1 , $\log(r(t_1))$, through the Brownian motion given by:

$$X(t) = \beta t + \sigma W(t) \quad (2.21)$$

that is $\log(r(t_2)) = \log(r(t_1)) + \beta \Delta t + \sigma W(\Delta t)$. This implies that $r(t_2)$ is dependent on the current state of the degradation; therefore, the model is a state-dependent model.

Let us denote x_{ij} as the depth of defect i at the time of j^{th} inspection, t_{ij} . The instantaneous growth rate at $t_{i,j-1}$ is defined as $r_{i,j-1} = (x_{ij} - x_{i,j-1})/\Delta t_{i,j-1}$, where $\Delta t_{i,j-1}$ is the time difference between current inspection j and previous inspection $j - 1$. $r_{i,j-1}$ is assumed to be constant within the inspection interval $\Delta t_{i,j-1}$. The inspection interval should not be too long, such as less than 5 years [37]. However, if only one inspection record is available, $\Delta t_{i,j-1}$ will be the time difference between current inspection and the commission date of the asset.

Following the assumptions mentioned above and (2.20b), $\log(r_{ij})$ can be characterised by a Brownian motion process as:

$$\log(r_{ij}) = \log(r_{i,j-1}) + \beta\Delta t_{i,j-1} + \sigma W(\Delta t_{i,j-1}). \quad (2.22)$$

The drift parameter β is the change rate between the logarithms of average growth rates corresponding to two consecutive inspection intervals; the standard Brownian motion process $W(t)$ characterizes the random noise in the change between $\log(r_{ij})$ and $\log(r_{i,j-1})$, and the diffusion parameter σ is a scaling factor that quantifies the uncertainty in the random noise.

The depth at the time of the j^{th} inspection, x_{ij} , can be calculated via

$$x_{ij} = x_{i,j-1} + r_{i,j-1}\Delta t_{i,j-1}. \quad (2.23)$$

2.5 Application

2.5.1 Conductor A

We chose one of the 84 conductors as an example of the application of extreme value analyses and called it Conductor A. Conductor A consists of four sections with a 30-inch diameter, and these sections of pipes are joined to form a single conductor pipe. The nominal thickness of sections one, two and three is 22mm each and section four is 28mm. Section 1 is above mean sea level while sections 3 and 4 are below mean sea level. Section 2 is in mean sea level or splash zone area where the external surface of the pipe is severely corroded. The corrosion level below mean sea level is observed to be minimal. For section 1 which is above mean sea level, only a few isolated pitting-like spots were detected. There are certain areas of Conductor A that could not be inspected due to the dead zone created by the conductor guide and because corrosion data of these areas are not available. For this case, we are

particularly interested in assessing section 2 where it is located in the sea water splash zone and external corrosion is found to be most severe in this area. To be more convenient, we called this section Conductor A2. Figure 2.10 shows the wall loss indication for Conductor A.

Conductor A2 has a total 1052463 thickness measurements. Out of this total number of measurements, there are 6664 thickness measurements with wall loss up to about 20%-40%; and 56 out of the total number of measurements with wall loss greater than 40% of nominal thickness.

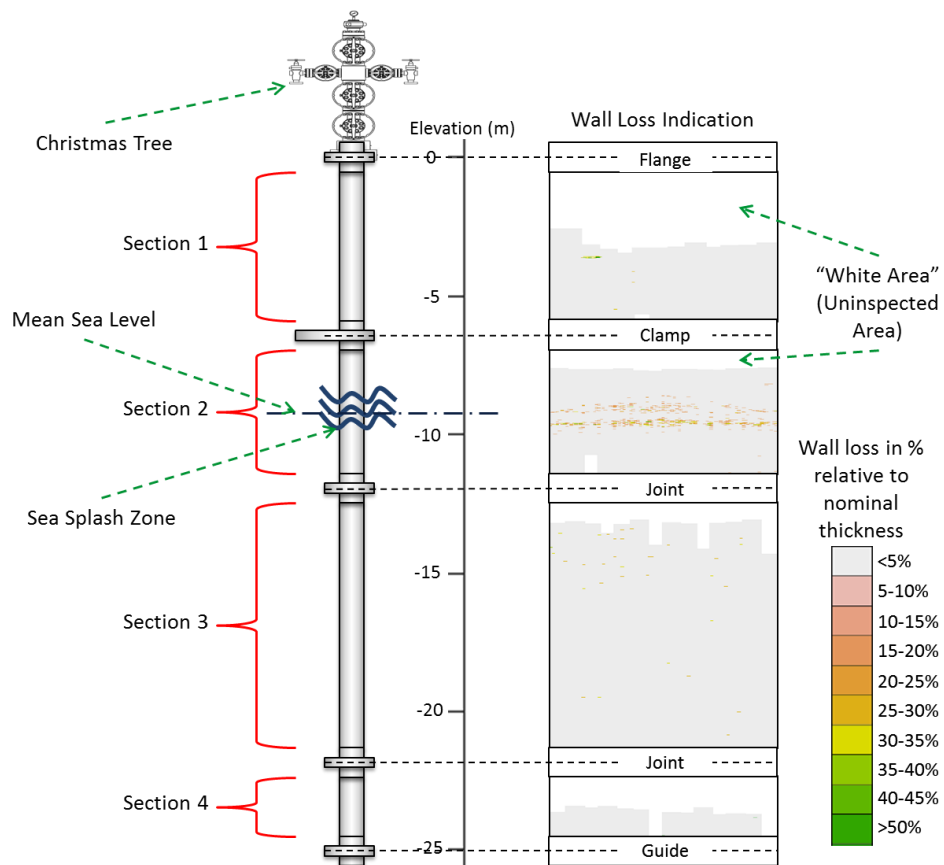


Figure 2.10. Conductor A consists of 4 sections, where section 2 is in sea splash zone and severe external corrosion is identified.

2.5.2 Block Maxima method

In this section, we will use the block maxima method with GEVD for the conductor's corrosion analysis. We focused on section 2 (shown in the Figure 2.10) as it is the most severely corroded area and used extreme value analysis to determine the maximum wall loss. We took the thicknesses measured between -8 metres to -10 metres elevated from the top of the conductor pipe where this area is the splash zone, and severe external corrosion is detected in this zone; this gives a total of 478239 readings. We assumed that the original thickness of Conductor A2 is its nominal thickness and obtained the defect depth by subtracting measured thickness from nominal thickness. We then mapped the defect depths onto a matrix. Figure 2.11 shows a surface plot of defects for Conductor A2.

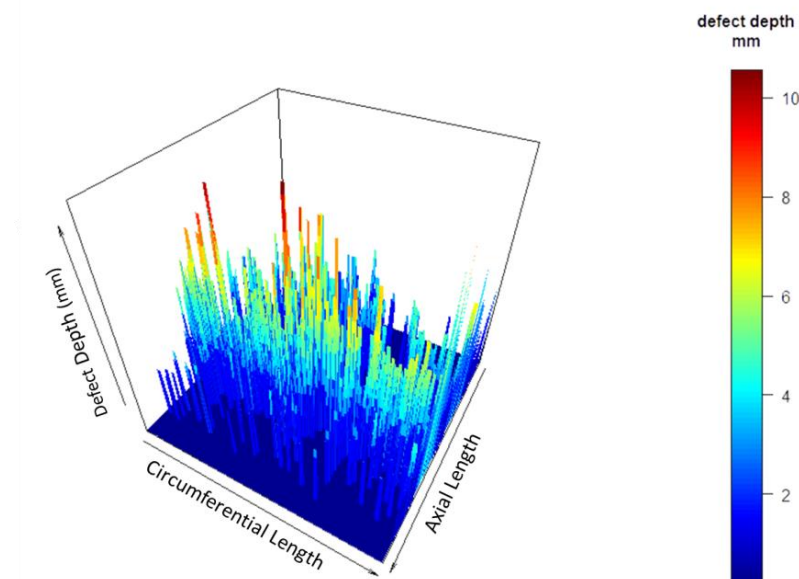


Figure 2.11. Defects of conductor A2 at mean sea level

The matrix consists of 239 columns and 2001 rows. The first step in our analysis is to define block maxima $X_{max_1}, \dots, X_{max_m}$ and specify how to read the maxima from the corrosion map. We need to make sure the block maxima definition covers as much maximum defect depth as possible.

Figure 2.12, Figure 2.13, and Figure 2.14 show the different methods the corrosion map is divided into equally-sized or approximately equally-sized blocks. As shown in Figure 2.12 and Figure 2.14, some blocks do not cover the corroded area, and these give a larger number of homogeneous readings of defect depth. Figure 2.13 shows blocks divided by the matrix columns, and most of the severely corroded areas are covered. We chose the method as shown in Figure 2.13 for block maxima.

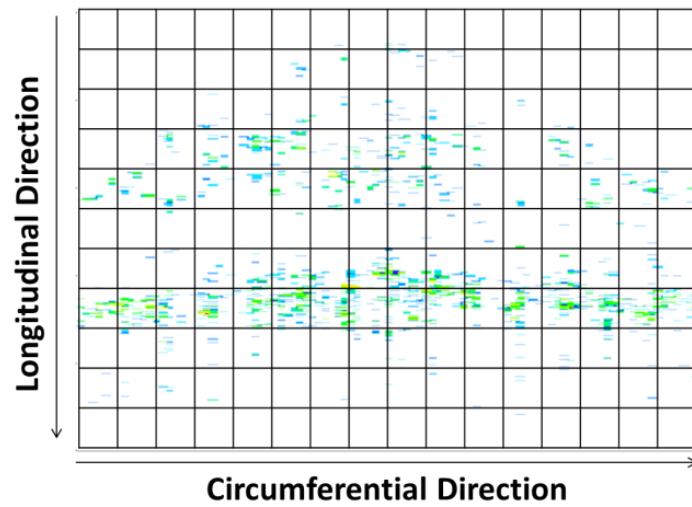


Figure 2.12. Block maxima that are divided equally into squares

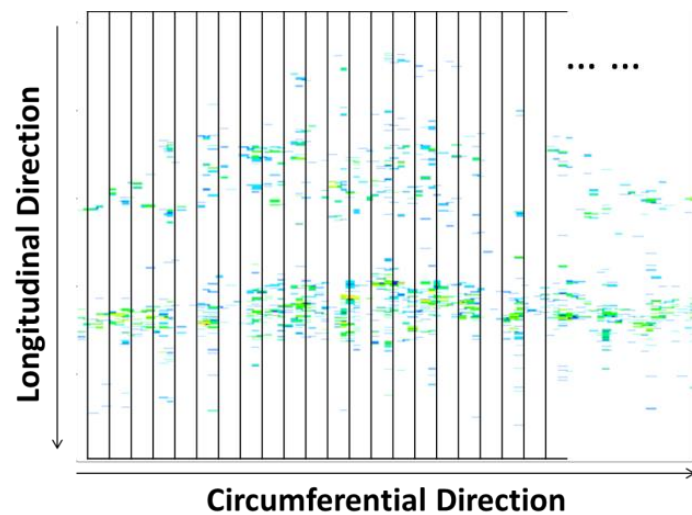


Figure 2.13. Block maxima divided by matrix columns

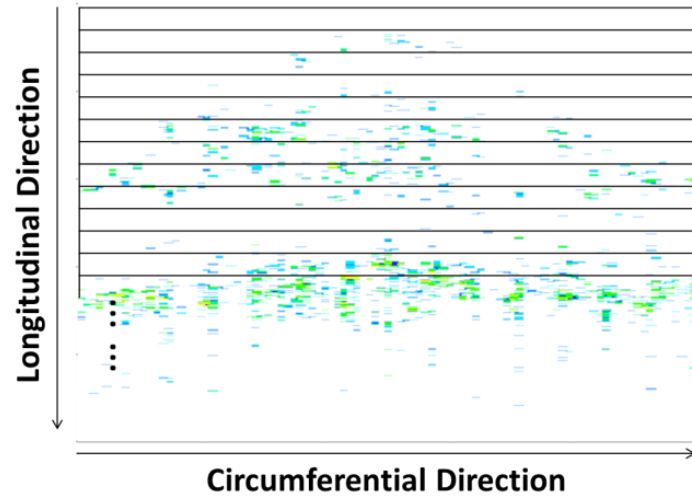


Figure 2.14 Block maxima divided by matrix rows

We know that the blocks should be divided by column, but we need to know also how many columns should be considered as one block. We conducted Kolmogorov-Smirnov tests (K-S tests) to determine the number of columns for one block. We fitted the different block maxima to a generalised extreme value distribution (GEVD) and tested their goodness of fit. A description of the K-S test can be found in Appendix C.

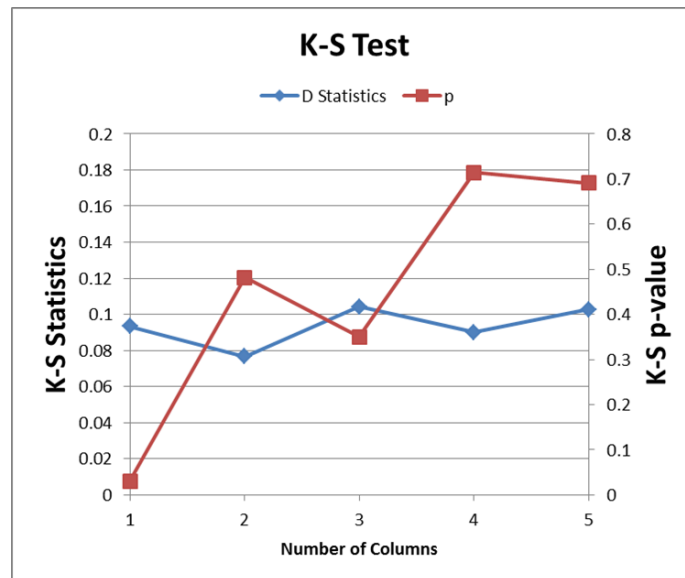


Figure 2.15. K-S test results for the GEVD for different block sizes

From Figure 2.15, we can see that 2-column blocks and 4-column blocks have smaller K-S D statistics. Smaller D statistics correspond with better fit of the distribution. Then we observed the significance level and p-value calculated in K-S test. The null hypothesis is that the data follow the extreme value distribution. The larger the p-value implies the higher the significance level and the p-value must be greater than 0.05 for the null hypothesis not to be rejected. The 4-column block has the larger p-value. By comparing the D statistics and p-values, we considered the 4-column block samples to be a better choice compared to others to fit with GEVD.

Parameters	Maximum Likelihood Estimate	Std. Error	95% Confidence Interval	
			Lower	Upper
μ	6.5086	0.2640	5.9912	7.0260
σ	1.8199	0.1925	1.4426	2.1972
ξ	-0.3081	0.1006	-0.5053	-0.1109

Table 2.1. Estimated parameters for GEVD fitting to block maxima

Having $\xi < 0$ implies that the support of the GEVD is bounded from above by $\mu - \frac{\sigma}{\xi} = 12.38mm$. The boundary value tells us that the predicted wall loss or defect depth should not be greater than 12.38mm regardless of the size of area to which we want to extrapolate the results.

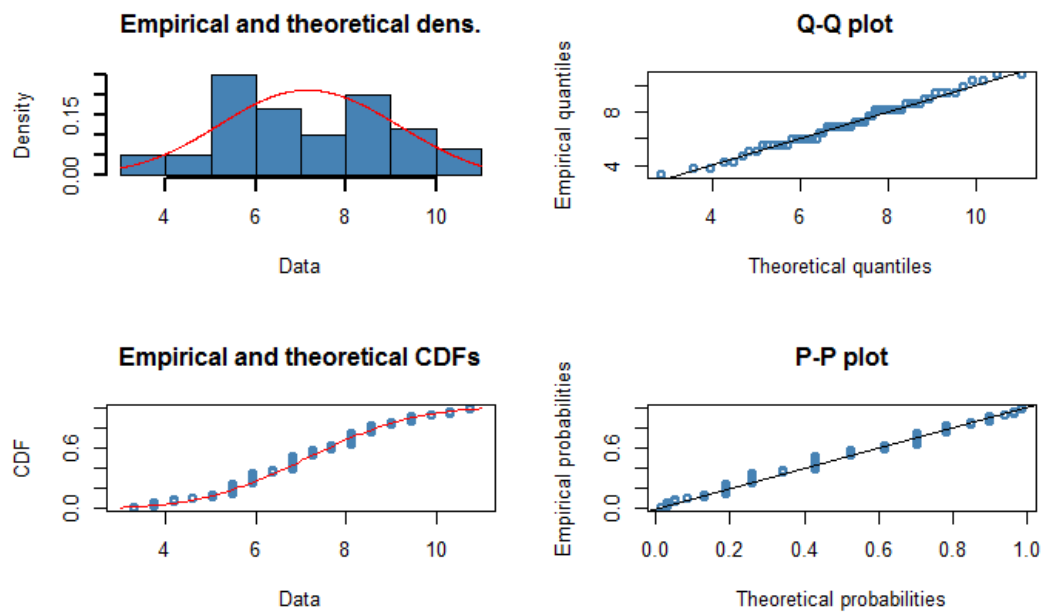


Figure 2.16 Block maxima fitted with GEVD

To extrapolate the maximum defect depth, we need to calculate the return level of the fitted GEVD. We need to determine the return period, T of the block maxima distribution before the return level, which is defined as:

$$\text{Return Period, } T = \frac{A_s}{B_s}$$

where A_s is the surface area of the section of conductor pipe or larger area of the corroded area where sample corrosion is used for block maxima analysis, and B_s is the area of the size of one block maximum. The definition of return period can be found in [14]. In this example we analysed only the splash zone where the area is between the elevation of -8 metres and -10 metres from the top of the conductor pipe. We denoted it by $dy = 2001mm$, and the entire Conductor A2 is from the elevation of -6.2 metres to -11.8 metres. The spacing between circumferential readings is 10mm, we denoted it by $dx \approx 10mm$ and we have 239 readings per row on the matrix. Therefore, $A_s = 2001mm \times 10mm \times 239 = 4782390mm^2$. We used 4 columns on the

matrix as one block, so the size of one block is $B_s = 2001mm \times 10mm \times 4 = 80040mm^2$:

$$T = \frac{4782390}{80040} = 59.75$$

We then calculated the $p = \frac{1}{M}$ value that is required to calculate the return level and the value obtained is $p \approx 0.0167$. Hence by using equation (2.4), the return level is

$$z_p = 6.51 - \frac{1.82}{(-0.31)} [1 - \{-\log(1 - 0.0167)\}^{-(-0.31)}] = 10.73mm.$$

Based on the extrapolated return level, we can say that the wall loss or defect depth that is expected to be exceeded on average once on the sample corroded area of section 2 of Conductor A is 10.73mm.

Using the same estimated GEVD parameters, we can then extrapolate the corrosion probability for the other area of section 2 including the uninspected areas. We calculated return period for another part of section 2 and denoted it by T_s :

$$\begin{aligned} T_s &= \frac{\text{Total Area} - \text{Sample Area}}{\text{Size of One Block}} \\ &= \frac{12566620mm^2 - 4782390mm^2}{80040mm^2} \\ &= 97.25 \end{aligned}$$

and p -value (denoted by p_s):

$$p_s = \frac{1}{T_s} = 0.0103.$$

Hence the maximum defect depth or return level for another area including uninspected area:

$$z_{p_s} = 6.51 - \frac{1.82}{(-0.31)} [1 - \{-\log(1 - 0.0103)\}^{-(-0.31)}] = 10.96mm.$$

2.5.3 Peak-over-threshold method

In this section, we will use the peak-over-threshold method together with the generalised-Pareto distribution to find the maximum defect depth of Conductor A2.

To determine a proper value of threshold u , we applied the methods discussed in 2.3.1 and plotted a mean residual life plot using the available defect depth data. We also plotted σ^* and ξ parameters against a range of threshold u . Figure 2.17 shows the mean residual life plot. It is initially linear but shows substantial fluctuations after $u = 2.8mm$; multiple linearities can be found after this threshold value. We want to select a threshold value which is as large as possible, but if the threshold value is too large, the number of exceedances may not be sufficient to make meaningful inferences. Figure 2.18 shows the scale and shape parameters plotted against a range of threshold values. In Figure 2.18, the change in pattern after threshold $u = 5mm$ was observed and the perturbations were increasing after this threshold value. Based on these two plots, we were given suggestions that the threshold value to use for inference is $u = 5mm$ or lower value. However, as a reminder, in 2.3.3 we discussed that the corrosion defects might be locally-dependent where the depth of a single defect may influence the depth of another. We need to run cluster analyses on these data and de-cluster them before fitting with GPD for deduction, and the number of maxima from those clusters must be large enough to fit with GPD for meaningful inferences.

Before running DBSCAN cluster analyses, we checked the clustering tendency of data for a range of thresholds. Figure 2.19 shows a plot of the Hopkins statistic against a range of thresholds. As discussed in 2.3.3.5, a data set is cluster-able if the Hopkins statistic is less than 0.5. The results show that all data sets that above the range of threshold values are cluster-able as they are much less than 0.5. Therefore, we conducted cluster analysis using DBSCAN for all the data set above the range of

threshold values. From every cluster, we select the maximum value and then fit with the GPD.

Figure 2.20 and Figure 2.21 show respectively the estimated scale and shape parameters against a range of threshold values. Threshold $u = 4.5mm$ was observed to have the minimum sampling errors, so we used this threshold value for our inference.

Figure 2.22 shows the $Eps = 10$ for DBSCAN clustering which we can see the sharp increase in distance. $MinPts = 4$ is used in this cluster analysis which is the original k value of DBSCAN algorithm, which is reasonable for most two-dimensional data sets. Figure 2.23 shows the defect data set before de-clustering and Figure 2.24 after de-clustering. With the data above threshold $4.5mm$, the DBSCAN gave 142 clusters; from each cluster, we calculate the maximum value and then fit with the GPD.



Figure 2.17. Mean residual life plot for defect depth data

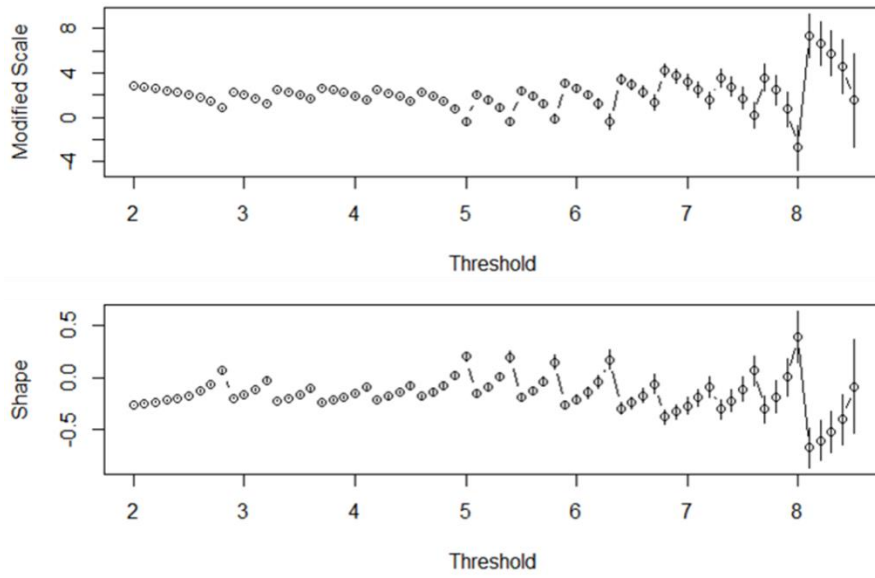


Figure 2.18. Parameter estimates against threshold for defect depth data

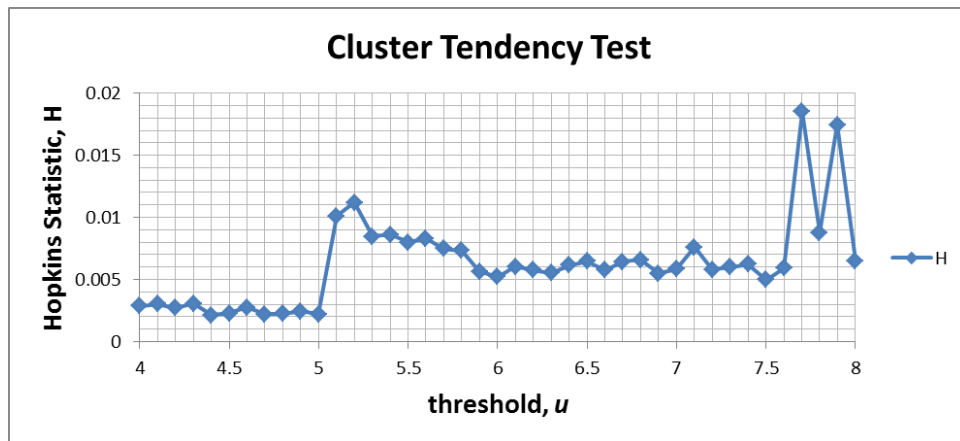


Figure 2.19. Hopkins Statistic against a range of threshold values.

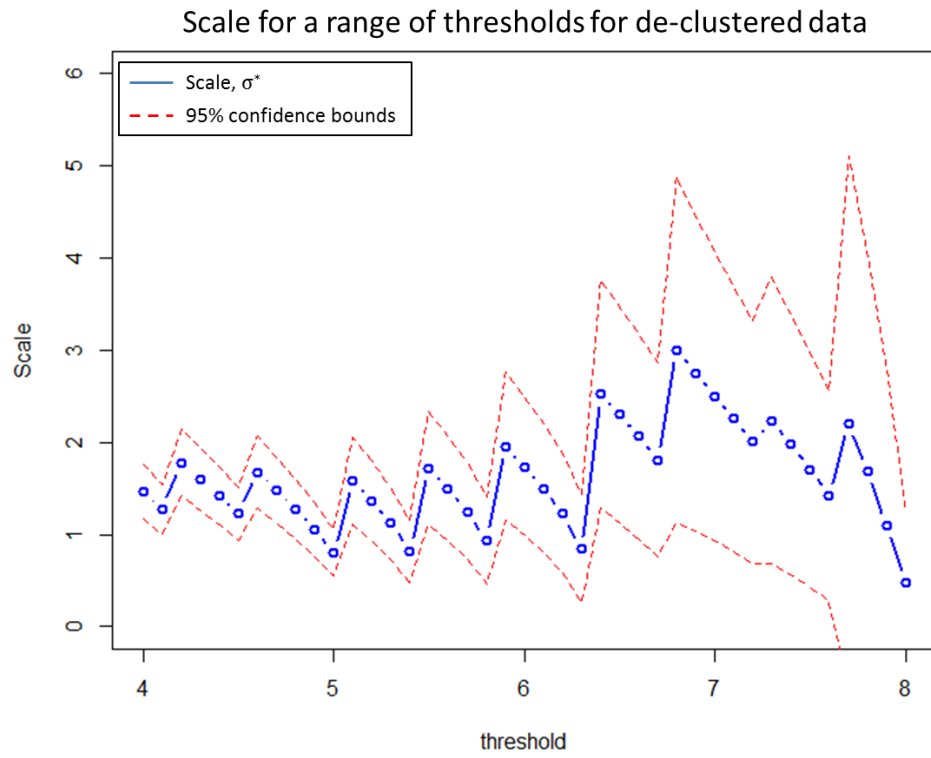


Figure 2.20. Scale parameters against a range of thresholds for de-clustered data.

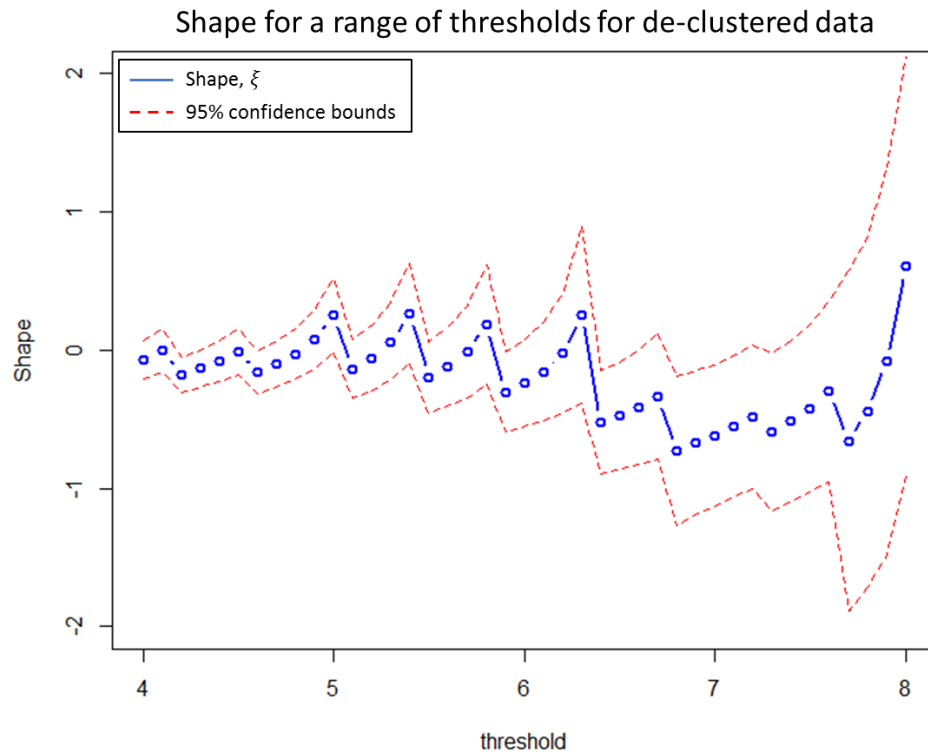


Figure 2.21. Shape parameter against a range of thresholds for de-clustered data.

The following table (Table 2.2) shows the results of fitting the GPD to excesses above the threshold $u = 4.5mm$:

Parameters	Maximum Likelihood Estimate	Std. Error	95% Confidence Interval	
			Lower	Upper
$\tilde{\sigma}$	1.2269	0.1467	0.9394	1.5144
ξ	-0.0070	0.0852	-0.1739	0.1600

Table 2.2 Estimated parameters for GPD fitting to defect data over threshold.

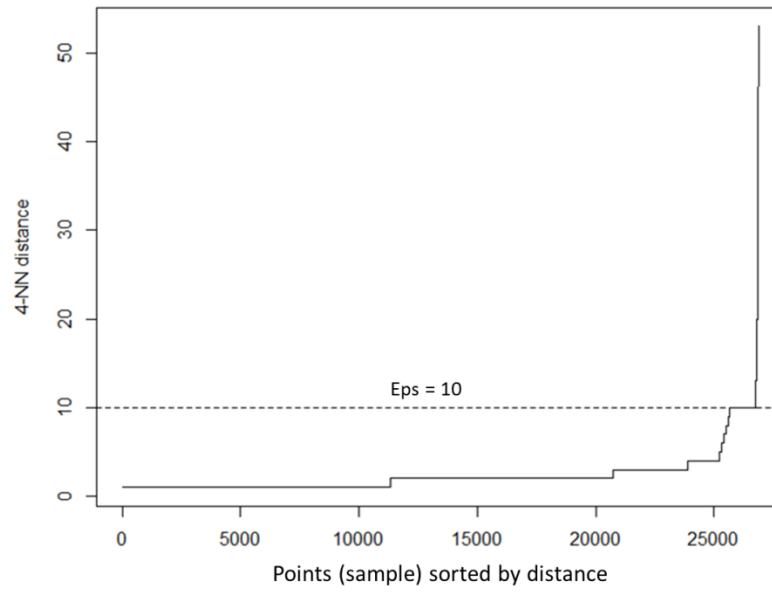


Figure 2.22. K-dist plot for defect depth data above threshold

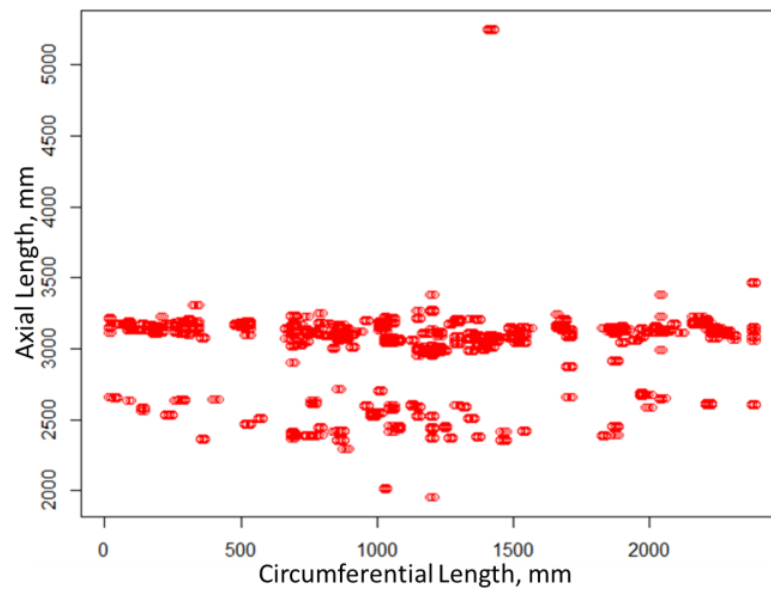


Figure 2.23. Defect depth data before de-clustering

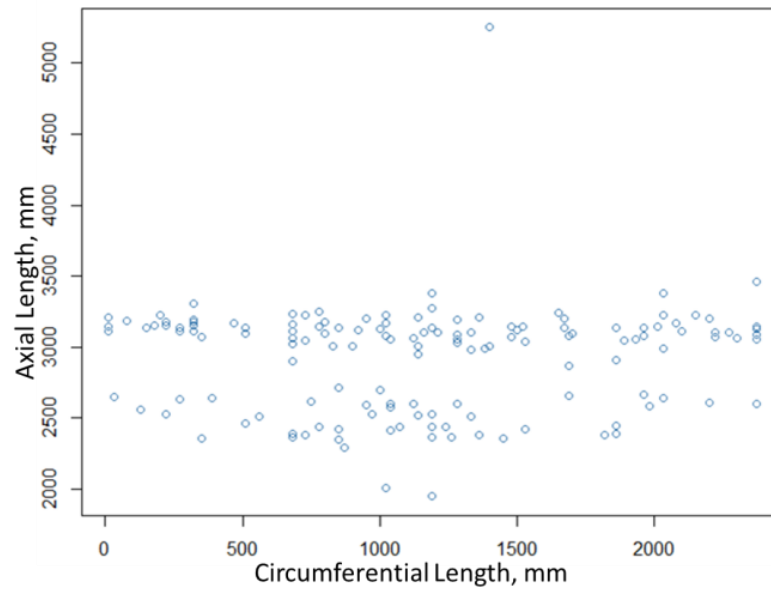


Figure 2.24. Defect depth data after de-clustering

To infer the maximum defect depths on the area of Conductor A2 with the inspection data we used for peak-over-threshold analysis, we need to calculate the return level of GPD using the estimated parameters as presented in Table 2.2.

If N_s denotes the number of threshold exceedances on the surface S then the expected value of N_s is given by:

$$E(N_s) = \lambda \times |S|$$

where λ is the rate of exceedances per unit area of S and $|S|$ is the size of S . We will need the rate of exceedances to extrapolate the maximum defect depth uninspected area. But firstly we will look at the inspected area we used for these parameters' estimation.

The total surface area of Conductor A2 is $12566620mm^2$ and the inspected area where this area has thickness measurements is $10524630mm^2$. Then the rate of exceedances is:

$$\hat{\lambda} = \frac{142}{10524630mm^2} \approx 0.0000135/mm^2$$

$$p = \frac{1}{E(N_s)} = 0.007042254.$$

Return level is then calculated using equation (2.11):

$$y_p = \frac{1.2269}{(-0.007)} [0.007042254^{-(-0.007)} - 1] = 5.976mm$$

The return level means that the wall loss expected to be exceeded on average once is $5.975mm$, and the maximum wall loss is $4.5mm + 5.976mm = 10.476mm$.

Using the same return level equation, we can extrapolate the maximum wall loss of the uninspected area. The expected number of wall loss defects on the other area of Conductor A2 is given by:

$$E(N_{unins}) = \hat{\lambda} \times (12566620mm^2 - 10524630mm^2) \approx 28$$

$$p_{uninsp} = \frac{1}{E(N_{uninsp})} = \frac{1}{280} = 0.03630$$

$$y_{p_{uninsp}} = \frac{1.2269}{(-0.007)} [0.03630^{-(-0.007)} - 1] = 4.022mm.$$

The maximum defect depth of uninspected area is $u + y_{p_{uninsp}} = 8.522mm$.

2.5.4 Defect depth simulation and prediction

Geometric Brownian motion (GBM) simulation is used to simulate the defect and predict the future growth rate. In extreme value analyses, we are interested in the maximum distribution of defects; hence we simulated the defect growth for each defect by using the block maxima x_1, \dots, x_m obtained in 2.5.2.

We have obtained only one single set of inspection data for these conductors. For Conductor A2, the inspection was conducted at the 32nd year start from its commissioning date; therefore the average growth rates between two inspections for defect cannot be estimated. In this case, we firstly treated the average growth rate for every defect in the block maxima as linear by averaging the defect depths into 32 years and got the constant rate for each defect; we denoted it by c_i :

$$c_i = \frac{x_i}{32 \text{ years}}, i = 1, 2, 3, \dots, m.$$

In corrosion analyses, this is considered inappropriate to use a constant corrosion rate to predict the future depth of defect, especially when the corrosion is in extreme behaviour.

Secondly, we need to estimate the growth rate of every year for 31 intervals. We defined the drift parameter and used β_{init} to denote it, as shown below:

$$\beta_{init} = \frac{c_i}{x_i - c_i} - \frac{\sigma_{init}^2}{2}$$

Parameter β_{init} is the average change rate and it is constant because the corrosion growth rate is constant. We assumed each defect started growing since the conductor was commissioned. We then simulated the defect depths using geometric Brownian motion from year zero to year $n = 31$ with the algorithm presented in Figure 2.25 for every defect x_i .

In Figure 2.25, we denoted $r_{i,t}$ as the defect growth rate at year t where $t = 0$ is the time when the instantaneous growth rate of defect occurs. We have only single inspection record, so instead of using $r_{i,j}$ as shown in equation (2.22) which is depending on inspection records, we replaced j as t , giving n numbers of growth rates per defect. We generated z random numbers for the diffusion parameter denoted by σ_k using the uniform distribution; the algorithm will loop through each σ_k then fit the simulated block maxima with GEVD and the diffusion which gives the closest simulated GEVD block maxima distribution is selected. In every loop of σ_k , there is

another loop that simulates the growth rates for 31 intervals using GBM, and every loop reduces the instantaneous growth rate $r_{i,0}$ by value f until the sum of $r_{i,t}, t = 0, \dots, n$ equals or is less than x_i . This is to ensure that the initial instantaneous growth rate is not too high and we assumed the defect depth is increasing throughout the service life of the Conductor A2. Figure 2.26 and Figure 2.27 show respectively the distributions for actual and simulated block maxima fitted with the GEVD, and the estimated parameters are shown in Table 2.3.

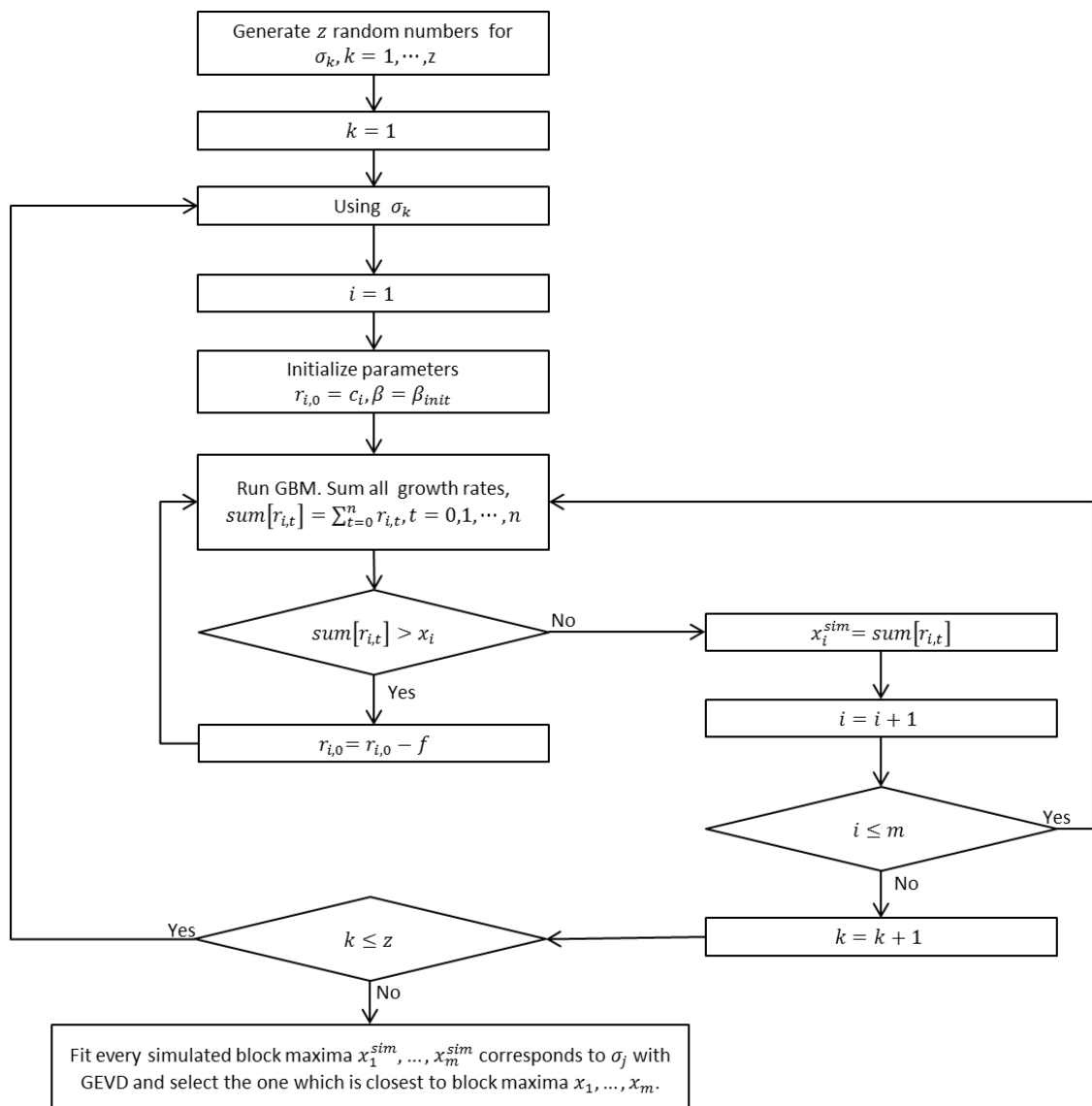


Figure 2.25. The algorithm to find diffusion parameter for the geometric Brownian motion and simulation of block maxima.

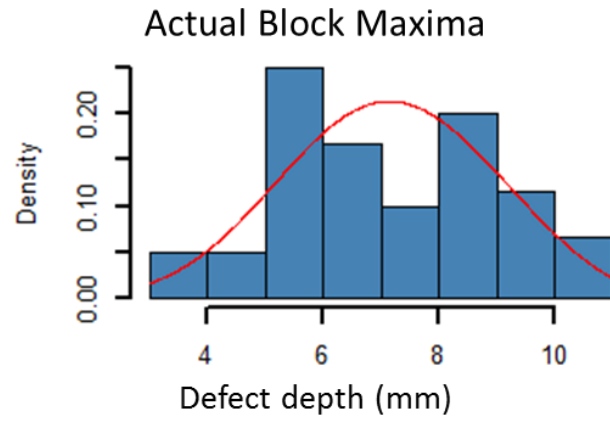


Figure 2.26. GEVD for actual defect depth block maxima from conductor A2.

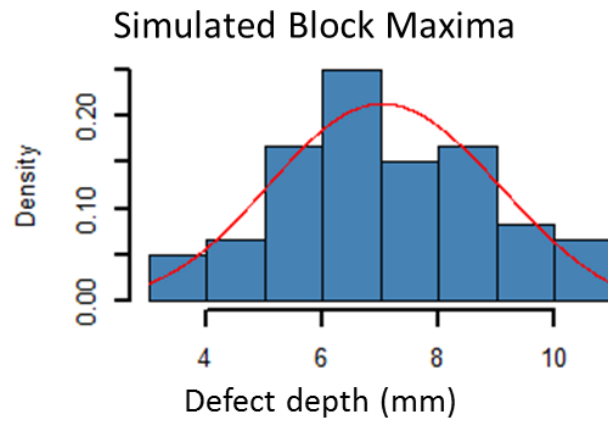


Figure 2.27. GEVD for simulated maximum defect depth block maxima.

Parameters	Maximum Likelihood Estimate		Std. Error		95% Confidence Interval			
					Lower		Upper	
	Actual	Sim.	Actual	Sim.	Actual	Sim.	Actual	Sim.
μ	6.509	6.451	0.264	0.264	5.991	5.882	7.026	6.9168
σ	1.82	1.829	0.193	0.193	1.443	1.443	2.197	2.1978
ξ	-0.308	-0.313	0.101	0.101	-0.505	-0.504	-0.111	-0.1079

Table 2.3 Estimated parameters for GEVD fitting to defect block maxima and simulated block maxima using the geometric Brownian motion.

Once the closest simulated block maxima are identified, we used the simulated growth rates for each defect to calculate the drift parameter β and diffusion parameter σ . With both of these parameters obtained, we can simulate and predict the future defect growth rates using equation (2.22) and defect depths of Conductor A2 using equation (2.23). We have plotted the future maximum defect depth as shown in Figure 2.28.

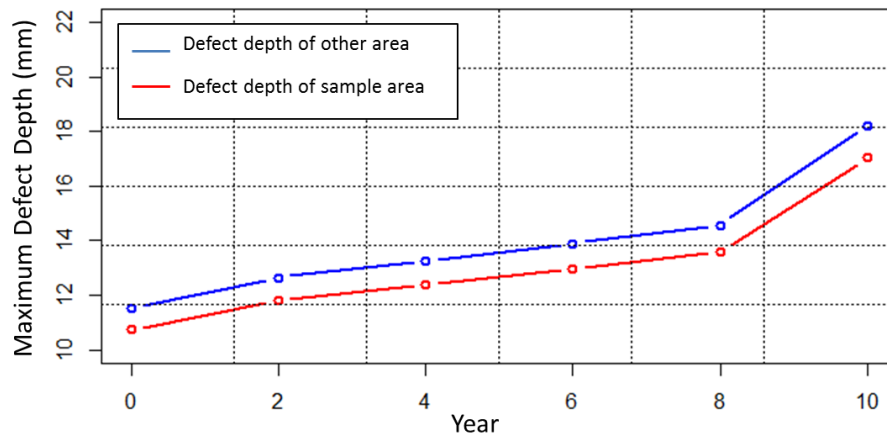


Figure 2.28. Simulated future maximum defect depth using the geometric Brownian motion.

2.6 Summary

We have presented two different methods of application of extreme value theory on offshore conductor pipes' inspection data; they are block maxima (BM) and peak-over-threshold (POT) methods. Both methods can give similar results in inferring the maximum defect depths distribution and extrapolating the defect depth of uninspected area. The question is always which method to use over the other. The BM method is easy to use with the assumption that the data are stationary and long-range dependence at extreme levels is weak. However, there may be cases where the data cannot be divided equally to cover most of the extreme data, for instance, as shown in Figure 2.29.

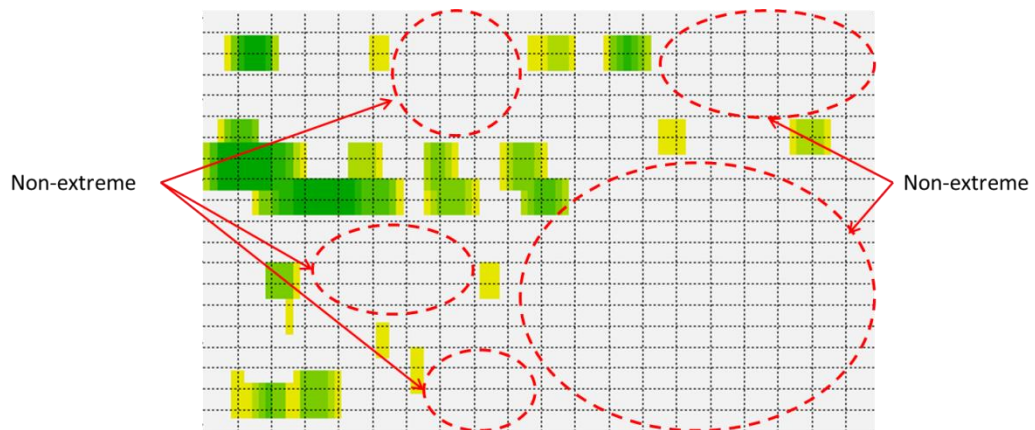


Figure 2.29. An example of block maxima that contains a large number of non-extrema.

Based on the same set of data, we use the POT method and consider only the data above a threshold value u , then fit these data that are above u to a generalised-Pareto distribution for inference. Although it takes more effort to use the POT method including de-clustering of data, it is more robust, and extrema are always the values considered for inference.

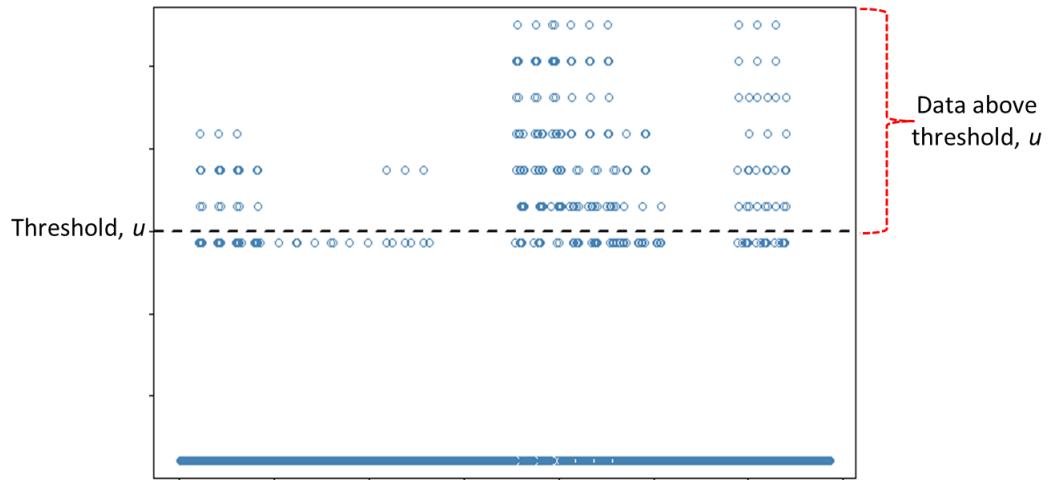


Figure 2.30. An example of peak-over-threshold extrema

Although we can apply the BM and POT methods to estimate the maximum defect depths of conductors, finding the distribution of maximum defect depth is not the end of the story; a prediction of future corrosion rate, given this information, is what integrity assessors would want to know. As it is recognised that localised corrosion or pitting can only be modelled stochastically, thus we used the geometric Brownian motion to model the defects and predict the future defect growths. This predictive method is essential for the plant operator or integrity engineer to make and take short-term and long-term mitigation plans for ensuring the fitness of these conductors for service, and to repair or replace the entire conductors.

2.7 Appendix

2.7.1 Appendix A: Stationary data

Some stationary sequences are important examples of dependent observations. To define the stationary sequences, we first need the definition of a condition known as the $D(u_n)$ dependence condition, which plays an important role, because the limit distributions for the maxima can be identified [16].

Definition 2A.1 (The $D(u_n)$ dependence condition). Let $\{u_n\}$ be a real sequence. The condition $D(u_n)$ is said to hold if for any set of integers $i_1 < i_2 < \dots < i_p$ and $j_1 < j_2 < \dots < j_q$ such that $j_1 - i_p \geq s \geq 1$, we have

$$\left| F_{i_1, i_2, \dots, i_p, j_1, j_2, \dots, j_q}(u_n) - F_{i_1, i_2, \dots, i_p}(u_n) F_{j_1, j_2, \dots, j_q}(u_n) \right| \leq \alpha_{n, s},$$

where $\alpha_{n, s}$ is non-decreasing in s and

$$\lim_{n \rightarrow \infty} \alpha_{n, [n\delta]} = 0, \quad \forall \delta > 0.$$

Note that for independent sequences, the dependence condition, (u_n) , holds trivially with $\alpha_{n, s} = 0$.

Definition 2A.2 (Stationary sequence). A sequence X_1, X_2, \dots of random variables is called stationary if

$$F_{i_1, i_2, \dots, i_k}(x_1, x_2, \dots, x_k) = F_{i_1+s, i_2+s, \dots, i_k+s}(x_1, x_2, \dots, x_k),$$

for every pair of integers k and s .

Theorem 2A.1 (Limit distributions of maxima: The $D(u_n)$ condition). Let $\{X_n\}$ be a stationary sequence and let $\{a_n\}$ and $\{b_n\}$ be two sequences of real numbers such that

$$\lim_{n \rightarrow \infty} \Pr(a_n + b_n X_{n:n} \leq x) = F(x),$$

where $F(x)$ is a cdf. If the $D(u_n)$ dependence condition holds for the sequence $\{U_n = a_n + b_n X_{n:n}\}$ for each x , then $F(x)$ is one of the limit distributions for the independence case.

2.7.2 Appendix B: Standard Brownian motion

A standard Brownian motion (or Wiener process) is a stochastic process with the following properties [38]:

- $W_0 = 0$
- With probability 1, the function $t \rightarrow W_t$ is continuous in t .
- The process $\{W_t\}_{t \geq 0}$ has stationary, independent increments.
- The increment $W_{t+s} - W_s$ has the normal distribution.

The term independent increments means that for every choice of non-negative real numbers

$$0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \dots \leq s_n < t_n < \infty,$$

the increment random variables

$$W_{t_1} - W_{s_1}, W_{t_2} - W_{s_2}, \dots, W_{t_n} - W_{s_n}$$

are jointly independent; the term stationary increments means that for any $0 < s, t < \infty$ the distribution of the increment $W_{t+s} - W_s$ has the same distribution as $W_t - W_0 = W_t$.

2.7.3 Appendix C: Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (K-S test) is one of the oldest goodness-of-fit tests, proposed by Kolmogorov (1933) and Smirnov (1939), to use D_n statistic based on the comparison between the hypothesised distribution function $F_0(x)$ and the empirical distribution function of the sample $S_n(x)$: $D_n = \sup_{-\infty < x < \infty} |S_n(x) - F_0(x)|$. If $F_0(x)$ is continuous and under the null hypothesis, the distribution of D_n is independent of $F_0(x)$ [39].

The Kolmogorov-Smirnov goodness-of-fit test involves the examination of a random sample from a one-dimensional and continuous random variable, in order to test if the data were extracted from a hypothesised distribution $F_0(x)$. The test is about the null hypothesis against a generic alternative:

$$\begin{cases} H_0: F(x) = F_0(x) \text{ for every } x \\ H_1: F(x) \neq F_0(x) \text{ for some } x \end{cases} \quad (2C.1)$$

where $F(x)$ is the true cumulative distribution function.

Let X be the random variable with the continuous cumulative distribution function

$$F(x) = Pr(X \leq x)$$

and let $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$ be the order statistic of the random sample $x_i \sim IID(F)$, $i = 1, 2, \dots, n$, so that $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$.

The empirical distribution function is defined as follows:

$$S_n(x) = \begin{cases} 0, & \text{for } x < x_{(1)} \\ k/n, & \text{for } x_{(k)} \leq x < x_{(k+1)} \text{ with } k = 1, 2, \dots, n-1 \\ 1, & \text{for } x \geq x_{(n)} \end{cases}$$

The statistic introduced by Kolmogorov (1933):

$$D_n = \sup_{-\infty < x < \infty} |S_n(x) - F_0(x)|$$

for which the critical region of size α to reject the null hypothesis in (2C.1) is

$$R = \left\{ D_n : D_n > D_{\alpha,n} = \frac{d_\alpha}{\sqrt{n}} \right\}$$

where d_α depends only on α .

Smirnov (1939) suggest an asymptotic distribution of the one-sided statistic:

$$D_n^+ = \sup_{-\infty < x < \infty} |S_n(x) - F_0(x)|$$

and Kolmogorov had given a proof of the theorem suggested by Smirnov.

2.8 References

1. L. E. Eiselstein and R. Huet (2011) Corrosion failure analysis with case histories, Uhlig's Corrosion Handbook, 3rd Edition, John Wiley & Sons, Inc.
2. M. Kowaka, H. Tsuge, M. Akashi, K. Masamura, and H. Ishimoto (1994) Introduction to Life Prediction of Industrial Plant Materials : Application of the Extreme Value Statistical Method for Corrosion Analysis. New York : Allerton Press.
3. T. Shibata (1994) Application of extreme value statistics to corrosion, Journal of Research of the National Institute of Standards and Technology, Vol 99, 4, 327.
4. P. A. Scarf and P. J. Laycock (1994) Applications of extreme value theory in corrosion engineering, Journal of Research of the National Institute of Standards and Technology, Vol 99, 4.
5. J. J. Vajo, R. Wei, A. C. Phelps, L. Reiner, G. A. Herrera, O. Cervantes, D. Gidanian, B. Bavarian, and C. M. Kappes (2002) Application of extreme value analysis to crevice corrosion, Corrosion Science, 45, 497-509.
6. M. Glegola (2007) Extreme value analysis of corrosion data, Master thesis, Delft University of Technology.
7. D. Rivas, F. Caleyó, A. Valor, J.M. Hallen (2008) Extreme value analysis applied to pitting corrosion experiments in low carbon steel: Comparison of block maxima and peak over threshold approaches, Corrosion Science, 50, 3193-3204.
8. M. Stone (2009) Wall thickness distributions for steels in corrosive environments and determination of suitable statistical analysis methods, 4th European-USA Workshop on Reliability of NDE, Sonomatic Ltd, UK.
9. C. Schneider (2009) Application of extreme value analysis to corrosion mapping data , 4th European-USA Workshop on Reliability of NDE, TWI Ltd, UK.

10. R. E. Melchers (2008) Extreme value statistics and long-term marine pitting corrosion of steel, *Probabilistic Engineering Mechanics*, 23, 482-488.
11. A. Valor, F. Caleyó, D. Rivas, and J.M. Hallen (2009) Stochastic approach to pitting-corrosion-extreme modelling in low-carbon steel, *Corrosion Science*, 52, 910-915.
12. D. Benstock and F. Cegla (2015) Sample selection for extreme value analysis of inspection data collected from corroded surfaces, *Corrosion Science*, 103, 206-214.
13. N. Kasai, S. Mori, K. Tamura, K. Sekine, T. Tsuchida, and Y. Serizawa (2016) Predicting maximum depth of corrosion using extreme value analysis and Bayesian inference, *Internal Journal of Pressure Vessels and Piping*, 146, 129-134.
14. R. Vezzoli, P. Mercogliano, and S. Pecora (2012) A brief introduction to the concept of return period for univariate variables, CMCC, RP0139.
15. S. G. Coles (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer, ISBN 1852334592.
16. E. Castillo, A. Hadi, N. Balakrishnam, and J. M. Sarabia (2005) *Extreme Value and Related Models with Applications in Engineering and Science*, John Wiley and Sons.
17. H. Saleh, A. L. Aouni (2011) Application of time-frequency analysis for Automatic hidden corrosion detection in a multilayer aluminum structure using pulsed eddy current, *NDT&E International*, 47, 20-79.
18. A. Yajima (2015) Assessment of soil corrosion in underground pipelines via statistical inference, PhD Dissertation, University of Akron.
19. H. Bi, Z. Li, D. Hu, I. Toki-Gyamerah, and Y. Cheng (2015) Cluster analysis of acoustic emission signals in pitting corrosion of low carbon steel, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.
20. H. Wang, A. Yajima, R. Y. Liang, and H. Castaneda. (2015) A Bayesian model framework for calibrating ultrasonic in-line inspection data and estimating actual external corrosion depth in buried pipeline utilizing a clustering technique, *Structural Safety*, 54, 19-31.

21. D. G. Harlow. (2011) Constituent particle clustering and pitting corrosion, The Minerals, Metals & Materials Society and ASM International.
22. L. Calabrese, G. Campanella, and E. Proverbio (2012) Noise removal by cluster analysis after long time AE corrosion monitoring of steel reinforcement in concrete, *Construction and Building Materials*, 34, 362-371.
23. F. Pashmforoush, M. Fotouhi, and M. Ahmadi (2012) Acoustic emission-based damage classification of glass/polyester composites using harmony search k-means algorithm, *Journal of Reinforced Plastics & Composites*, 31 (10), 571-680.
24. D. Birant and A. Kut (2007) ST-DBSCAN: An algorithm for clustering spatial-temporal data, *Data & Knowledge Engineering*, 60, 208-221.
25. P. N. Tan, M. Steinbach, V. Kumar (2006) Introduction to Data Mining, ISBN: 978-0321321367, Chapter 8.
26. A. Banerjee and R. N. Dave (2004) Validating clusters using the Hopkins statistic, *IEEE*, 0-7803-8353-8/04, 25-29.
27. U. R. Evans (1974) Localized corrosion, NACE, Houston, page 158.
28. D. E. Williams, C. Westcott, and M. Fleischmann (1984) Stochastic models of pitting corrosion of stainless steels, The Electrochemical Society, 1804.
29. T. Shibata (1911) Evaluation of corrosion failure by extreme value statistics, The Iron and Steel Institute of Japan International, 31, 115-121.
30. A. K. Sheikh, J. K. Boah, and D. A. Hansen (1990) Statistical modeling of pitting corrosion and pipeline reliability, *Corrosion* 46, Houston, 3.
31. H. P. Hong (1999) Inspection and maintenance planning of pipeline under external corrosion considering generation of new defects, *Structural Safety*, 21, 3, 203-222.
32. S. Zhang, W. Zhou, and H. Qin (2012) Inverse Gaussian process-based corrosion growth model for energy pipelines considering the sizing error in inspection data, *Corrosion Science*, 73, 309-320.
33. S. Zhang and W. Zhou (2013) Probabilistic characterization of metal-loss corrosion growth on underground pipelines based on geometric Brownian motion process, *Structural and Infrastructure Engineering: Maintenance, Management, Life-Cycle Design and Performance*, DOI: 10.1080/15732479.2013.875045.

34. F. Caleyó, J. C. Velázquez, A. Valor, and J. M. Hallen (2009) Markov chain modelling of pitting corrosion in underground pipelines, *Corrosion Science*, 51, 2197-2307.
35. J. C. Velázquez, J. A. M. V. D. Weide, E. Hernández, and H. H. Hernández (2014) Statistical modelling of pitting corrosion: extrapolation of the maximum pit depth-growth, *International Journal of Electrochemical Science*, 9, 4129-4143.
36. M. Guida and G. Pulcini (2013) The inverse gamma process: a family of continuous stochastic models for describing state-dependent deterioration phenomena, *Reliability Engineering and System Safety*, 120, 72-79.
37. A. N. K. Agyenim-Boateng (2014) Determination of corrosion rate and remaining life of pressure vessel using ultrasonic thickness testing technique, 3, 2, 43-50.
38. I. Karatzas and S. E. Shreve (1991) *Brownian Motion and Stochastic Calculus*, Springer-Verlag New York Inc, Chapter 2, Page 104.
39. S. Facchinetti (2009) A procedure to find exact critical values of Kolmogorov-Smirnov test, *Italian Journal of Applied Statistics*, 21, 3-4, 337-359.
40. M. R. Leadbetter, G. Lindgren and H. Rootzén (1983) *Extremes and Related Properties of Random Sequences and Processes*, Springer Series in Statistics, ISBN: 978-1-4612-5454-5.

Chapter 3. Application of Weibull Density Regression Analysis on Piping Deadlegs

Abstract

Deadlegs are segments in the piping systems that are continuously exposed to the process stream but with relatively different flow velocity where the process is stagnant. Deadlegs are prone to internal corrosion as a result of water separation from the primary product due to the low flow velocity. Estimating and predicting the corrosion rate of deadlegs in oil and gas process piping systems is a challenging task, due to multiple factors in the service environment causing internal corrosion, inspection being difficult due to inaccessibility or cost implications. It is important to make accurate estimation and prediction of the corrosion rates of deadlegs since they are usually connected to the main line, and their leakage can lead to the release of fluid in the piping system. Case studies of piping deadlegs are presented here in which a Weibull density regression is carried out to make inference about the corrosion rates of deadlegs in a piping system by treating operational or design factors as independent regression variables. The best linear unbiased estimators are derived for the estimation of the coefficients of the regression. Notably, a new Weibull density-based regression model is proposed and applied to predict the corrosion rates of piping for several oil and gas process piping systems and water injection systems. The corrosion rate data of these pipings are analysed given the presence of multiple factors simultaneously. The results are interpreted with using the combination of different factors. The regression method shown here is an alternative to other methods that are more approximate and involve asymptotic inference. To reduce the use of dummy variables for categorical data and also minimise the impact of multicollinearity among those factors, the Bayesian hierarchical model is used together with Weibull density regression.

Keywords: piping; deadleg; corrosion; density regression; regression analysis; Weibull distribution; best linear unbiased estimators (BLUE); Bayesian hierarchical model; Markov chain Monte Carlo

3.1 Introduction

3.1.1 Motivation and purposes

In oil and gas process piping systems, deadlegs are segments continuously exposed to the process but with abnormal flow velocity where the process stream is stagnant. These inactive segments usually connect to active pipes that are carrying the main petrochemical products. Some deadlegs examples include blanked branches, lines with normally closed block valves, lines with one end blanked, pressurised dummy support legs, stagnant control valve bypass piping, and instrument connections [1]. Deadlegs can be identified through piping and instrumentation diagrams (PI&D) and engineering judgement. Piping systems design typically avoids deadlegs, but they often form due to the changes of process. If deadlegs are unavoidable in a piping system, they can be minimised by eliminating dead ends in piping manifolds or providing drains [2]; the length of deadlegs must be as short as possible to prevent the fluid from being stagnant.

Deadlegs are prone to internal corrosion because of stagnant flow as the water emulsifies with crude drops out of the petrochemical solution and disperses onto the surface of the metal. Water is the primary corrosion agent of internal corrosion. The stagnant particles will deposit on the metal surface, and this causes pitting corrosion and consequently leads to leakage of the metal. In a gas system, the low-velocity pipe section will cause the accumulation of condensate, and due to the increase of humidity, the corrosion problem becomes even more accentuated [3]. Deadlegs corrosion in piping systems contributes to the major percentage of internal damages in the oil and gas process industry. According to the American Petroleum Institute Pipeline Performance Tracking System (PPTS) Operator Advisory 2009-5, 85% of

deadleg incidents over the period from 2003 to 2007 were caused by internal corrosion. As deadlegs are connected to main pipes, identifying deadlegs and preventing incidents due to their internal corrosion has become vital to maintaining the integrity of the piping system.

The challenge which most of the plant integrity engineers are facing is the inaccessibility of some of the deadlegs as they are commonly found at the dead end of the piping system, where thickness measurements cannot be obtained. Since it is hard to get at these deadlegs, monitoring of and maintaining these stagnant pipe segments are thus not easy. Therefore, a predictive numerical method for corrosion rate estimation is essential for those inaccessible deadlegs. We propose using Weibull density regression for inaccessible deadlegs corrosion extrapolation. The regression method models the relationship between field corrosion rates of deadlegs and corrosion rates of their main lines. The model also considers the influencing factors such as the geometry of deadlegs, the velocity of the main line, and deadleg length-to-diameter (L/D) ratio, as these factors are considered influential on the corrosion rate of deadlegs [5-9]. These factors are included in the Weibull density regression by taking them into account as independent variables of the model. Weibull distribution is used because it is flexible enough to model different types of datasets.

3.1.2 Previous works on deadlegs corrosion

Many kinds of research discuss deadlegs and their internal corrosion. Habib et al. [5-7] analysed the effect of geometry of deadlegs on corrosion in a crude oil and water solution piping system. They described how the size of the stagnant fluid region increases with the increase of length-to-diameter (L/D) ratio and is influenced by the geometry. Thermal-fluid analyses of deadlegs in pharmaceutical water systems were discussed in detail with the effect of various factors such as flow rate, material, length and size of the pipe [8-9]. Ding et al. [10] investigated the complex flow, oil/water separation and the relation between fluid flow and the water concentration of vertical deadlegs based on the solution of algebraic slip mixture model. A technique of the

Process Hazard Analysis (PHA) model is created for recognising and addressing deadlegs and related corrosion issues that can be utilised to refresh corporate PHA systems to be more effective in preventing corrosion related incidents [11]. Bensman [12] discussed the line flushing method that is one of the mitigation techniques used to manage internal corrosion of deadlegs and provided an overview of the available models to determine the minimum velocity for flushing. The authors in [13] show that a numerical model that has the capability to model flow and mixing zone in a deadleg that modelled mass transfer rate can be used to estimate corrosion of carbon steel due to dissolved oxygen and microorganisms in deadleg under mass transfer limited conditions.

3.1.3 Previous works on regression analyses for sample corrosion

Regression analysis has been used widely in corrosion applications for various plant assets, ranging from high-temperature equipment and concrete structures to oil and gas pipelines and offshore floating structures. There are some pieces of literature regarding both the mathematical model and regression analysis. Barrett [14] used multiple linear regression analysis to estimate hot corrosion attack for nickel-based cast turbine alloys; different methods of transforming independent variables were shown to generate a higher R-squared value for a better fit of the model to the corrosion test samples. A regression method was used by Kung et al. [15] to identify the relationship between the corrosion rates of several alloys that were commonly used in the lower furnace of utility boilers and the corrosion damage factors such as H₂S, Cr concentration in the alloy, and temperature. Garud et al. [16] evaluated flow-accelerated corrosion in power plants using a regression technique for the spatial and temporal evolution of wall loss which can predict the wear rates of components and their associated uncertainties. Wen et al. [17] proposed the use of a support vector regression method that combined with particle swarm optimization to predict the corrosion rate of 3C steel in different seawater environments; it has shown a promising result such that the model could be used for real-time corrosion tracking of steel in an uncertain seawater environment. Ossai [18] analysed the effect of operating parameters such as temperature, CO₂ partial pressure flow rate, etc. on the

corrosion rate of oil and gas wellheads by using multiple linear regression. Siamphukdee et al. [19] used regression methods for sensitivity analysis to assess the influencing factors for corrosion in reinforced concrete structures and found that those structures are highly sensitive to corrosion duration time, concrete resistivity, and chloride content. Suleiman et al. [20] investigated the process parameters that significantly influence the corrosion rate of mild steel in 0.5M sulphuric acid. Linear regression is used to analyse the effect of varied temperature, time, and concentration of inhibitor; corrosion rate was found to be reduced if the concentration of inhibitor and time were increased. A relationship between soil pH and soil resistivity on corrosion growth rate of carbon steel was studied and determined by Anyanwu et al. [21] using multiple regression analysis. Sodiki & Ndor [22-23] conducted corrosion experiments on steel, brass and aluminium that were exposed to certain external factors; the laboratory results were then fitted with a regression model where the response variable was the corrosion extent, and the external factors were taken as independent variables. Besides these, there are many studies about the corrosion rate modelling for pipelines using regression analyses. Qiu & Orazem [24] developed a model that coupled a boundary-element forward model with a nonlinear regression algorithm to interpret pipeline survey data. Mior et al. [25] used a power law to model the time dependence of metal loss in underground pipelines where parameters of the model such as metal loss constant and corrosion growth pattern were modelled using multiple linear regression. Velázquez et al. [26] proposed multivariate regression method for pitting corrosion in buried oil and gas pipelines that considered the age of pipeline, soil properties, and pipe properties as the covariates in the model. El-Abbasy et al. [27] developed models using regression technique for predicting the condition oil and gas pipelines based on actual inspection samples that account for the influence of various factors including corrosion.

3.1.4 Piping deadlegs corrosion data

We used data from various piping systems in an oil and gas plant that carry different products and have different design and operation properties. To protect the sensitive information of the source of data, the name of the company and locations are not

disclosed. These piping systems consist of the crude oil system, water injection system and fuel gas system. Thickness measurements were collected on a periodic basis as part of the preventive maintenance. An ultrasonic testing method (UT) is used to measure the wall thickness of identified critical areas or in another term, gauge point. Thickness was gauged at 3, 6, 9, and 12 o'clock for each gauge point and minimum measurement of each gauge point was recorded. Corrosion rates of those piping systems are estimated using the linear regression method [28] where the minimum thicknesses at gauge point at the last two inspections were taken, and the prediction is a linear extrapolation.

More than 100 deadleg locations were identified from P&IDs, but only 1/3 of those deadlegs are accessible for UT inspection. The company intended to eliminate or minimise as many deadlegs as they could with an optimised cost and to remove those deadlegs by prioritising each deadleg removal; removing deadlegs requires shutting down the piping system. The challenge they encountered is the inaccessibility of some of the deadlegs as thickness measurements are not obtained, and a predictive numerical method for corrosion rate is needed for those inaccessible deadlegs.

3.2 Weibull density regression

3.2.1 Weibull distribution

There are two forms of the Weibull distribution distinguished by the presence of either two or three parameters [29]; for our purposes we used the 2-parameter form for our density regression model. As its name implies, 2-parameter Weibull consists of 2 parameters, the scale parameter η and shape parameter λ .

The probability density function (pdf) is given by

$$f(y; \lambda, \eta) = \frac{\lambda}{\eta} \left(\frac{y}{\eta}\right)^{\lambda-1} e^{-\left(\frac{y}{\eta}\right)^\lambda}, y > 0 \quad (3.1)$$

where

$\eta = \text{scale parameter or characteristic life}$

$\lambda = \text{shape parameter}$

and the cumulative distribution function (CDF) is given by

$$F(y; \lambda, \eta) = 1 - e^{-\left(\frac{y}{\eta}\right)^\lambda}. \quad (3.2)$$

The quantile function of this Weibull distribution $Q(p; \lambda, \eta)$ is given by

$$Q(p; \lambda, \eta) = \eta \cdot [-\log(1 - p)]^{\frac{1}{\lambda}}. \quad (3.3)$$

The median \check{T} of this Weibull distribution is given by

$$\check{T} = \eta \cdot [\log(2)]^{\frac{1}{\lambda}}. \quad (3.4)$$

3.2.2 Gamma transformation proportional hazards (Gt-PH) model

We will introduce the Weibull regression via a general frame of density regression models which is based on the proportional hazards family. The regression is introduced via the scale parameter of the proportional hazards distribution [4]. Assume that the response variables Y have a probability distribution belonging to the proportional hazards family with baseline cumulative distribution function G . In the

method [30] based on proportional hazards family-based, the Weibull distribution is written as

$$F(y; \lambda, \theta) = 1 - [1 - G(y; \lambda)]^\theta = 1 - \exp(-\theta y^\lambda) \quad (3.5)$$

where θ is the proportionality parameter with $\theta = 1/\eta^\lambda$ and $G(y; \lambda)$ is dependent only on λ . For the 2-parameter Weibull distribution $G(y; \lambda) = 1 - \exp(-y^\lambda)$, and λ is the shape parameter.

In general, when covariate $x = (x_1, x_2, \dots, x_p)$ information is available, regression model is usually introduced via the scale parameter θ , see Cox [49]. The log linear model is used to model $\log(\theta)$ as the linear regression function of x , thus

$$\log(\theta(x)) = x^T \beta \quad (3.6)$$

where x^T is the transpose of x and β is the regression coefficient vector.

Coefficients β are estimated using a combination of transformation of gamma-distributed random variable and ordinary least squares estimate (OLS). This method is different from the conventional parametric regression fitting such as maximum likelihood estimate (MLE). The method derives the best linear unbiased estimators (BLUE) [31] of β with the least variance of the estimate among linear unbiased estimates. This method is called gamma transformation proportional hazards (Gt-PH) density regression.

Given the response Y and covariates x , $\{x_i, Y_i\}_{i=1}^n$: let $\theta_i \equiv \theta(x_i)$ and $S_i = -\log(\bar{G}(Y_i; \lambda))$ then $2\theta_i S_i \sim \chi^2(2)$. The distribution $\chi^2(2)$ is a special case of a gamma distribution for a gamma random variable, Γ . The gamma distribution is the maximum entropy probability distribution of Γ for which $E(\log(\Gamma)) = \psi(\text{shape} - \text{parameter}) - \log(1/(\text{scale} - \text{parameter}))$ is fixed, where $\psi(t) = d \log(\Gamma(t))/dt$ is the digamma function [4][32]. Therefore,

$$E[\log(S_i) + \log(\theta_i)] = \psi(1)$$

$$\text{Var}[\log(S_i) + \log(\theta_i)] = \psi'(1)$$

where $\psi'(t) = d^2 \log(\Gamma(t))/dt^2$, $\psi(1) = -\gamma$ with the Euler-Mascheroni constant $\gamma \approx 0.5772$ and $\psi'(1) = \frac{\pi^2}{6}$.

Let

$$U_i = -\log(S_i) - \gamma. \quad (3.7)$$

Then we have the standard linear regression function

$$E(U_i) = x^T \beta$$

$$\text{Var}(U_i) = \psi'(1).$$

It is required to estimate the shape parameter before we move on to the next step of the calculation. McCool [29] has provided a proven likelihood function for shape parameter estimation. Qiao and Tsokos explains McCool's theorem in [33]. They showed that, given the response data Y , the shape parameter λ of the Weibull distribution can be estimated by solving the equation (3.8) using the Newton-Raphson method:

$$\frac{\lambda}{n} \sum_{i=1}^n \log(y_i) \sum_{i=1}^n y_i^\lambda - \lambda \sum_{i=1}^n y_i^\lambda \cdot \log(y_i) + \sum_{i=1}^n y_i^\lambda = 0. \quad (3.8)$$

Once λ is known, by using the equation (3.7) we obtain the observed vector U from $U_i = -\log(\bar{G}(Y_i; \lambda)) - \gamma$, where $\bar{G}(y; \lambda) = e^{-y^\lambda}$, and $\gamma \approx 0.5772$. Then the coefficients of density regression β are estimated via a linear regression model with data (x, U) .

The density regression model (3.6) can be written as

$$\log(\theta)_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p, \quad i = 1, 2, 3 \dots n$$

where β_0 is the intercept. Once the regression for $\log(\theta)$ or θ is obtained, we move to obtain equations (3.2), (3.3), and (3.4) to find the regression functions for CDF, quantile function and median; we transformed the proportional parameter θ to become scale parameter with the equation (3.9)

$$\eta = \left(\frac{1}{\theta}\right)^{\frac{1}{\lambda}}. \quad (3.9)$$

3.3 Bayesian hierarchical regression analysis

The essential characteristic of Bayesian methods is their explicit use of probability to quantify uncertainty in inferences based on statistical data analysis [34]. The Bayesian approach also allows the model to take into account the historical information or expert judgement, where it is always called the prior information, to update the model to give a more accurate outcome. More detail about the advantages of applying the Bayesian method to data analysis can be found in [34-36].

Various statistical applications incorporate different parameters that can be seen as related or associated by one means or another by the structure of the problem, inferring that a joint probability model for these should mirror the dependence among them. For the most part, the objective of hierarchical modelling is to determine the extent to which factors measured at different levels or groups influence an outcome using a typical regression modelling framework. Ordinary regression, however, is inappropriate, because of the lack of independence of errors for observations within groups. Thus, an alternative model must be developed to compensate for this lack of independence.

Lynch [36] explains that, with few parameters in nonhierarchical models, they generally cannot fit large datasets accurately, whereas, with many parameters, they tend to overfit such data in the sense of producing models that fit the existing data well but lead to inferior predictions for new data. By comparison, hierarchical models

can have enough parameters to fit the data well, while using a population distribution to structure some dependence on the parameters, thereby avoiding problems of overfitting.

Bayes' Theorem is often expressed as:

$$p(\theta|data) \propto p(data|\theta) \times p(\theta)$$

where $p(\theta|data)$ is the posterior, $p(data|\theta)$ is the likelihood, and $p(\theta)$ is the prior. This equation itself uncovers a basic hierarchical structure in the parameters since it says that a posterior distribution for a parameter is equivalent to a conditional distribution for data under the parameter (second level) multiplied by the marginal (prior) probability for the parameter (first level). In other words, the posterior distribution is the prior distribution weighted by the observed information.

The hierarchical structure of these parameters does not need to be stopped at a higher level; it can, in theory, continue to infinity levels. Assume we have J observations inside each of G groups: $y_{11}, \dots, y_{J1}, y_{12}, \dots, y_{J2}, \dots, y_{1G}, \dots, y_{JG}$, and that the data are distributed within groups according to some distribution Q with parameter θ , but that each group has its own parameter (θ_g). Thus:

$$y_{ig} \sim Q(\theta_g).$$

We assume further that parameters θ_g are from a distribution W with parameter ξ , so:

$$\theta_g \sim W(\xi)$$

ξ is called a "hyper-parameter".

Lastly, assume ξ has some vague distribution, for example a uniform distribution:

$$\xi \sim U(-100, 100).$$

Then the posterior distribution for all unknown parameters would be:

$$p(\xi, \theta | y) \propto p(y | \theta, \xi) p(\theta | \xi) p(\xi).$$

We may be interested only in the marginal distribution for ξ rather than the posterior distributions for the group level parameters θ_g :

$$p(\xi | y) \propto \int_{-\infty}^{\infty} p(y | \theta, \xi) p(\theta | \xi) p(\xi) d\theta.$$

We perform Markov chain Monte Carlo (MCMC) simulation to solve this integration stochastically as we sample from the conditional posterior distributions for each parameter.

3.3.1 Markov Chain Monte Carlo (MCMC) methods

It is hard to use the analytic method to evaluate and analyse the solution of the posterior distribution. Markov chain Monte Carlo methods are utilised to simulate the posterior so that it can be analysed. The results can then be used to make inferences about the models and parameters. MCMC had been known for a considerable length of time before its implications for Bayesian statistical modelling were fully recognised. Formally, a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \dots$ forms a Markov chain if, for all t , the distribution of the $t + 1^{th}$ variable in the sequence is given by

$$X^{(t+1)} \sim p_{trans}(x | X^{(t)} = x^{(t)}), \quad (3.10)$$

that is, conditional on the value of $X^{(t)}$, the distribution of $X^{(t+1)}$ is independent of all other preceding values, $X^{(t-1)}, \dots, X^{(0)}$ [37]. The transitional distribution of the Markov chain $p_{trans}(x | X^{(t)} = x^{(t)})$ defines the conditional probability of current value of the chain to move to a specific new value. The marginal distribution of $X^{(t+1)}$ will converge to a unique stationary distribution as $t \rightarrow \infty$ under a fairly general regularity condition. In simple terms, this says that although each variable in the chain depends directly on the value that came before it, in the end we reach a point such that all subsequent values are distributed marginally according to the same fixed distribution,

which, extremely importantly, is independent of the start value $X^{(0)}$. Another way to say, in the end, overlooks where it began and fits in with an underlying "equilibrium" distribution.

3.3.1.1 Gibbs sampling

Gibbs sampling is one of the most simple and widely used MCMC methods used in Bayesian statistics. Its use in Bayesian statistics was introduced by Gelfand and Smith [48], although it was developed and used in physics before 1990. Gibbs sampling sweeps through each variable to sample from its conditional distribution by fixing the remaining variables to their current values to generate a posterior distribution. According to Lynch [36], the algorithm of generic Gibbs sampler is given by:

Algorithm 1 Gibbs sampler

Assign a vector of starting values, S , to the parameter vector: $x^{(0)} = S$

for iteration $i = 1, 2, \dots$ **do**

$$x_1^{(i)} \sim p\left(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \dots, X_D = x_D^{(i-1)}\right)$$

$$x_2^{(i)} \sim p\left(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \dots, X_D = x_D^{(i-1)}\right)$$

\vdots

$$x_D^{(i)} \sim p\left(X_D = x_D | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \dots, X_{D-1} = x_{D-1}^{(i)}\right)$$

end for

In Algorithm 1, sampling is not done directly from the posterior distribution itself. The samples are simulated by sweeping through all the conditionals of posterior with one random variable at a time. The algorithm is initialised with random values so the early iterations of the simulation based on the algorithm may not necessarily be representing the posterior distribution. Nevertheless, according to the theory of

MCMC, Algorithm 1 generates a stationary distribution of the samples which will be the target joint posterior of interest.

3.3.2 Hierarchical stages

The levels of a hierarchical Bayesian model are called stages. Regardless of the model's complexity, the first stage always is the likelihood that is closest to the data. The second stage of a hierarchical model consists of probability distributions for some or all of the parameters that appeared in the likelihood. The parameters of the priors at the second stage of a hierarchical model are more unknown than the first stage. The third (and last) stage specifies prior distributions for the parameters at the second stage, called hyper-priors. The use of "hyper" in the third stage for prior distribution of parameter is to differentiate the priors in the third stage from the second; they arise particularly in the use of conjugate priors.

3.3.2.1 First stage

The first stage of the model provides the distribution of the data given certain model parameters. Let u_{ik} represent the transformed observed value of U in equation (3.7) of category i , $k = 1, \dots, n_i$, and $q = 1, \dots, p$. The likelihood is expressed as:

$$u_{ik} | \alpha_{0i}, \alpha_{qi}, \tau_u^2 \sim N \left(\mu_{\alpha_i}, \frac{1}{\tau_u^2} \right) \quad (3.11)$$

where τ_u^2 is the precision of the points around the group-specific regression line, α_{0i} is the intercept, $\alpha_{1i}, \dots, \alpha_{pi}$ are coefficients for category i , μ_{α_i} is the expected value of regression function, n_i is the number of values for category i , and q is the number of coefficients. This model assumes that all categories share the same precision parameter.

3.3.2.2 Second stage

The second stage prior consists of independent normal densities for the category-specific intercepts and coefficients:

$$\alpha_{0i} | \beta_0^b, \frac{1}{\tau_{\alpha_0}^2} \sim N\left(\beta_0^b, \frac{1}{\tau_{\alpha_0}^2}\right)$$

$$\alpha_{qi} | \beta_q^b, \frac{1}{\tau_{\alpha_q}^2} \sim N\left(\beta_q^b, \frac{1}{\tau_{\alpha_q}^2}\right).$$

We assumed that the category-specific intercepts α_{0i} and α_{qi} are draws from normal densities. Here the precision parameter $\tau_{\alpha_0}^2$ captures the variability among intercepts for different categories. If this precision is small then the intercepts at different categories tend to be very different from one another. Similarly, the precision parameter $\tau_{\alpha_q}^2$ captures how different the slopes for different categories are. As expected in a hierarchical model, the parameters in the second-stage priors are all unknown quantities that we wish to estimate.

3.3.2.3 Third stage

The third stage consists of prior distributions for all the remaining unknown parameters. The conventional, semi-conjugate priors are as follows:

$$\beta_0^b \sim N\left(\mu_0, \frac{1}{\tau_0^2}\right)$$

$$\beta_q^b \sim N\left(\mu_q, \frac{1}{\tau_q^2}\right)$$

$$\tau_u^2 \sim G(a_u, b_u)$$

$$\tau_{\alpha_0}^2 \sim G(a_{\alpha_0}, b_{\alpha_0})$$

$$\tau_{\alpha_q}^2 \sim G(a_{\alpha_q}, b_{\alpha_q}).$$

The prior on τ_u^2 can be very vague, because the observed data values make estimation of this precision parameter quite easy. However, $\tau_{\alpha_0}^2$ and $\tau_{\alpha_q}^2$ are the precisions of unknown and unobservable parameters. Without prior knowledge, these parameters should be specified to make the hyper-prior vague (e.g. say $\mu_0 = 0, \frac{1}{\tau_0^2} = 10,000$; $\mu_q = 0, \frac{1}{\tau_q^2} = 10,000$) [36]. The hyper-prior distributions for the population variance and the error variances are gamma distributions, with parameters $a_u = 0.001, b_u = 0.001, a_{\alpha_0} = 0.001, b_{\alpha_0} = 0.001, a_{\alpha_q} = 0.001, b_{\alpha_q} = 0.001$.

3.4 Application

3.4.1 Weibull density regression

Here we applied the Weibull density regression method to the corrosion rate data of deadlegs. We identified the covariates for the regression model as the following (Table 3.1):

Covariates	Categories
System Type	Gas System
	Water System
	Oil System
Deadleg Design	T-joint
	Branch
Orientation	Horizontal
	Vertical
	Mix Orientation (Horizontal + Vertical)
Length	
Diameter	
Log(Flow Rate)	
Main Line Corrosion, CR ^{ml}	

Table 3.1. Covariates that are used in the Weibull density regression.

We modelled main line flow rate by transforming it into its natural logarithm because the variance of flow rate is large; it ranges from 2.5 to 600000m³/hr. The flow of the main line is influential to microbiological induced corrosion in the deadleg and tuberculation from deadleg opening to a certain length of deadleg [13]. We factorised system type, deadleg design and orientation into categories and introduced dummy variables for them [38]. For water system type, we assigned 1 to the water system and 0 to oil system; while for oil system we assigned 1 to oil system and 0 to water. One of the categories has to be the base category by dummy variable approach; hence we do not include the gas system as a coefficient in the equation. The same is done for deadleg design and orientation covariates. We believe there is a correlation between main lines and deadlegs corrosion rates; hence we included it into the regression model. There are three types of orientation: horizontal, vertical and mix orientation. Mix orientation is for the deadlegs which are oriented partially horizontal and vertical. Based on many pieces of research about deadlegs and corrosion [5-7], the size of a stagnant fluid region increases with the increase of deadleg length-to-diameter ratio L/D, and L/D has been considered an important factor in a stagnant

column of a pipe. However, we separate length of deadleg and diameter by making them two different independent variables in the regression equation in order to see the effect of diameter of deadleg on corrosion as well.

The target regression model for deadleg corrosion rate is as the following (3.12):

$$\begin{aligned} \log(\theta) = & \beta_0 + \beta_{1,1} \cdot [\text{Water System}] + \beta_{1,2} \cdot [\text{Oil System}] + \beta_2 \\ & \cdot [\text{Branch}] + \beta_{3,1} \cdot [\text{Vertical}] + \beta_{3,2} \cdot [\text{Mix Orientation}] \\ & + \beta_4 \cdot [\text{Length}] + \beta_5 \cdot [\text{Diameter}] + \beta_6 \cdot [\log(\text{Flow Rate})] \\ & + \beta_6 \cdot [\text{CR}^{ml}]. \end{aligned} \quad (3.12)$$

Given the deadleg corrosion rate data, the shape parameter λ estimated by solving the equation (3.8) using Newton-Raphson method is 2.364. The coefficients derived from the density regression are shown in the following table:

Coefficients		Value	Std. Error
Intercept, β_0		6.8038	0.650615
System Type	Water System, $\beta_{1,1}$	0.7330	0.583899
	Oil System, $\beta_{1,2}$	1.8565	0.776923
Deadleg Design	Branch, β_2	-0.6366	0.482866
Orientation	Vertical, $\beta_{3,1}$	-0.7128	0.488898
	Mix (Horizontal + Vertical) , $\beta_{3,2}$	-0.1236	0.552269
Length, β_4		-0.0013	0.003148
Diameter, β_5		-0.2638	0.179658
Log(Flow Rate) , β_6		0.1055	0.056706
Main Line Corrosion, β_7		-14.2714	2.272186

Table 3.2. Coefficients estimated in density regression

We plotted a chart to see the trend of the model corrosion rate against the field corrosion rate from gauge points as shown in Figure 3.1. Field corrosion rates are the corrosion rates of the deadlegs which are estimated using the linear regression

method where the minimum thicknesses at gauge points at the last two inspections were taken; a gauge point is the location where thickness measurement is gauged. We plotted the median and the 0.95 quantile against the field corrosion rate.

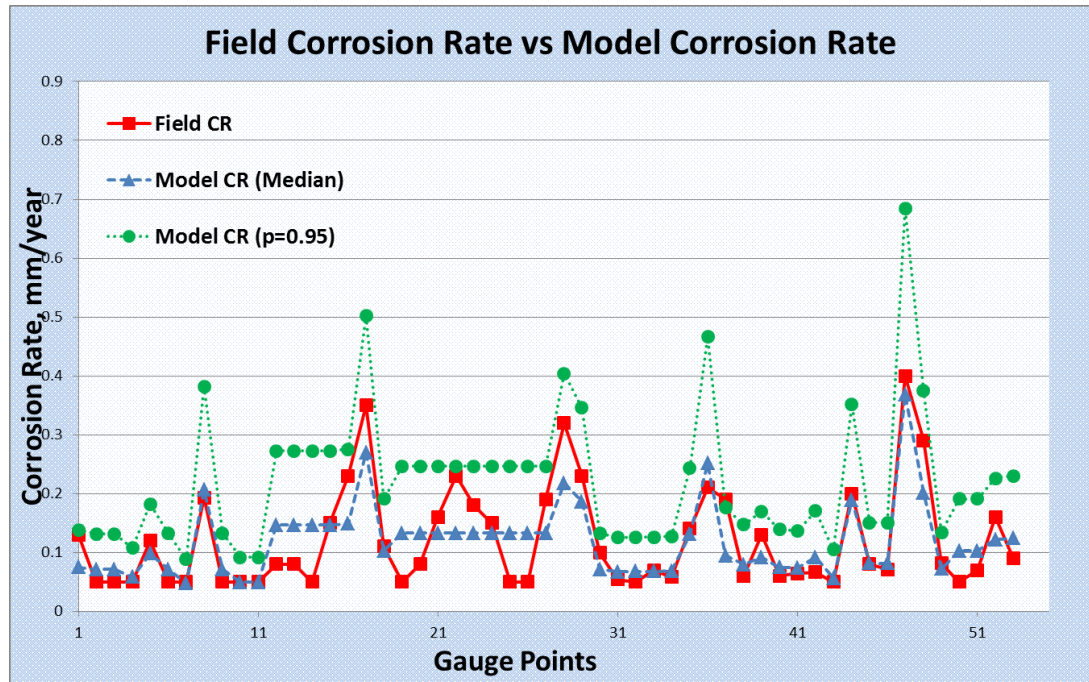


Figure 3.1. Corrosion rate from the field vs model

Figure 3.2 shows the effect of different influencing factors on the corrosion rate of deadlegs. We plotted median corrosion rate against main line flow rate by setting other factors as constant; then we plotted median corrosion rate against different geometry for the different piping system. The same are done for length and diameter of deadleg.

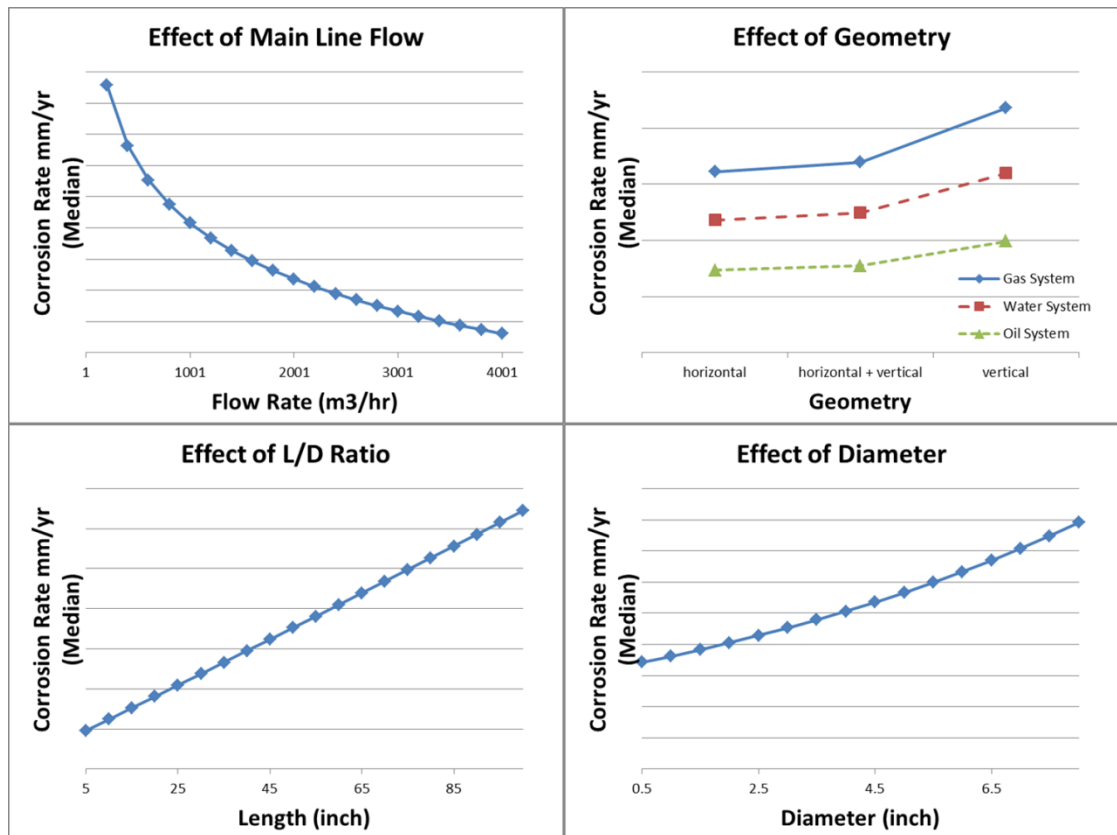


Figure 3.2. Effect of factors on deadleg corrosion

Based on the results of density regression analysis, we can check and compare the effect of each factor on the corrosion of deadlegs in different piping systems. In Figure 3.1, the model corrosion rate has a similar trend as the field corrosion rate. Median corrosion is close to the field corrosion, but nearly half of the corrosion rates are lower than the field corrosion rate. Most of the model values of the 0.95 quantile are higher than those from field data, but it may be overly conservative by using only the 0.95 quantile for the extrapolation of corrosion rates of all deadlegs. Different quantile values are used for different scenarios. An assumption is made here that corrosion rate of a deadleg is higher than the corrosion rate of its parent line. We used the 0.80 quantile for vertical deadleg as most of the modelled corrosion rates at this quantile are higher than main line and field deadleg corrosion rates; this is shown in Figure 3.3. For horizontal deadlegs, we used the 0.7 quantile to model their corrosion rates while for mixed-oriented deadlegs we used the 0.95 quantile as illustrated in Figure 3.4 and Figure 3.5 respectively. Based on the assumption that deadleg corrosion

should be higher than the main line corrosion rate, if the modelled corrosion rate is lower than main line's, higher quantiles will be used.

Figure 3.2 shows that an increase of flow rate decreases the rate of corrosion of deadlegs. Corrosion is found to be more severe in vertically-oriented deadlegs than other deadlegs. For the increase of deadleg length where its diameter is the same, we can see the increase of corrosion rate where with the diameter is unchanged, and increase of deadleg increases the length-to-diameter ratio. The diameter also plays a significant role in deadleg corrosion. The larger diameter of deadleg pipe will have a higher corrosion rate. These results are in line with the research in [5-9, 13].

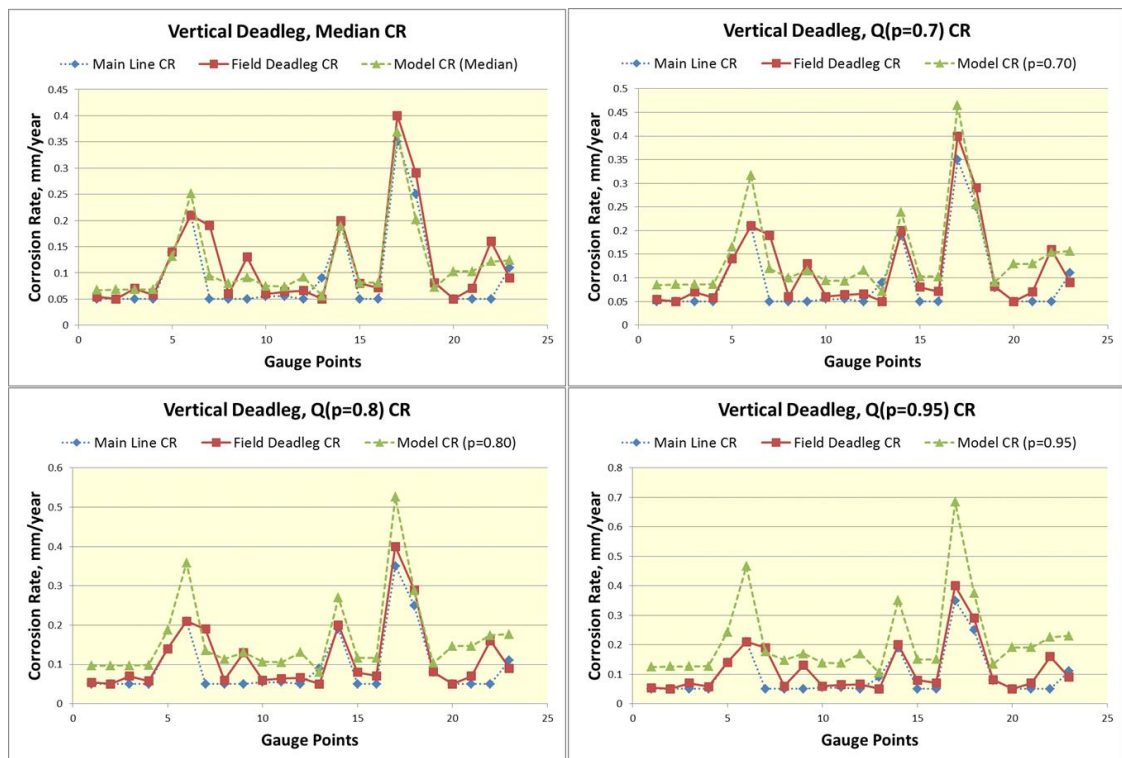


Figure 3.3. Model corrosion rates at different Weibull quantiles for vertical deadlegs.

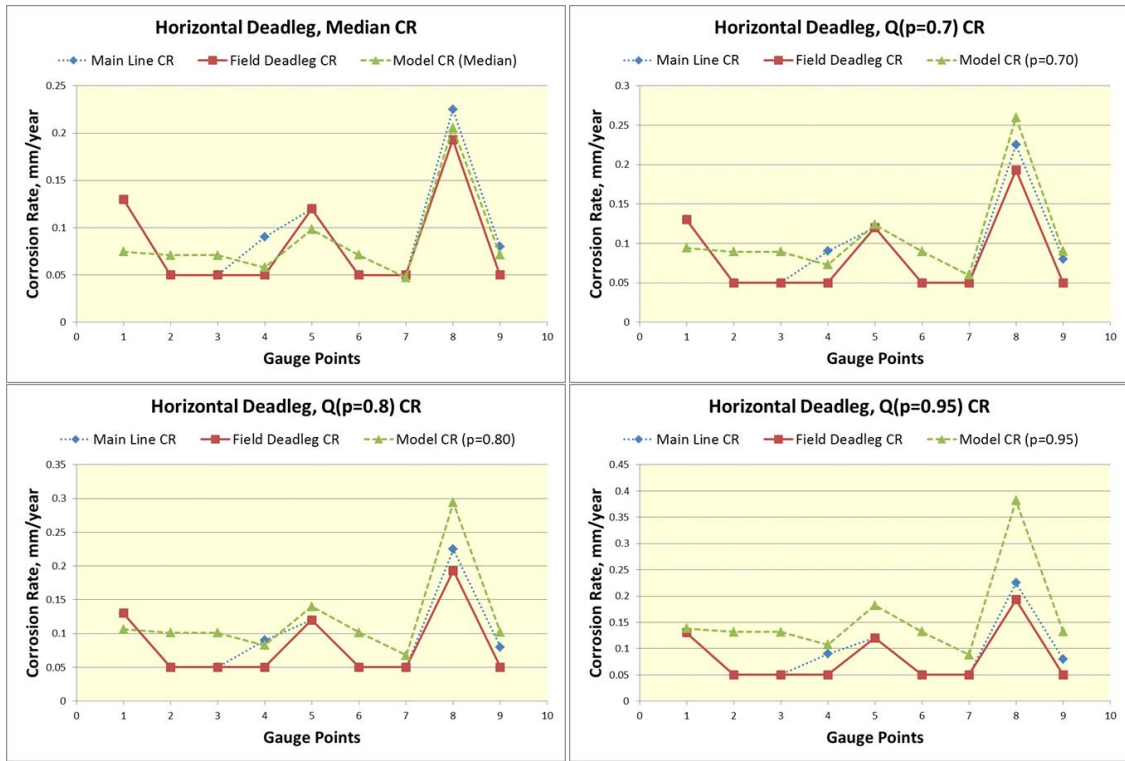


Figure 3.4. Model corrosion rates at different Weibull quantiles for horizontal deadlegs.

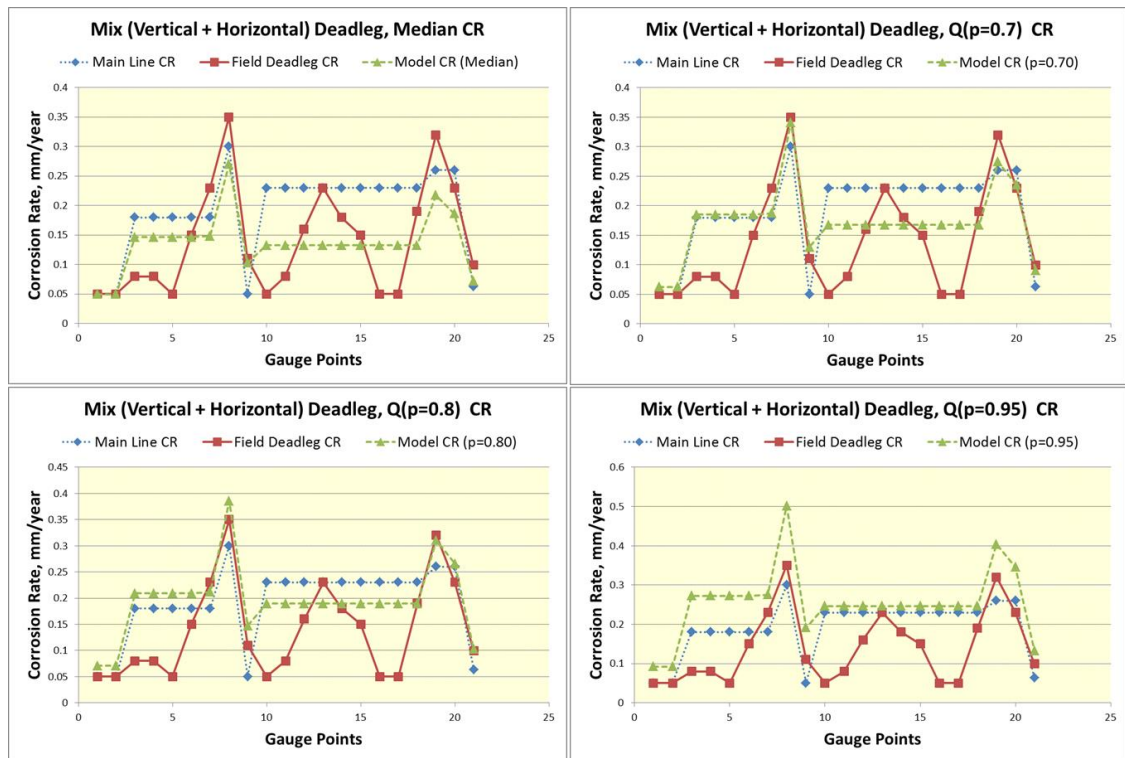


Figure 3.5. Model corrosion rates at different Weibull quantiles for mix (vertical + horizontal) deadlegs.

3.4.2 Bayesian hierarchical regression analysis

The application shown in section 3.4.1 is generally considered as frequentist approach where the conclusion is drawn from a sample of deadlegs corrosions by emphasizing the frequency or proportion of the original data. As previously discussed, the deadleg corrosion dataset consists of corrosion data from the different piping system; they are categorised into gas, oil, and water piping systems. We used dummy variables to handle the categories in the ordinary regression model. Although a dummy variables approach seems to do the work fine, it introduces several problems as well at the same time that may distort the entire regression analysis. According to Holgersson et al. [38], few reasons were stated of avoidance of using dummy variables:

- It is argued that the hypothesis tests for dummy variables are only valid if all different categories have equal variances. This is unlikely to be the case in real-world applications.
- Dummy variables may incur multicollinearity in the analysis thereby leading to a distorted result.
- The number of observations must be equal in all different categories for the dummy variable method to be invariant to the coding of zeros and ones (e.g. 50% men and 50% women).

Holgersson et al. proposed a separate regression for each category of data. However, due to the small corrosion dataset of deadlegs we obtained, we ruled out the option of splitting the dataset into multiple datasets based on types of piping system.

Due to the reasons mentioned above, we used the Bayesian methods to analyse our data. The nature of our data are in groups (gas, water, and gas piping deadlegs). We want to reduce the number of dummy variables in our model; therefore the common problems imposed by dummy variables as discussed by Holgersson et al. can be improved. Besides that, the Bayesian hierarchical models are flexible tools for combining information and partial pooling of inferences thus giving more efficient estimates of the parameters in each group. If given more data in the future, the model can be updated by taking the current estimates as the prior information. Hence the models can be evolved over time to become more accurate by taking new information into consideration.

We ran the Bayesian hierarchical regression analysis by using OpenBUGS software. The output from our OpenBUGS sampler for regression is presented below in Table 3.3.

We denoted the simulated coefficients for Bayesian regression by β^b to differentiate them from ordinary regression estimates. It is apparent that the MCMC sampling yielded smaller standard errors for all the parameters, which implicates that these parameters are relatively less spread in the sampling distribution. Each different piping system has its own respective intercept, $\beta_{0,1}^b$ for gas system, $\beta_{0,2}^b$ for water

system and $\beta_{0,3}^b$ for oil piping system. Other coefficients are shared for all different types of piping system:

For gas system:

$$\begin{aligned} \log(\theta) = & \beta_{0,1}^b + \beta_1^b \cdot [\textit{Branch}] + \beta_{2,1}^b \cdot [\textit{Vertical}] + \beta_{2,2}^b \\ & \cdot [\textit{Mix Orientation}] + \beta_3^b \cdot [\textit{Length}] + \beta_4^b \cdot [\textit{Diameter}] \\ & + \beta_5^b \cdot [\log(\textit{Flow Rate})] + \beta_6^b \cdot [\textit{CR}^{ml}]. \end{aligned} \quad (3.13a)$$

For water system:

$$\begin{aligned} \log(\theta) = & \beta_{0,2}^b + \beta_1^b \cdot [\textit{Branch}] + \beta_{2,1}^b \cdot [\textit{Vertical}] + \beta_{2,2}^b \\ & \cdot [\textit{Mix Orientation}] + \beta_3^b \cdot [\textit{Length}] + \beta_4^b \cdot [\textit{Diameter}] \\ & + \beta_5^b \cdot [\log(\textit{Flow Rate})] + \beta_6^b \cdot [\textit{CR}^{ml}]. \end{aligned} \quad (3.13b)$$

For oil system:

$$\begin{aligned} \log(\theta) = & \beta_{0,3}^b + \beta_1^b \cdot [\textit{Branch}] + \beta_{2,1}^b \cdot [\textit{Vertical}] + \beta_{2,2}^b \\ & \cdot [\textit{Mix Orientation}] + \beta_3^b \cdot [\textit{Length}] + \beta_4^b \cdot [\textit{Diameter}] \\ & + \beta_5^b \cdot [\log(\textit{Flow Rate})] + \beta_6^b \cdot [\textit{CR}^{ml}]. \end{aligned} \quad (3.13c)$$

Coefficients		Value	Std. Error
Intercept	Gas System, $\beta_{0,1}^b$	6.280769	0.00114
	Water System, $\beta_{0,2}^b$	6.692429	0.001628
	Oil System, $\beta_{0,3}^b$	7.484044	0.001878
Deadleg Design	Branch, β_1^b	-0.188750	0.0006321
Orientation	Vertical, $\beta_{2,1}^b$	-0.308220	0.0006847
	Mix (Horizontal + Vertical), $\beta_{2,2}^b$	0.034518	0.0005398
Length, β_3^b		-0.000739	0.000005645
Diameter, β_4^b		-0.137906	0.0002886
Log(Flow Rate), β_5^b		0.083375	0.0001044
Main Line Corrosion, β_6^b		-13.085336	0.004409

Table 3.3. Estimated coefficients in the Bayesian hierarchical modelling.

3.4.3 Convergence checking

In order to make sure the parameters estimated by sampling process in MCMC can be used for predictions, assessment of convergence is essential by using some diagnostic tools. As long as Markov chains of the sampling converge, it is safe to say that the multicollinearity issue is addressed and the estimated parameters can be used for the model. Cowles and Carlin [39] reviewed thirteen convergence diagnostics and described the basic theory and practical usage of each. We have selected three diagnosis methods for our model convergence checking; they are trace plots, autocorrelation function plots, and Gelman-Rubin diagnostics.

Figure 3.6 shows the trace plots for all the parameters from two chains run for 200000 iterations. First 50000 iterations have been removed which are regarded as "burn-in". Regularly, initial samples are not entirely valid because the Markov chain has not stabilised. With the absence of clear upward or descending patterns, the trace plots suggest the convergence of the simulation.

The Markov chain usually produces dependence in the samples. Autocorrelation is a quantitative measure of this dependence and autocorrelation function plots are used for these measurements. In Figure 3.7, the chains for all parameters are converged to 0 in the end. This is consistent with convergence of the Markov chains sampling.

Lastly, we performed the Gelman-Rubin convergence diagnostic. The Gelman-Rubin diagnostic evaluates MCMC convergence by analysing the difference between multiple Markov chains. The estimated between-chains and within-chain variances are compared to assess the convergence for each model parameter. Significant differences between these variances indicate non-convergence [40-41]. The Gelman-Rubin diagnostic returns a factor which is called the potential scale reduction factor (PSRF) that can be interpreted as a convergence diagnostic. We can safely conclude that each simulation is close to the target distribution and nearly converged if the PSRF value is close to 1 [40]. Table 3.4 presents the PSRF for each parameter. As the PSRFs for all parameters are equal to 1, this manifests the simulations for all parameters are approximately converged. Figure 3.8 shows the visual confirmation of PSRF.

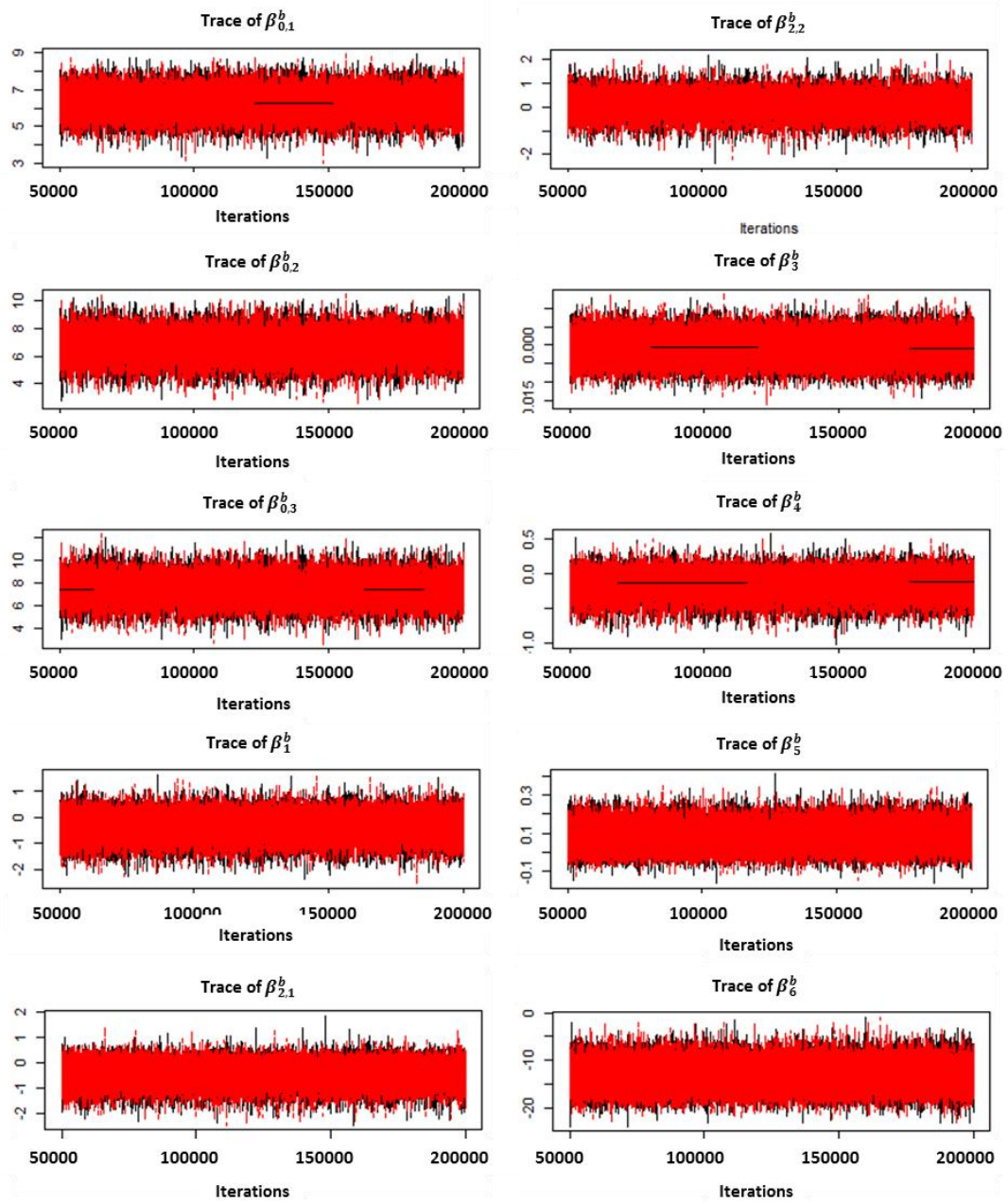


Figure 3.6. Trace plots for 200000 iterations of the model.

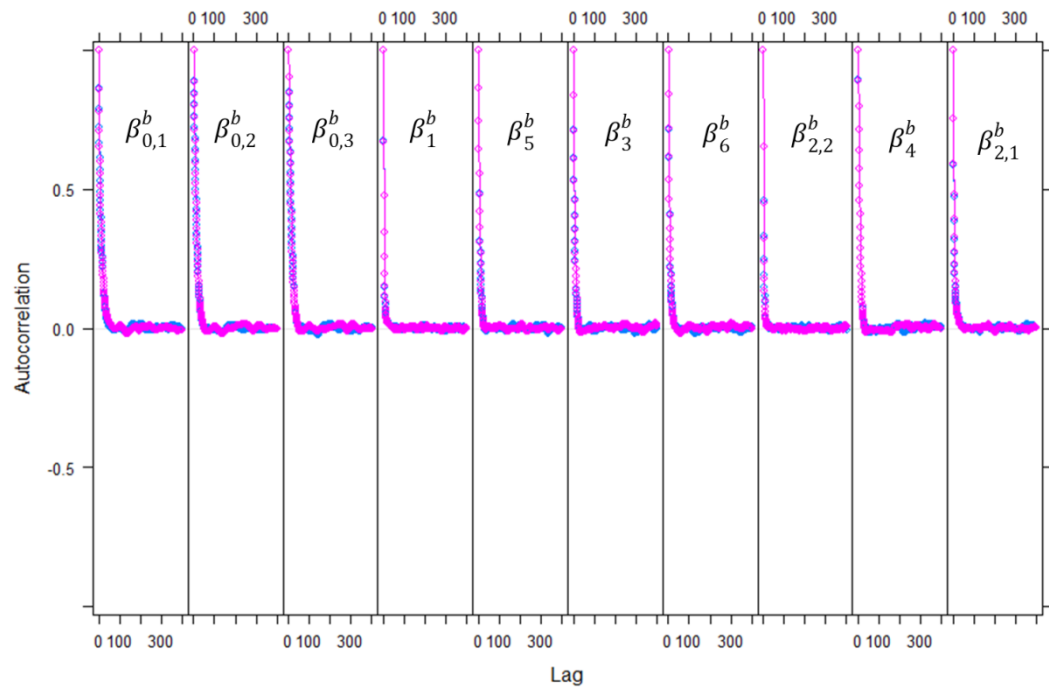


Figure 3.7. Autocorrelation plots for 200000 iterations of the model.

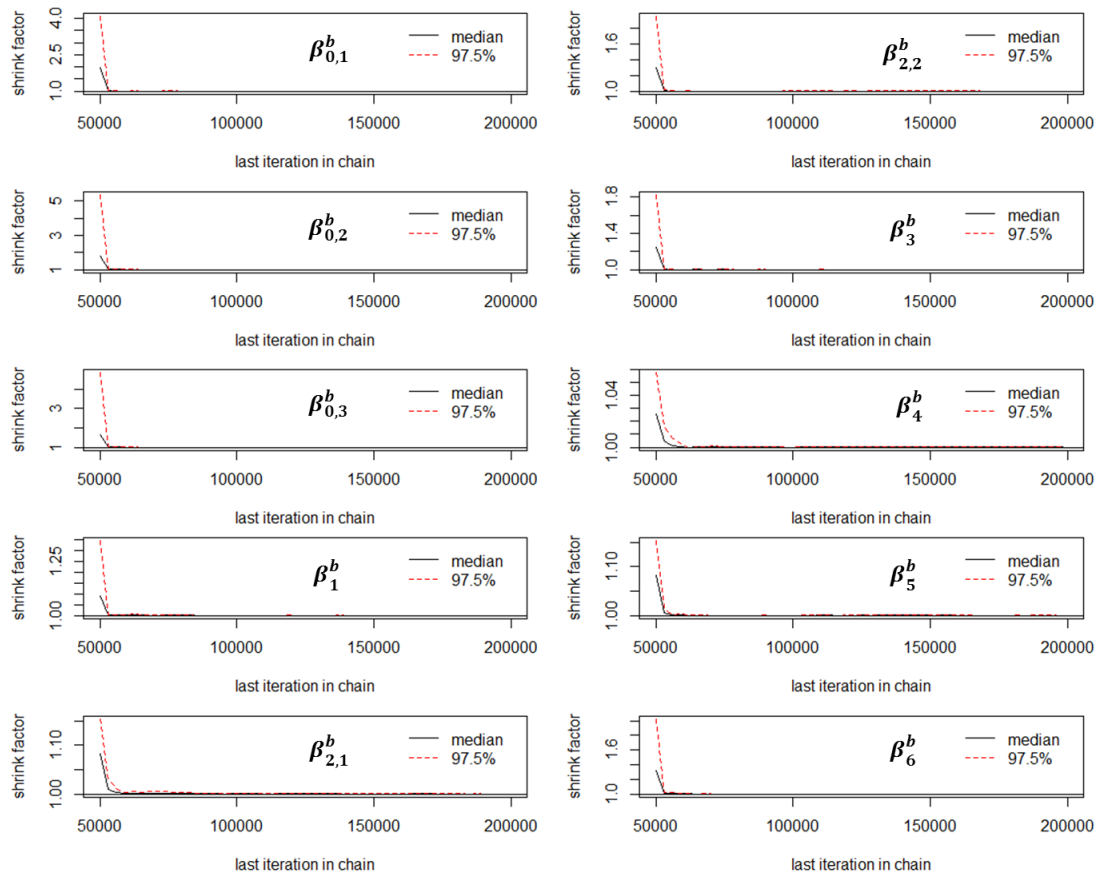


Figure 3.8. Gelman-Rubin diagnostic plots for 200000 iterations of the model.

Coefficients		PSRF
Intercept	Gas System, $\beta_{0,1}^b$	1
	Water System, $\beta_{0,2}^b$	1
	Oil System, $\beta_{0,3}^b$	1
Deadleg Design	Branch, β_1^b	1
Orientation	Vertical, $\beta_{2,1}^b$	1
	Mix (Horizontal + Vertical) , $\beta_{2,2}^b$	1
Length, β_3^b		1
Diameter, β_4^b		1
Log(Flow Rate) , β_5^b		1
Main Line Corrosion, β_6^b		1

Table 3.4. Potential scale reduction factor (PSRF) of Gelman-Rubin convergence diagnostic.

3.4.4 Mann-Whitney-Wilcoxon test

Looking at the results of both frequentist and the Bayesian hierarchical approaches, it is hard to tell which method is giving a better estimation, because frequentist and the Bayesian approach are based on different philosophy. However, we conducted a statistical test to determine which model generates a closer estimate compared to the field data, which are the measured deadlegs corrosion rates.

We used the Mann-Whitney-Wilcoxon (MWW) test for the statistical test for the Bayesian and frequentist analyses to compare the results of the estimation. The choice of this statistical test is on account of its ability to accommodate different statistical distributions without assuming the samples to follow the normal distribution [42].

We used the median of estimated Weibull distribution from both frequentist density regression and the Bayesian hierarchical model to compare with the deadlegs corrosion samples. We then observed the significance level and p-value calculated in

MWW test. The null hypothesis is that the field corrosion rates and median of estimated Weibull distribution are identical.

Approach	p-value
Frequentist	0.2441
Bayesian	0.1454

Table 3.5. Mann-Whitney-Wilcoxon test for frequentist Weibull median and the median of Weibull estimated using the Bayesian approach.

Table 3.5 shows the result of the MWW test for the two different analyses of the model. As the p-values of both methods turns out to be greater than the 0.05 significance level, this implies that there is no evidence to reject the null hypothesis and conclude that both the presented frequentist and Bayesian methods can be used for analysis of the model.

3.5 Summary

Corrosion rate can be obtained by the actual field inspection using NDT, but sometimes some areas of the piping system are difficult for the inspector to access to collect thickness measurements. There are other parts of the system where thickness measurements may simply not be available, requiring a prediction of the corrosion rate to be made using available data from comparable operating environments. Even in cases where thickness measurements are available, it is useful to supplement this information based on predictive methods. Indeed, the predictions themselves can be made more reliable by incorporating new data as and when they become available.

Weibull density regression is suitable for corrosion analyses where there exist needs to consider multiple factors at the same time in the estimation. The proposed approach demonstrates the effect of various factors on the rate of corrosion of deadlegs. The results are in line with previous studies on how those factors considered in the density regression model affect the internal corrosion of piping

deadlegs. By using the different quantile levels of the Weibull distribution, we can predict the corrosion rates of inaccessible piping deadlegs of different design and operation conditions on account of the assumption that deadlegs' corrosion rates are always greater than the corrosion rates of their parent lines.

The Bayesian analysis is also conducted on the Weibull density regression model. Both the frequentist and the Bayesian analysis result in similar regression coefficient estimates in terms of posterior mean. But, instead of the point estimate by frequentist method, the Bayesian analysis can provide a complete picture by posterior.

We also found that the Bayesian analysis is easy to cope with a regression model with overused dummy variables (for categorical data) in the independent variables. The Bayesian analysis enables the current analysis to be updated when future data is available. For example, the Bayesian analysis method can directly take current analysis as prior for any future data analysis of the same model and comparison.

The regression approach shown here complements other approaches including those that involve physics-based models and Bayesian belief networks [43-45]. A piping integrity management approach may have inputs for a predicted life model or structural reliability models that consider other damage mechanisms such as shown in [46], and risk-based decision support models that include the impact of consequential failure such as shown in [47].

3.6 References

1. API 570 (2009) Piping Inspection Code: In-service Inspection, Rating, Repair, and Alteration of Piping Systems, American Petroleum Institute, Third Edition.
2. V. S. Sastri (2015) Challenges in Corrosion: Costs, Causes, Consequences, and Control, Wiley Series in Corrosion, ISBN: 978-1-118-52210-3, Page 297.
3. S. Y. Kovalenko, A. O. Rybakov, A. V. Klymenko, and L. H. Shytova (2012) Corrosion of the internal wall of a field gas pipeline, *Materials Science*, 48, 2, 225-230.

4. W. Dang and K. Yu (2015) Density regression based on proportional hazards family, *Entropy* 17, 6, 3679-3691.
5. M. A. Habib, H. M. Badr, S. A. M. Said and I. Hussaini (2005), On the development of deadleg criterion, *American Society of Mechanical Engineers*, Vol. 127, 1, 124-135.
6. M. A. Habib, S.A.M. Said, H.M. Badr and I. Hussaini (2004) Effect of geometry on flow field and oil/water separation in vVertical deadlegs, *International Journal of Numerical Methods for Heat & Fluid Flow*,15, 4, 348 – 362.
7. M. A. Habib, S.A.M. Said, H.M. Badr and I. Hussaini (2005) Characteristics of flow field and water concentration in a horizontal deadleg, *Heat and Mass Transfer*, 41, 4, 315-326.
8. B. L. Austen (2005) *Pharmaceutical Water Systems: A Thermal-fluid Analysis of Pipe Dead-legs*, School of Mechanical and Manufacturing Engineering, Dublin City University, Ireland.
9. D. C. Coyle (2007) *A Thermal-fluid Analysis of Piping Dead-legs in High Purity Water Systems*, School of Mechanical and Manufacturing Engineering, Dublin City University, Ireland.
10. K. Ding, H. Zhu, J. Zhang, C. Wang and J. Hao (2011) A numerical study of flow field oil water separation in vertical deadlegs, *Applied Mechanics and Materials*, 121-126, 2465-2470.
11. R. Murata, J. Benaquisto and C. Storey (2015) A methodology for identifying and addressing dead-legs and corrosion issues in a process hazard analysis (PHA), *Journal of Loss Prevention in the Process Industries*, 35, 387-392.
12. L. Bensman (2016) Dead leg internal corrosion management, *Corrosion 2016*, 6-10 March, Vancouver, British Columbia, Canada, NACE-2016-7715.
13. E. D. Beaton, M. L. Serran and L. Sun (2010) Numerical simulation of mixing in process deadlegs in order to model microbiologically influenced corrosion and tuberculation at these locations, *NACE International Conference, Corrosion 2010*, Paper 10214.
14. C A. Barrett (1985) A multiple linear regression analysis of hot corrosion attack on a series of nickel base turbine alloys, *NASA Technical Memorandum*, 87020, 19850022971.

15. S. C. Kung (1997) Prediction of corrosion rate for alloys exposed to reducing/sulfidizing combustion gases, *Corrosion* 97, NACE International, Paper No. 136.
16. Y. S. Garud, P. Besuner, & M. J. Cohn (1999) Recent developments in measurement and evaluation of FAC damage in power plants, *Corrosion* 99, NACE International, Paper No. 353.
17. Y. F. Wen, C. Z. Cai, X. H. Liu, J. F. Pei, X. J. Zhu and T. T. Xiao (2008) Corrosion rate prediction of 3C steel under different seawater environment by using support vector regression, *Corrosion Science*, 51, 2, 349-355.
18. C. I. Ossai (2012) Predictive modelling of wellhead corrosion due to operating conditions: a field data approach, *ISRN Corrosion*, 2012, 237025.
19. K. Siamphukdee, F. Collins and R. Zou. (2012) Sensitivity analysis of corrosion rate prediction models utilized for reinforced concrete affected by chloride, *Journal of Materials Engineering and Performance*, ASM International, 22, 6, 1530-1540.
20. Y. Suleiman, O. B. Oloche, and S. A. Yaro (2012) The development of a mathematical model for the prediction of corrosion rate behaviour for mild steel in 0.5M sulphuric acid, *ISRN Corrosion*, 2013, 710579.
21. S. Anyanwu, O. Eseonu and H. U Nwosu (2014) Experimental investigation and mathematical modelling of corrosion growth rate on carbon steel under the influence of soil pH and resistivity, *IOSR Journal of Engineering*, 04, 10, 7-18.
22. J. I. Sodiki and M. V. Ndor (2016) Regression equations for predicting the corrosion of steel, *Innovative Systems Design and Engineering*, 7, 3, 26-33.
23. J. I. Sodiki, M. V. Ndor, and A. Sodiki (2016) Regression analysis for predicting the corrosion extent of brass and aluminum, *Innovative Systems Design and Engineering*, 7, 4, 78-87.
24. C. Qiu and M. E. Orazem (2004) A weighted nonlinear regression-based inverse model for interpretation of pipeline survey data, *Electrochimica Acta*, 49, 22-23, 3965-3975.
25. S. N. F. Mior, N. Yahaya, N. M. Noor, K. S. Lim and A. A. Rahman (2015) Underground corrosion model of steel pipelines using in situ parameters of soil, *Journal of Pressure Vessel Technology*, ASME, 137, 5, 1-6.

26. J. C. Velázquez, F. Caleyó, A. Valor and J. M. Hallen (2008) Predictive model for pitting corrosion in buried oil and gas pipelines, *NACE International, Corrosion* 2009, 65, 5, 332-342.
27. M. S. El-Abbasy, A. Senouci, T. Zayed, M. ASCE, F. Mirahadi and L. Parvizsedghy . (2014) Condition prediction models for oil and gas pipelines using regression analysis, *Journal of Construction Engineering and Management*, ASCE, 140, 6, 1-17.
28. TWI Limited (2002) *Guidelines for Use of Statistics for Analysis of Sample Inspection of Corrosion*, HSE UK, ISBN 0 7172 5540.
29. J. I. McCool (2012) *Using the Weibull Distribution*, John Wiley & Sons, Inc., Hoboken, New Jersey, Chapter 3, Page 73.
30. A. Asgharzadeh and R. Valiollahi (2009) Inference for the proportional hazards family under pProgressive type-II censoring, *JIRSS*, 8, 1-2, 35-53.
31. Y. Liu (2009) On equality of ordinary least squares estimator, best linear unbiased estimator and best linear unbiased predictor in the general linear model, *Journal of Statistical Planning and Inference*, 139, 4, 1522-1529
32. S. Y. Park and A. K. Bera (2009) Maximum entropy autoregressive conditional heteroskedasticity model, *Journal of Econometrics*, 150, 2, 219-230.
33. H. Qiao and C. P. Tsokos (1994) Parameter estimation of the Weibull probability distribution, *Mathematics and Computers in Simulation* 37, 1, 47-55.
34. A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin. (2004) *Bayesian Data Analysis*, A CRC Press Company, Chapman & Hall, Second Edition, Chapter 1, Page 3.
35. P. Congdon (2003) *Applied Bayesian Modelling*, John Wiley & Songs, ISBN 0-471-48695-7, Chapter 1, Page 1-2.
36. S. M. Lynch (2007) *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, Springer Science + Business Media, ISBN 978-0-387-71265-9, Page 88 – 89, Page 232 – 233.
37. P. Robert (2012) *The BUGS Book: A Practical Introduction to Bayesian Analysis*, CRC Press, Taylor & Francis Group, ISBN 978-1-4665-8666-6.
38. H.E.T. Holgersson, L. Nordström and Ö. Öner (2014) Dummy variables vs. category-wise models, *Journal of Applied Statistics*, 41:2, 233-241.

39. M. K. Cowles and B. P. Carlin (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association*, Taylor & Francis Ltd., 91, 434, 883-904.
40. S. P. Brooks and A. Gelman (1997) General methods for monitoring convergence of iterative simulations, *Journal of Computational and Graphical Statistics* 7: 434–455.
41. A. Gelman and D. B. Rubin (1992) Inference from iterative simulation using multiple sequences, *Statistical Science* 7: 457–511.
42. P. Wu, Y. Han, T. Chen and X. M. Tu (2014) Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics, *Statistics in Medicine*, 04/2014, 33, 8, 1261-1271.
43. N. Sridhar, F. Ayello and F. Candreva (2015) Predicting the impact of corrosion under insulation, *Managing Ageing Plants*. 2015:47-51.
44. S. Jain, F. Ayello, J. A. Beavers and N. Sridhar (2013) Probabilistic model for stress corrosion cracking of underground pipelines using Bayesian networks, NACE International Publication 2013.
45. F. Ayekki, N. Sridhar, G. Koch, V. Khare, A. W. Al-Methen and S. Safri (2013) Internal corrosion threat assessment of ipelines using Bayesian networks, NACE International Publication. 2013.
46. M. Lamb and T. Zimmerman (2000) The decision process for the use of risk-based design for onshore pipelines, *The onshore pipeline cost reduction conference organised by Pipes & Pipelines International and Clarion Technical Conferences*; 3-5 April, 2000; Amsterdam, Netherlands.
47. U. R. Bharadwaj, V.V. Silberschmidt and J. B. Wintle (2012) A risk based approach to asset integrity management, *Journal of Quality in Maintenance Engineering*, 18, 4, 417-31.
48. A. E. Gelfand and A. F. M. Smith (1990) Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, American Statisitcal Association, 85, 410, 398-409.
49. D. R. Cox (1972) Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 2, 187-220.

Chapter 4. The Use of Corrosion Data Analyses for Predicting Remaining life for Integrity Assessments

Abstract

In order to show how the analyses of raw inspection monitoring data feed into integrity assessments, prediction of remaining useful life of conductors and deadleg piping is discussed in this research. By using the estimated corrosion rate and growth model, we can predict the remaining useful life of piping and the structural reliability of the conductors. For piping, there is always a threshold value for remaining wall thickness that is determined from design data. Once the wall thickness is thinned beyond or defect depth grows above this threshold, the pipe should be repaired or replaced to maintain its fitness for operation in the entire piping system. For conductors, the situation is different; although leaking is a failure mode, the main failure mode and concern is the loss of the physical integrity of the entire system due to the wall thickness of an area reducing to less than a certain threshold. The discussion in this section of the thesis is centred on demonstrating the use of results from analyses of data for the purpose of piping and offshore conductors remaining life assessments. We employ the first-order reliability method (FORM) to evaluate the time-dependent system reliability for these two different components.

Keywords: Remaining useful life; first order reliability method; probability of failure; piping deadlegs; conductor pipes.

4.1 Introduction

The remaining useful life (RUL) of an asset is defined as the length from the present time to the end of the useful life. The idea of the RUL has been broadly applied as a part of operational research, reliability, and statistics literature with important applications in many fields such as material science, biostatistics and econometrics.

RUL estimation is one of the key factors in oil and gas asset integrity and maintenance management.

The degradation rate and growth estimation methods have been discussed in previous chapters, for piping deadlegs and oil and gas offshore conductor pipes. One of the approaches in determining the remaining life of a component is by measuring the remaining thickness of the component where the thickness of the component must be above a certain threshold value to ensure it is fit for service.

The remaining life of a component can be determined based on the computation of a minimum required thickness for the intended service conditions, thickness measurements from an inspection, and an estimate of the anticipated corrosion rate [1]. The general remaining life equation based on thickness is given by:

$$RL = \frac{\text{Current Thickness} - \text{Minimum Required Thickness}}{\text{Corrosion Rate}} \quad (4.1)$$

API579 is one of the standards that provide calculations for the minimum required thickness for different types of components and design codes [1]. For a newly installed component, the remaining life is the entire life of the component, where the nominal thickness is usually used as the current thickness.

4.1.1 Motivation and purposes

Measurement of wall thickness for inaccessible piping deadlegs is not available. However, the corrosion density is extrapolated by using the available information of other deadlegs such as thickness measurements, design factors, and operational information. For offshore conductor pipes, the maximum defect depth is estimated by using extreme value theory and the corrosion growth is simulated by using geometric Brownian motion. When the corrosion density and defect growth are calculated, the predicted remaining lifetime distribution of these components can be obtained. The purpose of this research is to predict the remaining useful life of existing inaccessible

pipings deadlegs and offshore conductor pipes; also to provide information that can be used to plan the future construction and maintenance of piping systems and offshore conductors. We propose that a reliability model be used as a prognosis method for remaining lifetime of these components. The first order reliability method (FORM) is used to estimate the current and target probabilities of failure for these components. It is feasible and efficient in evaluating the probability of failure and lifetime of the target components by using FORM.

4.1.2 Previous works on remaining useful life estimation with corroded assets

Kowaka and Tsuge [2] used extreme value theory to identify corrosion rates for different assets such as buried pipelines, storage tanks, and so on, and showed a pragmatic approach in predicting remaining life of those assets by observing the curve shift across the thickness axis in time on probability plots. Ahammed [3] developed a non-linear limit state model which incorporates multiple normal and non-normal random variables to probabilistically estimate the remaining life of a corroded pipeline. Caleyó et al. [4] presented different reliability assessment methods utilising a limit state function to predict the remaining useful life of pipelines with active corrosion defects. Li et al. [5] predicted the remaining life of underground pipelines by using a mechanically-based probabilistic model that takes account of random effects of corrosion in the pipelines. Ossai [6] estimated the mean time to failure (MTTF) for oil and gas pipelines by using Monte Carlo simulation and degradation models, and the survival function was determined with the Weibull distribution. Liu [7] computed the parameters in the degradation model by using a Bayesian approach to derive the failure time and remaining life distribution for circuits of pipelines. Zangenehmadar and Moselhi [8] used an artificial neural network (ANN) approach to predict the deterioration rate of water distribution networks along with remaining useful life estimation that included various important factors in the model. Son et al. [9-10] showed prognostic methods of assessing remaining useful lifetime for plant components based on stochastic processes. Yasserli and Mahani [11] illustrated a pragmatic way of determining the remaining useful life of pipelines by using a first

order reliability method; the method includes several influencing factors such as material properties, corroded area and degradation rates.

4.2 Remaining useful life assessment for inaccessible piping deadleg

In Chapter 3 we used Weibull density regression to model the corrosion rates of the piping deadlegs. Due to the inaccessibility of these deadlegs, thickness gauging cannot be performed. Because of this reason, we used the nominal thickness and corrosion density to estimate the probabilities of failure of these deadlegs.

The ultimate failure of a pipe occurs when the thickness of an area or point is completely worn out, in other words, $thickness = 0$. However, any pipe may fail when its toughness, strength, and chemical resistance capabilities are below the allowable thresholds. In the material design standard for piping such as ASME B31.3, the thickness of these materials must be maintained above a required level of thickness; this threshold is commonly known as the minimum required thickness. If $thickness \leq minimum\ required\ thickness$, fitness-for-service (FFS) will be conducted to examine whether they are still safe to continue to operate, otherwise repair or replacement would be carried out.

We employ the first-order reliability method to evaluate the time-dependent system reliability of these inaccessible piping deadlegs. By using the equation (4.1), we calculated the the predicted remaining lifetime distribution with nominal thickness and minimum required thickness respectively. Let $X_{nom}^{deadleg}$ be the nominal predicted remaining lifetime distribution and $X_{min}^{deadleg}$ be the minimum required predicted remaining lifetime distribution:

$$X_{nom}^{deadleg} = \frac{Nominal\ Thickness - 0}{Corrosion\ Rate\ Density}$$

$$X_{min}^{deadleg} = \frac{\text{Minimum Required Thickness} - 0}{\text{Corrosion Rate Density}}.$$

The limit state for the piping deadleg life is $X_{nom}^{deadleg} - X_{min}^{deadleg}$.

4.2.1 First order reliability method

The first order reliability method has been broadly used in corrosion applications for reliability-based assessment [4, 12-17] due to its effectiveness in dealing with the numerous uncertainties that are either inherent to the corrosion process or involved in the inspection and evaluation of corrosion defects.

Given the n -dimensional vector of random variables $X = (X_1, X_2, \dots, X_n)^T$, with continuously differentiable distribution $F_X(x)$, $G(x)$ is the function that defines the limiting state such that $G(x) \leq 0$ is the unsafe domain and $G(x) > 0$ represents the safe domain. Figure 4.1 shows the probability integration in FORM, failure zone, and safe domain. To make the shape of the integrand $f_x(x)$ regular, all the random variables X are transformed to standard normal variables $U = (U_1, U_2, \dots, U_n)$. The mean and standard deviation of the standard normal variable are 0 and 1 respectively.

The condition for the CDFs of the random variables to remain the same before and after the transformation from X to U must be met. This type of transformation is called a Rosenblatt transformation [18], which is expressed by:

$$F_X(x) = \Phi(u) \tag{4.2}$$

in which $\Phi(\cdot)$ is the CDF of the standard normal distribution.

The transformed standard normal variable is then given by:

$$U = \Phi^{-1}[F_X(x)]. \tag{4.3}$$

After the transformation, the probability integration becomes:

$$p_f = P\{g(U) < 0\} = \int_{g(U) < 0} \phi_U(U) du. \quad (4.4)$$

The reliability index β is the shortest distance from origin to failure $g(U) = 0$ in U -space. The point $U^* = (u_1^*, u_2^*, \dots, u_n^*)$ on the failure surface is the so-called probable point (MPP). The reliability index β and MPP U^* can be computed by the solution of a constrained optimization in U -space as follows:

$$\begin{cases} \min \|U^*\| \\ \text{subject to } g(U) = 0. \end{cases} \quad (4.5)$$

The HLRF-BGFS method is used to solve the constraint optimisation problem [19].

If U^* , with $U^* \neq 0$, is the solution point of the optimisation problem in Equation (4.5), then the following equation (4.6) holds for each distribution parameters of random vector X :

$$U^{*T} \nabla_u g(U^*) + \|U^*\| \|\nabla_u g(U^*)\| = 0 \quad (4.6)$$

where

$$\nabla_u g(U^*) = \left[\frac{\partial g(U^*)}{\partial U_1} \frac{\partial g(U^*)}{\partial U_2} \dots \frac{\partial g(U^*)}{\partial U_n} \right]^T \quad (4.7)$$

$$\|U^*\| = \sqrt{U_1^{*2} + U_2^{*2} + \dots + U_n^{*2}} \quad (4.8)$$

$$\|\nabla_u g(U^*)\| = \sqrt{\left[\frac{\partial g(U^*)}{\partial U_1} \right]^2 + \left[\frac{\partial g(U^*)}{\partial U_2} \right]^2 + \dots + \left[\frac{\partial g(U^*)}{\partial U_n} \right]^2}. \quad (4.9)$$

From equation (4.6), the reliability index β can be expressed as follows:

$$\beta = \|U^*\| = \frac{-U^{*T} \nabla_u g(U^*)}{\|\nabla_u g(U^*)\|}. \quad (4.10)$$

The probability of failure can be approximately calculated by:

$$p_f = 1 - \Phi(\beta). \quad (4.11)$$

More information about FORM can be found in [20].

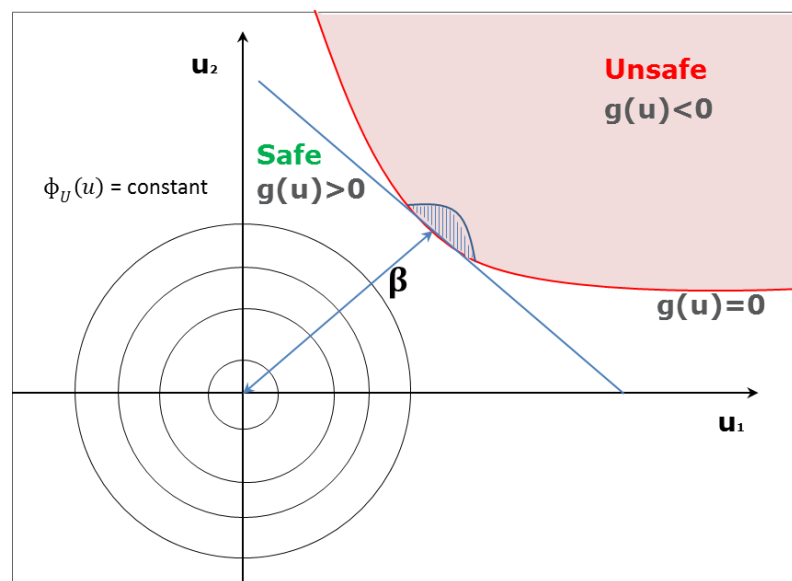


Figure 4.1. Probability Integration in FORM

4.2.2 Case study

We used example data to show how the remaining useful life using FORM works. Assume the corrosion density we obtained is $CR \sim Weibull(shape = 2.364, scale = 0.25)$, the nominal thickness of the deadleg pipe is 15.2mm and its minimum required thickness is 8.6mm; the pipe has been in service for 15 years. Table 4.1 shows the example data for this case:

Parameters		Values
Corrosion Distribution (Weibull Density), mm/year	Shape	2.364
	Scale	0.25
Nominal Thickness, mm		15.2
Minimum Required Thickness, mm		8.6
Current Years in service		15

Table 4.1. Example data of piping deadleg for a demonstration of the probability of failure estimation using FORM.

The predicted remaining lifetime distribution from the nominal thickness to 0:

$$X_{nom}^{deadleg} = \frac{Nominal\ Thickness - 0}{CR \sim Weibull(2.364, 0.25)}$$

and the predicted remaining lifetime distribution from the minimum required thickness to 0:

$$X_{min}^{deadleg} = \frac{Minimum\ Required\ Thickness - 0}{CR \sim Weibull(2.364, 0.25)}.$$

The performance function for the deadleg is given by:

$$g(X) = X_{nom}^{deadleg} \geq X_{min}^{deadleg}.$$

It is important to identify which probability distribution $X_{nom}^{deadleg}$ and $X_{min}^{deadleg}$ belong to before running FORM. The two distributions are not symmetric and have positive skewness; hence we used lognormal, Weibull, Gumbel, and also Normal distribution to fit to these lifetime data and observed which probability distribution has the best fit to the data. The fitting is done using maximum likelihood estimation [21].

We conducted a statistical test on all the distribution fittings by using the Kolmogorov-Smirnov (KS) test [22] and Anderson-Darling (AD) test [23]. We observed the statistics calculated in KS-test and AD-test; the smaller the statistics' values, the better the fit of the data to the lifetime distribution. Table 4.2 and Table 4.3 show the

estimated parameters of all our fitted distributions and the statistics of KS test and AD test for the lifetime distributions, respectively for $X_{nom}^{deadleg}$ and $X_{min}^{deadleg}$.

Based on the KS and AD test, it is suggested that both lifetime distributions are fitted the best with the lognormal distribution among all the selected distributions. So the life distribution of this component using its nominal thickness is $X_{nom}^{deadleg} \sim LN(4.3535, 0.545)$ and the life distribution using minimum required thickness is $X_{min}^{deadleg} \sim LN(3.7839, 0.545)$.

Statistical Distribution	Parameters	KS-Test Statistic	AD-Test Statistic
Lognormal	Mean log = 4.3535; SD log = 0.545	0.0767	122.2
Weibull	Shape = 2.1; Scale = 102.3	0.1481	577.9
Normal	Mean = 93.533; SD = 87.996	0.2314	1105.9
Gumbel Max	Location = 53.93; Scale = 68.61	0.2427	862.2

Table 4.2. KS and AD tests for $X_{nom}^{deadleg}$

Statistical Distribution	Parameters	KS-Test Statistic	AD-Test Statistic
Lognormal	Mean log = 3.7839; SD log = 0.545	0.0767	122.2
Weibull	Shape = 2.1; Scale = 57.874	0.1481	577.9
Normal	Mean = 52.92; SD = 49.787	0.2314	1105.9
Gumbel Max	Location = 30.513; Scale = 38.819	0.2427	862.2

Table 4.3. KS and AD tests for $X_{min}^{deadleg}$

The life distribution $X_{nom}^{deadleg}$ is shown in Figure 4.2.

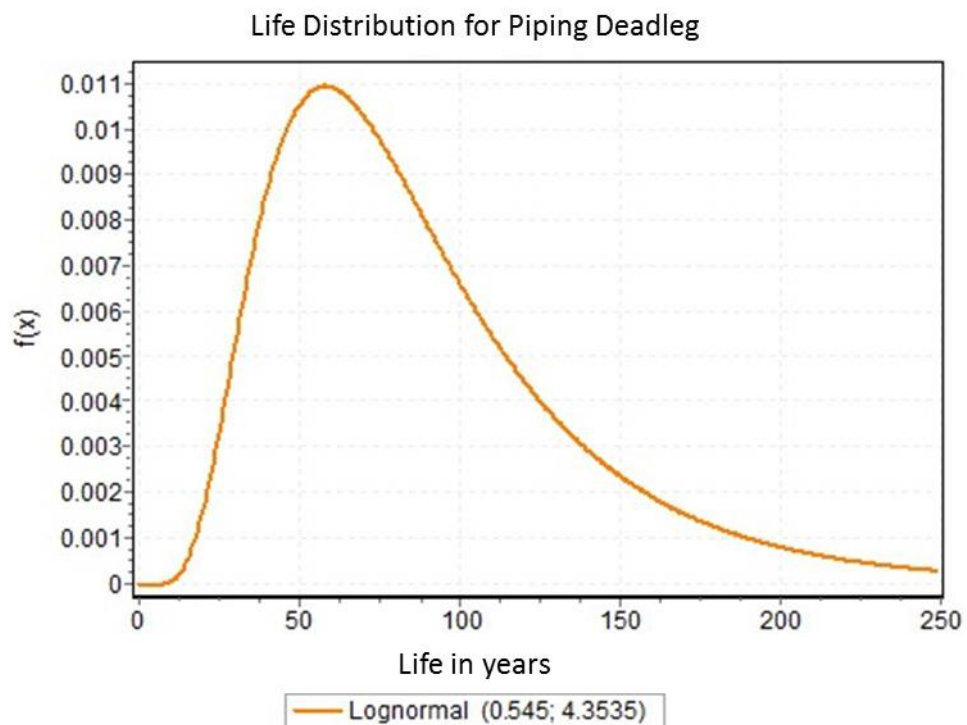


Figure 4.2. Life distribution of piping deadleg

Given lognormally distributed variable X with expected value μ and standard deviation σ , the Rosenblatt transformation is expressed by:

$$F_X(x) = \Phi\left(\frac{\log x - \mu_l}{\sigma_l}\right)$$

where $\sigma_l = \sqrt{\log\left(\frac{\sigma^2}{\mu^2} + 1\right)}$ and $\mu_l = \log\mu - \frac{1}{2}\sigma_l^2$.

The transformed standard normal variable then becomes:

$$U = \Phi^{-1}[F_X(x)] = \frac{\log x - \mu_l}{\sigma_l}$$

$$X = \exp(\sigma_l u + \mu_l).$$

And the performance function is:

$$g(U) = \exp(\sigma_{l_{nom}} \cdot u_{nom} + \mu_{l_{nom}}) - \exp(\sigma_{l_{min}} \cdot u_{min} + \mu_{l_{min}})$$

where $\exp(\sigma_{l_{nom}} \cdot u_{nom} + \mu_{l_{nom}})$ is the transformed standard normal distribution for $X_{nom}^{deadleg}$ and $\exp(\sigma_{l_{min}} \cdot u_{min} + \mu_{l_{min}})$ for $X_{min}^{deadleg}$.

We can obtain the probability of failure for the deadleg component with current year in service through the cumulative function of $X_{nom}^{deadleg} \sim LN(4.3535, 0.545)$. By using FORM, we obtained the target failure probability of 0.23 for which the thickness of the component will be thinned below its acceptable thickness at approximately year 52. The difference in year between the current probability of failure and the target failure probability is the remaining useful life of this component. Figure 4.3 shows the probability of failure for the current year in service, target probability, and theoretical remaining useful life on a cumulative function plot.

Nonetheless, the calculated failure probability is not an absolute figure; it is left to the judgement of the integrity engineer to decide if any further adjustment is of any value. Usually, the target POF should be set lower than the 0.23 in the preventive maintenance routine. Once the POF exceeds a certain level of threshold, the integrity

engineer should take action to come out with a plan to survey these deadlegs, to see whether they need further maintenance.

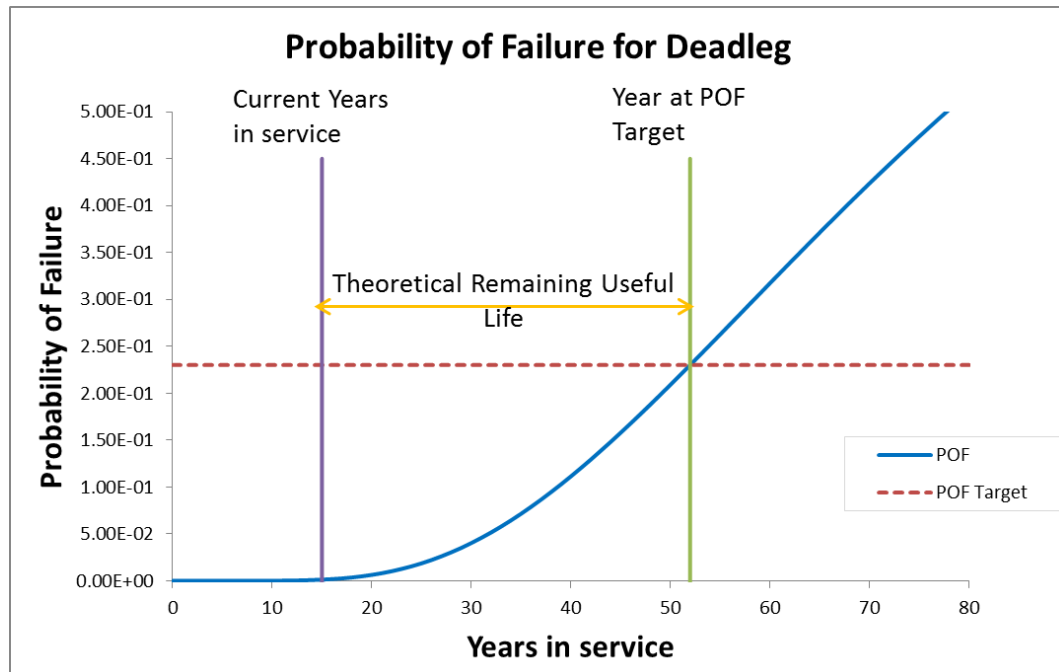


Figure 4.3. Remaining useful life and probability of failure of piping deadleg

4.3 Remaining useful life assessment for offshore conductor pipe

The oil and gas operators encounter a critical challenge in maintaining the ageing offshore well conductors assets as most of these assets have been operated beyond their original design life. The limited information about the oil well construction, unknown operational conditions, and poor through-life maintenance further aggravates the problems. These conductor pipes provide the structural foundation for oil wells and prevent the corrosion effect due to sea splash to the external surface of the inner casing of the entire conductor asset.

Unlike pipeline or piping systems that carry the oil and gas product, leakage of the conductor does not cause immediate failure to the system. However, the thickness of the conductor pipe must be maintained above a certain level of threshold, which is known as minimum required thickness, to continue providing the support to the

conductor system. There are not many industrial guidelines for computing the minimum required thickness for conductor pipes. However, Ramasamy et al. have shown a method of calculating the minimum required thickness for offshore conductors in [24].

In Chapter 2, extreme value theory and the geometric Brownian motion are used to predict the corrosion growth of the conductor pipes. The geometric Brownian motion model is a state-dependent model that assumes the defect depth continues to grow deeper or larger over time and calculates the maximum defect depth for every year by using a return period method. To estimate the remaining life of the conductor pipe, we can observe the defect growth and identify the year when defect depth \geq maximum allowable defect depth. Maximum allowable defect depth is the difference between the nominal thickness and minimum required thickness of a conductor pipe. Figure 4.4 shows the allowable defect depth and remaining useful life of a conductor pipe. Recall that we extrapolated the maximum defect depth for the non-sample area of the conductor pipe surface as well, which is why there are two curves plotted in Figure 4.4. We can see that the remaining life for this conductor is about 4 years since last inspection, based on the defect growth on other areas, as the growth of this area reaches the minimum required thickness first.

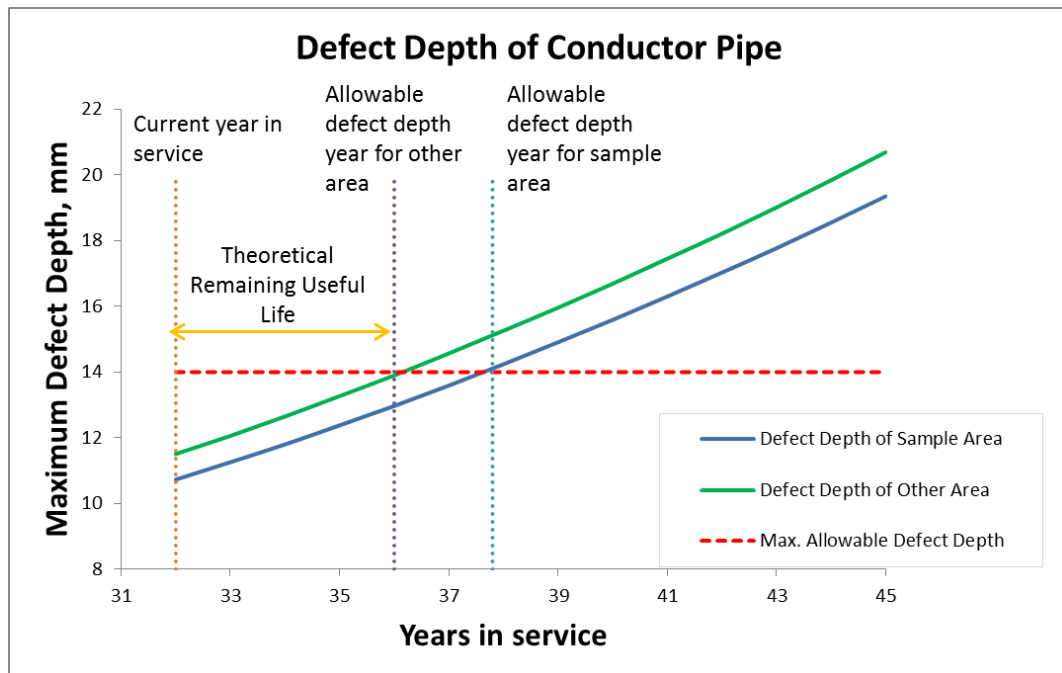


Figure 4.4. Allowable defect depth of conductor pipe

4.4 Summary

In oil and gas exploration, production and refinery plants, there are many components in the asset systems that require adequate and sufficient maintenance to enable the system to function properly. Each element in the system has a design life limit that the manufacturer has come out with during the laboratory testing. Some components may operate longer than their original design life, whereas some may fail earlier than their design life; it is due to various factors that can determine the degradation and deterioration of the components.

It is important to conduct remaining useful life assessments for these components by taking various influencing factors into account, to allow the operator to know the remaining life of the assets after they have been put in service for years. Appropriate preventive actions can be taken based on the outcome of remaining useful life assessment to hold back from severe system failure that could result in high consequential cost and fatality.

We have presented two different models for conducting remaining life assessment in two different components using the latest taken thickness measurements for piping deadlegs and offshore conductor pipes. The geometric Brownian motion is used for conductor pipe remaining life assessment while FORM is used for piping deadlegs. Depending on the availability of data, the models can be used for different assets and components, not only limited to the components presented above. Our approach may have inputs for risk-based decision support models that include the impact of consequential failure.

It is important to note that defining the probability of failure and risk targets is the responsibility of the owner-user. The targets should be developed based on owner-user internal guidelines and overall risk tolerance. Owner-users often have corporate risk criteria defining acceptable and prudent levels of safety, environmental, and financial risks. These owner-user criteria should be used when making risk-based inspection (RBI) decisions since acceptable risk levels, and risk management decision-making will vary among companies [25].

It is argued that to know the remaining thickness of an asset for remaining life estimation, more frequent non-destructive testing (NDT) should be conducted on those assets instead of using probabilistic and statistical analyses for remaining life estimation. It is true that the inspection gives more information on the fitness of service of those assets; however, over inspection will increase the total budget in integrity assessment and maintenance. Besides this, statistical analyses and risk-based approaches enable integrity engineers to prioritise the inspection on higher risk areas.

4.5 References

1. American Petroleum Institute (2016) Fitness-For-Service, API 579-1/ASME FFS-1, 3rd Edition.

2. M. Kowaka and H. Tsuge (1994) Introduction to Life Prediction of Industrial Plant Materials: Application of the Extreme Value Statistical Method for Corrosion Analysis, Allerton Press Inc, New York, 1994, ISBN 0 89864 073 3, Chapter 5.
3. M. Ahammed (1997) Probabilistic estimation of remaining life of a pipeline in the presence of active corrosion defects, International Journal of Pressure Vessels and Piping, 75, 4, 321-329.
4. F. Caleyó, J. L. González and J. M. Hallen (2001) A study on the reliability assessment methodology for pipelines with active corrosion defects, International Journal of Pressure Vessels and Piping, 79, 1, 77-86.
5. S. X. Li, S. R. Yu, H. L. Zeng, J. H. Li and R. Liang (2009) Predicting corrosion remaining life of underground pipelines with a mechanically-based probabilistic model, Journal of Petroleum Science and Engineering, 65, 3-4, 162-166.
6. C. I. Ossai (2013) Pipeline corrosion prediction and reliability analysis: a systematic approach with Monte Carlo simulation and degradation models, International Journal of Scientific and Technology Research, 2, 3, 58-69.
7. S. Liu (2013) Statistical Methods for Extreme Values and Degradation Analysis, Graduate Theses and Dissertations, Paper 13351, 48-54.
8. Z. Zangenehmadar and O. Moselhi (2016) Assessment of remaining useful life of pipelines using different artificial neural networks models, American Society of Civil Engineers, Journal of Performance of Constructed Facilities, 30, 5, 1-7.
9. K. L. Sona, M. Fouladirada, A. Barrosa, E. Levratb and B. Iungb (2013) Remaining useful life estimation based on stochastic deterioration models: a comparative study, Journal of Reliability Engineering and System Safety, 112, 165-175.
10. K. L. Sona, M. Fouladirada and A. Barrosb (2016) Remaining useful lifetime estimation and noisy gamma deterioration process, Journal of Reliability Engineering and System Safety, 149, 76-87.
11. S. F. Yasserli and R. B. Mahani (2016) Remaining useful life (RUL) of corroding pipelines, Conference Paper, ISPOE 2016.
12. C. Gong & W. Zhou (2017) First-order reliability method-based system reliability analyses of corroding pipelines considering multiple defects and failure modes, Structure and Infrastructure Engineering, 1573-2479, 1-11.

13. A.C.W.M Vrouwenvelder (2014) Reliability-based structural design, *Safety, Reliability and Risk Management – Steenbergen & van Gelder (eds)*, Taylor & Francis Group, London, 45-53.
14. O.S. Lee, D. H. Kim, and N. H. Myoung (2011) Reliability estimation of buried pipelines using first-order reliability method, *Journal of Thermoplastic Composite Materials*, 24, 3, 303-315.
15. S. Zhang and W. Zhou (2014) An efficient methodology for the reliability analysis of corroding pipelines, *Journal of Pressure Vessel Technology*, 136, 4, 041701.
16. H. Yamini and B. J. Lence (2010) Probability of failure analysis due to internal corrosion in cast-iron pipes, *Journal of Infrastructure System*, ASCE, 16, 1, 73-80.
17. A. P. Teixeiraa, C. G. Soaresa, T. A. Nettob and S. F. Estefenb (2007) Reliability of pipelines with corrosion defects, *International Journal of Pressure Vessels and Piping*, 85, 4, 228-237.
18. M. Rosenblatt (1952) Remarks on a multivariate transformation, *The Annals of Mathematical Statistics*, 23, 3, 470-472.
19. G. A. Perıçaroa, S. R. Santosa, A. A. Ribeirob and L. C. Matiolib (2014) HLRFBFGS optimization algorithm for structural reliability, *Journal of Applied Mathematical Modelling*, 39, 7, 2025-2035.
20. Z. Yanfang, Z. Yanlin, and Z. Yimin (2011) Reliability sensitivity based on first-order reliability method, *Journal of Mechanical Engineering Science*, 225, 9, 2189-2197.
21. I. J. Myung (2002) Tutorial on maximum likelihood estimation, *Journal of Mathematical Psychology*, 47, 1, 90-100.
22. S. Facchinetti (2009) A procedure to find exact critical values of Kolmogorov-Smirnov test, *Italian Journal of Applied Statistics*, 21, 3-4, 337-359.
23. S. Engmann and D. Cousineau (2011) Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test, *Journal of Applied Quantitative Methods*, 09/2011, 6, 3, 1-17.
24. R. Ramasamy M. Aljaberi and H. Aljunaibi. (2014) Ageing offshore well structural integrity modelling, assessment and rehabilitation, *SPE International*, Abu Dhabi International Exhibition and Conference, 10-13 November 2014.

25. American Petroleum Institute (2016) Risk-based Inspection Methodology, API Recommended Practice 581, 3rd Edition, Page 23.

Chapter 5. General Conclusion

In this thesis, we established statistical methods to estimate the integrity and reliability for piping deadlegs and offshore conductors. The thesis consists of three technical chapters. In Chapters 2 and 3, we discuss the modelling of defects for offshore conductor pipes and piping deadlegs respectively; whereas in Chapter 4 we studied the remaining useful life of these components by making use of the corrosion density and growth rate derived from the previous chapters.

In Chapter 2, we present two different methods of application of extreme value theory on offshore conductor pipes' inspection data; they are Block Maxima (BM) and Peak-over-Threshold (POT) methods. Both approaches can give similar results in inferring the maximum defect depths distribution and extrapolating the defect depths of uninspected areas. The BM method is easy to use with the assumption that the data are stationary and long-range dependence at extreme levels is weak. Our studies show that the POT method is more robust than BM especially for the cases where the data cannot be divided into equally-sized blocks for the BM method, and de-clustering of data reduces the chance of local dependency in the sample. It is recognised that localised corrosion or pitting can only be modelled stochastically. Thus we use geometric Brownian motion to model the defects and predict the future defect growths. This predictive method is essential for a plant operator or an integrity engineer to make and take short-term and long-term mitigation plans for ensuring the fitness of these conductors for service, and when to repair or replace the entire conductors.

In Chapter 3, we use a Weibull density regression to extrapolate the corrosion density of inaccessible piping deadlegs. The proposed approach demonstrates the effect of various factors on the rate of corrosion of deadlegs. The results presented are in line with previous studies on how those factors considered in the density regression model affect the internal corrosion of piping deadlegs. By using different Weibull quantiles levels, we can predict the corrosion rates of inaccessible piping deadlegs of different design and operation conditions on

account of the assumption that deadlegs' corrosion rates are always greater than the corrosion rates of their parent lines. Both the frequentist and the Bayesian methods are used for the analysis of the regression model; they result in similar coefficient estimates but the Bayesian analysis can provide a complete picture by posterior. We found that the Bayesian analysis is easy to cope with a regression model with overused dummy variables (for categorical data) in the independent variables, and it enables the current analysis to be updated when future data is available.

In Chapter 4, we present two separate models for assessing remaining life assessment in two different components using the latest recorded thickness measurements from a piping system and offshore conductor pipes. The models use the geometric Brownian motion and limiting state functions with first order reliability method to estimate the components' remaining useful life. Depending on the availability of data, the models can be used to predict the remaining useful life for different assets and components, not limited only to piping deadlegs and offshore conductors. Our approach may have inputs for risk-based decision support models that include the impact of consequential failures.

In future, when new inspection data are made available to these components, the models demonstrated can be updated and enhanced on account of the new information given. The models can be evolved over time to become more accurate by taking new information into consideration. Improvements to the models can be made by applying various statistical approaches: Bayesian updating theorem, artificial neural network and machine learning techniques, or any other statistical methods. Besides the first-order reliability method, other numerical integration methods such as crude Monte Carlo and importance sampling can be used to find the limit state of a component.