A statistical approach to determine Microbiologically Influenced Corrosion (MIC) Rates of underground gas pipelines.

by

Lech A. Grzelak

A thesis submitted to the Delft University of Technology in conformity with the requirements for the degree of Master of Science.

Delft, The Netherlands

Abstract

A statistical approach to determine Microbiologically Influenced Corrosion (MIC) Rates of underground gas pipelines.

by Lech A. Grzelak

Chairperson of Graduate Committee: *Prof. dr Roger M. Cooke*
Department of Mathematics

Graduate Committee: *Prof. dr Roger M. Cooke*
Prof. Jolanta Misiewicz
Ir Paul M. Wesselius
Dr Dorota Kurowicka
Ir Daniel Lewandowski

N.V. Nederlandse Gasunie is the leading Dutch gas transportation company. Its main aim to manage, maintain the gas transport system. The grid of gas pipelines belonging to Gasunie consists of over 11.600 km steel pipelines (diameters from 4 to 48 inch) constructed in the 60s.

Integrity management is based on the ability of the pipeline operator to predict the growth of defects detected in inspection programs. The predictions of the corrosion and defect rates can be based on environmental input parameters. Accurate predictions allow interventions/re-inspections to be scheduled in order to eliminate defects which pose a high potential risk.

This thesis investigates three main issues. Firstly, it shows an appropriate tool for the corrosion rate modeling when data from in-line inspections are available. A low number of inspections contribute to high uncertainty about the corrosion rate estimation. In many cases, a poor dataset combined with high uncertainty about the measurements cause corrosion estimates that are not agreeable with reality; for example corrosion is decreasing in time. The outputs from the corrosion rate model are incorporated as input to the second section, where analysis is focused on investigating parameters influencing Microbiologically Induced Corrosion (MIC) rates. The last part of the thesis presents the design and results of the defect rate.

Keywords: Corrosion, Corrosion rate modeling, In-line inspection, MIC, factors influencing MIC, defect rate

Acknowledgments

The author wishes to express a gratitude to Roger M. Cooke, Jolanta Misiewicz and Dorota Kurowicka for the opportunity to spend unforgettable two years in the Risk and Environmental Modeling group.

The author would also like to acknowledge Thomas Mazzuchi and Daniel Lewandowski for coordination while doing the thesis project; moreover; the author would like to thank to the people at Gasunie for a given opportunity to work within their organization, especially to Paul M. Wesselius, Karine Kutrowski and Robert Kuik.

Many people have inspired, guided and helped the author during the two years at TUDelft, and the writer would like to thank them all for a great fun and experience.
The author owes a huge debt of thankfulness to his best friends: Poorwa, Misha and Gosia, for their support, humor throughout not only the courses of M.Sc. program but all the way through entire time.

Last but not the least; the author would like to express his appreciation to his family for the support and love they provided through author's entire life.

List of contents

List of tables

List of figures

Introduction

N.V. Nederlandse Gasunie is the main gas-transportation company in the Netherlands. Its gas pipeline network consists of approximately 11.600 km of steel pipelines (diameters from 4 to 48 inch) that was largely constructed in the period of the mid sixties to early seventies.

The network is split into a high pressure part (HTL) (5600 km, 66-80 bar) and a medium pressure part (RTL) (600 km, 40 bar). The high pressure network is possible to inspect whereas the medium pressure part does not possess required inspection facilities.

Until 1999 Gasunie had no indications whatsoever that there was a corrosion problem on one of its pipelines. Both regular Cathodic Protection (CP) measurements as well as observations during excavations or reroutings indicated that no significant corrosion problem existed.

Nevertheless Gasunie policy was to verify pipeline integrity periodically by inspecting one of its high pressure lines on average once every 5 years since 1979. The results of these inline inspections confirmed the existing opinion that no corrosion problem existed.

In-line inspections have been part of the verification of pipeline integrity since the late seventies in N.V. Nederlandse Gasunie. The discovery of external microbial corrosion (MIC[1]) in 1999 in one of the high pressure pipelines changed the inspection policy from inspection of a randomly selected pipeline once every 5 year to an inspection program for the whole high pressure grid (approximately 5.600 km) to be completed in 10-12 years.

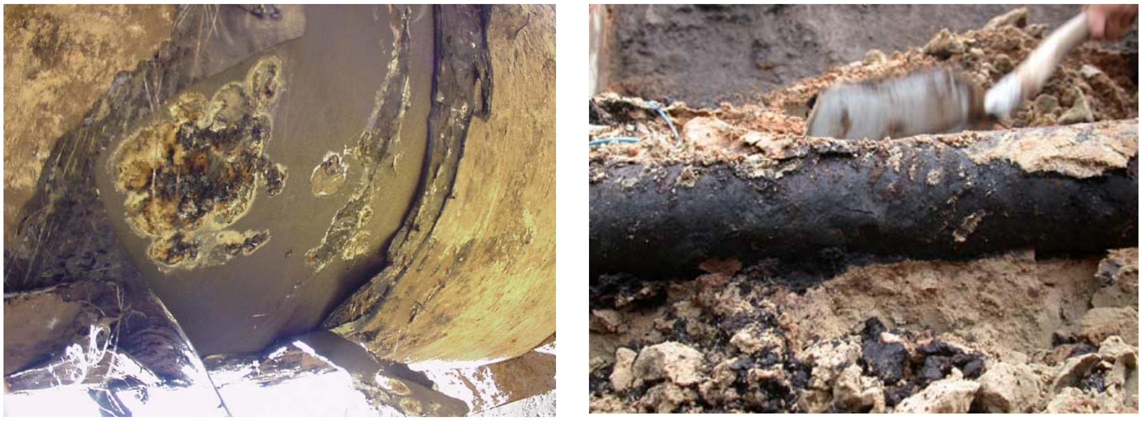


Figure 0-1: deteriorating gas pipelines

External Microbial Induced Corrosion is a type of corrosion where the corrosion rate is influenced by the activity of bacteria, especially Sulfate Reducing Bacteria (SRB). It can be found in many environments. Within Gasunie it is found as external corrosion on gas pipelines. The chemical and microbial processes are complex and can therefore depend on many parameters. Based on experience, Gasunie believed that MIC is found in certain areas more than in others and that therefore the occurrence of MIC is related to soil type or other soil parameters.

[1] Microbiologically Influenced Corrosion

MIC is initiated at locations of disbonded coating (usually at field applied coating) when the bacteria and also the pipeline surface are shielded from the cathodic protection system. Even well maintained cathodic protection systems cannot protect against deteriorated, disbonded coating. MIC is not only dangerous due to incapacity of protecting, but also because of relatively high corrosion rate, which is higher than for galvanic corrosion.

TU Delft was commissioned by Gasunie to make a statistical analysis of the available data on three high pressure pipelines where MIC was recognized.

The goals of the research

The main issue of the thesis is the analysis of the MIC based on the data delivered by N.V. Nederlandse Gasunie. One of the MIC influenced lines was used to qualify MFL-pig[2] (Magnetic Flux Leakage) from different suppliers. In the time period 1999-2004 four different MFL qualification runs have taken place, resulting in 18 excavations. After the fourth pigrun (5 years after the first) Gasunie decided to determine the corrosion and defect rate and investigate influencing factors.

The goal of this thesis is to discuss to Microbiologically Induced Corrosion and its effects on the safety and maintenance issues.

The main objectives of Gasunie are:

- Developing models describing the Microbial Induced Corrosion and Defect rates
- Finding a number of factors influencing both estimated rates.

These models and results can be used for prioritizations of pipelines for inspection and determination of inspection intervals.

The structure of corrosion modeling is presented in the Figure 0-2 below.

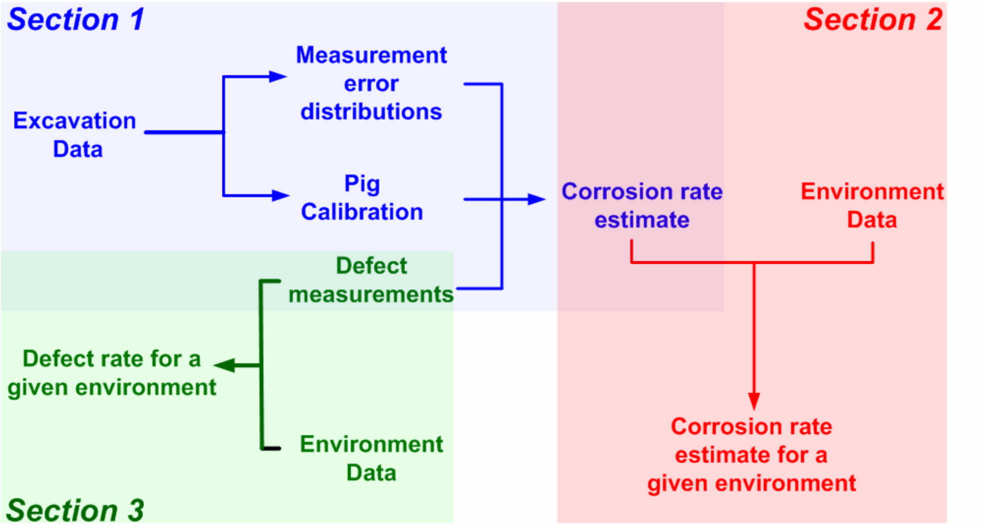


Figure 0-2: corrosion modeling schema

First part of the schema indicated by blue color refers to the first section of the thesis where the corrosion rate model is presented. The results from this study are the input for the part number two where the parameters influencing the microbiologically induced corrosion rate are investigated. The final section shows the results of defect rate modeling of pipeline affected by MIC. This constitutes a detailed statistical description of a connection between environment measurements and the reported corrosion events. The main research is directed to find factors (if there exist) that influence both: the corrosion rate and the defect rate.

The main assumptions in the analysis are:

- Coating condition is assumed uniform and is not governing defect rate.
- Pigrun feature type and stationing indications are fully correct
- Estimates of the corrosion rate (previous study) are certain
- Nominal wall thickness is assumed to be real

[2] Also called: “Intelligent Pig”, more detailed description is presented in Chapter no. 2

Outline of the thesis

The thesis consists of three main parts.

The first section aims to model the corrosion rate. Firstly, Chapter 1 describes available corrosion data and gives a small overview on inspection procedures and associated inspection tools. Later on, the data will be used as an input for corrosion rate model. The model for corrosion rate will be introduced and discussed in Chapter 2. Chapter 3 shows number of numerical approaches in order to get estimate of the corrosion rate. The analysis is carried out starting from the simplest to more sophisticated models. The first and the second approach are simply based on the unconstrained regression analysis. The third model is based on the unbiased measurements and pigs' accuracies introduced previously in Chapter 2.

Second part of the thesis begins with Chapter 4 investigating historical data about the potential from the test-posts. Chapter no. 5 incorporates the results from Part 1, on-potential analysis and bio-data in order to get number of parameters influencing the corrosion rate. The last section is dedicated to defect rate analysis. The section starts with pipeline characteristics (Chapter 6) where general information about gas pipelines is presented. Chapter 7 shows procedures of collecting and analyzing the soil data from geotechnical surveys. The purpose of this chapter is to associate the defect rate of the pipelines induced by MIC with soil compositions. The parameter which is investigated in Chapter 8 is a groundwater step level data analysis. Chapter no. 9 combines all the available soil data, pipelines features and groundwater levels, and treats them as inputs for the regression model describing the defect rate.

Part I

Corrosion Rate Modeling

Chapter 1

1 Description of the corrosion data

1.1 Introduction

In 1999 a 70 kilometer long pipeline (36 inch) in the north of the Netherlands was inspected by MFL pig as part of the verification program. The results of the pigrun indicated, as an unpleasant surprise, 65 external corrosion features but no internal corrosion. After a thorough defect assessment by the safety department of Gasunie it was decided that 6 indications had to be excavated and repaired. The first excavation by the end of 1999 led to a second unpleasant surprise. At the excavation the appearance of the corrosion defect turned out to be totally different from the few corrosion defects that Gasunie had experienced in the past. Analysis of the corrosion products and the appearance of the defect led to the conclusion that the corrosion was influenced by sulphate reducing bacteria (SRB's). Similar experience was obtained at three other excavations. At the end the conclusion was that at four out of the six excavations the corrosion was Microbially Influenced Corrosion (MIC). An additional excavation in 2000 showed also MIC.

From the results Gasunie concluded that it was no longer safe to assume that other lines were free of this corrosion problem. Therefore the inspection policy was revised. It was decided to start an in-line inspection program for all high pressure lines to be completed in a time frame of approximately ten years. It was also decided that pig suppliers had to qualify before they could inspect Gasunie pipelines. Because of gas-transport reasons and because of the fact that some defects on the line had been repaired by clock spring[3] or coating repair and thus can be used as reference points for MFL pigs, the inspected line was appointed as a qualification line. When starting up the inspection program Gasunie realized that a reliable corrosion rate is paramount to determine a re-inspection time interval for its pipelines.

[3] Clock Spring is a composite sleeve used to repair external defects in high-pressure pipelines

1.2 Magnetic Flux Leakage (MFL)-pig

In-line inspections are performed by so-called MFL-pigs also called “intelligent pigs” that locate and characterize mechanical damage in pipelines. It is a common approach to the management of external corrosion in the pipeline industry. Inspections followed by excavations of extreme deep defects minimize potential risk.

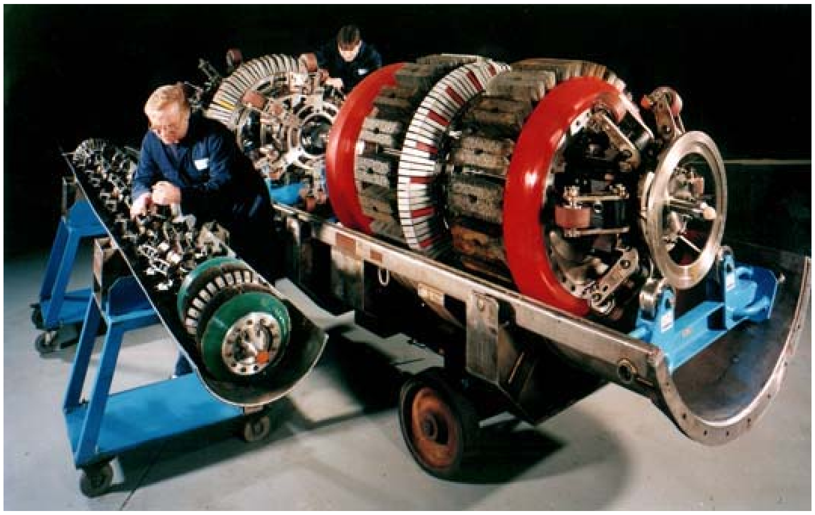


Figure 1-1: small and large diameter MFL-pig

1.3 Matching of the reported defects

In order to compare reported defects from two pigruns the defects will have to be matched. This can be done by using the reported log-distances from the pig, pipeline lengths, clock position of the defect, and distances to reference points like welds, valves. Within Gasunie this process can be automatically done within the PIMS[4] software.

In this software the matching process visualization of the defects on a pipeline segment is possible. Figure 1-2 shows a typical example of reported defects from two pigruns.

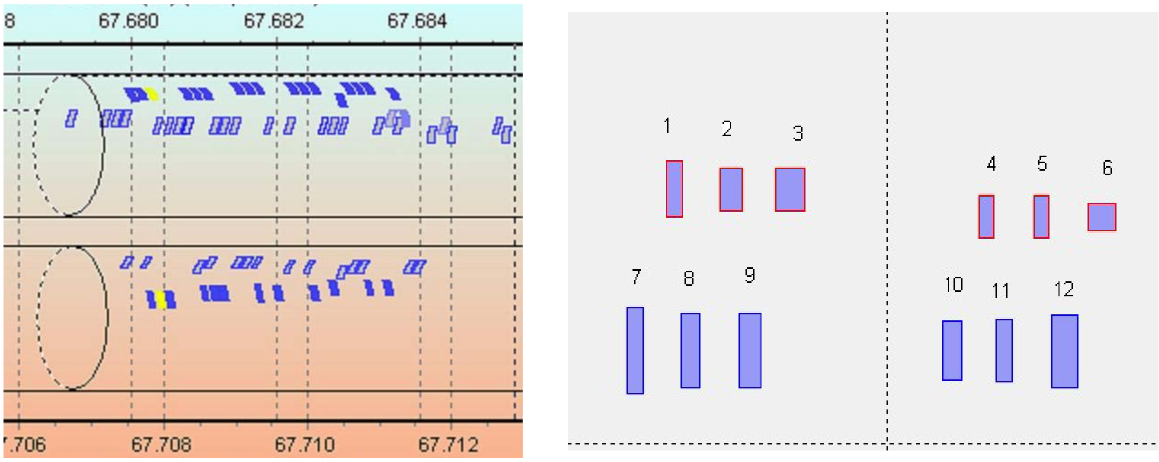


Figure 1-2: visualization of matched defects
(defects 1 to 6 are from run 1 and 7 to 12 from run 2)

[4] Pipeline Integrity Management System

As can be seen from Figure 1-2 it is not always clear after the matching of defects by the program, which defects belong together (7 to 1 or 8 to 1?). In the program however the user has the possibility to define certain areas around a defect. If a defect falls within such an area then the two defects are considered to be the same defect. The advantage of the software is that when many defects are close together the user can optimize the area sizes to get the best possible matching of defects.

Factors that complicate the matching of defects are differences in terminology of suppliers or differences in interpretation of defects: external corrosion defect, internal corrosion defect, mill defect etc. Suppose that the software has matched a defect from run 1 to a defect from run 2. Supplier 1 calls the defect “corrosion” whereas supplier 2 identifies it as “mill defect”. The question is then if this matched defect should be taken for the determination of the corrosion rate. It was decided to work with two scenarios to find out whether it was critical for the determination of the corrosion rate if only the defects with the same identification were used (only indicated as “External Corrosion”) or also defects with different identification (“External Corrosion” and “External metal loss, possible mill defect”). It turned out that the corrosion rates that were calculated for both situations were almost the same. In order to keep the uncertainties in the process as small as possible it was decided to use only the matched defects with the identification of “External Corrosion” in all pigruns.

1.3.1 Number of matched defects

In the matching process different categories of matching defects originated: defects from run A that could not be matched or that could be matched only once (to B, C or D), twice (to B and C, C and D or B and D) or three times (to B, C and D). A similar result was obtained for the matching of defects from the other runs.

For the final data set used for the calculation of corrosion rate it was decided to use only the defects that had been detected by three or more pigs. This resulted in a data set of 52 matched defects with a subdivision as indicated below.

	A	B	C	D
A	-	30	46	45
B	-	-	36	35
C	-	-	-	51
D	-	-	-	-

Table 1: number of matched defects per combination of pigruns

As can be seen from the table, the number of matching defects within this subset was smallest for the comparison of run A and B: only 30 defects matched there.

Altogether a number of 29 defects were reported by all and only four suppliers whereas 23 defects were reported by three suppliers.

1.4 Available data

For the analysis of the data two data sets are available:

1. reported defect dimensions from the pig supplier
2. defect dimensions as determined at the excavation and repair of the defect

1.4.1 Reported defects

Although the claimed performance of the different pigs is comparable (see Table 2) the number of reported external corrosion defects is different per supplier as can be seen in the table below.

Supplier	Sizing accuracy[5]	Number of external corrosion	Date of inspection
A	+/- 10% w.t.	65	October 1999
B	+/- 10% w.t	72	April 2001
C	+/- 10% w.t	1708[6]	June 2002
D	+/- 10% w.t	441	March 2004

Table 2: sizing accuracy of the pigs and number of reported defects

As it is generally known the process of defect recognition consists of three steps: detection of the defect, sizing of the defect and identification of the defect. The experience of Gasunie is that most of the differences between suppliers arise from differences in identification. The distinction between a mill defect and a corrosion defect seems to be troublesome for the suppliers in quite a number of situations. This accounts for part of the differences in the reported number of defects.

Another explanation for the difference in numbers is time related: in general the performance of pigs has improved over the last 5 years and corrosion defects that had a defect depth under the reporting threshold 5 years ago may have grown to a defect depth above the reporting threshold.

1.4.2 Excavation data

After the first pigrun 7 excavations have been performed, in which 17 separate defects have been repaired. All of the defects that have been repaired were manually measured in the ditch by the usual gouges. The defects that have been repaired by welded sleeves could not be used as reference points for the pigruns B, C and D whereas the defects that were repaired by clock spring or coating could be used as reference points.
Table 3 indicates the number of reference points that have been detected by the different suppliers.

Supplier	Number of available reference points from excavation
A	17
B	9
C	11
D	10

Table 3: number of reference points per pigrun

The fact that the number of reported reference defects varies between B, C and D is due to the fact that the applied POF- interaction rules [1] lead to clustering of defects. Differences in defect sizes or distances between defects will inevitably lead to different numbers of reported defects.

[5] for defect depth of general corrosion with 80% confidence level (w.t. = wall thickness)
[6] of which 576 are below 10% of wall thickness

Figure 1-3: clock spring installation (left), gas pipeline excavation (right)

Figure 1-4 shows reported defect depths compared to the real metal loss measured at the excavation (all pig measurements are oscillating around a dashed line- which indicates a linear relationship between real metal loss, and pigs' measurements).

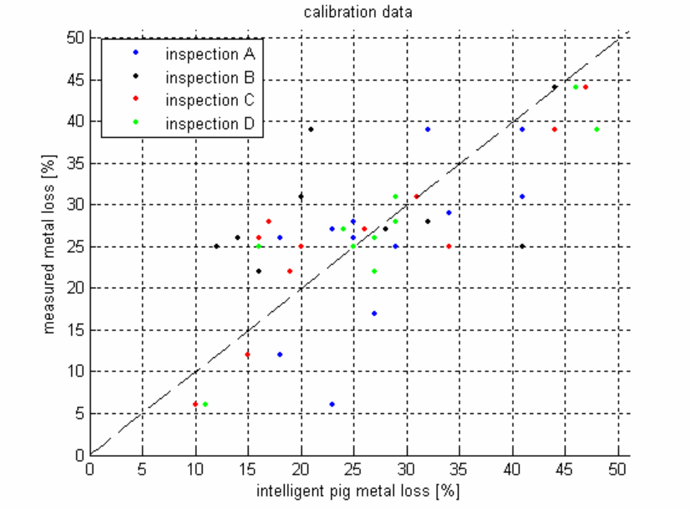


Figure 1-4: reported vs. measured defect depth

Because the excavations were performed in a relatively short time period after the pigrun, it is assumed that defect depth has not significantly changed between the time of the pigrun and the time of excavation.

1.4.3 Measurement data from matched defects

All the matched defects come from different segments with varying wall thickness. In 10 cases the defects are from a segment with a wall thickness of 11.2 [mm], while the rest of the defects are from the pipe with a wall thickness of 12.86 [mm] (These values are nominal wall thicknesses). The measurement dataset of matched defects is presented in Figure 1-5.

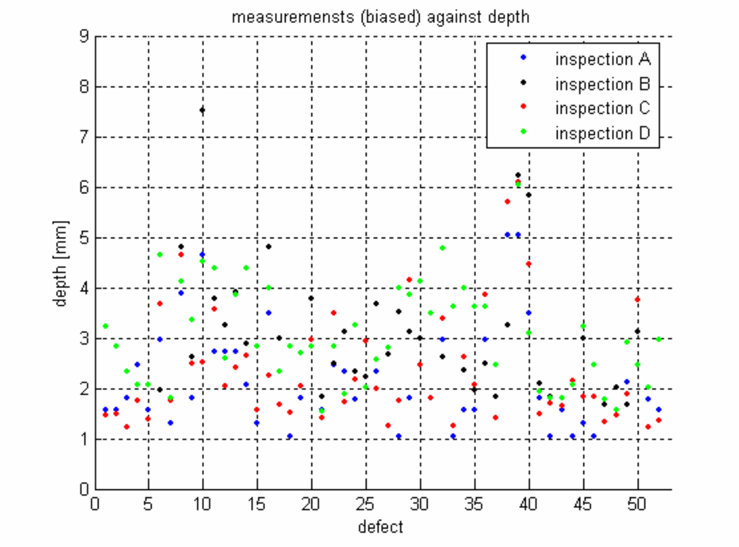


Figure 1-5: reported defect depth

Since the inspections are chronologically ordered (A, B, C and D), it is clear from the above figure that some of the defects are improving in time (corrosion depth is decreasing), which is physically impossible.

Chapter 2

2 Description of the corrosion rate model

2.1 Introduction

Since the corrosion is reported by intelligent pigs, it is very important to know what the accuracy of the pig is. Such information can be obtained from the calibration data collected during pipeline excavations. Given that all data gathered by pigs is not certain it is reasonable to combine the measurements with error distributions obtained from the calibrating procedure. If the excavation indicates that for a certain pig the measurement error is significantly smaller than for the other pigs, then the model should also take this information into account. Another demand which model has to satisfy is to deal with the missing data – if some of corrosion spots haven't been registered in all inspections; so this information also needs to be taken into account and to give physically unreadable estimates.

This chapter shows ways of dealing with uncertainty about measurements, estimating reasonable corrosion rate(s) (i.e. non- decreasing in time) and dealing with the missing corrosion data.

2.2 Data calibration

The data measured at each pigrun can be calibrated by removing the bias. There can be different reasons for this bias in defect measurements. Two of the most important reasons are: a measurement error associated with the measurement technique of a MFL-pig, and an effect caused by the clustering of defects. The result of clustering of several individual defects can be caused by that only the deepest points are compared they may not be related to the same individual defect.

The analysis starts with the measurements calibration for a possible bias. This is one of the most important actions because the calibration influences all the collected measurements.

Due to the small number of excavations, it is difficult to investigate how well a given pig is calibrated for deeper or shallower defects. However, it is possible to check the pigs' accuracies, assuming homogeneity between defects for each pig. By comparing the calibration results, a conclusion on which pig is the best, with respect to a bias spread of the measurement errors can be obtained. The calibration algorithms are in the next paragraph.

2.2.1 Data calibration procedures

Algorithm 1 (Calibration data algorithm)

In order to perform data calibration we need to follow the procedure:

- *take X as a vector of metal loss registered by an intelligent pig*
- *define an actual metal loss vector Y*
- *define the bias vector* $Z = X - Y$
- *calculate the bias by taking the expectation of* $Z\ (EZ)$

This procedure allows to measure (by means of the average value) how "far" is the registered by a pig metal loss from actual the metal. The expectation is equivalent to the measure of bias, and indicates how pig measurements are consistent with actual data.

Algorithm 2 (Measurement error distribution algorithm)

This algorithm presents the procedure for estimating the distribution of measurement error.

- *Use the* ***Calibration algorithm*** *and find* EZ
- *Define the corrected (unbiased) pig calibration measurements as:* $X' = X - EZ$
- *Define the residual random variable* $\varepsilon = X' - Y$
- *Find the distribution of* ε *(using techniques introduced in the background chapter- appendix A1)*

From Algorithm 1 and 2:

$$EX' = E(X - EZ) = EX - EZ = EX - E(X - Y) = EX - EX + EY = EY$$

So, the expectation of "unbiased pig" equals the expected value of actual metal loss.

2.2.2 Data calibration results

The calibration procedure showed that all of the pigs have a bias. All the measurements require removing the bias. The bias for all pigs did not exceed a value of 0.6 [mm], and on average had a level of 0.1 [mm]. Two MFL-pigs led to underestimation and two led to overestimation of defect depth. Table 4 presents the results of the calibration procedure applied to the excavation dataset.

insp.	*no. of calibr. samp.*	*bias [mm]*	*conclusion*
A	*17*	***0.12***	overestimated
B	*9*	***-0.55***	underestimated
C	*11*	***-0.05***	underestimated
D	*10*	***0.12***	overestimated

Table 4: calibration results

Removing the bias can be done by subtracting it from the measurements reported by the corresponding MFL pig. When all the measurements are unbiased, the second stage of the pig calibration can be initiated, namely- the measurement error analysis.

Since none of the MFL pigs reports measurements without error (see Figure 1-4), all the pigs have their own measurement error distributions. The example of the measurement error histogram for pig A with the corresponding theoretical curve is presented in Figure 2-1.

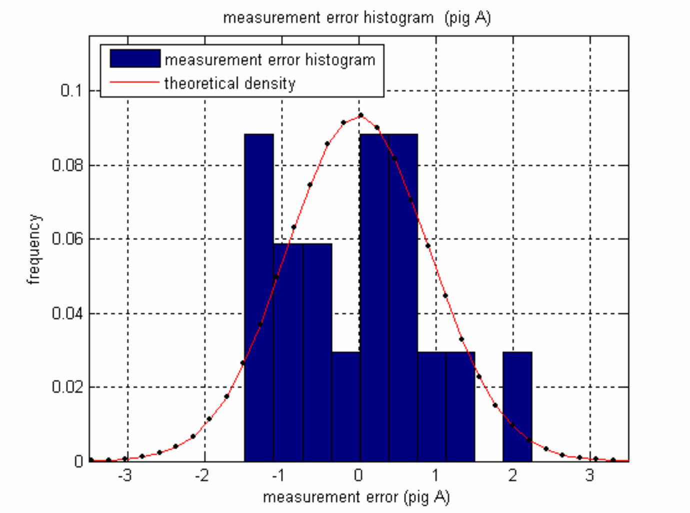


Figure 2-1: error distribution for pig A

When the measurement error distributions for all the pigs are known, the conclusion about the pig's accuracy can be drawn. It depends on two factors. Firstly: on the level of the bias and secondly: on the standard deviation of the measurement error. The analysis showed that for all pigs, under the null hypothesis, the measurement errors are normally distributed cannot be rejected (a significance level is customarily chosen to be 0.05). The standard deviations of the measurement errors are tabulated and presented beneath in Table 5.

inspection	distribution[7]
A	*N(0,0.93)*
B	*N(0,1.34)*
C	*N(0,0.77)*
D	*N(0,0.63)*

Table 5: standard deviations of the measurement errors

The results from the table indicate that pig D (the last inspection) has the smallest spread of the measurement error. The worst one is pig B, which has less than half the accuracy of D. Since the uncertainty about the measurements reported by pigs is known, it is advisable to take this information into account for corrosion rate modeling.

[7] N stands for a Normally distributed random variable with two parameters, mean and standard deviation

2.2.3 General model for data calibration

Let's assume that we have carried out *n* inspections- done by *n* different pigs. Each pig is biased depending on the registered defects depth. This kind of situation requires special treatment, which is the main part of this chapter. Proposed procedure shows ways to avoid (reduce) the correlation between reported defect depth and measurement error.

Suppose, we take a certain pig *i* for which, measurements and actual metal loss can be presented as follows:

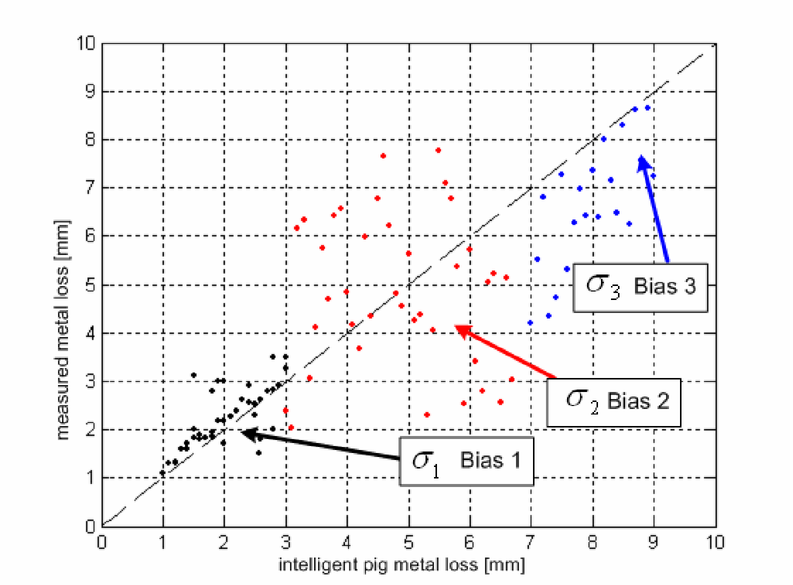


Figure 2-2: example of defects clustering

Figure 2-2 shows that for three different defect populations three different biases can be specified, and three associated error deviations. Of course, the decision about combining defects (*clustering* [8]) into subpopulations generally can be subjective. However the choice of clustering also can be done in mathematical manner. Mathematical tools that work with this problem are so called "*Clustering algorithms*". Literature available on the topic of the clustering is introduced in references [8], [9], [10] and [11]. Given that *i'th* pig measurements are presented in Figure 2-2 above, it is possible to recognize three different defects groups: small defects (*black dots*) where the bias is negative (pig gives lower values than actual loss) with small standard deviation, second- where observations oscillate around actual values but with high spread, and third subpopulation (*blue dots*) where the bias is positive. This observation motivates to distinguish groups of shallow, middle, and deep defects. Such groups should be calibrated separately. Presented situation, might not the case of real measurement; but it is important to know that if such situation occurs then needs to be taken into account in the modeling.

According to previous notation, we have *n* pigruns and each of them can be biased for different clusters of defects. This leads to more general procedure of data calibration than the one introduced before.

[8] "The process of organizing objects into groups whose members are similar in some defined way"

2.2.4 General calibration procedure:

1. *Take calibration data for pig i where* $i = 1, \ldots, n$
2. *Check whether pig's measurements are homogenous, if not, then find number* n_i *of defects clusters (i.e. subgroup of defects, where members are similar in some way)*
    a. *For each* $j = 1, \ldots, n_i$ *apply the algorithm 1 and get the bias for group j.*
    b. *Remove the bias from all measurements obtained by pig i.*
    c. *Apply the algorithm 2 and get the distributions for* $\varepsilon_{i,j}$ *where* $j = 1, \ldots, n_i$

The effects of applied general procedure are:

- All measurements done by pigs can be calibrated according to pig accuracy for different defects depth.
- We have $\sum_{i=1}^{n} n_i$ distributions functions of measurement error, which will be applied in order to estimate the corrosion rate.

In the previous part it was checked that the measurement error distributions for all pigruns are normally distributed. In the general model, if both: the assumption about the same population for all the errors and normal distribution are satisfied then in order to estimate the corrosion rate a linear regression can be applied. On the other hand the least squares errors approximation without imposed any constrains can produce best estimate which for ex. indicates decreasing corrosion rate.
Next paragraph presents the general corrosion rate model and numerical results for describing corrosion growth as a functional dependence on time (inspections).

2.3 General corrosion rate model

Let's assume that according to dataset *n* distinguishable defects in time were observed. Suppose that defect i was observed in n_i inspections. The task is to find the best function of time, which describes the corrosion growth for specific defect i. The first idea is to apply linear regression to all observed values of defect i. This is a reasonable guess, but has significant drawbacks:

- From the collected data it is clear that in many cases inspections report the depths for which the best linear estimate is:
    - decreasing in time- what is unacceptable
    - the slope of a function is too high- it means that the corrosion according to the function grows too fast, and indicates leakage- but such leakage in pipeline was not observed
    - corrosion according to the best estimated function starts before pipeline installation or even pipeline production
- The standard regression estimation can only be applied to the model if it is assumed that the errors are normal, come from the same distribution and are uncorrelated. However, in the case when the calibration procedure is applied, it is clear that the normality might not be the case; furthermore, it can happen that measurements error don't come from the same population (distribution).

The model that has none of introduced drawbacks and according to delivered data gives a functional description of the corrosion rate is presented beneath.
The idea behind the model is to give an estimate which takes into account information about the measurement error distribution for each specific pig (if the case then also clusters for each pig).

First, let's define:

- $f_i : (T_j, \alpha_i^0, \dots, \alpha_i^l) \to R^+$ - theoretical model function with *l+1* parameters, associated with *i'th* defect,
- T_j - time since pipeline installation at *j'th* inspection
- $d_{i\,j}$ -unbiased depth of defect *i,* measured at *j'th* inspection
- w_i - nominal wall thickness where *i'th* defect is observed
- m - total number of inspections
- $P_{i,j}$ - measurement error density function of defect *i* at *j*'th inspection

The function f_i associated with *i'th* defect needs to satisfy following optimization task:

$$\begin{array}{ll} \textit{maximize} \;\; : \; L_i = \prod_{j=1}^{m} P_{i,j}\left(f_i\left(T_j, \alpha_i^0, \dots, \alpha_i^l\right) - d_{i,j}\right) & \\ \textit{subject to :} & f_i\left(T_m, \alpha_i^0, \dots, \alpha_i^l\right) - w_i \le 0 \\ & -\, f_i^{-1}\left(0, \alpha_i^0, \dots, \alpha_i^l\right) \le 0 \\ & f_i \;\; \textit{is non decreasing} \end{array}$$

The first restriction imposed on function f_i says that a value of the function at the last inspection cannot be higher than the pipeline wall thickness where defect *i* was observed. The second condition rejects situations where corrosion initiation according to data is before pipeline installation (if we want to find the corrosion initiation time, we need to find a *t*, for such $f_i(t, \alpha_i^0, \dots, \alpha_i^l) = 0$ i.e. corrosion level at initiation time is exactly equal to zero, hence it is equivalent with $f_i^{-1}(0, \alpha_i^0, \dots, \alpha_i^l) = t$). The third and last constraint says that the function associated with defect's growth cannot be decreasing in time.

2.3.1 Example

Suppose that:

- In three inspections one defect *i* was observed.
- Each time, the measurements were done using different pigs.
- From calibration procedure it is known that all three pigs have nonhomogenous measurement error i.e. parameters (or distribution) are different for different pigs.

Assuming linearity of defect's growth, the model has to find such estimators of α_i^0 and α_i^1 for which the Likelihood function L_i is maximum. The function: $f_i = \hat{\alpha}_i^0 + \hat{\alpha}_i^1 t$ is a function that describes the corrosion growth in time for specific defect *i*. The schema of this procedure is presented below in Figure 2-3.

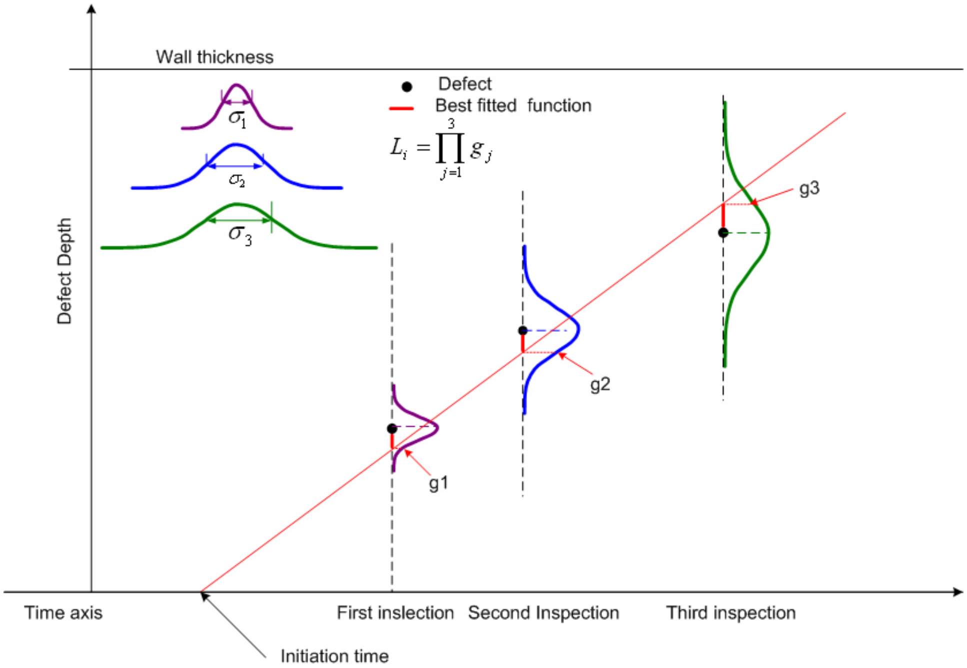


Figure 2-3: corrosion rate determination- general model

It is clear that L_i gets the higher value if the estimated line is closer to real measurements. In the case when the line goes through all observed measurements, then this line is the best, and the likelihood is maximal, hence this approach agrees with natural expectation.

Remark
The optimization process can be performed by applying techniques introduced in a field of optimization as "multidimensional constrained non- linear programming". The results presented in the report are obtained by using Matlab[9] optimization toolbox. Furthermore because of the computational complexity of introduced non-linear task it is worth to transform the task by using logarithm transformation[10]. The implementation of the formulated problem is presented in the appendices.

2.3.2 Dealing with missing data

Many of defects were observed only in three inspections (while total number of inspections is four). The model assumes that if depth of certain defect was not reported during inspection, then the measurement error density function for this defect is uniform on the interval bounded by the minimum and maximum observed defect's depth.
Suppose that defect *i* was not observed at third inspection, then in optimization problem the measurement density function for unmeasured depth is $P_{i,3} = 1_{[\text{min depth of i'th defect ,max depth of i'th defect}]}$. This means that if a certain defect was not registered, then the function describing corrosion growth is derived by using only reported defects. The draft of such situation is presented on the Figure 2-4.

[9] Matlab 7.0.0.19920 (R14)
[10] Any monotonic transformation of a function doesn't change its extremes (mA1, min).

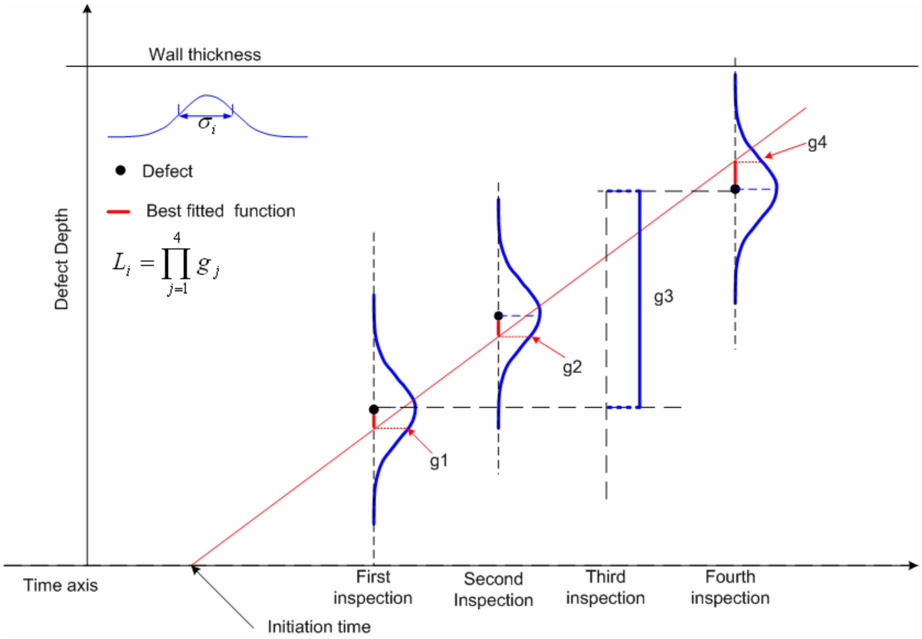


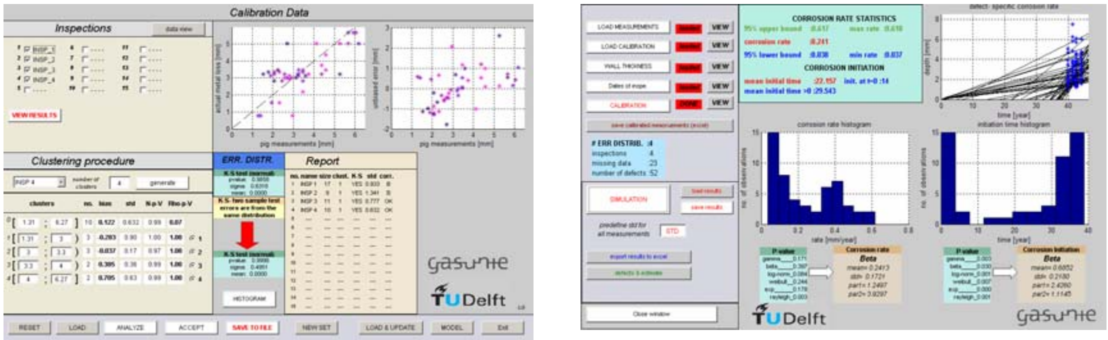
Figure 2-4: corrosion rate modeling- dealing with missing data

2.4 Conclusions

The model presented in this section gives the corrosion rate estimate when low number of the defect measurements is available. Very often standard regression model doesn't give reliable and acceptable results; hence alternative is required. For many defects the regression estimates are negative or indicate defect initiation before pipeline installation. The General corrosion rate model solves all these drawbacks, and also takes into account information on pigs' accuracies.

2.5 Implementation (CoroGas 1.0v)

The theory introduced in this chapter has been implemented in CoroGas, software package developed by the author. This program analyzes the excavation data, unbiases the measurements, assesses the weights for MFL-pigs and gives estimate of the corrosion rate. The program has implemented algorithms for predefining the clusters for the calibration.



CORROSION RATE STATISTICS
95% upper bound 0.617 max rate 0.618
corrosion rate 0.241
95% lower bound 0.038 min rate 0.037
CORROSION INITIATION
mean initial time -22.157 init. at t=0 :14
mean initial time >0 :29.543

ERR DISTRIB. :4
inspections 4
missing data :23
number of defects :52

Figure 2-5: calibration data & optimizer window of CoroGas

Appendix B describes CoroGas and all available functions.

Chapter 3

3 Numerical results

This chapter is mostly based on the article "Determination of the corrosion rate of MIC influenced pipeline using 4 consecutive pigruns" by Lech A. Grzelak & Giorgio G.J. Achterbosch published in "International Pipeline Conference" (IPC06-10142)

3.1 Introduction

Three different approaches for corrosion rate modeling will be presented. The analysis is carried out starting from the simplest to more sophisticated models. The first and the second approach are simply based on the unconstrained regression analysis. The third and the last model, is based on the unbiased measurements and pigs' accuracy described in the Chapter 2.

3.2 Approach 1

In the first approach all the defects are pooled in 1 dataset and no corrosion rate is calculated for individual defects but only for the dataset as a whole.

The first approach starts with verification if the hypothesis that the measurement errors for all MFL pigs are from the same population and are normally distributed can be accepted. This was the case. According to the maximum likelihood estimation for the measurement error the parameters are 0 (mean) and 0.91 (for a standard deviation). If it is assumed that for all the errors, there is no correlation between them, then the task of finding a corrosion rate associated with all the measurements is equivalent to a Gauss Markov regression model.

Let’s define necessary matrixes *X* and *Y* in following way:

- $X = \begin{bmatrix} 1_{m\times 1} & \begin{bmatrix} (T_1)_{m_1\times 1} \\ \vdots \\ (T_n)_{m_n\times 1} \end{bmatrix} \end{bmatrix}_{m\times 2}$
- $Y = \left[d(T_1)_{m_1\times 1} \quad d(T_2)_{m_2\times 1} \quad \cdots \quad d(T_{n-1})_{m_{n-1}\times 1} \quad d(T_n)_{m_n\times 1} \right]^T_{1\times m}$

  where:

  - $d(T_i)_{m_i\times 1} = \left[\underbrace{d(T_{i,1}) \quad d(T_{i,2}) \quad \cdots \quad d(T_{i,m_i-1}) \quad d(T_{i,m_i})}_{m_i \text{ times}} \right]^T$ -a vector of unbiased depths measured by pig at time T_i, second index indicates defect’s number
  - n total number of inspections
  - m_i - number of defects at *i’th* inspection
  - *m* is total number of observed defects at *n* inspections ($m = m_1 + \ldots + m_n$)
  - $1_{m\times 1} = \left[\underbrace{1 \quad 1 \quad \cdots \quad 1 \quad 1}_{m \text{ times}} \right]^T$
  - $(T_i)_{m_i x 1} = \left[\underbrace{T_i \quad T_i \quad \cdots \quad T_i \quad T_i}_{m_i \text{ times}} \right]^T$ where T_i is a time of *i’th* inspection

If additionally, it is assumed that the corrosion rate is uniform over time (i.e. corrosion growth is linear), then an application of the Least Squares Error (LSE) method gives a linear description of the corrosion growth in the following form: $y = \hat{\alpha}^0 + \hat{\alpha}^1 t$ and the remaining issue is to find the estimator $\hat{\beta} = \left[\hat{\alpha}^0 \quad \hat{\alpha}^1 \right]^T$ for the linear function.

Standard calculations give that the estimators for unknown parameters are:

$$\hat{\alpha}^0 = -2.42 \text{ and } \hat{\alpha}^1 = 0.12$$

Coefficient $\hat{\alpha}^1$ is equivalent to the measure of the corrosion rate [mm/yr], so the LSE model estimated a corrosion rate for the calibrated measurements of **0.12 [mm/yr]** with 95% confidence interval **[0.05, 0.20].**

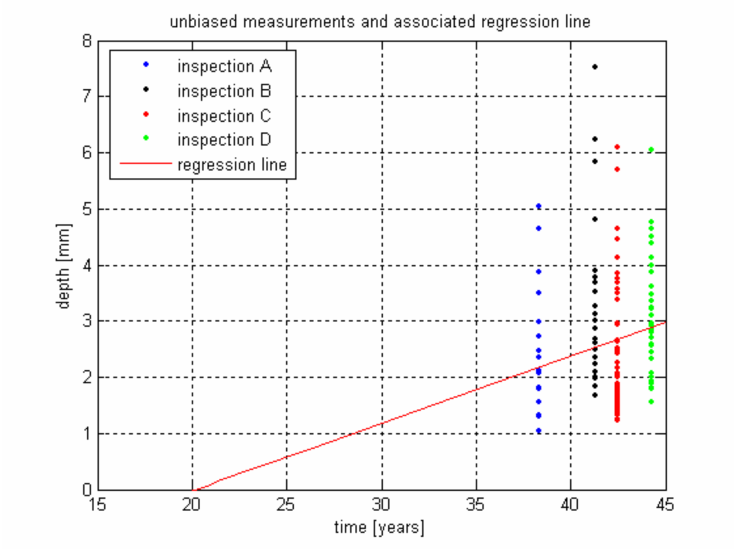


Figure 3-1: defect depth in time

To check how well the model fits the data, a determination coefficient is calculated. A goodness of fit measure resulted in $R^2 = 0.04$. This is poor because it indicates low relative predictive power of the model. According to the model, the initiation time for corrosion is **20** years [yrs since pipeline installation].

Even though the estimated parameters are in an acceptable range, this approach has significant drawbacks:

- the model does not distinguish defects
- it does not take into account that some of the defects are improving in time (decreasing defect's depth which is physically impossible), or for some the defects initiation time is before pipeline installation
- the model assumes that all the defects have one corrosion rate

3.3 Approach 2

As was pointed out in the previous section, the first approach has significant drawbacks. The second approach, proposes a way of dealing with some of the enumerated disadvantages. Like before it is assumed that corrosion growth is linear in time.

The second model checks what the corrosion rate is, if the defects are analyzed individually i.e. the model does not assume any more that there is only one corrosion rate for all the defects but it calculates a corrosion rate per defect. A simple regression analysis applied to each unbiased defect gives the following graph.

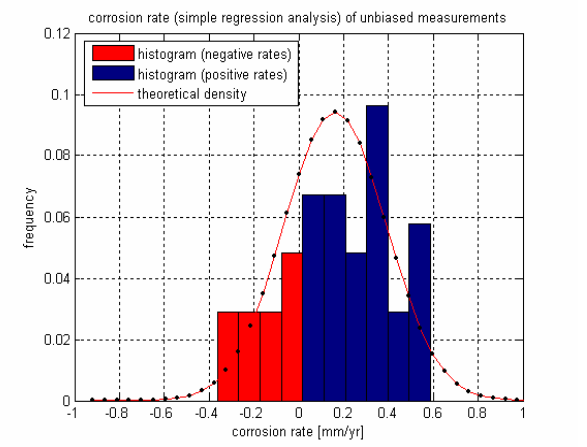


Figure 3-2: corrosion rate distribution

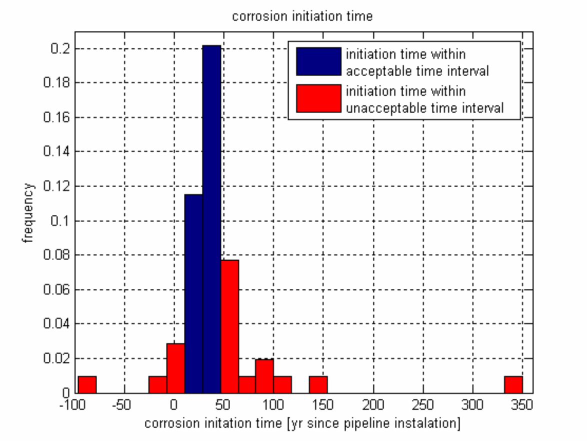


Figure 3-3: distribution of initiation time

From the histogram presented in Figure 3-2 it is clear that in many cases, simple regression analysis applied to each defect results in negative corrosion rates. The mean corrosion rate according to this model is **0.16** [mm/yr] which is close to the result obtained before, however the 95% confidence interval for the corrosion rate is quite different: **[-0.31, 0.54]**. The 95% confidence interval comprises negative values.

The number of the defects indicating either negative corrosion rate or corrosion initiation time before the pipeline installation is 16. One way of dealing with a negative corrosion rate is to remove all the outliers from the dataset. However, such treatment is undesirable since the dataset consists of 30% bad defects. Further investigation confirmed that the corrosion rate follows a normal distribution. The initiation time of the corrosion is presented above, also in the form of a histogram. The red bars in the picture indicate an initiation time outside the interval determined by the time of pipeline installation (t=0 [yr]) and the time of the last inspection (t=44.25 [yr]). A summary of the results obtained from the second approach is presented in Table 6.

results		corrosion rate	init. time
mean		**0.16 [mm/yr]**	**44.56 [yr]**
95% conf. int.	**Lower bound**	-0.31 [mm/yr]	-28.77 [yr]
	Upper bound	0.54 [mm/yr]	179.49 [yr]

Table 6: corrosion rate and initiation time for approach 2

Still the Least Squares Errors approach produces negative corrosion rates or initiation times before pipeline installation. Therefore an alternative model for the presented models is presented: approach 3[11].

3.4 Approach 3

This approach is based on the General corrosion rate model introduced in previous chapter.

The model takes into account information about the measurement error distributions for each specific pig and according to these distributions assigns weights to the measurements. The weights are chosen in the following way: a pig which is accurate influences the final results stronger than a pig with a lower level of accuracy.

[11] Approach 3 is simplified version of General corrosion rate model.

Assuming that the function associated with *i'th* defect is linear in time ($f_i(t,\alpha_i^0,\alpha_i^1)=\alpha_i^0+\alpha_i^1 t$), let's define simplified version of general corrosion rate model in following way:

- $f_i:(T_j,\alpha_i^0,\alpha_i^1)\rightarrow R^+$ - theoretical linear function, associated with *i'th* defect (*number of defects is 52*)
- T_j - time since pipeline installation at *j'th* inspection
- $d_{i\,j}$ -unbiased depth of defect *i,* measured at *j'th* inspection
- w_i - nominal wall thickness where *i'th* defect was observed (*two cases: 11.2 for 10 defects and 12.86 for 42 defects*)
- m=4 - total number of inspections
- $P_{i,j}$ - measurement error density function of defect *i*[12] observed at *j'th* inspection

The likelihood estimation is optimal when the following function is maximized:

$$maximize: L_i = \prod_{j=1}^{m} P_{i,j}(\alpha_i^1 T_j + \alpha_i^0 - d_{i,j})$$

$$subject\ to: \qquad \alpha_i^1 T_m + \alpha_i^0 - w_i \le 0$$

$$\alpha_i^0 / \alpha_i^1 \le 0$$

$$-\alpha_i^1 \le 0$$

The results of the corrosion rate obtained by the model are presented in the Figure 3-4.

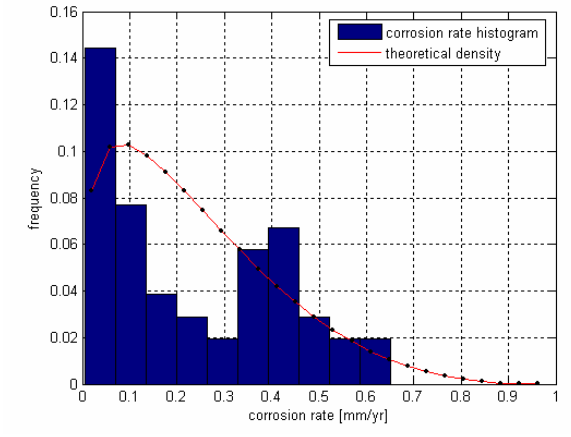


Figure 3-4: corrosion rate distribution for approach 3

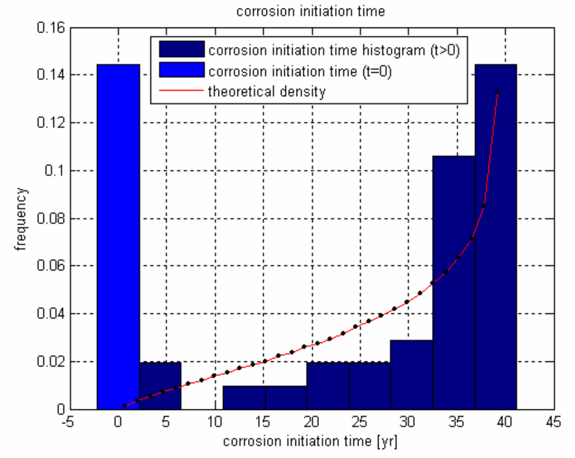


Figure 3-5: distribution of initiation time for approach 3

Because of the constraints the model's output is in harmony with the physical corrosion properties: none of the corrosion rates are negative.

[12] A measurement error density function for each defect can depend on the cluster, from which the defect comes. I.e. defects from one pig can have a few distributions- one per individual cluster, however, here in the analysis because of low number of calibrating data each pig has one distribution

In the other two approaches the corrosion rate distribution was normally distributed, now it is quite different—there is no basis to reject the hypothesis that the corrosion rate is Beta distributed (with parameters 1.25 and 3.92).

The histogram of the initiation time is presented in the Figure 3-5. The model says that all the initiation times are in acceptable time intervals. However, 14 of the defects initiate at the time of pipeline installation- this is indicated with a lighter blue color in the picture above. The initiation time distribution is a composition of two distributions, discrete in 0 and continuous elsewhere. The continuous part is also Beta distributed (with parameters 1.60 and 0.50). A small summary of the results of the corrosion rate and the corrosion initiation time is presented below.

results		**corrosion rate**	**initiation time**
mean		**0.24 [mm/yr]**	**22.16 [yr]**
95% conf. int.	**Lower bound**	0.038	0 [yr]
	Upper bound	0.62	38.36 [yr]

Table 7: corrosion rate and initiation time for approach 3

As is shown in the Table 7, all the results obtained from the model are acceptable because they are physically possible. The model even for a bad dataset gives reasonable and acceptable results. The results from Table 7 show that the model gives a corrosion rate that is approximately 0.05 higher than the previous ones. The reason for this is the following: in the previous cases, the result of corrosion rate was an outcome of all corrosion rates (including negative values), which resulted in a lower average value.

3.5 Depth influence on the corrosion rate

In order to investigate whether the defects' depths have a significant influence on the corrosion rate, the dataset of 52 unbiased measurements is divided into two subsets. First, the weighted average for each defect was calculated. The weights were associated with the measurement errors' standard deviations i.e. an accurate pig has the highest weight etc. Then, the dataset was divided into two subsets, namely "shallow" and "deep" defects, in such a way that the MFL pig with the lowest number of observations (pig B) has an equal number of measurements in both sets. For both subsets, the General corrosion rate model was applied. The mean corrosion rate for deep defects turned out to be 0.25 [mm/y], and for shallower ones the rate is 0.23 [mm/yr]. However statistically, there is no basis to reject the null hypothesis that these mean values are the same. Hence it follows that statistically the corrosion rates for deep and shallow defects are not significantly different. The histograms of the rates for shallow and deep defects are presented beneath.

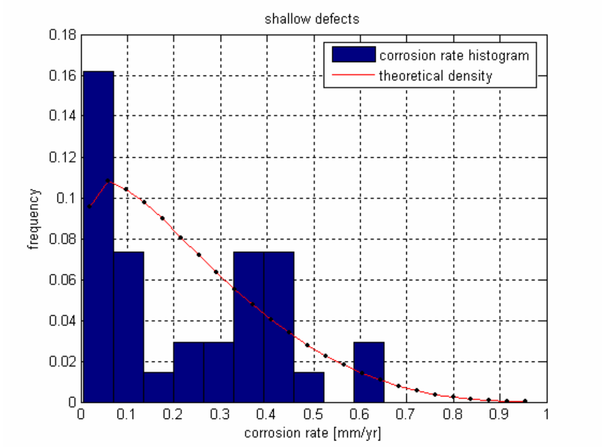


Figure 3-6: distribution of corrosion rates for shallow defect

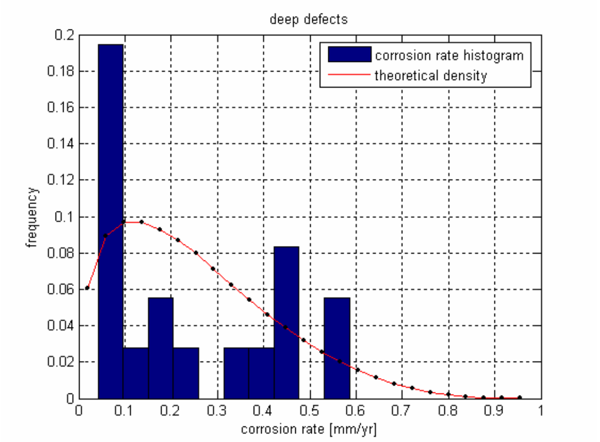


Figure 3-7: distribution of corrosion rates for deep defects

In both cases (shallow and deep defects) there is no basis to reject the hypothesis that the corrosion rates are from a beta distribution.

Although the number of measurements of pig B is equal in both subsets, the number of defects is not the same in these two sets. The number of shallow defects, according to the presented criterion, is 34. The rest (18 defects) are deep defects. Analysis showed that if the set of unbiased measurements is divided in such a way that the number of shallow and deep defects is equal in both subsets, it leads to the same conclusion that there is no significant difference in corrosion rates.

3.6 Conclusions and recommendations

In order to determine reliably the bias for a MFL-pig it is crucial to have multiple reference defects in the pipeline for which the dimensions are well known. These can then be used to calibrate the MFL reported values. For the described pipeline the number of available reference points was limited but still made it possible to estimate the bias for every pig. Because the measurement uncertainty of the MFL-tool is dominant compared to the corrosion growth in the time period between the pigruns, it is very difficult to determine a reliable corrosion rate per defect. However, by assuming a similar corrosion process (MIC) for each defect, based on evaluation of the MFL signals, historical CP-measurements and results of excavations, in combination with the assumption of a linear corrosion growth, it was possible to calculate a realistic corrosion rate for this pipeline. Depending on the approach that was used a value for the average corrosion rate of 0.12 mm/yr (approach 1), 0.16 mm/yr (approach 2) or 0.24 mm/yr (approach 3) was obtained. The numbers for the 95% upper bound values were respectively 0.20 mm/yr, 0.54 and 0.62 mm/yr. The results from the first two models are clearly underestimating the corrosion rate since the final result is an average over positive and negative corrosion rates. The idea of the third approach is quite different: the estimate of the corrosion rate is an outcome of all the defects' growths. Firstly, the model describes each defect separately as time dependent function. A function that needs to satisfy imposed constrains derived from physical features (i.e. function cannot be decreasing in time etc.). When all these functions are known, then the information about corrosion rate associated with defects' population is a product of all functions derived for all defects.

The estimates of the third approach are input data for the section 2 where influencing factors are investigated.

Part II

Parameters influencing microbiologically induced corrosion rate

Chapter 4

4 Potential analysis

4.1 Introduction

Cathodic protection (CP) systems are used to protect buried steel pipelines. The exceptions might be instances the pipelines are installed in fairly non- corrosive soil and where regulations do not require such CP systems. According to regulatory agency requirements, a pipe-to-soil potential is to be at least -0.85 millivolts with reference to a copper-copper sulfate reference electrode[13]. More negative potential protects more against galvanic corrosion, but on the other hand too negative potentials may damage the coating protection[14]. The idea behind cathodic protection is to ensure a current flow towards the pipeline opposed to a corroding current away from the pipeline.

This chapter analyzes if there is any relationship between the potentials measured at test-posts and the corrosion rate. Moreover, the results from this chapter will be applied in the regression analysis in Chapter 5 (Microbial data analysis).

Available "potential" dataset delivered by Gasunie is not a set of potentials associated with real potentials at the coating defects but so called on-potentials measured at test posts. On-potentials contain an IR-drop component. This is the potential drop in the soil between the location of the reference electrode (somewhere at ground level) and the steel/soil interface at coating defects. The IR-drop is caused by CP- and stray currents in the soil. These on-potentials are only an indication of the general status of the Cathodic Protection system. Usually on-potentials were collected during a certain time (about 5, 15, 60 minutes) - during this time maximum and minimum potential were recorded.

[13] Criterion: NEN-EN 12954

[14] The criteria are set to prevent corrosion. This does not necessarily mean that corrosion will occur when the criteria are not met. It is of course not the case that a pipeline corrodes at –849 mV and does not corrode at –851 mV. The potentials are with reference to a $Cu/CuSO_4$ reference electrode.

On-potentials can vary continuously due to e.g. interference from other currents in the soil. These variations are superimposed upon the CP potential of the pipeline. Because potentials are recorded over a certain time-interval (5 minutes, 60 minutes or 24 hours) and only min/max values were recorded the exact on-potential is not always known.
It is also not clear if values indicated as max or min were measured only 1% time or a large part of time. The assumption imposed on the measurements is that variation of the CP on- potentials in certain local time interval is limited. To account for outliers in the measurements a smoothing procedure was applied.

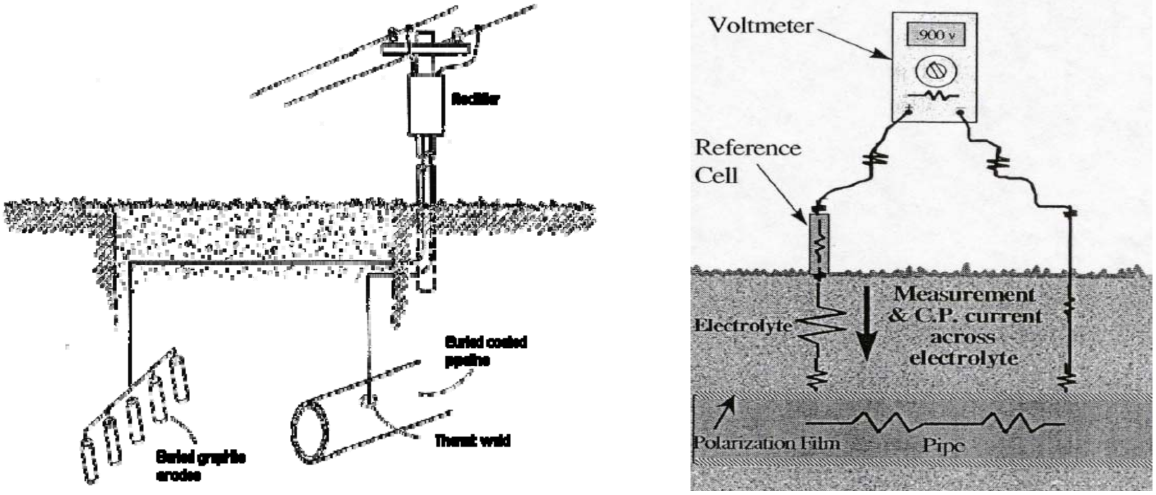


Figure 4-1 cathodic protection for a gas pipeline (left), voltage drops in a measuring circuit (right)

Figure 4-2: Cu/CuSO$_4$ reference electrode

The data which will be analyzed in this chapter was collected for the pipeline for which the estimates of corrosion rates for 52 distinguishable defects are available. Let’s call this pipeline A3.

4.2 Measurements and smoothing method

In analysis one of the smoothing methods called moving averages was applied. The smoothing algorithm minimizes local variability of measurements, allowing to spot trends. The moving average is one of the simplest and oldest analytical tools around. Some patterns and indicators can be somewhat subjective, where analysts may disagree on if the pattern is truly forming or if there is a deviation that might be an illusion. The moving average is more of a cut-and-dry approach to analyzing potential changes and predicting performance. The mathematical formulae for moving average are presented in appendix A.

The on-potential measurements were collected at 67 test posts randomly distributed along the pipeline. It was possible to distinguish 302 different dates when the measurements were collected. Picture below presents how measurements from the test posts were distributed over time along the pipeline.

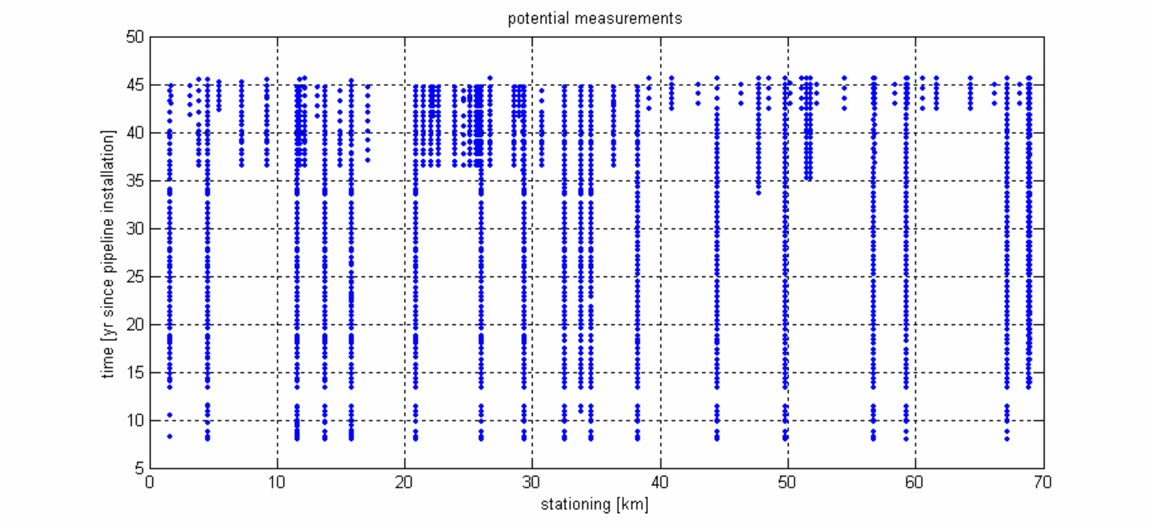


Figure 4-3: Test posts distribution over time

From the Figure, it is clear that number of test posts increased over time from only 18 test posts in the 60's to 67 today. Firstly, a grid which presents on-potentials variability wrt stationing and time has to be defined. The idea is to reconstruct potentials given partial available data. The grid will consist of 302x67 points, where each point will present the measured average potential (since only max and min are measured the analysis will be carried out wrt to average of these measurements). The uncertainty of the measurements over time per test post is relatively high- it is indicated by high variation of the potentials over a small period of time.

The Figure 4-4 below presents potentials at certain stationing before and after smoothing method applied (spam equal to 3 was applied).

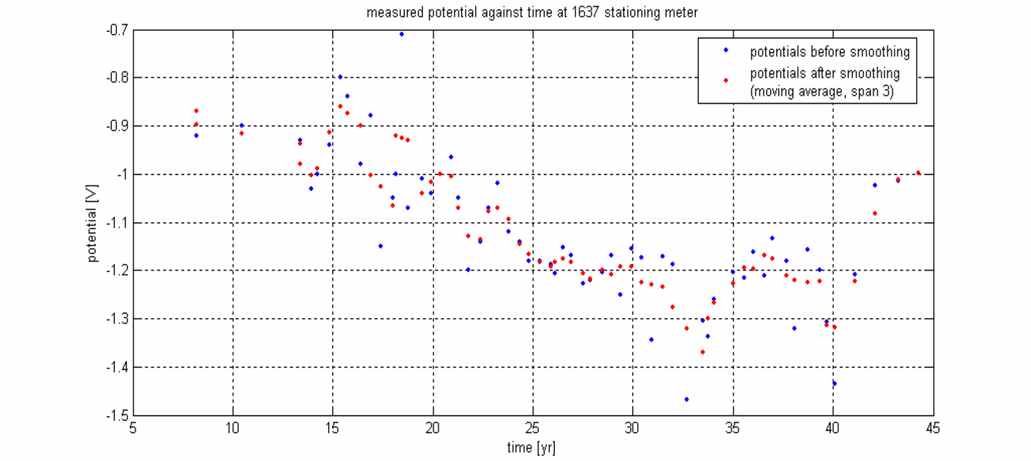


Figure 4-4: on-potentials measurements and results of applied smoothing method

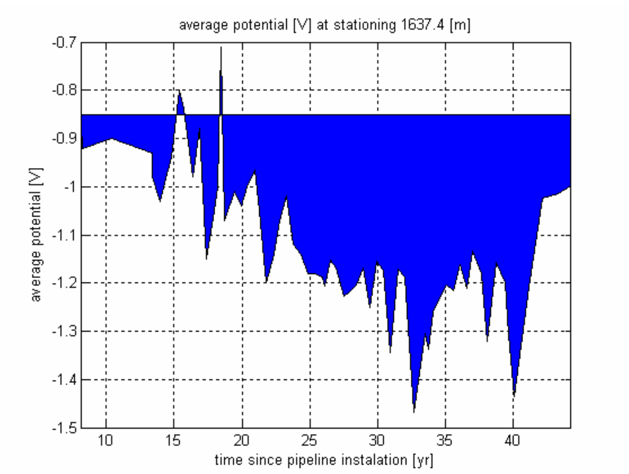


Figure 4-5: on-potentials before smoothing

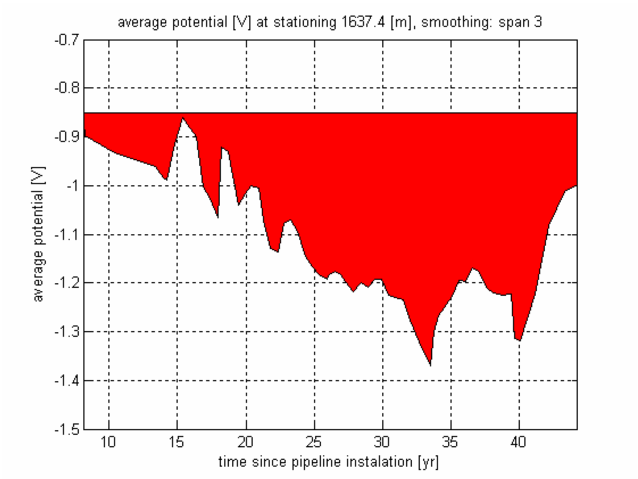


Figure 4-6: on-potentials after smoothing

Figure 4-5 and Figure 4-6 present how the applied smoothing procedure reduced the number of outliers in the dataset. The pattern of potential change over time can be recognized.

Through out the analysis it is assumed that the change of potential between the measurements is linear, it is sensible since there is no additional information available.

4.3 On-potential Grid construction

In order to define the potential grid of the pipeline, all the measurements from the test posts are used.

Suppose that:

- $S_1, S_2, S_3, ..., S_r$ - stationing of the test posts,
- for each test post it is possible to define a function $g_i : t_{i,j} \rightarrow V_{i,j}$ where $t_{i,j}$ - indicates the time since pipeline installation at time *j* at stationing S_i
- $V_{i,j}$ - is a average potential at test post S_i at time $t_{i,j}$

In order to formulate pipe-to-soil potential grid with respect to time and stationing, following procedure is defined:

1. Apply smoothing procedure for each test post (smoothing with respect to time), span of smoothing procedure should be chosen in a such way that the average potential from data corresponds to physical potential phenomenon i.e. local change of average potential cannot change too sharply.

2. Define $\tilde{T}$ as **ordered times** of collected measurements (for all the test posts) $\tilde{T} = \{t_{1,1}, t_{1,2}, ..., t_{i,m_i}, ..., t_{r,1}, ..., t_{r,m_r}\}$ where $\forall i, j, k, l \quad t_{i,j} \neq t_{k,l}$ and m_i - indicates the last measurement recorded at stationing S_i. $\tilde{T}$ consists of ordered dates of all collected measurements. Since none of test posts was observed for all $\tilde{T}$,

the interpolation for all the test posts is required. Assuming linearity between measurements, interpolation can be done in following way:

- Suppose that two different measurements for one test post were collected $V_{i,j}$ and $V_{i,k}$ where $k > j$, then convex combination for $V_{i,l}$ (for $t_{i,l} \in \tilde{T}$) can be presented in following way: $\forall l < k \quad \forall l > j$ $V_{i,l} = (1-\alpha)V_{i,j} + \alpha V_{i,k}$ where $\alpha = (t_{i,l} - t_{i,j})/(t_{i,k} - t_{i,j})$, but still, the interpolation can only be applied if $t_{i,j}$ and $t_{i,k}$ are defined[15].

3. Because of an increasing number of test posts over time, it is possible to generate the measurements for a given stationing (even between the test posts). This can be done by using a linear interpolation between test posts for each time from $\tilde{T}$. If one takes one test post *i* then $\forall t_j \in \tilde{T} \quad V_{i,j} = (1-\alpha)V_{p,j} + \alpha V_{q,j}$ where p and q indicate the closest monitored test post, where $\alpha = (S_{i,j} - S_{p,j})/(S_{q,j} - S_{p,j})$, as before $V_{i,j}$ can be calculated if $V_{p,j}$ and $V_{q,j}$ are defined.

4. In the case when interpolation cannot be carried out, because of missing boundaries, then the simplest way is to generate these data by using linear regression approach applied to each test post.

Beneath, the results of the applied technique are presented. In the smoothing technique, span 3 was chosen. Investigation showed that changing the spam for moving average doesn't change significantly final results.
Application of the introduced technique produced the following contour and 3D plot of the average potentials over the last 45 years along the pipeline.

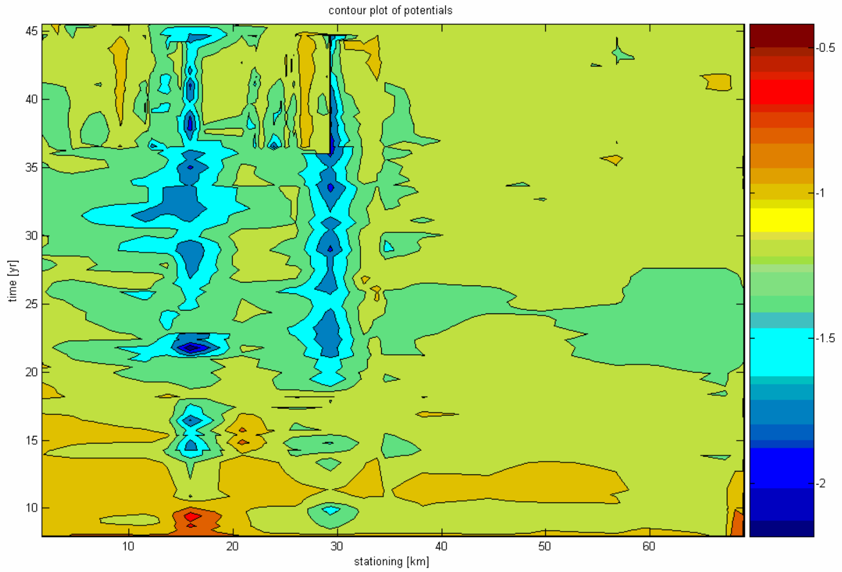


Figure 4-7: on-potential contour plot

[15] It is possible to generate missing data- the procedure for that will be introduced later on in text.

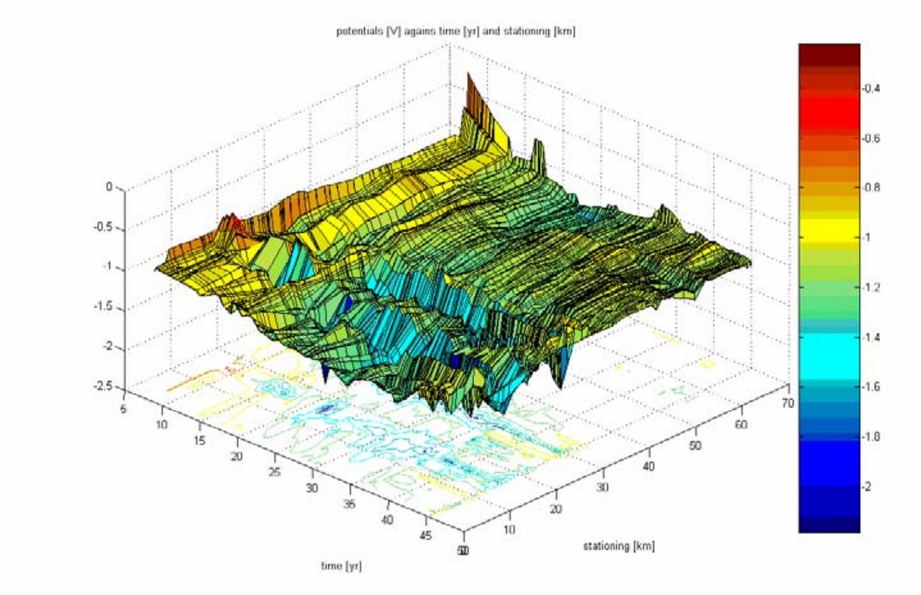


Figure 4-8: on-potential grid

The Figure 4-7 and Figure 4-8 show the average potentials registered at the test posts during last 45 years. The contours present how potentials were changing over time. Plot indicates lack of the cathodic protection for the first pipeline kilometers for a first few years (the potential is significantly higher than the one registered after 20 years. From these figures it looks like CP increased up to 30 years then stabilized more or less for about ten years and then started to decrease. The plots show that in the first 20 km section of the pipeline the potential has significantly decreased after first 20 years. If one compares obtained potential with places where 52 defects are distributed, a blurry pattern can be recognized.

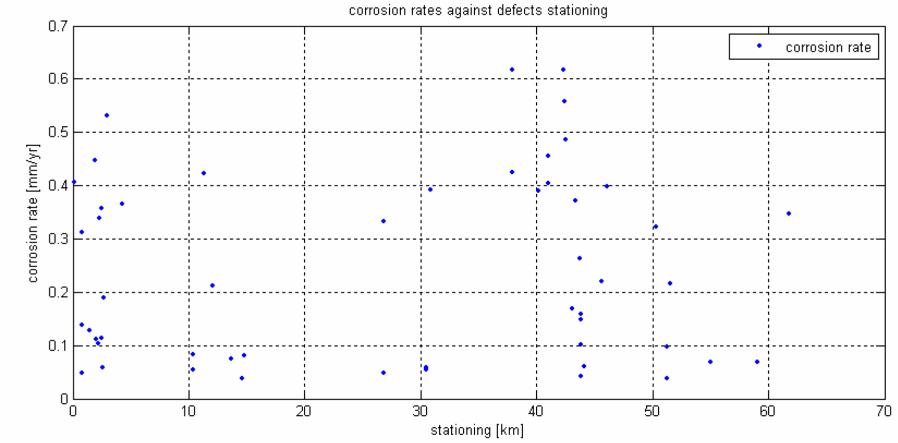


Figure 4-9: 52 distinguished defects and estimated corrosion rates

For the estimated on-potential grid a correlation between corrosion rates and average potentials has to be calculated. For each of the 52 distinguished defects for the pipeline A3 the potentials can be easily obtained using techniques introduced before i.e. for each stationing it is easy to find neighboring test posts and interpolate the potentials over time. Since the estimated corrosion rate of the 52 defects is based on 4 inspection and measurements are carried out within 5 years the average level of on-potentials should be taken only from that period. The results are presented below.

4.4 Correlation between the corrosion rate and average potentials recorded between first and last inspection.

4.4.1 Approach 1

First, let's define:

- $\tilde{V}_i = \left[V_{i,T_1}, \ldots, V_{i,T_m}\right]^T_{1xl}$, where T_1 and T_m indicates time of first and last inspection, $V_{i,j}$ - is on-potential at stationing *i* at time $t_{i,j}$, $i = 1,...n$, *n*- number of the defects and *l* indicates number of available measurements (real and interpolated between first and last inspection)
- $R = \left[r_1, \ldots, r_n\right]^T_{1xn}$ - vector of corrosion rates for the corresponding defects
- $V = \left[\overline{V}_1, \ldots, \overline{V}_n\right]^T_{1xn}$ where $\overline{V}_i = \frac{1}{l}\sum_{k=1}^{l}\tilde{V}_i(k)$

The correlation between corrosion rate and average potential is defined as: $\rho(R,V)$. Small summary of applied techniques is tabulated below. The p-value[16] presented in the last column corresponds to the hypothesis:

$$H_0 : \rho(R,V) = 0 \text{ against } H_1 : \rho(R,V) \neq 0$$

No. of samples	type of correlation	ρ	p-value for the hypothesis that correlation is insignificant
52 defects, l =61	**Pearson**	***0.20***	*0.15*
	Spearman	*0.17*	*0.23*
	Kendall	*0.12*	*0.22*

Table 8: correlation between the corrosion rate and average on-potential recorded between first and last inspection

The normality condition for product moment correlation is satisfied, for a chosen significance level $\alpha = 0.05$ the p-value suggests that there is no basis to reject the null hypothesis that the correlation is insignificant. For a significance level 0.2 there are basis to reject the null hypothesis. Overall conclusion is that there is a weak positive correlation between average potentials and the corrosion rate of defects.

4.4.2 Approach 2

Second approach presents another way of looking at the connection between corrosion rate for the defects and average soil-to-pipe potentials. Suppose that we observe *n* distinguishable defects, and each of the defects has given an estimated corrosion rate r_i $i = 1,...,n$. For each defect it is easy to find (from the generated grid) the average potential vector $V_i = \left[V_{i,T_1},...,V_{i,T_m}\right]^T_{1xl}$ where T_1 and T_m indicate time of the first and last inspection (pigrun). The total number of possible different pairs of defects is $C_n^2 = \sum_{i=1}^{n-1} i$.

For each pair of defects *(i,j)* verify the hypothesis:

[16] Definitions and interpretations are included into appendix A

$$H_0 : \frac{1}{l}\sum_{k=1}^{l}\left(V_{i,k} - V_{j,k}\right) = 0 \text{ against } H_1 : \frac{1}{l}\sum_{k=1}^{l}\left(V_{i,k} - V_{j,k}\right) \neq 0$$

Under the assumption that $\left(V_i - V_j\right) \sim N\left(\hat{\mu}, \hat{\sigma}\right)$- parameter $\hat{\mu}$ is an unbiased maximum likelihood estimator. The estimator for σ is assumed to be unknown. Under the normality condition the cases a) and b) are counted:

a) $\overline{V_i} > \overline{V_j}$ & $r_i > r_j$ or $\overline{V_i} < \overline{V_j}$ & $r_i < r_j$ b) $\overline{V_i} > \overline{V_j}$ & $r_i < r_j$ or $\overline{V_i} < \overline{V_j}$ & $r_i > r_j$

The first point a) is equivalent to concordance of corrosion rate and average potential, and the second one to their discordance. In both cases there is a statistically defendable difference between averages of potentials.

Results:

1. total number of pairs for 52 defects is 1326
2. in 1174 cases the null hypothesis was rejected
3. in 564 the null hypothesis was rejected and normality condition was satisfied

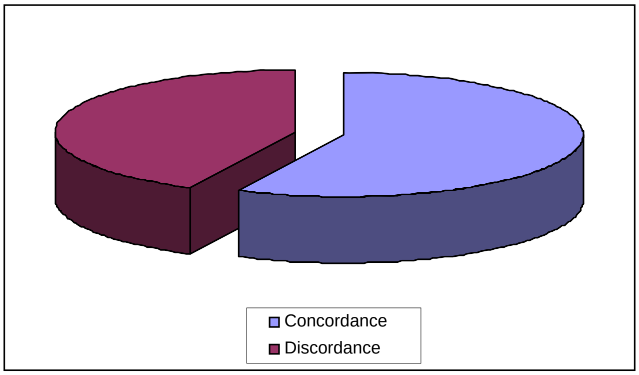


Figure 4-10: concordance and discordance

In 57 % cases higher corrosion rate is accompanied by higher average potential (less negative) and in 43% cases is the other way around. This approach also shows weak positive correlation between average on-potentials and the corrosion rate.

4.5 Correlation between the corrosion rate and potentials standard deviation recorded between first and last inspection.

First, let's define:

- $\tilde{V}_i = \left[V_{i,1}, \ldots, V_{i,T_m}\right]^T_{1xl}$, where T_1 and T_m indicates time of first and last inspection, $V_{i,j}$ - is on-potential at stationing *i* at time $t_{i,j}$, $i = 1,\ldots n$, n- number of the defects and *l* indicates number of available measurements (real and interpolated between first and last inspection)
- $\overline{V_i} = \frac{1}{l}\sum_{k=1}^{l}\tilde{V}_i(k) \ \forall i = 1, \ldots, n$ and $R = \left[r_1, \ldots, r_n\right]^T_{1xn}$ - corrosion rates for *n* defects
- $V^S = \left[V_1^S, \ldots, V_n^S\right]^T_{1xn}$ where $V_i^S = \frac{1}{l-1}\sum_{k=1}^{l}\left(\tilde{V}_i(k) - \overline{V_i}\right)^2$

Then the correlation between corrosion rate and calculated standard deviation of potential is defined as: $\rho(R,V^S)$.

no. of samples	Type of correlation	ρ	p-value for the hypothesis that correlations are insignificant
52 defects, l=61	**Pearson**	***-0.15***	*0.29*
	Spearman	*-0.13*	*0.37*
	Kendall	*-0.08*	*0.43*

Table 9: correlation between the corrosion rate and on-potential standard deviation recorded between first and last inspection

Table 9 shows weak negative correlation between corrosion rate and standard deviation of the on-potential.

4.6 Conclusions

Corrosion processes take time and are therefore governed by a number of circumstances. Unfortunately some information is missing: the measurements were made and only min/max were recorded. The data used in this chapter is not accurate. It is difficult to say precisely in how on-potentials describe real pipeline potential. Introduced methodology didn't give defendable results i.e. clear massage about the connection between corrosion rate and average on-potential.

The analysis showed certain patterns of weak[17] correlations between corrosion rate for the 52 registered defects along the A3 and level and on-potentials and their variability. The results from this chapter will be applied as an input into regression in the next section. Further analysis of the on-potentials has to be carried out in order to check how on-potentials are interacting with other variables potentially influencing the corrosion rate.

[17] read: statistically insignificant

Chapter 5

5 Microbial Data analysis

5.1 Introduction

The dataset analyzed in this chapter is delivered and collected by one company specializing in bio-analysis, a company which collected soil samples at the stationing of certain group of defects from the pipeline A3. The analysis of the collected soil samples was done in order to assess the circumstances which can influence the growth of bacteria involved in corrosion processes. Here, the bio-analysis is used as an input for the corrosion rate regression model.

The data collection and analysis was carried out with respect to qualitative and quantitative description of the environment where defects were reported. The defects under analysis (18 defects) were chosen from the set of 52 defects recognized. For each of these defects linear corrosion (constant corrosion rate) was assumed. The corrosion rate for these defects was calculated using corrosion rate estimation model described in first part of this thesis. The estimation was based on measurements from four consecutive inspections done by intelligent pigs.

The analysis will be carried out in order to find the connection between the environment data and the corrosion rate estimated using the corrosion rate model presented in Part 1 of the thesis.

5.2 Microbiologically Influenced Corrosion

According to chemists: abundant in natural environments Sulphate Reducing Bacteria (SRB) are the most influential in MIC processes. SRB are anaerobic bacteria utilizing sulfate as a terminal electron acceptor and organic substances as carbon sources. It is shown that although SRB are strictly anaerobic, some subpopulations tolerate oxygen and are even able to grow at low oxygen concentrations. SRB has ability to reduce sulfate produced carbonate which neutralizes acids and sulfide, which chemically stabilizes toxic metal ions as solid metal sulfides. Experiments showed that ph levels supposed to increase in presence of SRB metabolism[18]. The soil analysis from Texas to New Jersey has shown that number of bacteria is living in the soil at or near protective coatings. In paper of Joseph L. Pikas [23], author suggests that if one compares two environments with the same soil type but one at or near to ditch and second undisturbed, then higher number of SRB is expected to be in the first environment. J. O. Harris [24] in his notes says that since conditions of the soil do not remain static whether the soil is close to surface or near to a pipeline at the bottom of a ditch-mostly due to water fluctuations- bacterial populations in the soil consist of many types of different species. Moreover, the interrelationship between different types of bacteria of microorganisms contributes to changes that occur in the soil.

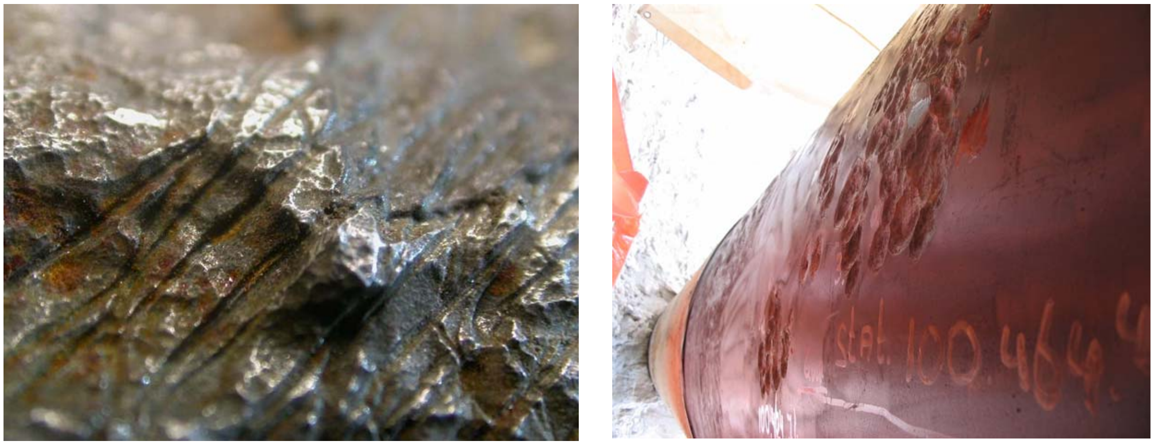


Figure 5-1: microbiologically influenced corrosion on gas pipelines

Microorganisms can be grouped into few types presented in the table below.

Prerequisite	Provider	Kind of growth
Energy source	Light Chemical substances	Phototropic Chemiotrophic
Carbon source	CO_2 Organic substances	Autotrophic Heterotrophic
Electron donor (to be oxidized)	Inorganic substances Organic substances	Lithotrophic Organotrophic
Electron acceptor (to be reduced)	Oxygen NO_2^- , NO_3^- SO_4^{2-} , CO_2	Aerobic Anoxic Anaerobic

Table 10: groups of microorganisms

Interesting result is that a very good place to live for an anaerobic organism is below an active colony of aerobic organisms as these consume the oxygen and create anaerobic areas which serve as habitats for the anaerobics. As a result, anaerobic organisms like SRB can be found next to aerobic organisms that protect anaerobic bacteria which can easily grow and multiply. The oxidation of sulfide, which can be performed sulfur oxidizing bacteria results in decreasing pH value is typical example important for MIC.

[18] P. Frank, UC Berkeley Department of Environmental Sciences

Decreasing pH value is equivalent with transformation of weak acid in a strong one. Chemical reaction of this transformation is shown as follows.

$$S^{2-} + 2O_2 \rightarrow SO_4^{2-}$$

Over last decades many different models have been proposed to explain the mechanisms by which SRB can influence the corrosion of the steel. These models were concentrated on analysis based on cathodic depolarization by the enzyme hydrogenase, anodic depolarization, production of corrosive iron sulphides, release of exopolymers capable of binding Fe - ions, sulphide-induced stress- corrosion cracking, and hydrogen-induced cracking or blistering. All the models showed that there is not only one predominant factor influencing MIC and many different factors are involved.

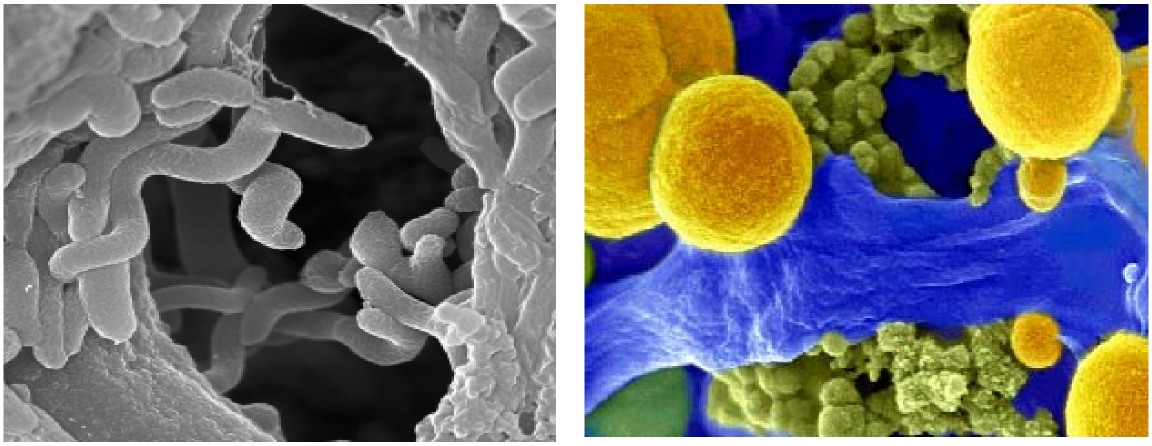


Figure 5-2: image of a sulphate-reducing bacterial culture with a carbonate precipitate, the bacteria on the left are about 6-8 μm long and 2 μm in diameter

In order to confirm MIC it is essential to check presence of microorganisms by obtaining samples of the natural environment surrounding the metal.

5.3 Dataset description

The dataset consists of two kinds of variables, independent which are used as an input to the model (the variables which may influence the corrosion rate) and dependent variable- often called variable of interest or criterion variable which is associated with the output which is the corrosion rate.

5.3.1 Independent variables

The main bio-analysis of the samples was performed in order to give Multi Criteria Analysis (MCA) for each specific environment. The MCA basically gives score relative to chance of getting MIC corrosion for specific environment. The formula for the score is based on five factors: **redox**, **availability of a carbon**, **availability of nutrition**, **degree of acidity (pH)** and **conductivity**. The MCA formula is presented as follows:

$$MCA_i = 3 \cdot S_i^{redox} + 2 \cdot S_i^{carbon} + 1 \cdot S_i^{nutr} + 1 \cdot S_i^{pH} + 1 \cdot S_i^{EC}$$

where: *i* indicates *i'th* measurement (place where defect was registered) and scores S_i^j are assigned according to the table below. Scores are assigned by experts.

Factor	Classification	Score
S_i^{redox}	*Aerobic*	**1**
	Nitrate reducing	**1**
	Iron reducing	**2**
	Sulphate reducing	**3**
	Lack of methane	**2**
S_i^{carbon}	*0-20 mg/l TOC*[19]	**1**
	20-40 mg/l TOC	**2**
	>40 mg/l TOC	**3**
S_i^{nutr}	*N-tot < 1 mg/l P-tot < 0.05*	**1**
	N-tot > 1 mg/l P-tot < 0.05	**2**
	N-tot < 1 mg/l P-tot > 0.05	**2**
	N-tot > 1 mg/l P-tot > 0.05	**3**
S_i^{pH}	*pH > 5.5*	**3**
	pH < 5.5	**1**
S_i^{EC}	*0-500* μS	**1**
	> μS	**3**

Table 11: factors and weights for MCA score

Other factors investigated and measured in the field by bio-company are:

Variable	Notation/units	Description	
Oxygen	oxygen $[mg/l]$	Amount of oxygen can indicate existence of anaerobic/ aerobic bacteria in soil, delivered data in half cases are beyond the detection limit, so the oxygen variable will be treated as a "dummy" variable[20].	
Redox	redox potential $[mV]$	The redox potential is the reduction/ oxidation potential of a compound measured under standard conditions against a standard reference half- cell.	
cond.	conductivity $[\mu S]$	Conductivity is a measure of a material's ability to conduct an electric current.	
pH	$[pH]$	pH is a measure of the activity of hydrogen ions in a solution and, therefore, its acidity or alkalinity	
TOC	$[mg/l]$	The amount of carbon bound in organic compounds.	
Fe	iron $[\mu g/l]$	Iron is a chemical element with the symbol Fe (L.: Ferrum) and atomic number 26.	
S1	S^{2-} $[mg/l]$	In both cases, similarly to oxygen measurements, approximately half of the measurements are indicated as "beyond the detection limit", so the dummy variable is applied[21].	
S2	sulfate SO_4^{2-} $[mg/l]$		
Methane	$[\mu g/l]$	The simplest hydrocarbon, methane, is a (natural) gas with a chemical formula of CH_4.	
SRBA	SRB- A $[N]$	Sulphate reducing bacteria: type A	In the measurements, boundaries of the possible interval of number of bacteria were given. In the analysis use middle of this interval.
SRBB	SRB- B $[N]$	Sulphate reducing bacteria: type B	
water	water level [m]	Water levels with respect to the top of the ground- here the water (fluctuations are not taken described).	
D	[m]	depth of cover	

[19] TOC- Total organic concentration

[20] A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in a study. In research design, a dummy variable is often used to distinguish different treatment groups, in the simplest case takes values 0 or 1

[21] Idem

PN	[m]	pipeline wrt NAP level
NAP	[m]	NAP level of the ground
AP[22]	[mV]	average of pipe-to-soil potential measured at the test posts reported between first and last inspection
SP	[mV]	standard deviation of the on potentials measured between first and the last inspection
WD=**D-W**	[m]	amount of water wrt the top of the pipeline

Table 12 Microbial data description

5.3.2 Dependent variable

The analysis will be based on the 16 defects for which it was possible to associate the corrosion rate- only for 16 defects the detailed environment data was delivered. Small summary of corrosion rate data is presented below.

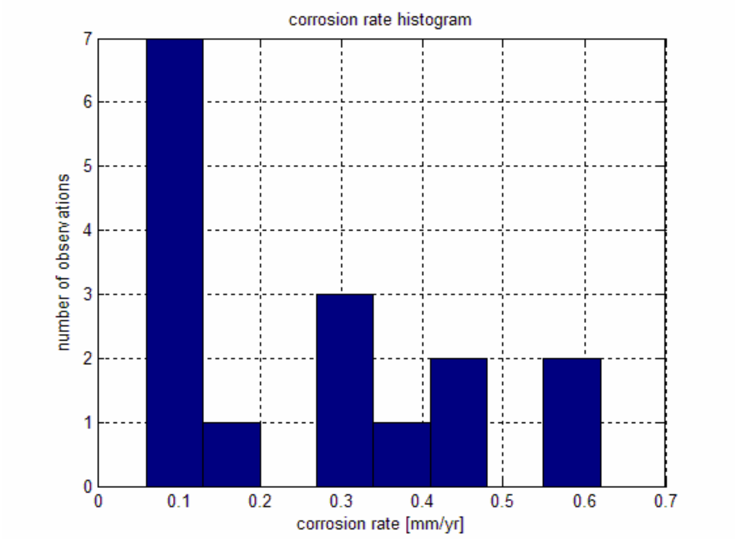


Figure 5-3 Corrosion rate summary for 16 measurements indicated by bio-analysts

No. meas.	Mean corrosion rate:	Std corrosion rate:	Upper 95% conf. int.	Lower 95% conf. int.
16	0.26	0.20	0.62	0.06

Table 13: corrosion rate summary for 16 measurements indicated by bio-analysts

5.4 Correlation analysis

Since all the measurements collected by bio-analysts were measured once, after the last pigrun, so it is impossible to check the connection between changes of defect's depths and change of the soil measurements (wrt e.g. groundwater fluctuations, amount of oxygen etc.) However it is possible to check the connection between estimated corrosion rate and reported bio measurements. Firstly, the analysis will be investigating the correlation between the corrosion rate and all the variables.

Due to low number of measurements the predictive model will not be very reliable. However it is enough to indicate patterns and relationships. Table below presents correlations and *p-values* associated with following hypothesis testing:

[22] The detailed analysis of the average pipe-to-soil potentials was introduced in “Potential analysis” chapter.

$$H_0 : \rho(X,Y) = 0 \text{ against } H_1 : \rho(X,Y) \neq 0$$

where: X- predictor variable, Y- criterion variable (corrosion rate), ρ_P - stands for Pearson, ρ_S for Spearman and ρ_K for Kendall correlation coefficients.

Name	no. of samp.	ρ_P	p-value	ρ_S	p-value	ρ_K	p-value
oxygen	***9***	-0.46	0.07	-0.37	0.16	-0.31	0.17
MCA	18	-0.18	0.50	-0.13	0.62	-0.08	0.71
redox	16	***0.52***	0.04	0.38	0.24	0.25	0.21
cond.		-0.20	0.46	-0.39	0.13	-0.31	0.10
pH		***-0.56***	0.02	-0.52	0.04	-0.38	0.05
TOC		***-0.34***	0.20	-0.30	0.31	-0.17	0.39
Fe		-0.20	0.48	-0.28	0.29	-0.20	0.30
methane		-0.20	0.47	-0.17	0.53	-0.10	0.65
SRBA	14	0.05	0.87	0	1	0	1
SRBB		-0.24	0.40	0	1	0	1
water	17	***-0.34***	0.18	-0.19	0.47	-0.12	0.53
S1	18	***0.15***	0.58	0.22	0.41	0.19	0.43
S2		***0.13***	0.64	0.18	0.51	0.15	0.53
WD	15	***-0.001***	0.99	-0.13	0.64	-0.09	0.69
D	18	***-0.40***	0.10	-0.49	0.04	-0.35	0.04
PN		***0.34***	0.16	0.30	0.23	0.25	0.16
AP		***0.33***	0.18	0.35	0.15	0.25	0.18
SP		***-0.44***	0.07	-0.45	0.06	-0.3	0.1
NAP		***0.34***	0.16	0.30	0.23	0.25	0.16

Table 14 correlation between corrosion rate and the variables, cell with red background indicate that normality assumption doesn't hold, yellow rows indicate variables for which the null hypothesis was rejected for significance level 0.2

Graphically, the correlations between the predictor variables and criterion variable can be expressed in the form of radar graph. Each of the variables from the table corresponds to a ray in the graph below. The variable with the highest correlation is plotted furthest from the center, and the variable with the lowest respectively closest to the center.

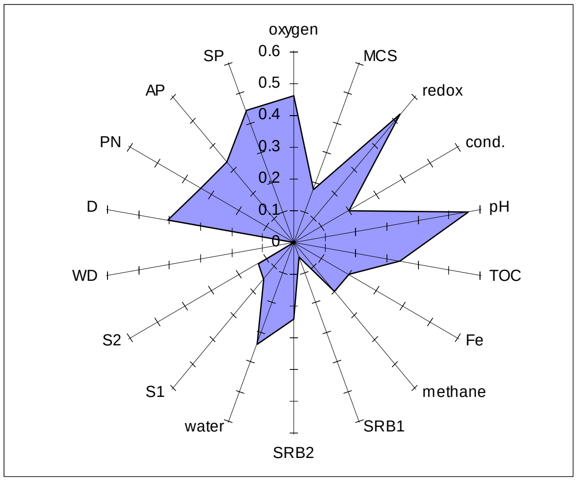


Figure 5-4 radar graph- Pearson product moment correlation coefficient

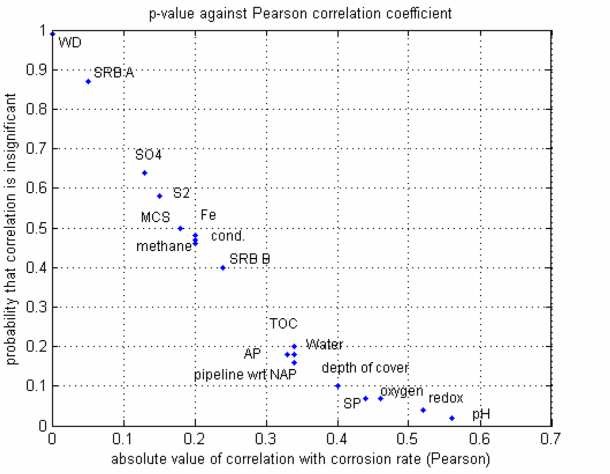


Figure 5-5 correlations and associated p-values (ordering the variables)

The results are promising, for significance level α =0.2 eight of variables are significantly correlated with corrosion rate, for a level 0.05 only two variables have significantly high correlation. The most correlated with corrosion rate is **pH (negative correlation)** and as second one is **redox potential (positive correlation)**. If one looks only at two most correlated variables it is clear that the lowest corrosion rate is obtained for alkaline soil samples with low redox potential. Interesting is that none of SRB has significant correlation with the corrosion rate. If the significance level is changed to α=0.2, then additional 6 variables indicate significant correlation with the rate. In four cases the correlation is negative: TOC, water level, depth of cover and standard deviation of on potential, and in two positive: pipeline with respect to NAP and average on-potential.

From the data one can observe that ***the best scenario for a low corrosion rate is when: oxygen is detectable, redox potential is low, soil is alkaline, high level of TOC, pipeline is dried (high values of W mean that water is deeper under the cover), pipeline is deep under the ground level, pipeline is low wrt. NAP level, the on-potential from rectifier is low (more negative) and the standard deviation of the potential is relatively high***. The suggested “best scenario” is based only on simple correlation analysis; however, it doesn't give quantitative results, and it doesn't take into account correlations amongst the predictor variables.

The key of the analysis is to find the predictor variables which are significant for the multiple regression model.

5.5 Multiple regression analysis

This section is dedicated to the multiple regression model. Multiple regression is a statistical technique which allows predicting variables of interest (sometimes called: dependent or criterion variables) on basis of scores of several other variables (these variables are customary called: independent or predictor variables). The main point in the modeling is to explain the level of the variance on the basis of the level of one or more other variances. The regression analysis should be based on the predictor variables that might be (highly) correlated with the criterion variable, but not strongly correlated which each other. In reality correlations amongst the predictor variables are not unusual. Multicollinearity[23] can cause problems when trying to find the relative contribution of each predictor variable to the modeling. When there is a substantial multicollinearity in a regression model, it is possible to have the full model account for a substantial amount of the variability in the dependent variable without any tests of its individual parameters being significant. One of the ways to avoid this problem is to apply so called stepwise regression algorithms.

5.5.1 Analysis structure

The analysis is focused on finding a set of most influential predictor variables wrt corrosion rate. Investigation has to deal with one very relevant issue, namely: **missing data** (two variables have missing observations). Because of low number of the collected measurements, the analysis has to be done in a way that the final result is based on number of measurements as large as possible. When the proper model is defined then included variables should be ordered with respect to importance for the corrosion rate. Figure 5-6 below presents general idea of the regression modeling of the corrosion rate.

[23] Multicollinearity (collinearity)- the term is used to describe the situation when a high correlation is detected between two or more predictor variables.

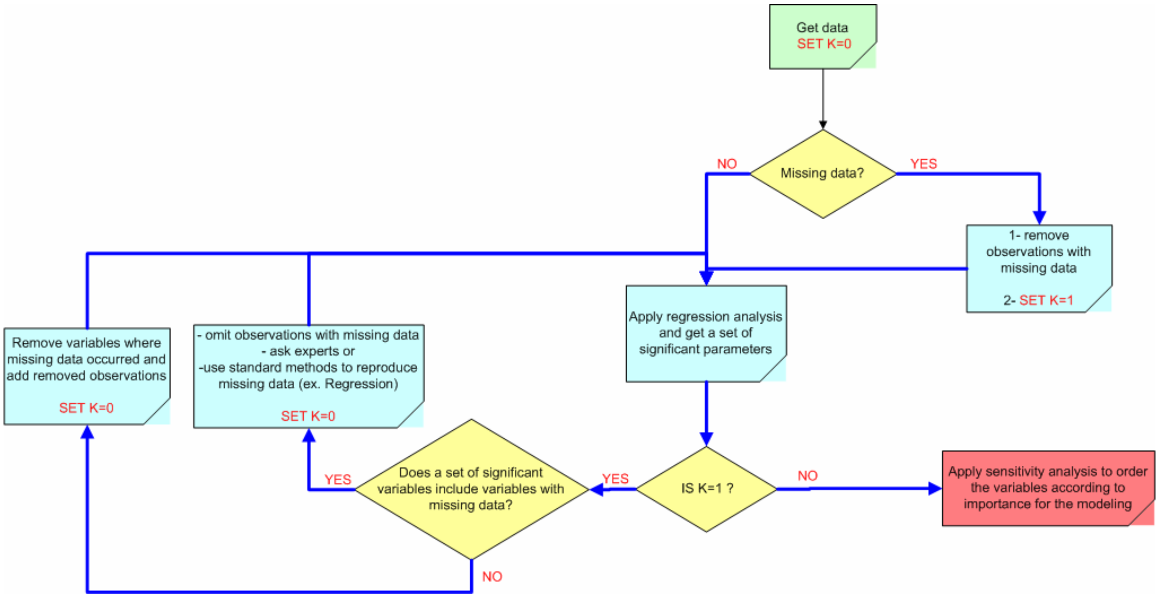

Get data
SET K=0
Missing data?
NO
YES
1- remove observations with missing data
2- SET K=1
Remove variables where missing data occurred and add removed observations
SET K=0
- omit observations with missing data
- ask experts or
-use standard methods to reproduce missing data (ex. Regression)
SET K=0
Apply regression analysis and get a set of significant parameters
YES
Does a set of significant variables include variables with missing data?
YES
IS K=1 ?
NO
Apply sensitivity analysis to order the variables according to importance for the modeling
NO

Figure 5-6 regression modeling schema for the corrosion rate

5.5.2 Missing data

Two variables **SRB-A** and **SRB-B** have missing observations at the stationings 2225 [m] and 2451 [m]. Due to these missing data, firstly the analysis has to verify if these two variables are significant for the analysis. If they are not significant then the problem with missing data for these two variables doesn't exist anymore (i.e. then these two variables can be easily removed and the study can be carried out for all the observations), if they are significant, then missing observations have to be reproduced or removed (the missing data for one of the variables would imply removing the corresponding values for all the other input variables).

In order to check if SRB-A and SRB-B are significant, let's

- remove the missing observations for all the variables (total number of remaining observations is now 14)
- in order to get a number of the most relevant parameters apply stepwise regression[24].

Stepwise regression will be applied to the following variables (each of the variables has 14 observations):

- Y- variable of interest- defect rate [mm/yr]
- X_i - independent variables:

1	MCA	**5**	pH	**9**	S2	**13**	water	**17**	SP
2	oxygen	**6**	TOC	**10**	methane	**14**	depth	**18**	WD
3	redox	**7**	Fe	**11**	SRBA	**15**	PN	**19**	NAP
4	cond	**8**	S1	**12**	SRBB	**16**	AP		

- n is associated with the number of variables which is equal to 19
- X_iX_j - $\forall i, j \in N$ where $i \neq j$ product of centered independent variables

The tables below describe following hypothesis testing:

- t-statistics and p-value for t-test are associated with following hypothesis testing:

$$H_0 : \hat{\beta}_i = 0 \text{ against alternative } H_1 : \hat{\beta}_i \neq 0$$

- F-statistics and p-value for F-test are associated with:

[24] See chapter- Analysis methods and interpretation, appendix A

$$H_0 : \hat{\beta}_0 = \hat{\beta}_1 = \ldots = \hat{\beta}_n = 0 \text{ against alternative } H_1 : \exists \hat{\beta}_i \neq 0$$

Two main models will be considered:

5.5.2.1 Model without interactions between variables

Model 1

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

Variables	Coefficients		t- statistics	p-value for t-test
	$\hat{\beta}_i$	Std. Error		
(Constant)	2.03	0.41	4.98	ε [25]
pH	-0.21	0.62	-3.42	0.006
Depth of cover	-0.27	0.09	-3.01	0.012

Table 15: parameters and associated statistics

And model statistics

MODEL	R^2	Adjusted R^2	Statistics	
			F statistics	p-value for F test
	0.66	0.59	10.5	0.003

Table 16: regression model standard description

5.5.2.2 Model with interactions between variables

Model 2

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \beta_{n+1} X_1^2 + ... + \beta_{2n} X_n^2 + \beta_{1,2} X_1 X_2 + ...$$
$$\beta_{1,n} X_1 X_n + ... + \beta_{2,3} X_2 X_3 + ... + \beta_{i,j} X_i X_j + \beta_{n-1,n} X_{n-1} X_n + \varepsilon$$

Variables	Coefficients		t- statistics	p-value for t-test
	$\hat{\beta}_i$	Std. Error		
(Constant)	0.27	0.02	16.3	ε
PN*water	-0.07	0.01	-6.1	ε
PN	0.09	0.007	12.88	ε
Oxygen*pH	0.71	0.08	8.79	ε
MCA*MCA	-0.01	0.003	-4.34	ε

Table 17: parameters and associated statistics

And the model statistics:

MODEL	R^2	Adjusted R^2	Statistics	
			F statistics	p-value for F test
	0.97	0.96	78.7	ε

Table 18: regression model standard description

[25] ε stands for number less than 0.0001

Conclusions (based on 14 measurements):

From the applied stepwise regression to the set of 19 variables with 14 observations we have that:

1. For the model without interactions:
    - There is a negative correlation between pH and the corrosion rate
    - There is a negative correlation between depth of cover and corrosion rate
    - Both **SRB-A** and **SRB-B** are insignificant for the corrosion rate modeling, hence can be easily removed

2. For the model with interactions:
    - Also in this case analysis did show that both **SRB-A** and **SRB-B** are insignificant (even when interacting with other variables)
    - Only one main effect is significantly affecting the corrosion rate- pipeline wrt. NAP

For both models R^2 and adjusted R^2 indicate that the corrosion rate is quite well described by the proposed models. Analysis showed that errors from the both models are normally distributed and uncorrelated. High adjusted R^2 indicated well defined model; however the number of the measurements is not high enough to make such conclusion.

5.5.3 Stepwise regression for included variables

In the previous subsection it was shown that SRB-A and SRB-B can be removed from the analysis since they are insignificant for the corrosion rate modeling.

Here, the analysis will be based on the remaining variables. The study will be based on two models introduced before: model 1 (without interactions) model 2 (with interactions). These two models are applied to the following dataset:

- Y- variable of interest- defect rate [mm/yr]
- X_i - independent variables $i = 1,..,17$:

1	MCA	**5**	pH	**9**	S2	**13**	PN	**17**	NAP
2	oxygen	**6**	TOC	**10**	methane	**14**	AP		
3	redox	**7**	Fe	**11**	water	**15**	SP		
4	cond	**8**	S1	**12**	depth	**16**	WD		

5.5.3.1 Model without interactions

Suppose that the corrosion rate is modeled only by main effects- according to the Model 1. Stepwise regression resulted that only one main effect may be an influential parameter for the corrosion rate.

Variables	Coefficients		t- statistics	p-value for t-test
	$\hat{\beta}_i$	Std. Error		
(Constant)	0.25	0.04	5.92	ε
PN	0.06	0.02	2.66	0.02

Table 19: parameters and associated statistics (model without interactions)

The determination coefficient-R^2 from the Table 20 shows model fit. This means that Model 1 may be too simple to give reliable estimate of the parameters influencing the corrosion rate.

MODEL	R^2	Adjusted R^2	Statistics	
			F statistics	p-value for F test
	0.34	0.29	7.09	0.02

Table 20: regression model standard description (model without interactions)

5.5.3.2 Model with interactions

- The model 2 presents much more complicated situation, where except main effects (17 variables), all the possible combinations (136 variables). Stepwise regression gave the following results.

Variables	**Coefficients**		**t- statistics**	**p-value for t-test**
	$\hat{\beta}_i$	**Std. Error**		
(Constant)	0.25	0.007	34.15	ε
Redox*water	-0.001	0.0001	-7.00	ε
PN	0.10	0.003	35.51	ε
Oxygen*pH	0.76	0.03	27.58	ε
TOC*PN	0.001	0.0002	9.62	ε
MCA*MCA	-0.01	0.001	-6.96	ε
Methane*SP	0.0001	ε	4.50	0.001

Table 21: parameters and associated statistics (model with interactions)

MODEL	R^2	Adjusted R^2	Statistics	
			F statistics	p-value for F test
	0.997	0.985	471.21	ε

Table 22: regression model standard description (model with interactions)

Now, variables included in the model are different than for the case without interactions. The final model (consisting only of significant variables) includes 6 variables (one main effect and 5 interaction effects). High value of R^2 indicates very good fit. It can, however indicate over-fit because of limited number of the measurements. As before both models the errors are normally distributed and uncorrelated.

5.6 Sensitivity analysis of the parameters influencing the corrosion rate

A sensitivity analysis is a process of investigating influences of model inputs on outputs. If a small change in a parameter results in relatively larger changes in the outcomes, then the outcomes are said to be sensitive to that parameter. This may mean that the parameter has to be determined very accurately. Basically, a sensitivity analysis is a study of how the variation in the output of a model can be apportioned, qualitatively or quantitatively, to different sources of variation of the input. One of the most popular and easiest sensitivity methods is so called Correlation Ratio (CR)[26] which detailed is presented in the appendix A1. According to this method certain the level of the polynomial of $E(Y \mid X_i)$ has to be assumed. If we calculate the CR for all variables, then as before it is possible to order the variables according to the correlation ratio wrt importance. The results of the applied technique are following:

Order	Predicted variable	Correlation ratio	Degree of polynomial for conditional expectation
1	PN	0.3556	2
2	Redox * W	0.3522	2
3	TOC* PN	0.2024	2
4	Oxygen * PH	0.1857	2
5	Methane * SP	0.1645	2
6	MCA* MCA	0.0964	2

Table 23 sensitivity analysis of the significant parameters

All the variables in the Table 23 are ordered with respect to importance for the model. Applied sensitivity analysis showed that the most influential/ important for the corrosion rate modeling is variable- pipeline wrt nap level, then interacting redox potential with groundwater step level etc.

[26] See appendix A1- Analysis methods and interpretation

5.7 Conclusions and recommendations

The analysis showed that the most relevant (statistically significant) parameters for the corrosion rate are:

- **Pipeline with respect to NAP level** (positive correlation)
- Interactions between:
    - **Redox** with **water level** (negative correlation)
    - **TOC** with **Pipeline wrt NAP level** (positive correlation)
    - **Oxygen** with **pH** (positive correlation)
    - **Methane** with **standard deviation of on-potentials** (positive correlation)
    - **MCA** with **MCA** (negative correlations)

The presented variables are ordered with respect to level of the correlation with the corrosion rate (according to sensitivity analysis) i.e. pipeline wrt NAP level is the most important, second is interaction between redox and water level etc.

The analysis showed that number of sulphate reducing bacteria of type A and B (SRB-A and SRB-B) are insignificant for the analysis. The study proved that there is not only one predominant factor influencing MIC and many different interacting parameters are involved. It was shown that number of observations strongly influences the number of statistically significant variables. Depending on the approach different sets of variables are important for the model. First approach with only 14 observations resulted in two models with and without interactions. First one consisted of two main effects: depth of cover and pH level, and the second one with one main effect**: pipeline wrt NAP level** and remaining interactions: **pipeline wrt NAP interacting with water level**, **oxygen with pH** and **MCA with MCA**. Both models showed common three parameters as the most important- pipeline wrt NAP level, oxygen interacting with pH and MCA square. Because of high measurement error additional observations are required. The models with interactions illustrate very good fit to the real measurements. Because of lack of the measurements this perfect fit may indicate existing overfiting problem which can be reduced by supporting the model with higher number of the environmental measurements.

Recommendations

- In the study it was assumed that estimates of the corrosion rate from first section are certain. This assumption is unlikely to be realistic. The future investigations should be also aimed to take the estimate errors into account.
- Crucial in the modeling is to have large and accurate dataset.
- The analysis was based on certain number of possible influencing parameters; however those parameters do not exhaust all the possible influencing factors- ex. lack of data about groundwater fluctuations etc.
- Some of the defects indicating decreasing corrosion rate, the corrosion rate model imposed certain number of constrains, it is likely that this phenomenon came out as a result of the clustering procedure. As a consequence association defect with the environment parameters doesn't guarantee good results. The clustering/matching defects and associated errors should be deeper investigated.
- Poor dataset is strongly influenced by unusual of observations; since missing data occur it is important to know reasons of that- perhaps the defects are in unusual environment.
- Non-linear relationships should be investigated.

Part III

Parameters influencing microbiologically induced defect rate

Chapter 6

6 Pipeline characteristics

In this chapter three high-pressure pipelines: A1, A2 and A3 are under the analysis. The investigations in the previous two sections concerned only the pipeline A3 for which the set of 52 distinguishable defects was collected. Since this section is dedicated to defect rate modeling, it is not required any more to be restricted to one pipeline (for which it was possible to give an estimate of the corrosion). Two additional high pressure pipes were chosen according to following features: the existence of MIC recorded during the excavations, -the age of all three pipelines -the coating and applied technology.

6.1 Defect distribution

The pictures presented below show the pipelines profile, depth of cover and defects distributed along the pipelines. In the cases of A1 and A2 number of defects is much lower than for A3, although the pipelines were installed about the same time in the 60s. In all three cases the installation customs and applied coating (bitumen) were generally the same. This may indicate that the difference between the numbers of defects can be caused by environmental factors. Also in the case of “parallel” pipelines A1 and A2, the profiles are very similar; however number of the defects for A2 is significantly higher.
The analysis starts with verifying the hypothesis about defects random (uniform) distribution along the pipeline. In all three cases there are statistical basis to reject the hypothesis that the defects are uniformly distributed along the pipeline.

Pipeline A1

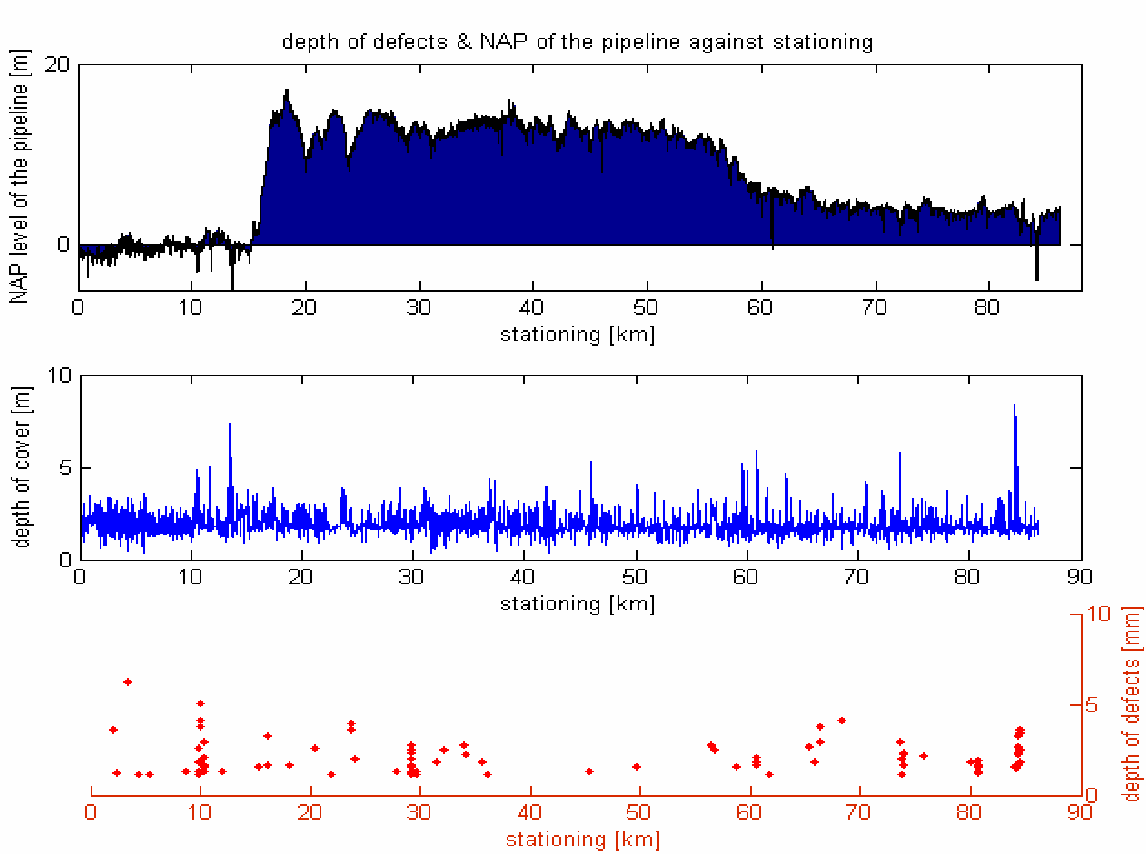


Figure 6-1: Pipeline A1, profile, depth of cover, defect distribution

Pipeline A2

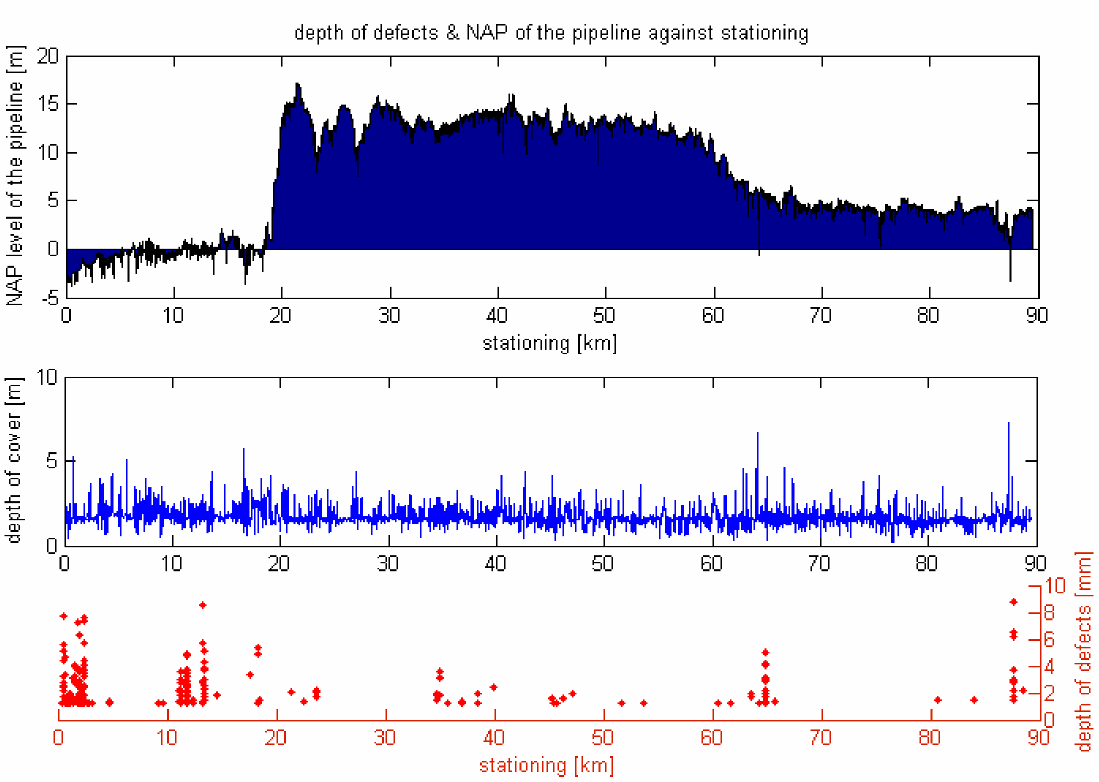


Figure 6-2: pipeline A2, profile, depth of cover, defect distribution

Pipeline A3

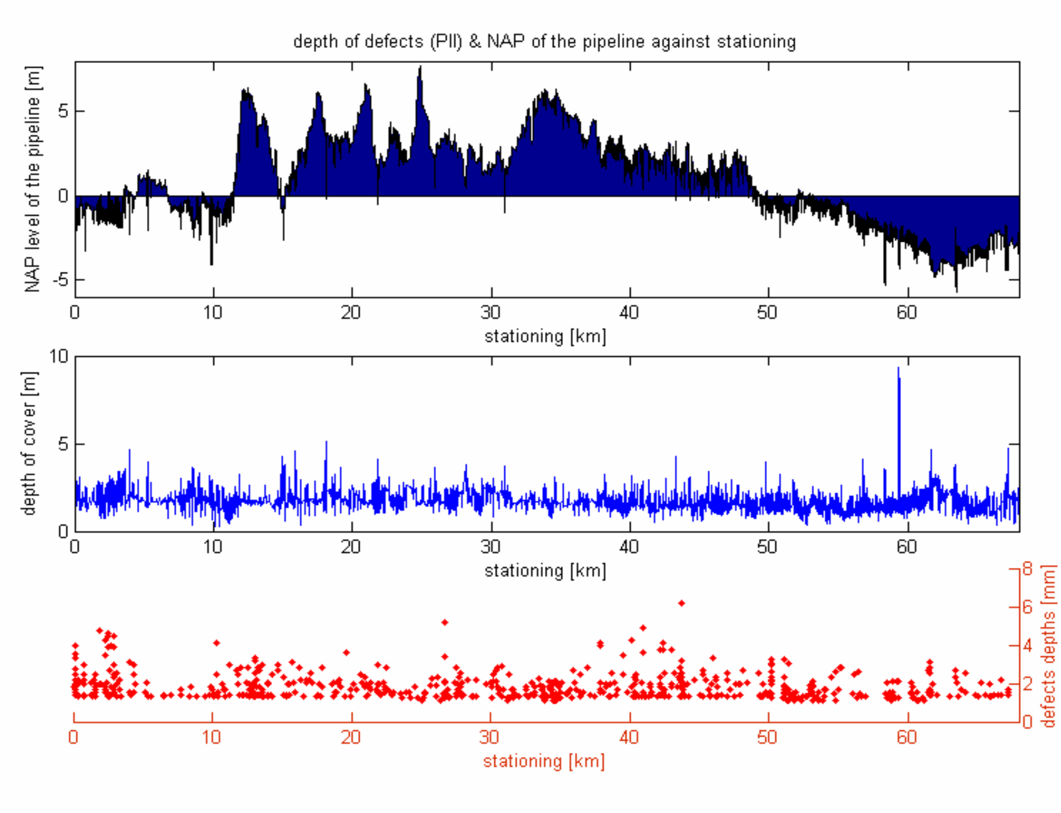


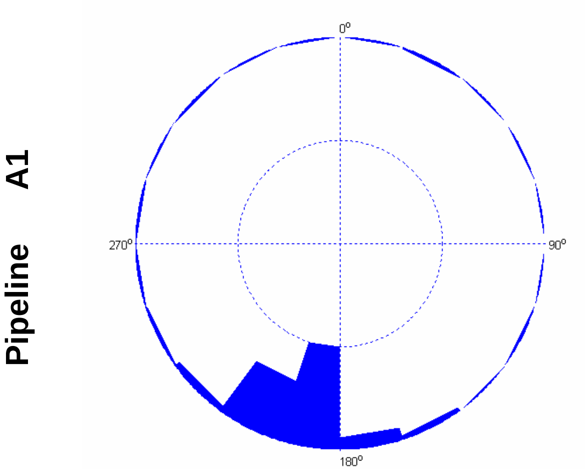
Figure 6-3: pipeline A3, profile, depth of cover, defect distribution (PII)

6.2 Depth of cover

Figures presented above don't depict that there is any connection between the defects existence and depth of cover. Moreover for the similar pipelines profiles of A1 and A2 the pattern of the defects distributions is similar wrt number of defects and their stationings.

6.3 Pipeline elevation

For the pipeline A3 the profile indicates that the pipeline is laid in lowland (max elevation is about 5-6 meters); whereas for A1 and A2 pipeline profiles changes about 20 meters within few hundred meters. The distribution of the defects may be caused by the groundwater levels (pipeline which is higher wrt NAP level is less likely to be in wet environment than pipeline closer to the reference NAP level). Graphs below present how the defects are distributed wrt pipeline circumference. In all the cases most of the defects are concentrated in the bottom of the pipeline. It is not certain that three pipelines under analysis are induced by MIC; however the number of excavations showed that this is really the case. Moreover, most of excavated defects with MIC were on the bottom of the pipeline.



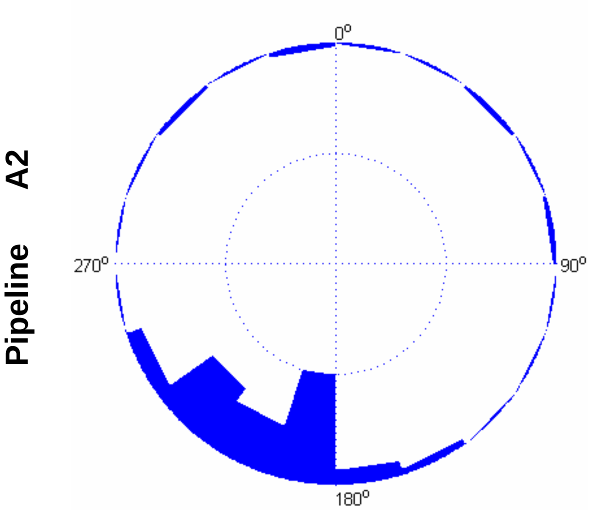


Figure 6-4: pipeline A1 defect distribution wrt pipeline circumference

Figure 6-5: pipeline A2 defect distribution wrt pipeline circumference

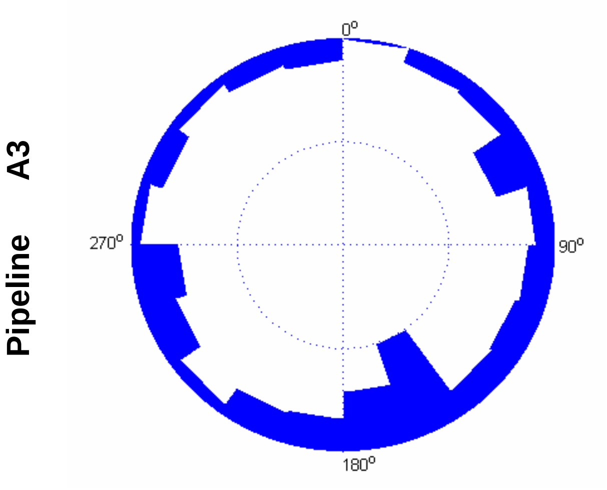


Figure 6-6 pipeline A3 defect distribution wrt pipeline circumference

Chapter 7

7 Soil data analysis

7.1 Introduction

This chapter is focused on finding the relation between the soil composition and corrosion defect rate for the high pressure underground gas pipelines. The soil data analysis is performed on three pipelines where microbiologically influenced corrosion was detected. In all three cases the soil data was collected from a geotechnical surveys performed before pipelines construction.

7.2 Description of available dataset

During inspections, an intelligent pig reports defects and associated defects' stationing. Due to technological drawbacks the stationing of defects and defect feature type are not accurate. The analysis in this chapter is aimed at statistically significant number of environmental parameters influencing the defect rate. Hence the data about environment has to be incorporated. The most reliable information about the pipeline natural environment is one obtained from pipeline geotechnical surveys and presented on route maps. The data about the pipeline soil type was collected before pipeline installation. Each route map represents certain part of the pipeline route. Usually the length of the route maps is about 1-1.5 km. These maps present ground elevation (ref. NAP) where the pipeline is placed, but not the pipeline's profile. A pipeline's profile is available on PiMS[27] so match can be easily done. Each map presents about 5-25 soil samples spread within a route map. In most cases it is impossible to find exact stationing of the soil samples within a map[28]. However the approximation can be done based on the fact that each map is spited in few parts (usually 5-10) and the stationings of the boundaries are given.

[27] Pipeline Integrity Management System

[28] Exact stationing of the measurements was given only for the pipeline A1 and A3

The stationings of the soil samples collected from the maps are not certain. The main reasons of the uncertainty are following:

- re-routing of the pipeline i.e. in some places, because of the infrastructure, some pipelines had to be rerouted
- the "starting point" for a pig and "zero meter" of the first route map are not the same so the calibration is always required

Each of collected samples presented in the route maps is in the form where soil layers can be distinguished. However, according to the installation customs of the 60s a soil layer during backfill of the pipeline trench were mixed, hence there is no point of analyzing influence of the soil layers on corrosion. There may still be an influence of soil layers, but due to lack of exact data such analysis won't be carried out[29].

7.3 Soil data collection

Essential in a defect rate modeling is to specify environments along the pipeline. As it was mentioned, each of the route maps shows certain number of measurements. Since the soil samples are distributed along the pipeline, certain assumptions about the environments between the measurements have to be introduced.

Suppose that at a certain part of the pipeline two soil samples were collected. The assumption that requires to be imposed is about the soil type in between the collected measurements. If an intelligent pig reports a defect somewhere in between the collected samples, then the problem is to decide in which environment defect is observed. Figure 7-1, Figure 7-2 and Figure 7-3 below present example of proposed procedure. Each of pictures presents a route map, where the samples and some inner stationings (not of soil sample) are reported.

Each route map is divided into sections for which she stationing is known. Along each section soil samples are presented. The problem is that the stationing of the samples is unknown.

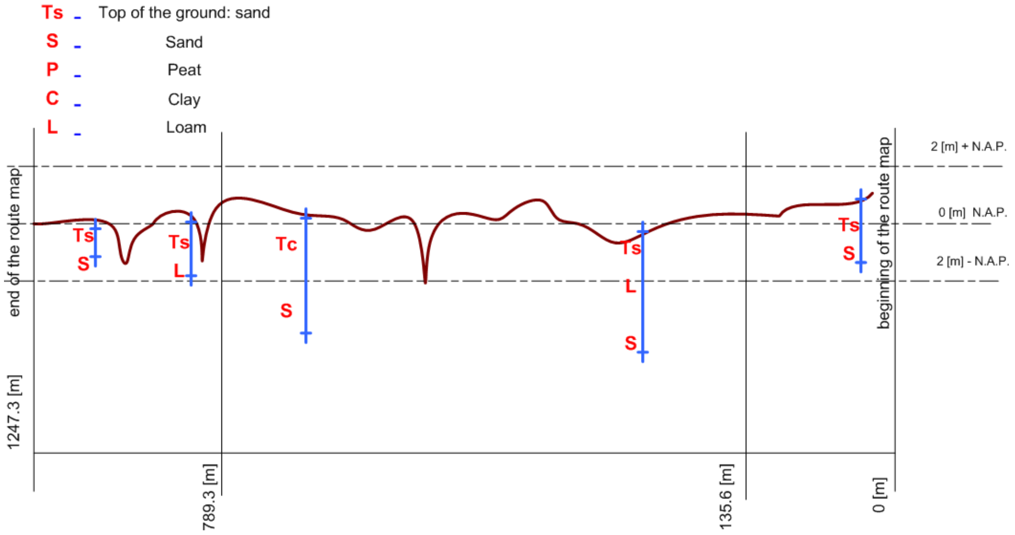


Figure 7-1: route map, geotechnical data

Step 1

Since the stationing of the measurements is unknown within the sections, so let's assume that the measurements are equal-distance distributed within each section. And so, for the first section 0 [m]- 135.6 [m] the distance between the measurements is equal

[29] Soil layers from the route maps do not show today's layers state.

to “c”, and for example for the section two 135.6 [m]- 789.3 [m] the distance is equal to “b”. Since now, each of the measurements will have specified stationing (before this was unknown).

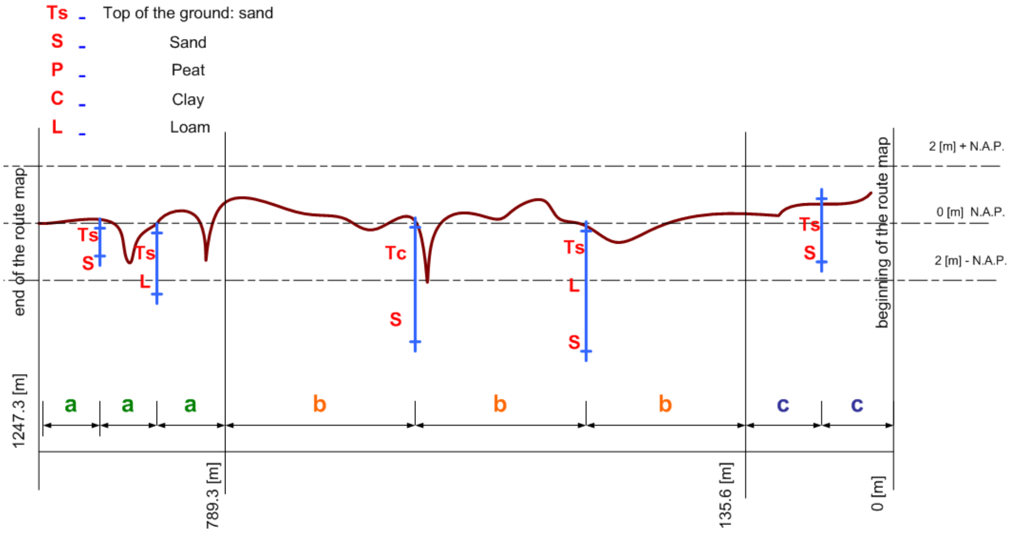


Figure 7-2: route map, geotechnical data- STEP 1

Step 2

Second and the final step, defines the environments (also called soil clusters). Since the stationings of the measurements are specified (at step 1) the sections have to be combined. Sequentially, to get the environment map of the whole pipeline the boundary measurements from each section have to be defined as well. In this procedure, it is assumed that for each two neighboring sections the boundary between the closest samples for these sections is in the middle of them. On the schema the clusters are indicated by different colors.

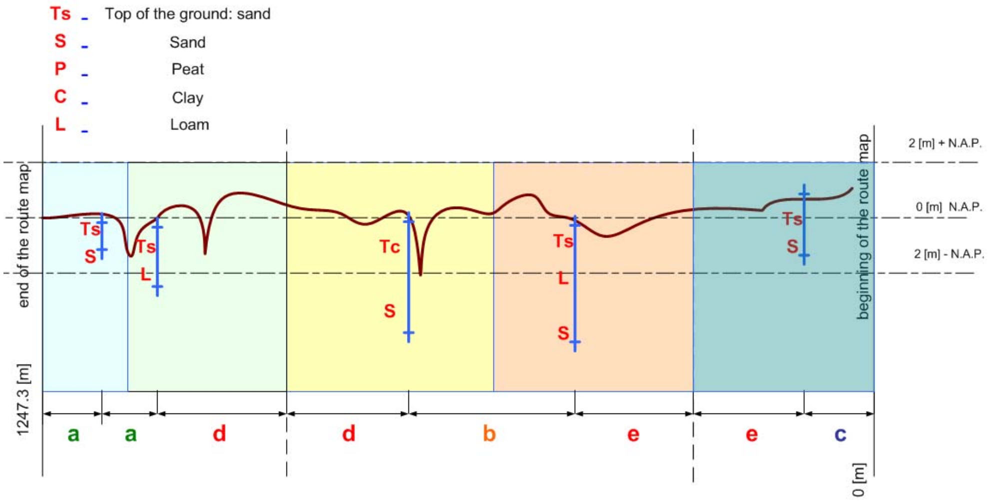


Figure 7-3: route map, geotechnical data- STEP 2

Application of the presented procedure defines soil type for any given stationing. Because of the lack of quantitative measurements it is impossible to avoid sharp boundaries between the sections/clusters.

The number of measurements varies from one pipeline to another; small summary of available data is presented in the Table 24 below.

pipeline	no. of soil measurements	no. of route maps	length of the pipeline	average cluster length	number of measurements per km
A1	298	50	86 [km]	280 [m]	3.5
A2	861	50	89 [km]	100 [m]	10
A3	287	47	69 [km]	240 [m]	4

Table 24: general pipelines description

First glance at the Table 24 shows that the highest accuracy for soil samples is obtained for the pipeline A2—average cluster length is 100 meters and it is more than two times less than for other pipelines. The accuracy of the obtained results for a pipeline A2 is much higher than for the others.
The data available, doesn't allow analyzing the soil samples quantitatively i.e. it is impossible to say if in the soil sample is more one component or another.

Remark
The route maps are connected by using the same methodology as for combining the section within a route map. So for step 2 the route map boundaries are ignored.

Introduced procedure allows describing a soil composition for each of the pipelines using the data presented by geotechnical data from the route maps. Pictures below show results of applied algorithm.

Soil type pipeline A1

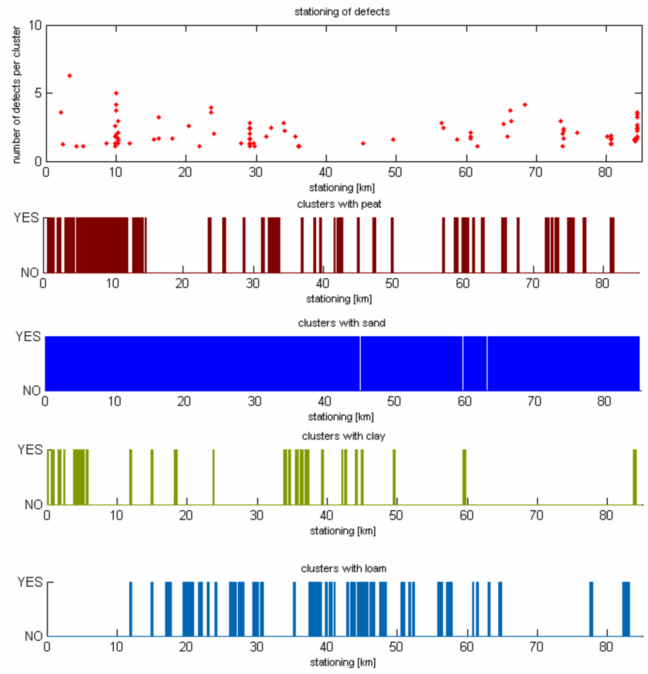


Figure 7-4: Soil composition for the pipeline A1

Soil type pipeline A2

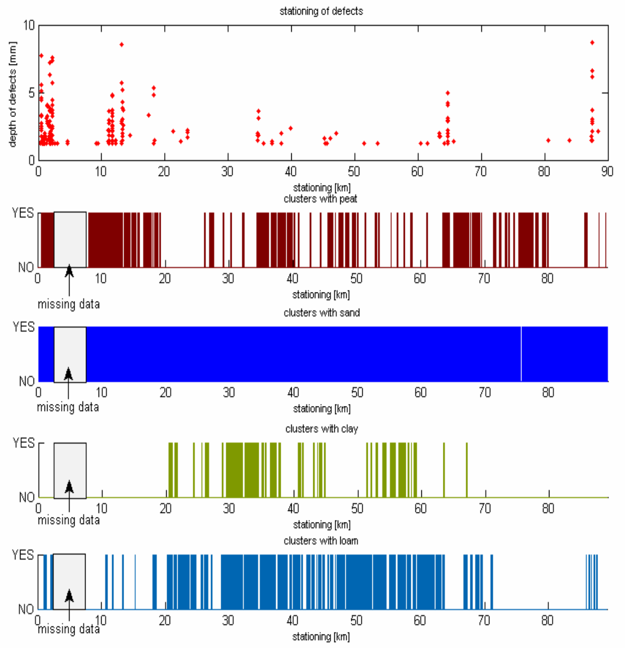


Figure 7-5: Soil composition for the pipeline A2

Soil type pipeline A3

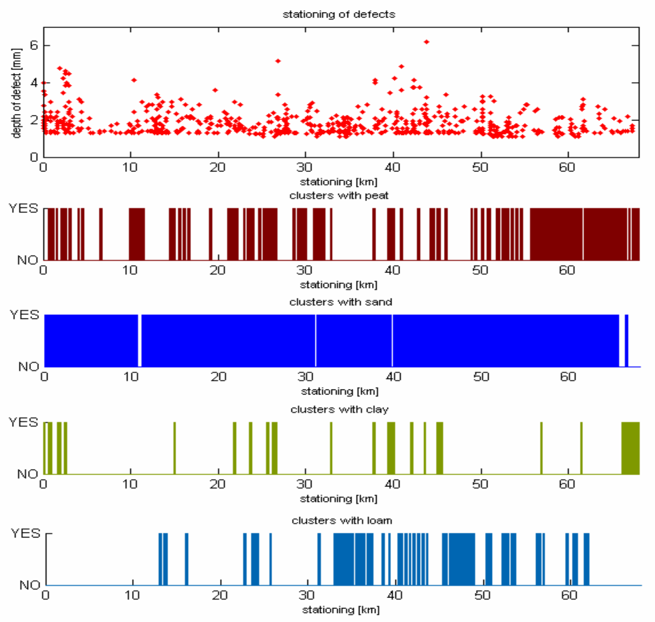


Figure 7-6: Soil composition for the pipeline A3

All three pictures Figure 7-4, Figure 7-5, and Figure 7-6 present both: defects distribution and existence of the each soil component[30]. The top plot presents the stationing of the defects reported by an intelligent pig and the depth of the defects. Each of the pictures beneath is associated with the soil component. Red color indicates existence of peat, blue of sand, light green of clay and light blue loam. Each of the vertical lines indicates a sample cluster (defined before) where presence of peat, sand, loam or clay was pointed out. The figures illustrate that almost whole the pipeline is laid in sand mixed with other elements. It reasonable to conclude that for the pipelines A1 and A2 the most of the defects are in the area where soil type is a mixture of sand and peat. For the pipeline A3 there is no clear pattern of relationships. Another outline is that in the middle of the pipelines A2 and A1 high concentration of clay and loam with sand is associated with relatively low number of defects.

From the collected data it is clear that four presented soil types do not exhaust all the possibilities. All the possible types of the soil which can be observed solely from the geotechnical data are: peat, sand, clay, loam, peat-sand, peat-clay, peat-loam, sand-clay, sand-loam, clay-loam, peat-sand-clay, peat-sand-loam, peat-clay-loam, sand-clay-loam and the last one peat-sand-clay-loam. Each of the soil types has to be analyzed wrt influence on the defect rate.

Remarks

- *Pipeline A2: missing soil data for stationing 2112m- 6723 m*
- *Pipeline A3: missing soil data for the first 3135 m- this missing data is recovered from the A1 which is parallel to A3*

7.4 Soil type influence on defect rate

The main point of the study in this subchapter is to check what the defect rate for whole the pipeline is and to verify if the defect rate depends on the soil type. The plan is to

[30] Main components are: sand, peat, clay and loam. The soil samples consists either from main components or components combinations.

count the number of the defect associated with each possible environment (soil type) and divide this by the total length where each specific soil type was observed. The overall defect rate for the pipelines is presented beneath in the Table 25.

pipeline	no. of defects	length of the pipeline [km]	Overall defect rate [def/km]
A1	92	86	1.1
A2	267	89	3
A3	657	69	9.5

Table 25 Overall defect rate for the pipelines

Two different approaches on defects rate wrt soil type are presented beneath. Each of the approaches presents different point of view on modeling.

7.4.1 Defect rate- Approach 1

The first approach associates the defect rate with each possible soil type. The assumption which is going to be imposed on the analysis is of following form:

Assumption
Assume that quality of a coating of any pipeline is deteriorated at the same level i.e. condition of it is uniform along the whole pipeline length.

Table 26 shows the results for all three pipelines.

	Soil type	Length of pipeline exposed to the soil type	Percentage of pipeline exposed to the soil type	Number of defects	Defect RATE
		[km]	[%]		[no. of defects per km]
Pipeline A1	Peat	0.09	0.11	0	0
	Sand	**38.10**	**45.42**	**50**	**1.31**
	Clay	0.30	0.35	0	0
	Loam	0.27	0.32	0	0
	Peat- Sand	**19.70**	**23.48**	**26**	**1.32**
	Peat- Clay	0.12	0.14	0	0
	Peat- Loam	0	0	0	0
	Sand- Clay	**4.21**	**5.02**	**7**	**1.66**
	Sand- Loam	**16.78**	**20**	**6**	**0.36**
	Clay- Loam	0	0	0	0
	Peat- Clay- Loam	0.26	0.3	0	0
	Peat- Sand- Clay	**2.67**	**3.18**	**2**	**0.75**
	Peat- Sand- Loam	0.96	1.14	0	0
	Sand- Clay- Loam	0.19	0.22	0	0
	Peat- Sand- Clay- Loam	**0.27**	**0.32**	**1**	**3.75**
	TOTAL	**83.9 [km]**	**100%**	**92 defects**	-
Pipeline A2	Peat	0.56	0.66	0	0
	Sand	**35.80**	**42.36**	**72**	**2.01**
	Clay	0	0	0	0
	Loam	0	0	0	0
	Peat- Sand	**19.87**	**23.51**	**104**	**5.23**
	Peat- Clay	0	0	0	0
	Peat- Loam	0	0	0	0
	Sand- Clay	0	0	0	0
	Sand- Loam	**13.94**	**16.5**	**15**	**1.08**
	Clay- Loam	0	0	0	0
	Peat- Clay- Loam	0	0	0	0
	Peat- Sand- Clay	0	0	0	0
	Peat- Sand- Loam	**4.49**	**5.31**	**38**	**8.46**
	Sand- Clay- Loam	**7.51**	**8.88**	**6**	**0.80**
	Peat- Sand- Clay- Loam	**2.35**	**2.77**	**2**	**0.85**
	TOTAL	**84.5 [km]**	**100%**	**237 defects**[31]	-
Pipeline A3	**Peat**	**1.45**	**2.11**	**15**	**10.33**
	Sand	**25.5**	**37.18**	**228**	**8.92**
	Clay	**0.54**	**0.79**	**5**	**9.26**
	Loam	**0.23**	**0.33**	**1**	**4.35**
	Peat- Sand	**19.27**	**28.05**	**168**	**8.72**
	Peat- Clay	**2.15**	**3.13**	**7**	**3.25**
	Peat- Loam	0	0	0	0
	Sand- Clay	**0.65**	**0.94**	**23**	**35.52**
	Sand- Loam	**10.84**	**15.78**	**131**	**12.08**
	Clay- Loam	0	0	0	0
	Peat- Clay- Loam	0	0	0	0
	Peat- Sand- Clay	**3.20**	**4.65**	**35**	**10.95**
	Peat- Sand- Loam	**3.50**	**5.1**	**27**	**7.71**
	Sand- Clay- Loam	**0.70**	**1.01**	**12**	**17.27**
	Peat- Sand- Clay- Loam	**0.63**	**0.92**	**4**	**6.35**
	TOTAL	**68.6 [km]**	**100%**	**656 defects**	-

Table 26 defect rate for each specific soil type

[31] Number of defects in the environment where soil data is missing is 30.

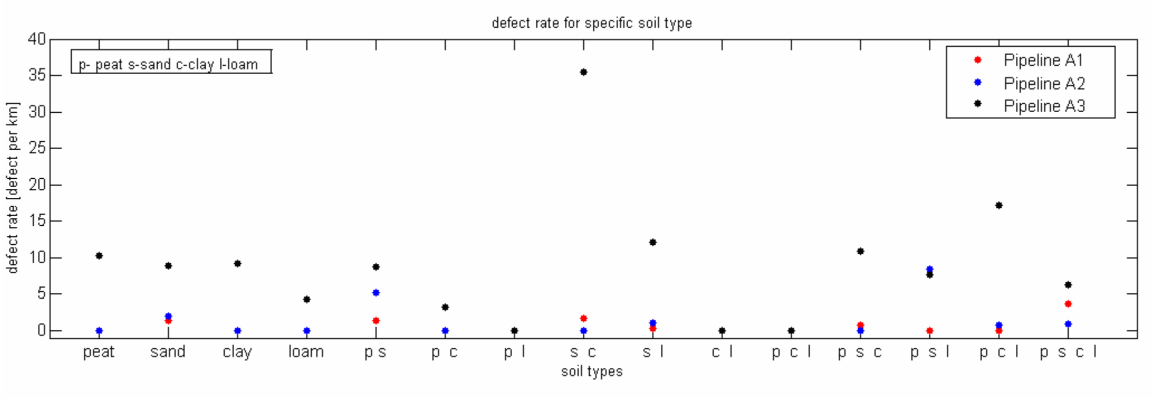


Figure 7-7: Defect rate associated for each possible soil composition

Figure 7-7 presents the results from the Table 26. Due to high number of the defects distributed along the pipeline A3 the defect rate for each of the soil types is much higher than for the remaining pipelines. In four cases: sand, peat-sand, sand-clay, and peat-sand-clay- loam both pipelines A1 and A2 have positive defect rate. The highest defect rate is obtained: A1 for peat-sand-clay-loam, A2 for peat-sand-loam, A3 for sand-clay. In classify and check if the estimated defects rates are reliable or not, it is required is to test how the soil type is distributed in potentially defective and no-defective environment.

Suppose that the soil type A occurs in two different parts of the pipeline, but defects are only observable in one of them (the second has none of defects). Then the defect rate won't be a reliable tool to say which soil type is more likely to be more defective. The additional information that has to be delivered is the information about changes of percentage content of the soil type A where the defects are observable and where they are not. To check how much of each soil type is in potentially defective environment and how much is not, the assumption about a potentially defective environment has to be impressed.

Suppose that defect *j* was registered at a certain stationing $S_j\ [km]$. The environment for this defect is defined as the area surrounding the defects within predefined radius *r [km]* and thus the environment for defect is $L_j = \lfloor S_j - r, S_j + r \rfloor$. The radius surrounding the defects is chosen to be equal to 250 [m]. This length is motivated by average length of the cluster for all the pipelines. The total length of the environments where all the defects were registered is then defined in following way:

- If defects' clusters L_j are not disjoint then total length of potentially defective environment is:

$$|T| = |L_1| + |L_2| + ... + |L_n| - |L_1L_2| - |L_1L_3| - ... - |L_1L_n| - |L_2L_3| - ... - |L_2L_n| - ... - \\ - ... - |L_{n-1}L_n| + |L_1L_2L_3| + ... + ... \pm |L_1L_2...L_n|$$

$\pm$ *next to the last terms means: "-" if n is even, and "+" if n is odd.*

Table 27 shows length of "potentially defective and not defective environments" according to introduced method.

pipeline	total length of the pipeline [km]	Potentially defective environment [km]	Potentially not detective environment [km]
A1	86	20.5	65.5
A2	89	20	64
A3	69	63.5	5.5

Table 27: description of the potentially defective and not defective environments

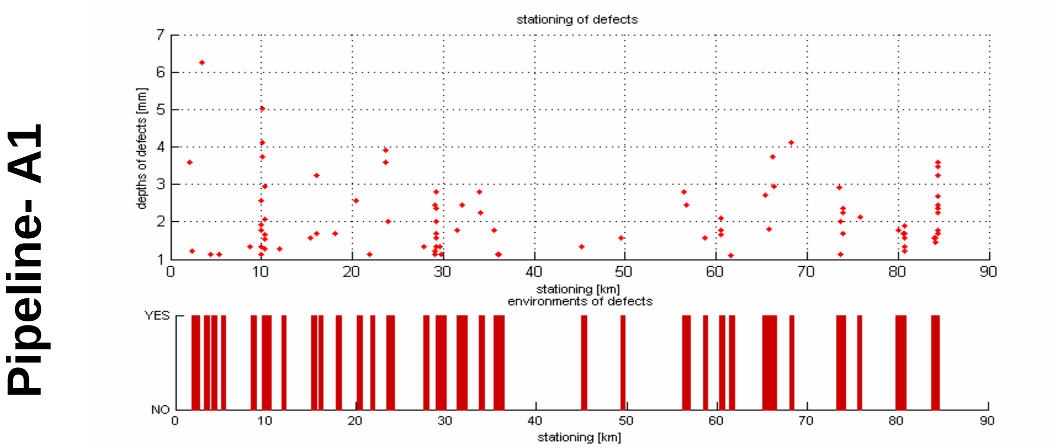


Figure 7-8: pipeline A1, potentially defective clusters, or clusters with “bad” coating

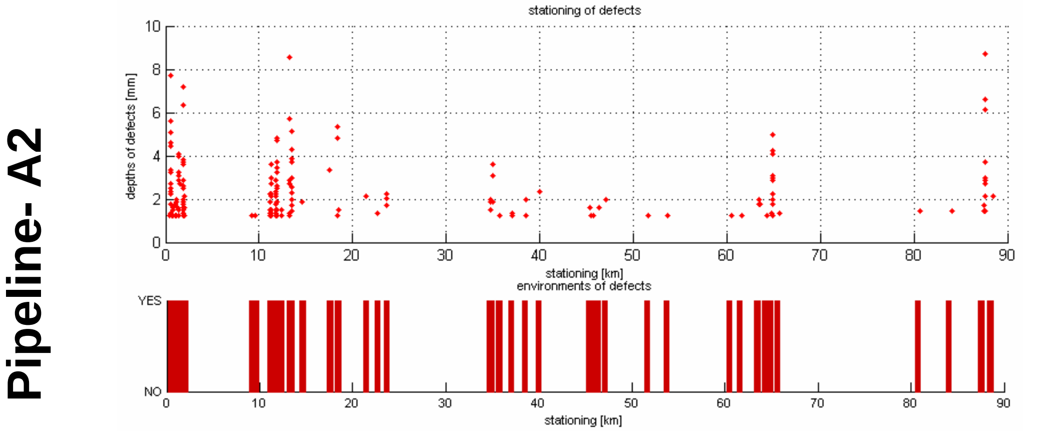


Figure 7-9: pipeline A2, potentially defective clusters, or clusters with “bad” coating

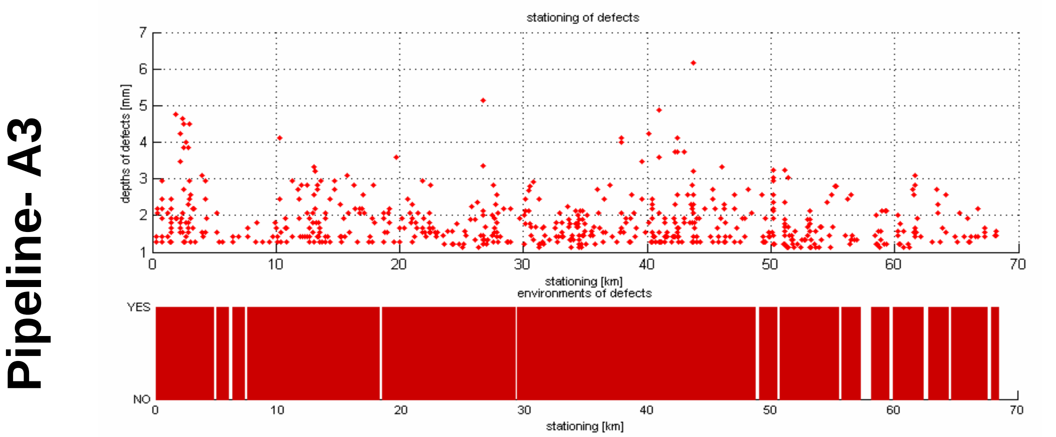


Figure 7-10: pipeline A3, potentially defective clusters, or clusters with “bad” coating

In order to compare the difference in soil composition between potentially defective and not defective environment, let’s define:

- D - potentially defective area (on the Figure 7-8, Figure 7-9 and Figure 7-10 *D* it is a set of intervals indicated by red bars)
- D^c - area defined as area complementary to *D*

- A - soil type A, then:

$$Difference = \frac{|A \cap D|}{|D|} - \frac{|A \cap D^c|}{|D^c|}$$

Table 28 below shows percentage content of these two environments:

	Soil type	DEFECTIVE- Environment		NOT DEFECTIVE- Environment		Difference
		%	[km]	%	[km]	%
Pipeline- A1	Peat	0	0	0.14	0.09	-0.14
	Sand	40.45	8.31	47.7	30.17	-7.25
	Clay	0	0	0.47	0.3	-0.47
	Loam	0	0	0.43	0.27	-0.43
	Peat- Sand	27.61	5.67	21.57	13.64	6.04
	Peat- Clay	0.58	0.12	0	0	0.58
	Peat- Loam	0	0	0	0	0
	Sand- Clay	9.41	1.93	4.02	2.54	5.39
	Sand- Loam	14.7	3.02	21.76	13.76	-7.06
	Clay- Loam	0	0	0	0	0
	Peat- Clay- Loam	0.64	0.13	0.2	0.12	0.44
	Peat- Sand- Clay	4.11	0.84	2.3	1.45	1.81
	Peat- Sand- Loam	1.2	0.25	1.12	0.71	0.08
	Sand- Clay- Loam	0	0	0.29	0.19	-0.29
	Peat- Sand- Clay- Loam	1.3	0.27	0	0	1.3
	TOTAL	**100**	**20.4**	**100**	**63.5**	**-**
Pipeline- A2	Peat	0.8	0.16	0.62	0.4	0.18
	Sand	33.98	6.83	44.59	28.49	-10.61
	Clay	0	0	0	0	0
	Loam	0	0	0	0	0
	Peat- Sand	38.25	7.68	19	12.14	19.25
	Peat- Clay	0	0	0	0	0
	Peat- Loam	0	0	0	0	0
	Sand- Clay	0	0	0	0	0
	Sand- Loam	13.52	2.72	17.57	11.23	-4.05
	Clay- Loam	0	0	0	0	0
	Peat- Clay- Loam	0	0	0	0	0
	Peat- Sand- Clay	0	0	0	0	0
	Peat- Sand- Loam	8.32	1.67	4.41	2.82	3.91
	Sand- Clay- Loam	1.49	0.3	11.28	7.21	-9.79
	Peat- Sand- Clay- Loam	3.64	0.73	2.53	1.61	1.11
	TOTAL	**100**	**20.1**	**100**	**63.9**	**-**
Pipeline- A3	Peat	2.28	1.45	0	0	2.28
	Sand	37.71	23.98	24.91	1.41	12.8
	Clay	0.85	0.54	0	0	0.85
	Loam	0.36	0.23	0	0	0.36
	Peat- Sand	26.89	17.1	34.62	1.96	-7.73
	Peat- Clay	2.46	1.56	21.08	1.19	-18.62
	Peat- Loam	0	0	0	0	0
	Sand- Clay	0.93	0.59	6.5	0.37	-5.57
	Sand- Loam	16.39	10.42	7.43	0.42	8.96
	Clay- Loam	0	0	0	0	0
	Peat- Clay- Loam	0	0	0	0	0
	Peat- Sand- Clay	5.03	3.2	0	0	5.03
	Peat- Sand- Loam	5.02	3.19	5.47	0.31	-0.45
	Sand- Clay- Loam	1.09	0.69	0	0	1.09
	Peat- Sand- Clay- Loam	0.99	0.63	0	0	0.99
	TOTAL	**100**	**63.6**	**100**	**5.7**	**-**

Table 28: comparison of potentially “defective” and “not defective” environments

Pipeline- A1

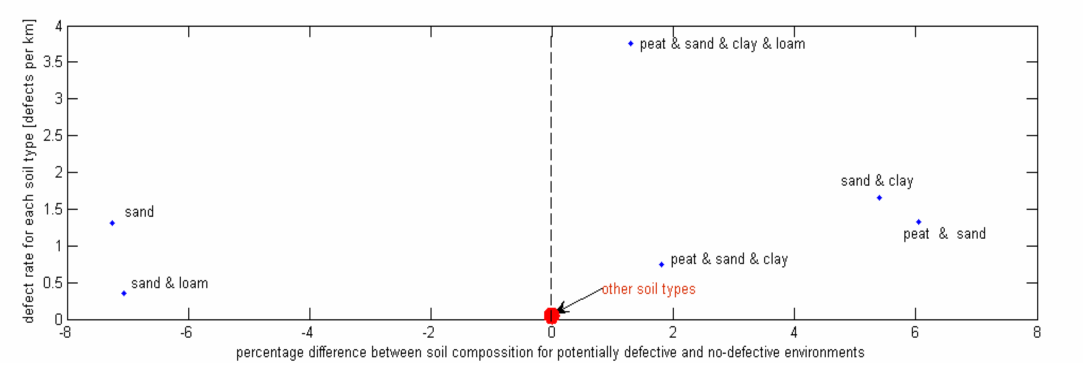


Figure 7-11: percentage difference between soil composition vs. defect rate (pipeline A1)

Pipeline- A2

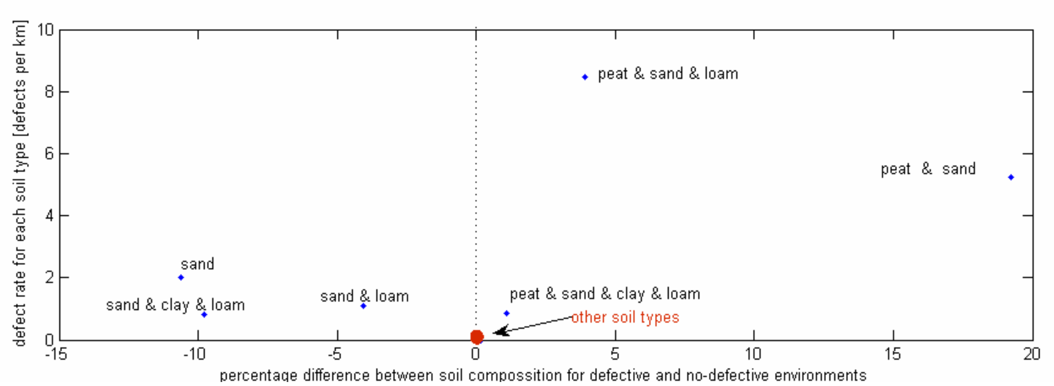


Figure 7-12: percentage difference between soil composition vs. defect rate (pipeline A2)

Pipeline- A3

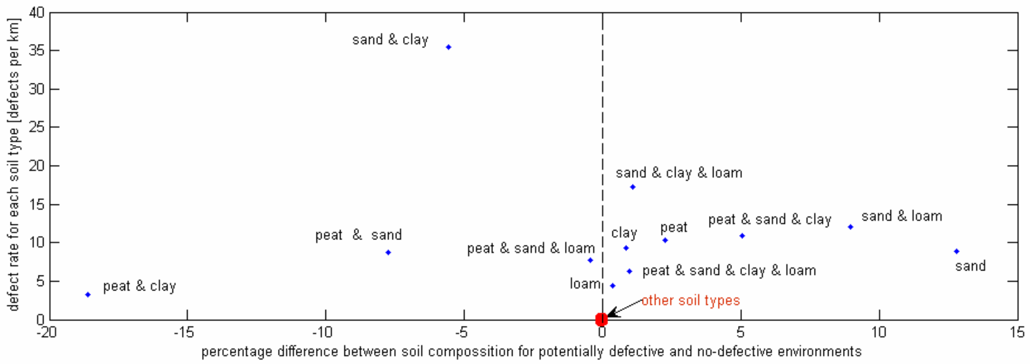


Figure 7-13: percentage difference between soil composition vs. defect rate (pipeline A3)

The Figures Figure 7-11, Figure 7-12 and Figure 7-13 show the relation between differences in soil composition for potentially defective and not defective environments. “Other soil types” indicates all the remaining soil types. According to the established technique, the general conclusion is that it is difficult to find clear pattern combining data from all the pipelines. However, certain matches are visible: for both A1 and A2 there is much more peat-sand in potentially defective environments and much more sand and sand-loam in potentially not-defective environment. There is no clear match between A3 and the others.

7.4.2 Defect rate- Approach 2

Before it was assumed that environment without defects is potentially not corrosive environment. This might be not the case. Many of excavations showed that existence of microbial corrosion was associated with coating damage. Here, different way of looking at the defect rate estimation is presented. In the previous section uniform coating along the pipeline was assumed. Now, the assumption is following:

Assumption

Assume that the pipeline consists of two different qualities of the coating “good” (without defects) and “bad” (with defects). Bad coating of the pipeline is defined in the same way as for “potentially defective environment” from previous section.

The second approach tries analyzing the pipelines defect rate only in the sections where bad coating was applied.

.

Below in the Table 29, summary of the results is presented.

	Pipeline A1		Pipeline A2		Pipeline A3	
Soil type	Exposure of “bad” coating	Def. rate for “bad” coating	Exposure of “bad” coating	Def. rate for “bad” coating	Exposure of “bad” coating	Def. rate for “bad” coating
	[km]	**[def/km]**	**[km]**	**[def/km]**	**[km]**	**[def/km]**
Peat	0	**0**	0.16	**0**	1.45	**10.34**
Sand	8.31	**6.02**	6.83	**10.54**	23.98	**9.51**
Clay	0	**0**	0	**0**	0.54	**9.26**
Loam	0	**0**	0	**0**	0.23	**4.35**
Peat- Sand	5.67	**4.59**	7.68	**13.54**	17.1	**9.82**
Peat- Clay	0.12	**0**	0	**0**	1.56	**4.49**
Peat- Loam	0	**0**	0	**0**	0	**0**
Sand- Clay	1.93	**3.63**	0	**0**	0.59	**38.98**
Sand- Loam	3.02	**1.99**	2.72	**5.51**	10.42	**12.57**
Clay- Loam	0	**0**	0	**0**	0	**0**
Peat- Clay- Loam	0.13	**0**	0	**0**	0	**0**
Peat- Sand- Clay	0.84	**2.38**	0	**0**	3.2	**10.94**
Peat- Sand- Loam	0.25	**0**	1.67	**22.75**	3.19	**8.46**
Sand- Clay- Loam	0	**0**	0.3	**20**	0.69	**17.39**
Peat- Sand- Clay- Loam	0.27	**3.7**	0.73	**2.74**	0.63	**6.35**
TOTAL	**20.4**	**-**	**20.1**		**63.6**	

Table 29: defect rate for sections where bad coating was applied

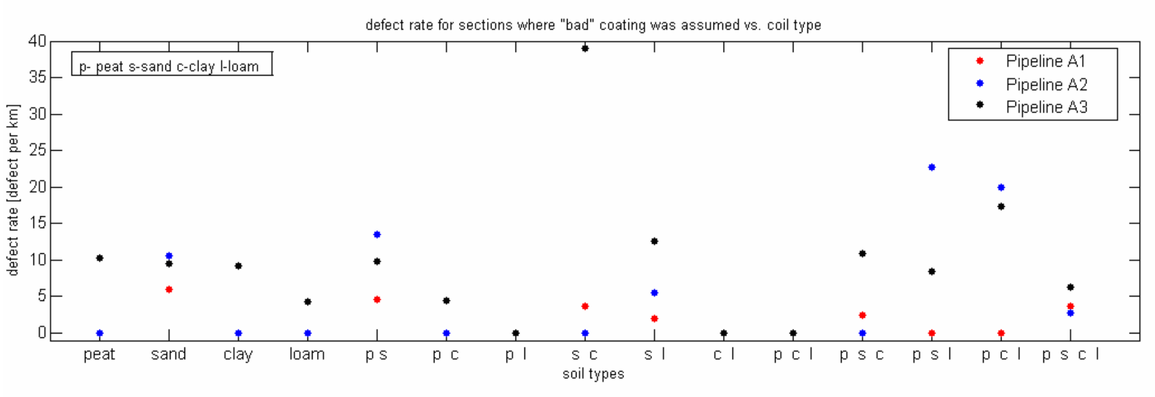


Figure 7-14 defect rate for sections where “bad” coating was applied vs. soil types

The highest defect rate for pipeline A1 is obtained for sand, A2 sand-clay-loam and A3 for sand-clay.

7.4.3 Correlation analysis

Techniques introduced before allow checking what the correlations between all the pipelines wrt soil types and corrosion defects are.

7.4.3.1 Correlation between soil exposures

Let's define:

- X_{A1}, X_{A2}, X_{A3} - percentage of the pipeline exposed to every soil type for the pipelines A1, A2 and A3, (i'th coordinate of the vector $X_{A...}$ describes percentage of the pipeline exposed to i'th soil type, number of i's is 13[32])

Pipeline	ρ_P (Pearson corr)[33]	p-value	ρ_S (Spearman corr)	p-value
A1-A2	0.96	0.001	0.47	0.1
A1-A3	0.97	0.001	0.58	0.03
A2-A3	0.96	0.001	0.73	0.005

Table 30: correlation between soil compositions of the pipelines

p-value in the table is associated with following null hypothesis

$$H_0 : \rho(X_i, X_j) = 0 \text{ against } H_1 : \rho(X_i, X_j) \neq 0$$

7.4.3.2 Correlation between defect rates wrt soil types

Let's define:

- X_{A1}, X_{A2}, X_{A3} - -defect rate vectors for the pipelines A1, A2 and A3 (i'th coordinate of the vector $X_{A...}$ describes defect rate for i'th soil type, number of i's is 13[32])

Pipeline	ρ_P (Pearson corr)[33]	p-value	ρ_S (Spearman corr)	p-value
A1-A2	0.004	0.98	0.32	0.28
A1-A3	0.21	0.50	0.27	0.37
A2-A3	-0.11	0.72	0.01	0.95

Table 31: correlation between defect rates of the soil types for pipelines

7.4.3.3 Correlation between defect rates wrt soil types where bad coating was assumed

- X_{A1}, X_{A2}, X_{A3} - -defect rate vectors for the pipelines A1, A2 and A3 for the areas of the pipelines where "bad" coating was assumed (i'th coordinate of the vector $X_{A...}$ describes defect rate for i'th soil type within the pipeline where bad coating was used, number of i's is 13[32])

Pipeline	ρ_P (Pearson corr)[33]	p-value	ρ_S (Spearman corr)	p-value
A1-A2	0.08	0.79	0.27	0.38
A1-A3	0.3	0.31	0.32	0.28
A2-A3	0.04	0.90	0.22	0.47

Table 32: correlation between defect rates of the soil types for pipelines, in the clusters where bad coating was applied

[32] The number of all possible component combinations is 15; however peat-loam and clay-loam are not registered at all.

[33] Normality condition for Pearson correlation coefficient is satisfied

7.5 Conclusions and recommendations

The analysis in this chapter was based on three high pressure pipelines for which excavations showed existence of MIC.

Conclusions
Overall conclusions from presented approaches are following.

- there is no significant correlation between defect rates and soil types
- there is no significant correlation between defect rates and soil types in the places where “bad” coating was assumed
- the highest defect rates are
    - all the pipeline
        - pipeline A1: in mixture of peat-sand-clay-loam (3.75 [def/km] it is about 41% of all the corrosion defect rates)
        - pipeline A2: in mixture of peat-sand-loam (8.46 [def/km] -46% of total corrosion defect rates)
        - pipeline A3: in mixture of sand-clay (35.52 [def/km]- 26%)
    - pipeline in the area with “bad” coating was assumed
        - pipeline A1: in sand (6.02 [def/km]- 27%)
        - pipeline A2: in mixture of peat-sand-loam (22.75 [def/km]- 30%)
        - pipeline A3: in mixture of sand-clay (38.98 [def/km]- 27%)
- For both the pipelines A1 and A2 there is much more peat-sand, and peat-sand-clay-loam and much less sand and sand-loam in the areas where “bad” coating was assumed
- Pipeline A3 doesn't show any pattern wrt the other pipelines

An explanation for lack of correlations (high uncertainty) may be due to

- high uncertainty of the soil measurements
- lack of data on coating quality- certain assumptions about “good” and “bad” coating had to be imposed
- lack of data about quantitative amount of soil components in the soil samples

Recommendations:

- The analysis was based on three pipelines, if this is the case, all the pipelines where MIC was reported should be analyzed in order to get better pattern.
- Deeper investigation of assumption about “good” and “bad” coating is required.

Chapter 8

8 Water table analysis

8.1 Introduction

This chapter is dedicated to groundwater table (level) analysis. The data on water levels were collected from the “soil map of the Netherlands”. The maps show the “average min and max ground water levels”. Unfortunately, the data are not accurate, mainly because of following features: only rough estimate is presented (groundwater level step classification data), maps do not present exact water level for a specified stationing. The data were not delivered in a digital form but were collected directly from the water table map by the author.

The legend in the maps is following:

Groundwater level step classification data							
Groundwater step	***I***	***II***	***III***	***IV***	***V***	***VI***	***VII***
Average of the highest groundwater level in cm below the ground	*-*	*-*	*<40*	*>40*	*<40*	*40-80*	*>80*
Average of the lowest groundwater level in cm below the ground	*<50*	*50-80*	*80-120*	*80-120*	*>120*	*>120*	*>120*

Table 33: ground water step levels (legend)

The data was collected simply by projecting the pipeline profile on the groundwater level map and the values of groundwater step levels were collected. The data shows that it is quite difficult to indicate precisely what the groundwater level for a given stationing is. This is mostly due to pipeline shape (it is difficult to calculate pipeline’s length using only the map), and the map itself (maps are constructed based on contour plots).

For simplification the pipelines were divided into 2.5 [km] long sections. For each of the section average groundwater class was calculated. The analysis is aimed to check the relationship between the number of defects and the water level for the associated sections.

8.2 Approach 1

The figures below: Figure 8-1, Figure 8-2 and Figure 8-3 consist of two parts. The plot on the top shows the number of defects per each segment of length 2.5 [km], and the bottom plot presents the groundwater step level for the corresponding section. From the available maps it was difficult to associate groundwater step levels for narrower sections.

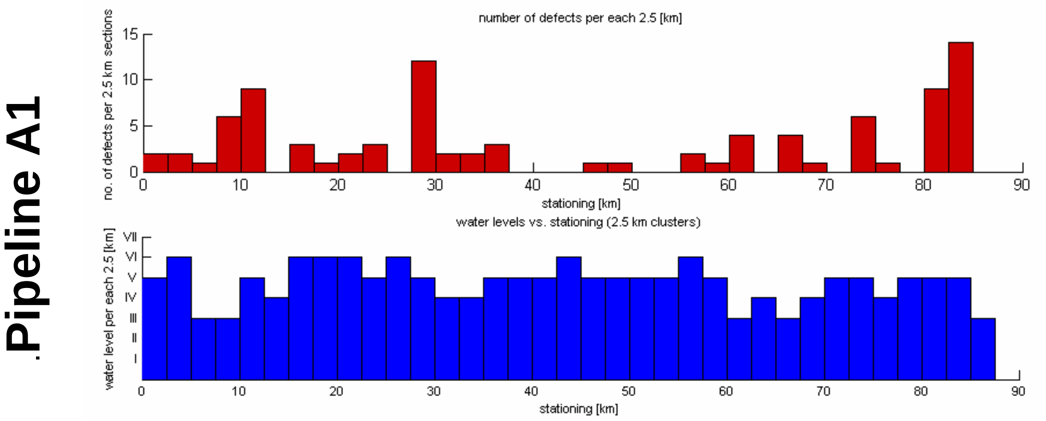


Figure 8-1: A1- no. of defects vs. water level per 2.5 km long sections

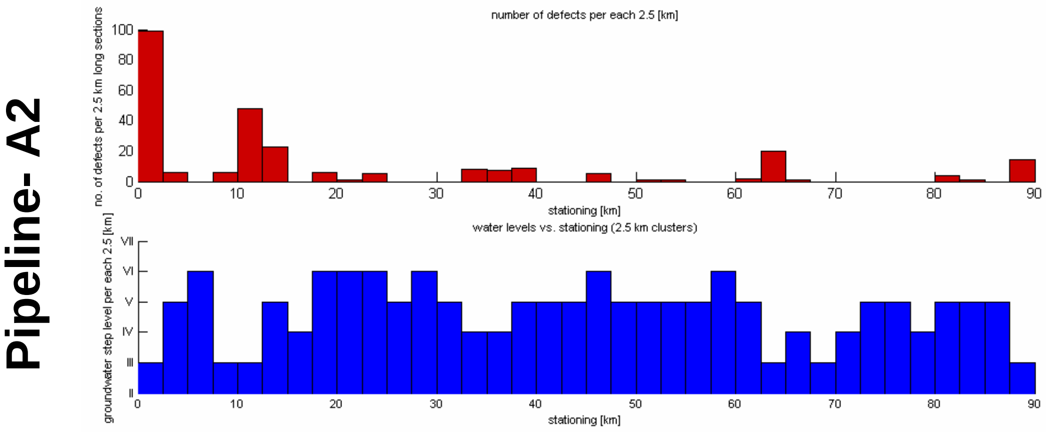


Figure 8-2 A2- no. of defects vs. water level per 2.5 km long sections

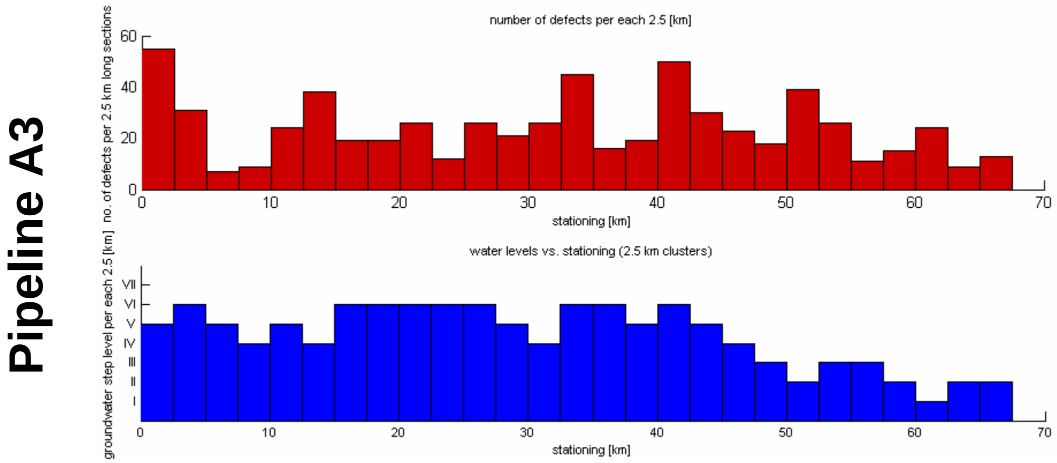


Figure 8-3: A3- no. of defects vs. water level per 2.5 km long sections

Table 34 shows summary of available groundwater level data.

Pipeline	Number of samples	Minimum	Maximum	Average
A1	35	3	6	5
A2	36	3	6	5
A3	28	1	6	4

Table 34: groundwater step level summary per each section of 2.5 km

Pipeline	Number of samples	Minimum	Maximum	Mean	Std. Deviation
A1	35	0	5.6	1.1	1.4
A2	36	0	39.6	2.88	7.2
A3	28	2	22	9.37	5.1

Table 35 defect rate summary- per each section of 2.5 km

Figure 8-1 and Figure 8-2 show that in the places where groundwater level is high (it is indicated by low step number) the number of the defects is also high. In order to verify this pattern for all the pipelines let's define:

- X- defect rate per each section of 2.5 [km]
- Y- average groundwater step level for each 2.5 [km] long section

The hypothesis that has to be tested is of the following form:

$$H_0 : \rho(X,Y) = 0 \text{ against } H_1 : \rho(X,Y) \neq 0$$

Table 36 below shows the results for correlations: ρ_P - Pearson correlation, ρ_S - Spearman correlation and ρ_K - Kendall correlation, and also associated p-values for all the pipelines.

Pipeline	ρ_P [34]	p-value	ρ_S	p-value	ρ_K	p-value
A1	*-0.1*	*0.7*	*-0.12*	*0.55*	*-0.09*	*0.5*
A2	*-0.32*	*0.05*	*-0.24*	*0.15*	*-0.19*	*0.17*
A3	*0.30*	*0.12*	*0.31*	*0.1*	*0.22*	*0.13*

Table 36 correlation between defect rate and water level for sections of 2.5 km

The results presented in the table confirm the suspicion. The groundwater step level for the pipelines A1 and A2 is negatively correlated with the defect rate. However this correlation is not statistically significant. For A3 there is no significant relationship between the defect rate and groundwater step level.

Overall conclusion is following: according to the introduced methodology there is weak correlation between defect rate and groundwater step level.

8.3 Approach 2

For the pipelines A1 and A2 it seems that there are many areas where zero or only one defect was reported. So the pipeline will be analyzed with respect to these sections. The first feature will be indicated as presence of the defects, and the second by defects absence. Let’s define:

- X- groundwater step level vector for area where only 0 or 1 defect was observed
- Y- groundwater step level vector for area where more than 1 defects were observed

The task is to check if the difference between averages of groundwater levels for defined variables is statistically significant. So define the hypothesis as:

$$H_0 : \overline{X} - \overline{Y} = 0 \Leftrightarrow \overline{X} = \overline{Y} \text{ against the alternative } H_1 : \overline{X} \neq \overline{Y}$$

pipeline	Variables	df	t-statistic	p-value for t-test	Lower 95% bound (for difference)	Upper 95% bound (for difference)
A1	X-Y	32	0.18	0.85	-0.59	0.71
A2	X-Y	34	1.55	0.13	-0.16	1.17

Table 37: statistics and confidence bounds for estimate (pipeline A1 and A2)

The results show that for a significance level $\alpha = 0.05$ the null hypothesis cannot be rejected. It means that averages of groundwater step levels of the environments with and without defects are not statistically different.

Beneath in the Table 38 small description of the groundwater levels for “defective” and “no defective” environments is presented.

pipeline	Descriptive Statistic	Number of samples	Min. groundwater step	Max. groundwater step	Mean groundwater step
A1	X	17	3	6	5
	Y	17	3	6	5
A2	X	21	3	6	5
	Y	15	3	6	4

Table 38: descriptive statistics (pipelines A1 and A2)

[34] The analysis showed that the normality assumption for the variable “number of defects per cluster” is not satisfied.

The results from the presented idea did show insignificant relationships.

8.4 Approach 3

Next approach which can be applied to the water table analysis is carried out by coding the groundwater steps levels. The Table 33 shows that first five steps (I, II, III, IV, V) can be associated with high groundwater level (and coded as 0) and other two (VI and VII) with low groundwater level (and coded as 1). The Figure 8-4, Figure 8-5, and Figure 8-6 below present how coded average groundwater step levels are associated with the number of defects for each 2.5 km long clusters.

Pipeline A1

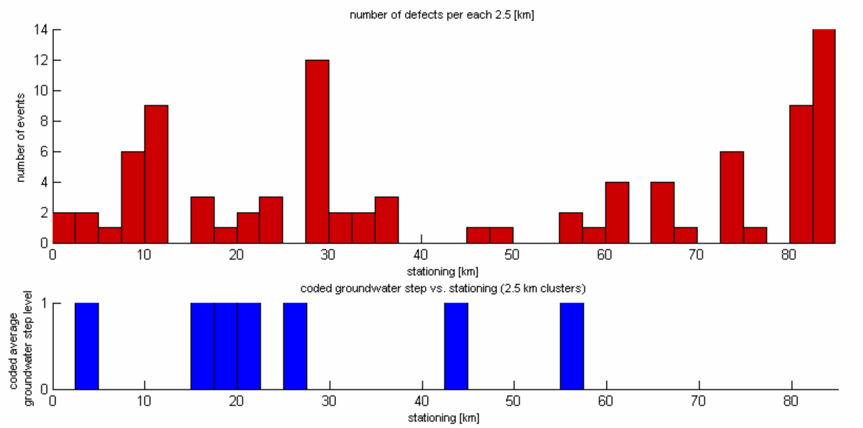


Figure 8-4: number of defects vs. coded groundwater level (pipeline A1)

Pipeline- A2

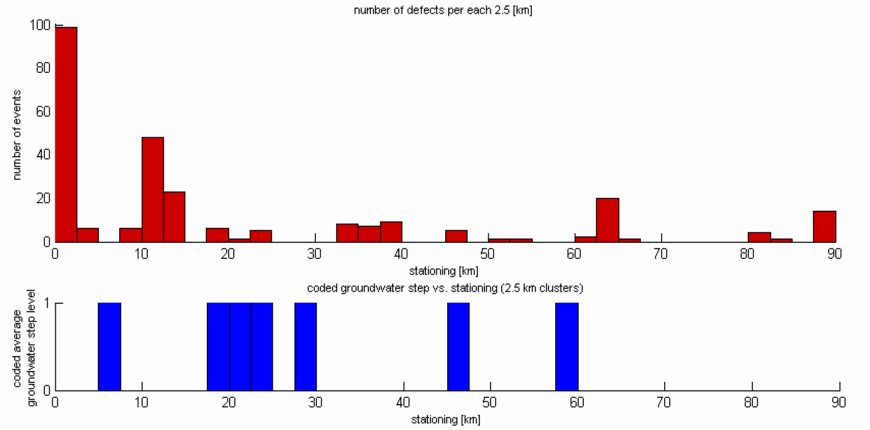


Figure 8-5: number of defects vs. coded groundwater level (pipeline A2)

Pipeline A3

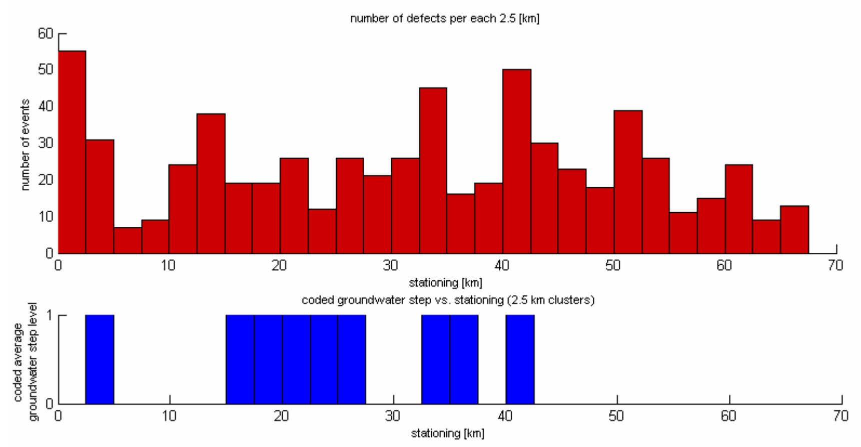


Figure 8-6: number of defects vs. coded groundwater level (pipeline A3)

The hypothesis which has to be verified is following: the average number of defects for average groundwater step level coded by 0 and 1 is significantly different against alternative that it is not. Necessary definitions are following:

- X- number of defects where average groundwater step level is coded as 0
- Y- number of defects where average groundwater step level is coded as 1

A mathematical formulation of hypothesis is following:

$$H_0 : \overline{X} - \overline{Y} = 0 \Leftrightarrow \overline{X} = \overline{Y} \text{ against the alternative } H_1 : \overline{X} \neq \overline{Y}$$

pipeline	Variables	df	t-statistic	p-value for t-test	Lower 95% bound (for difference)	Upper 95% bound (for difference)
A1	X&Y	32	1.07	0.29	-1.45	4.67
A2	X&Y	34	0.76	0.45	-9.97	21.96
A3	X&Y	26	-1.05	0.30	-16.0	5.16

Table 39: statistics and confidence bounds for estimate (pipelines A1, A2 and A3)

pipeline	Descriptive Statistic	Number of samples	Min. no. of defects	Max. no. of defects	Average no. of defects
A1	X	27	0	14	3.04
	Y	7	1	3	1.42
A2	X	28	0	99	8.42
	Y	7	0	6	2.42
A3	X	19	5	55	21.68
	Y	9	12	50	27.11

Table 40: descriptive statistics (pipelines A1, A2 and A3)

The Table 39 and Table 40 show that for A1 and A2 the average number of defects where groundwater step level is 0 (groundwater level is high) is higher than for the area where groundwater step level is 1 (groundwater is relatively lower); however this difference is not statistically significant[35]. In the case of A3 the there is no evidence to distinguish between the numbers of defects for coded groundwater levels.

Also this approach showed weak correlation between groundwater level and defect rate.

[35] This is due to high variability of the number of defects for groundwater step level coded as 0.

8.5 Conclusions and recommendations

The analysis didn't show that the groundwater levels in a statistically significant way influence the number of defects. The results may be not accurate since the pipelines were divided in sections of 2.5 [km]. However, the pipeline division in smaller segments is a challenging task since the data is not available in an electronic form, but available only directly from the maps. Analysis showed that for pipelines A1 and A2 higher groundwater level is positively correlated with the number of defects (more water - more defects); however these results according to statistical evidence are not strong enough. The Table 41 below shows small overall summary of the results.

Pipeline	Total number of defects	average groundwater step level
A1	93	5
A2	267	5
A3	657	4

Table 41: defects and average groundwater level

All three pipelines were installed in the 60s however the number of defects for each of them is very different. Comparison of the number of defects and the average groundwater step level shows that the highest number of defects for the A3 is associated with the highest groundwater level (lowest groundwater step level).

Recommendations

Deeper investigation of the groundwater levels is required. Because of the pipeline profile it was too difficult to collect the groundwater level for any given pipeline stationing. Precise approximate of the groundwater level for any given stationing would significantly increase the accuracy of the results.

Chapter 9

9 Factors influencing defect rate

9.1 Introduction

This chapter proposes the way of the defect rate modeling given all available measurements. Factors which are available are:

- groundwater level,
- soil types,
- NAP level of the ground,
- NAP level of the pipeline and
- depth of cover

The methodology used in this chapter is based on the regression analysis where the dependent variable is defect rate and independent variables are all the factors which may influence the defect rate. The regression analysis is based on a number of observations which describe the variable of interest. In order to define such observations a pipeline discretization is required. The question which has to be answered is in how long segments the pipeline should be divided. Here, the same as in the previous chapter – water table analysis -- the pipeline will be divided in 2.5 [km] long sections (according to the groundwater level measurements). In this study two general approaches will be presented.

First, let's define two models.

- model without interactions

Model 3

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

- model with interactions up to second degree

Model 4

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \beta_{n+1} X_1^2 + ... + \beta_{2n} X_n^2 + \beta_{1,2} X_1 X_2 + ...$$
$$\beta_{1,n} X_1 X_n + ... + \beta_{2,3} X_2 X_3 + ... + \beta_{i,j} X_i X_j + \beta_{n-1,n} X_{n-1} X_n + \varepsilon$$

where:

- Y- variable of interest- defect rate [no. of defects per km] calculated for each section of 2.5 [km]
- X_i - independent variables:
  - peat, sand, clay, loam, peat-sand, peat-clay, peat-loam, sand-clay, sand-loam, clay-loam, peat-clay-loam, peat-sand-clay, peat-sand-loam, sand-clay-loam, peat-sand-clay-loam which ***are described as percentage content of each section for each 2.5 [km] (each value says –"percentage of each soil type is in each section")***
  - groundwater step level (calculated as average for each section of 2.5 [km])
  - NAP level of the ground (average for each section)
  - NAP level of the pipeline (average for each section)
  - depth of cover (average for each section)
- n is associated with the number of variables which is equal to 19 (according to number of available variables)

9.2 Model without interactions

For the defined Model 3, in order to get the best possible combination of parameters which describe defect rate a stepwise regression[36] is applied.

The Model 3 is concerns each of the pipelines to check which parameters for each pipeline are significantly influencing the defect rate. Moreover, the model also is used to the pipelines combinations. The idea behind such combinations is to find the common parameters influencing the defect rate for all the pipelines. Table 42 and Table 43 below show standard statistics of the estimated coefficients. Stepwise regression output is a set of statistically significant variables influencing criterion variable which is defect rate. All the variables not included in the tables are insignificant in the modeling.

For the pipeline A1 there are no significant parameters describing the defect rate. For the A2 the most relevant variables are percentage amount of **peat-sand** and **peat**, and for A3, **sand-clay** and **sand-loam**. For A1 combined with A2 the most relevant is **peat-sand** (it means that **peat-sand** is a common factor influencing the defect rate for these two pipelines). And for combination of three pipelines A1, A2 and A3 the most relevant variables are percentage amount of **sand-loam** and **NAP level of the ground** (these variables are common variables for all the three pipelines).

[36] See chapter: analysis methods and interpretation- appendix A

	The parameters which are significant for the defect rate modeling	Coefficients		t- stat.	p-value for t-stat.	95% Confidence Interval for β_i	
		$\hat{\beta}_i$	Std. Error			L. Bound	U. Bound
A1[37]	**None of parameters is included**						
A2[38]	**(Constant)**	*-0.16*	*0.76*	*-0.21*	*0.84*	*-1.71*	*1.4*
	Peat-sand	*0.11*	*0.03*	*4.23*	ε	*0.06*	*0.16*
	peat	*-0.67*	*0.28*	*-2.34*	*0.03*	*-1.25*	*-0.08*
A3[39]	**(Constant)**	*7.846*	*0.97*	*8.03*	ε	*5.83*	*9.86*
	Sand-clay	*0.55*	*0.16*	*3.39*	*0.002*	*0.21*	*0.88*
	Sand-loam	*0.08*	*0.03*	*2.35*	*0.03*	*0.01*	*0.15*
A1-A2[40]	**(Constant)**	*0.45*	*0.46*	*0.98*	*0.33*	*-0.46*	*1.36*
	Peat-Sand	*0.044*	*0.01*	*3.21*	*0.002*	*0.02*	*0.07*
A1-A2-A3[41]	**(Constant)**	*6.93*	*0.76*	*9.1*	ε	*5.4*	*8.4*
	NAP level of the ground	*-0.60*	*0.09*	*-6.36*	ε	*-0.79*	*-0.41*
	Sand-loam	*0.08*	*0.03*	*3.15*	*0.002*	*0.03*	*0.13*

Table 42 Stepwise regression estimates (model without interactions)

The Table 43 below associates the estimated models and the multiple correlation coefficient (R^2), which describes level at which the variance of dependent variable is described.

In all the cases R^2 is relatively low- it doesn't exceed the level of 40%. This can be caused by uncertainty about the measurements or by the variables which are relevant but were not included in the model.

PIPELINE	Number of obs.	R^2	Adjusted R^2	Statistics	
				F statistics	p-value for F test
A1	34	-	-	-	-
A2	32	0.38	0.34	9.01	0.001
A3	27	0.39	0.33	7.54	0.003
A1-A2	66	0.14	0.13	10.3	0.002
A1-A2-A3	93	0.31	0.30	20.26	ε [42]

Table 43: standard model statistics (model without interactions)

[37] A1: The following variables are constants or have missing correlations, so will be deleted from the analysis: peat-loam, clay-loam

[38] A2: The following variables are constants or have missing correlations, so will be deleted from the analysis: Clay, Loam, peat-clay, peat-loam, sand-clay, clay-loam, peat-clay-loam, and peat-sand-clay.

[39] A3: The following variables are constants or have missing correlations, so will be deleted from the analysis: peat-loam, clay-loam, peat-clay-loam

[40] A1-A2: The following variables are constants or have missing correlations, so will be deleted from the analysis: peat-loam, clay-loam

[41] A1-A2-A3: The following variables are constants or have missing correlations, so will be deleted from the analysis: peat-loam and clay-loam

[42] ε stands for number smaller than 0.0001

9.3 Model with interactions

Second model is a model with interactions; it means that except all main effects, also interactions between effects are taken into account. This is idea is motivated by a well known fact that MIC corrosion is not only influenced by closed number of main effects.

	The parameters which are significant for the defect rate modeling	Coefficients		t- stat.	p-value for t-stat.	95% Confidence Interval for β_i	
		$\hat{\beta}_i$	Std. Error			L. Bound	U. Bound
A1	**(Constant)**	***0.76***	*0.27*	*2.82*	*0.01*	*0.21*	*1.31*
	Depth of cover * depth of cover	***38.525***	*7.79*	*4.94*	ε	*22.6*	*54.5*
	(sand-clay-loam) * depth of cover	***-3.74***	*0.92*	*-0.76*	ε	*-5.61*	*-1.90*
	Sand * Sand	***-0.001***	ε	*-2.22*	*0.03*	*-0.001*	$-\varepsilon$
A2	**(Constant)**	***1.56***	*0.06*	*2.48*	*0.02*	*0.27*	*2.86*
	Peat*(NAP level of the pipeline)	***0.18***	*0.05*	*3.82*	*0.001*	*0.08*	*0.28*
	Peat-sand	***0.07***	*0.03*	*2.39*	*0.024*	*0.01*	*0.121*
	(peat-sand) * (NAP level of the ground)	***-0.01***	*0.01*	*-2.54*	*0.02*	*-0.02*	*-0.003*
A3	**(Constant)**	***11.56***	*0.92*	*12.5*	ε	*9.65*	*13.5*
	(sand-clay) * (sand-clay-loam)	***-0.43***	*0.12*	*-3.59*	*0.002*	*-0.67*	*-0.18*
	Sand * sand	***-0.002***	*0.001*	*-3.38*	*0.003*	*-0.004*	*-0.001*
	(peat-clay-sand-loam) * (peat-clay-sand-loam)	***-0.08***	*0.03*	*-2.39*	*0.03*	*-0.14*	*-0.01*
A1-A2	**(Constant)**	***1.35***	*0.29*	*4.66*	ε	*0.77*	*1.9*
	(Peat-sand) * (groundwater step level)	***0.01***	*0.002*	*4.64*	ε	*0.006*	*0.015*
	Peat-sand	***0.05***	*0.01*	*3.87*	ε	*0.02*	*0.07*
A1-A2-A3	**(Constant)**	***5.02***	*0.59*	*8.45*	ε	*3.84*	*6.20*
	NAP level of the ground	***-0.61***	*0.09*	*-6.89*	ε	*-0.79*	*-0.43*
	Sand-loam	***0.08***	*0.02*	*3.27*	*0.02*	*0.03*	*0.13*
	(Sand-loam) * (NAP level of the pipeline)	***-0.01***	*0.004*	*-2.24*	*0.03*	*-0.02*	*-0.001*
	(Peat-sand) * (groundwater step level)	***0.008***	*0.003*	*2.78*	*0.007*	*0.002*	*0.01*
	Sand * sand	***-0.001***	*0.001*	*-2.30*	*0.024*	*-0.002*	$-\varepsilon$

Table 44 Stepwise regression estimates (model with interactions)

The Table 44 above shows which of the variables from the Model 4 are significant in the defect rate modeling. Interesting is that for the most of the cases (the pipelines and pipeline combinations) the most relevant variables included in the model are interactions. Only in two cases main effects were included in the model.

The Table 45 below shows that now R^2 is higher than before for model with only main effects. However, the R^2 still is very low- what indicates poor correlation.

PIPELINE	Number of obs.	R^2	Adjusted R^2	Statistics	
				F statistics	p-value for F test
A2	32	0.55	0.50	11.28	ε
A1	34	0.47	0.42	9.82	ε
A3	27	0.60	0.55	11.4	ε
A1-A2	66	0.36	0.34	17.56	ε
A1-A2-A3	93	0.43	0.39	13.0	ε

Table 45: standard model statistics (model with interactions)

9.4 Conclusions and recommendations

The analysis showed that much more relevant for the modeling the defect rate is looking at interactions than on main effects.
According to the introduced methodology the significant parameters (with strong correlation) which describe the defect rate are:

- For the pipeline A1:
    - **sand** (insignificant negative correlation)
    - interactions between
        - **sand-clay-loam** with **depth of cover** (insignificant negative correlation)
        - **depth of cover with depth of cover** (mid. positive correlation with the defect rate)
- For the pipeline A2:
    - Interactions between:
        - **peat** and **NAP level of the pipeline** (insignificant positive correlation)
        - **peat-sand** and **NAP level of the ground** (mid. negative correlation)
    - **Peat-sand** (mid. positive correlation)
- For the pipeline A3:
    - Interactions between:
        - **Sand-clay** with **sand-clay-loam** (mid. negative correlation)
        - **Sand** with **sand** (mid. negative correlation)
        - **Peat-clay-sand-loam** with **peat-clay-sand-loam** (insignificant negative correlation)
- Common factors for A1 and A2
    - **Peat- sand** (mid. positive correlation)
    - Interaction between
        - **Peat-sand** with **peat-sand-loam** (mid. positive correlation)
- Common factors for A1, A2 and A3
    - **NAP level of the ground** (mid. negative correlation)
    - **Sand-loam** (insignificant correlation)
    - Interactions between:
        - **Sand-loam** with **NAP level of the pipeline** (weak negative correlation)
        - **Peat-sand** with **groundwater step level** (weak positive correlation)
        - **Sand** with **sand** (insignificant negative correlation)

Recommendations

The analysis showed that the parameters included into regression model do not fully describe defect rate - this is indicated by low R^2. Two main reasons which have to be deeper investigated are:

- measurement error- there is a lot of uncertainty in the data, the uncertainty can be reduced by analyzing other pipelines where MIC was reported in order to get more general common factors influencing MIC defect rate
- other influencing factors which were not taken into account (this should be verified as well)
- pipelines were divided in 2.5 km long sections according to groundwater data, further investigation should be aiming at getting more narrower sections

Chapter 10

10 Conclusions

The thesis conducts an analysis on corrosion modeling for underground gas pipelines in the Netherlands. Each of the divisions incorporates certain knowledge about corrosion. Furthermore all the parts combined together deliver information about the whole process of corrosion rate/defect modeling.

First part showed the procedure of the corrosion rate modeling when low number of inspections is available. Implemented model shows that in order to determine reliable uncertainty and bias about MFL-pigs it is very important to have multiple reference defects in the pipeline, defects for which the real dimensions are well known (ex. from excavations). Because of the measurement uncertainty of the MFL-tool it dominant compared to the corrosion growth in the time period between the pigruns, it is very difficult to determine a reliable corrosion rate per defect. Model proposed in first section showed the way to calibrate all the inspecting pigs and how to derive physically acceptable functional description of corrosion growth. Assuming constant corrosion rate the model estimated average corrosion rate of level 0.24 [mm/yr] with upper bound value of 0.62 [mm/yr].

The assumption that the corrosion process is linear was underpinned by the analyses of two subsets: deep and shallow defects. For both subsets a similar average corrosion rate was calculated: 0.23 mm/yr and 0.25 mm/yr. The difference is not significant.

Since the corrosion rates have been determined, Gasunie will have to decide how to use these values for the calculation of a re-inspection interval for this line. The calculation will probably be done in a deterministic way for every defect, taking into account measurement uncertainties and the uncertainty of the corrosion rate.

Second section of the thesis incorporated results from the previous one in order to find factors influencing the corrosion rate. The analysis was performed based partially on bio-assessed measurements and partially using pipeline integrity management system.

The study based on regression analysis showed importance of having additional measurements. It has been shown that much more influencing for the corrosion rate are interactions between the variables than main effects. For incorporated set of 19 variables, two of them (SRB-A and SRB-B) were insignificant and due to missing data were removed from further analysis. According to applied sensitivity measures the most influential for the corrosion rate is variable describing pipeline wrt NAP level (positive correlation), then accordingly to importance: interactions between redox and water level (negative correlation), TOC and pipeline wrt NAP level (positive correlation), oxygen and pH (positive correlation), methane and SP (positive correlation) and final one MCA squared (negative correlation). The number of available observations plays crucial role in the modeling. Most analysts recommend that one should have at least 10 to 20 times as many observations as one has variables, otherwise the estimates of the regression are probably very unstable and unlikely to replicate if one were to do the study over. In the study number of included variables was 5 and 16 was a number observations. So, clearly final results cannot be used as reliable predictive tool.

Third and the last section demonstrated the ideas of defect rate modeling for a MIC influenced pipelines. Three pipelines affected by MIC were analyzed. Firstly, the soil type analysis was performed. Two of the pipelines A1 and A2 showed that in the areas where “bad” coating was assumed is much more peat-sand, and peat-sand-clay-loam and much less sand and sand-loam. Third pipeline A3 didn't show any significant pattern since the a whole pipeline is affected by corrosion. Applied correlation analysis didn't show significant correlations between soil types and defect rates for all the pipelines. Because of lack reliable measurements also groundwater analysis didn't show any significant correlations, between groundwater levels and number of defects. However, all the data combined together and applied to regression analysis showed certain patterns. Similarly like for the corrosion rate, the defect rate is much stronger influenced by interactions than by main effects. The defect rate modeled in this chapter showed that common factors influencing all three pipelines, A1, A2 and A3 are: NAP level of the ground (negative correlation), and sand mixed with loam (positive correlation) then interactions between: mixture of sand-loam and NAP level of the ground (negative correlation), mixture of peat-sand and groundwater step level (positive correlation) and sand squared (negative correlation). It was showed that pipeline A3 is much more different from the others. The parameters influencing the defect rate for both remaining pipelines A1 and A2 are: peat mixed with sand (positive correlation) and interaction between mixture peat-sand and groundwater step level (positive correlation).

Bibliography

[1] POF specifications

[2] Worthingham R.G., Fenyvesi L.L., Morrison T.B., Desjardins G. J., “Analysis of corrosion rates on a gas transmission pipeline”, NACExpo/Conference 2002 Technical Paper

[3] Bhatia A., Mangat N.S, Morrison T, “Estimation of measurement errors“.

[4] Desjardins G., “Corrosion rate and severity results from in-Line-Inspection data”, ACE Corrosion 2001 paper 01624

[5] Worthingham R.G., Morrison T, Mangat N.S., Desjardins G., “Bayesian estimates of measurement errors for In-line inspection and field tools”, IPC2002-27263

[6] Fenyvesi L, Miller S., “Determining corrosion growth”, Pipeline and Gas technology October 2005, page 22-30.

[7] Grzelak L. A., Achterbosch G. G. J., “Determination of the corrosion rate of a MIC influenced pipeline using 4 consecutive pigruns”, International Pipeline Conference, IPC06-10142

[8] Bijen, E.J., Cluster Analysis, Tilburg University Press 1973

[9] Späth, H., Cluster Analysis Algorithms for Data Reduction and Classification of Objects, Chichster, Horwood, 1980

[10] Hartigan, J.A., Clustering Algorithms, New Your, Wiley, 1985

[11] Clason, R. Finding Clusters: An Application of the Distance Concept, April 1990

[12] Edwards, A.L, An Introduction to Linear Regression and Correlation”, second edition, University of Washington, 1984

[13] McCullagh, P., Nelderd J.A. Generalized Linear Models, second edition, Department of Statistics, University of Chicago, Department of Mathematics, Imperial College of Science and Technology, London, 1989

[14] Miles, J., Shevlin M., Applying Regression & Correlation, SAGE Publications, 2001

[15] Muhlbauer, W.K., Pipeline Risk Management Manual- A Systematic Approach to Loss Prevention and Risk Assessment, Gulf Publishing Company, 1992

[16] Krysicki, W. Bartos, J., Dyczka W., Krolikowska, K., Wasilewski, M. Rachunek Prawdopodobienstwa I Statystyka Matematyczna w Zadaniach- 2, Wydawnictwo Naukowe PWN, Warszawa 2000

[17] Dziechciarz, J., Ekonometria, Wydawnictwo Akademii Ekonomicznej, Wroclaw 2003

[18] Varaiya, P.P., Lecture Notes on Optimization, Berkeley, California 1998 (online available)

[19] Han, S.P., A Globally Convergent Method For Nonlinear Programming", Journal of Optimization Theory and Applications, Vol. 22, p.297, 1977

[20] Powell, M.J.D., A Fast Algorithm for Nonlinear Constrained Optimization Calculus", Numerical Analysis, ed. G.A. Watson, Lecture Notes in Mathematics, Vol. 630, Springer Verlag, 1978

[21] Cegielski, A. "Programowanie nieliniowe" (lecture notes), 2004/2005,

[22] Lewandowski D. „Gas pipelines corrosion data analysis and related topics", Delft, The Netherlands, 2002

[23] Joseph L. Pikas, Case Histories of External Microbiologically Influenced Corrosion Underneath Disbonded Coatings, The NACE International Annual Conference and Exposition 1996, paper no. 198

[24] Harris, Dr. J. O., Soil Becteria and Corrosion, Appalachian Underground Short Course (1961)

Appendix A

A: Analysis methods and interpretation

Definition 1 (Random Variable)
Let F be a $\sigma -$algebra and Ω the probability space.
A function $X:(\Omega,F)\to R$ is a *random variable* if for every subset $A_k=\{\omega : X(\omega)\le k\}$ where $k\in R$, the condition $A_k\in F$ satisfies.

Definition 2 (The Likelihood Function)
Let $X=(X_1,\ldots,X_n)$ be a *random vector* (random variable on n components) and $\{f_X(x\,|\,\theta):\theta\in\Theta\}$ a statistical model parameterized by $\theta=(\theta_1,\ldots,\theta_k)$, the parameter vector on the parameter space Θ.
The likelihood function is a map $L:\Theta\to[0,1]\in R$ given $L(\theta\,|\,X)=f_X(x\,|\,\theta)$. The likelihood function is functionally the same in form as a probability density function (the emphasis is changed from *X* to the θ).

Definition 3 (A maximal Likelihood estimate)
The parameter $\hat{\theta}$ for such $L(\hat{\theta}\,|\,X)\ge L(\theta\,|\,X)\quad\forall\theta\in\Theta$ is called a *maximal likelihood estimate* (MLE) of θ.

Remark 1

> Many of density functions are smooth functions (exponential), hence it is very comfortable to transform them to the *log-likelihood function* (any strictly monotonic transformation preserves function’s extremes).

Definition 4 (The Normal distribution)

We say that X is normally distributed random variable with mean μ and standard deviation σ if its distribution function is following:

$$F(t;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

The density curve is symmetrical, centered about its mean, with its spread determined by its standard deviation.

- ***Maximal Likelihood estimation of parameters***

  Suppose that $X = (X_1,\ldots,X_n)$ is random vector and $X_1,\ldots,X_n$ are i.i.d. (independently and identically distributed), normally distributed random variables with the expectation μ and variance σ^2. In order to find estimators of unknown parameter $\theta = (\mu,\sigma^2)$ we apply the maximal likelihood estimate method.

$$L(\theta \mid X) = f_X(X \mid \theta) \overset{iid}{=} \prod_{i=1}^{n} f_{X_i}(X_i \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sigma^n (2\pi)^{n/2}}\right) e^{-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}}$$

The log-likelihood function is:

$$l(\theta \mid X) = \log(L(\theta \mid X)) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi)$$

In order to find extremes we compute gradient $l(\theta \mid X)\,|_{\theta}^{'}$ what gives:

$$\frac{\partial l(\theta \mid X)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) \text{ and } \frac{\partial l(\theta \mid X)}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{n}{2\sigma^2}$$

Setting $l(\theta \mid X)\,|_{\theta}^{'} = 0$ (first order condition) we get:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \text{ hence we get that:}$$

$$\hat{\theta} = (\hat{\mu},\hat{\sigma}^2) = (\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2)$$

Finally, to know whether $\hat{\theta}$ is indeed the MLE we need to check that second order derivatives are negative.

Estimated parameter $\hat{\theta} = (\hat{\mu},\hat{\sigma}^2)$ is indeed the MLE estimator.

Definition 5 (The Beta distribution)

We say that X is *standard beta distributed* a random variable with parameters α and β if its distribution function is following:

$$F(t;\alpha,\beta)=\int_0^t \frac{(1-x)^{\beta-1}x^{\alpha-1}}{B(\alpha,\beta)}dx$$

where:

Beta function: $B(\alpha,\beta)=\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and where $\Gamma(\alpha)=\int_0^\infty e^{-x}x^{\alpha-1}dx$ is a gamma function.

- ***Maximal Likelihood estimation of parameters***
  Using the same procedure as before we can easily derive MLE estimator $\hat{\theta}=(\hat{\alpha},\hat{\beta})$ for the standard Beta distribution, the results are shown below:

  $$\hat{\alpha}=\overline{X}\left(\frac{\overline{X}(1-\overline{X})}{1/n\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}-1\right) \text{ and } \hat{\beta}=\left(1-\overline{X}\right)\left(\frac{\overline{X}(1-\overline{X})}{1/n\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}-1\right)$$

  where $\overline{X}$ -customary means mean.

Definition 6 (The Gamma distribution)

We say that *X* is *gamma distributed* a random variable with parameters α and β if their distributions function is following:

$$F(t;\alpha,\beta)=\int_0^t \frac{\alpha^{\beta}}{\Gamma(\beta)}x^{\beta-1}e^{-\alpha x}dx$$

where $\Gamma(\alpha)$ is a gamma function introduced in previous definition.

- ***Maximal Likelihood estimation of parameters***
  Also in a case of a Gamma distribution we can easily derive MLE estimator of unknown parameter $\hat{\theta}=(\hat{\alpha},\hat{\beta})$, the results is following:

  $$\hat{\alpha}=\frac{\overline{X}}{1/n\sum_{i=1}^{n}(X_i-\overline{X})^2} \text{ and } \hat{\beta}=\left(\frac{\overline{X}}{1/n\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^2$$

Definition 7 (Expected value)

If *X* is a random variable defined on a probability space (Ω,F,P) then the expected value of *X* (denoted as *EX*) is defined in following way:

$$EX=\int_\Omega XdP$$

where integral is in the meaning of Lebesgue.

In case when random variable X admits a probability density function $f(x)$ then the expected value is:

$$EX=\int_{-\infty}^{+\infty}xf(x)dx$$

When *X* is a discrete random variable with values $x_1,\dots x_n$ and corresponding probabilities $p_1,\dots,p_n$ then:

$$EX = \sum_{i=1}^{n} x_i p_i$$

Definition 8 (P-value)

The *p-value* is the probability that sample could have been drawn from the population(s) being tested given the assumption that the null hypothesis is true. A p-value of 0.2, for example, indicates that we would have a 20% chance of drawing the sample being tested if the null hypothesis was actually true.

Definition 9 (Kolmogorov- Smirnov – two samples test)

The *Kolmogorov-Smirnov two-sample test* is a test of the null hypothesis that two independent samples have been drawn from the same population (or from populations with the same distribution). The test uses the maximal difference between cumulative frequency distributions of two samples as the test statistic. The main idea behind the test is to compare the proportion of the values less than certain level *x* between two sample sets. The test checks what the maximal difference between proportions is. The test doesn't require that samples are the same size. According to Kolmogorov- Smirnov the test is reasonably accurate for sample sizes n_1 and n_2 when $\frac{n_1 n_2}{n_1 + n_2} \geq 4$.

The procedure is following: for random vectors X_1 and X_2 with respectively number of samples n_1 and n_2

$$\forall x \;\; D = \max\{| F_{X_1}(x) - F_{X_2}(x) |\}$$

The usual way of carrying out the two-sample test is to compute the p-value directly from the test statistic, with no need to compare it to a critical value. This is an example presented in "Nonparametric Statistical Methods" by Hollander & Wolfe. So, the idea is to Compute the asymptotic *p-value* approximation and accept or reject the null hypothesis on the basis of the p-value. The direct formula for a two sided test p-value is expressed in the following way:

$$p - value = 2\sum_{k=1}^{\infty}(-1)^{k-1} e^{-2\lambda^2 k^2}$$ where $$\lambda = \max((\sqrt{n} + 0.12 + 0.11/\sqrt{n})D, 0)$$ and

$$n = \frac{n_1 n_2}{(n_1 + n_2)}$$

Definition 10 (Kolmogorov- Smirnov - goodness of fit test)

The *Kolmogorov- Smirnov test (K-S test)* can be used to decide whether a sample comes from a population with specified distribution. An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the K-S test has several important limitations:

1. It only applies to continuous distributions.

2. It tends to be more sensitive near the center of the distribution than at the tails.

The K-S test procedure of testing is following:

1. Specify distribution F_0 - associated with theoretical distribution function
2. Order samples in non-decreasing manner
3. Calculate $D = \max\left(\max_{1\le i\le n}\left(\frac{i}{n} - F_0(X_i)\right), \max_{1\le i\le n}\left(F_0(X_i) - \frac{i-1}{n}\right)\right)$
4. Calculate critical and p-value
   a) for small samples (less than 20), we use the direct values from tables
   b) for sample size larger than 20 Miller's formula is applied: $C = \sqrt{-0.5\ln(\alpha)/n}$
   c) the direct formula for p-value is given by the same formula as in **definition no. 9**

Definition 11.1 (Product moment correlation)

The correlation $\rho_{X,Y}$ between two random variables *X* and *Y* with expected values μ_X and μ_Y respectively, and standard deviations σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}.$$

The correlation is the measure which takes values from [-1,1] and is associated with the strength of the relationship between two variables. If correlation coefficient ρ takes value 1- then it means that there is a perfect linear relationship, in case when $\rho = -1$ the perfect linear relationship is negative, but when $\rho = 0$ - then there is no relationship between variables.

- ***Test of Pearson's correlation***

  Let suppose that we have already computed a correlation coefficient ρ. Now, we would like to verify the hypothesis that:

  $H_0 : \rho = 0$ against $H_1 : \rho <> 0$

  First, we need to calculate the probability of obtaining a statistic as different from or more different from the parameter specified in the null hypothesis as the statistic obtained in the experiment. The probability value is computed assuming the null hypothesis is true. If the probability value is below the significance level then the null hypothesis is rejected. To get a p-value we have to calculate:

  $t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$ where *n* is number of samples, now the analyst has to check In *t-Table* the probability value for a t, and to compare to significance level α. If the p-value is less than significance level, then the correlation is significant.

Definition 11.2 (Spearman's rho correlation coefficient)

General idea of rho correlation coefficient can be expressed in following way: instead of quantitative measures on each of *n* pairs of variables, we assign ranks a_i on the first variable (population characteristic) and a set of rankings b_i on the second one. Each of sets $\{a_i,...,a_n\}$ and $\{b_i,...,b_n\}$ is some perturbation of the integers *1,2,…,n*. The

Spearman correlation coefficient can be expressed as correlation between ranks instead of observations *X* and *Y* in following way:

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y - Y_i)^2}} \quad \Longrightarrow \quad \rho = \frac{\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2 \sum_{i=1}^{n}(b - \overline{b})^2}}$$

A basic algebra calculus will show that above formula can be reduced to:

$\rho = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^{\,2}$ where $d_i = a_i - b_i$

Definition 11.3 (Kendall's tau correlation coefficient)

Let's define variables: *"P"* and *"-Q"*, which corresponds to number of positive scores (concordant), and negatives (discordant) respectively. A linear relationship of variables is defined in following way:

$\tau = (P+Q)/S$

where *S* is maximal available positive score.

If we consider two rankings of exactly *n* components, then basic calculus gives that, number of possible pairs is *n(n-1)/2*. Hence Kendall's τ coefficient has the following form:

$\tau = \frac{S}{C_n^2} = \frac{2S}{n(n-1)}$ but $P + Q = C_n^2$ hence $\tau = \frac{4P}{n(n-1)} - 1$

Definition 12 (Quantiles)

Let's take random variable *X*. A *p* quantile is such an *q* that $P(X \le q_p) = p$

Definition 13 (correlation ratio)

Correlation ratio in statistics is a measure of the relationship between the statistical variability within individual categories and the dispersion of whole population or sample. The aim is to order the variables $X_1, \ldots, X_n$ (included in the model) according to influence on the criterion variable. The quantity of interest is $E(Var(\hat{Y} \mid X_i))$[43]. Since the equation:

$$E(Var(Y \mid X_i)) + Var(E(Y \mid X_i)) = Var(Y)$$

is well known if follows that $E(Var(Y \mid X_i)) \to 0$ indicates higher importance of the variable X_i. As a consequence the correlation ratio for the variable i'th is defined as:

$$CR_i = \frac{Var(E(Y \mid X_i))}{Var(Y)}$$

[43] The quantity that should be considered is $Var(\hat{Y} \mid X_i = x_i^*)$ (change of predictor variable if one quantity is said to be constant), however since x_i^* is unknown, the idea is to calculate change of the variance overall the values of X_i

Intuitively: higher value of the CR follows higher the share of the variance decomposition of the variable X_i.

Method 1 Linear Regression

The important task in statistics is concerned with determining functional relationships between a given set of variables. Linear least squares error technique is very important and applicable modeling method. Basically, the method gives the estimators for predefined function in order to get a function, which is best, fitted to the data in the meaning of least squares errors.

Standard assumptions of the multiple regressions are:

1. The model is defined as $Y = X\beta + \varepsilon$, where *Y* is the vector of outputs and *X* is a matrix of covariates
2. The number of observations is grater than number of parameters to be estimated
3. $E\varepsilon_i = 0 \quad Var(\varepsilon_i) = \sigma^2 \quad Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$

The idea to get the best estimator of β we need to solve optimization problem: $\min_{\beta} \| Y - X\beta \|^2 = \min_{\beta} \| \varepsilon \|^2$.

Simple calculations show that the optimal estimator of β is expressed in following way:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- ***Significance of estimated coefficients***

  Important fact in linear regression is that, we are allowed to check whether estimated parameter vector $\hat{\beta}$ can be neglected or not. To test it we need to verify the null hypothesis:

  $H_0 : \hat{\beta} = 0$ against alternative: $H_1 : \exists \beta_i$ *such that* $\beta_i \neq 0$

  To verify hypothesis above we need to calculate p-value for given significance level α. The p-value associated with introduced hypothesis is:

  $$p - value = \frac{\hat{\beta}^T X^T Y - \overline{Y}^2}{\left(Y^T Y - \hat{\beta} X^T Y\right)\left(p-1\right)/\left(n-p\right)} \sim F(p-1, n-p)$$

  For introduced p-value we have basis to reject null hypothesis when $p - value > F(p-1, n-1, 1-\alpha)$ otherwise we do not have such a basis. (The *n* is a number of samples; *p* is a number of columns in covariate matrix *X*; and F is *F- distribution*[44]).
- ***Calculation of confidence interval estimates for individual coefficients***

  The $(1-\alpha)$ confidence interval for estimated parameter $\hat{\beta}_i$ is given by:

  $$\left[\hat{\beta}_i - t(n-p, 1-\frac{\alpha}{2})\sqrt{(X^T X)^{-1}_{i,i} s^2}, \hat{\beta}_i + t(n-p, 1-\frac{\alpha}{2})\sqrt{(X^T X)^{-1}_{i,i} s^2} \right]$$

  where $s^2 = \left(1/\left(n-p\right)\right)\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$; $\hat{Y}_i = X_i \hat{\beta}_i$; where $(X^T X)^{-1}_{i,i}$ is i'th value from diagonal of matrix $(X^T X)^{-1}$; and where *t*- is *t-student*[45] distribution function

[44] F-distribution- Fisher-Snedecor Distribution

[45] t-student distribution- Gosset's distribution

- **Goodness of fit measure**

  To check whether created model is well fitted to data we introduce fit measure:

  $$R^2 = \frac{(Y-\overline{Y})^T(\hat{Y}-\overline{Y})}{(Y-\overline{Y})^T(Y-\overline{Y})(\hat{Y}-\overline{Y})^T(\hat{Y}-\overline{Y})}$$ which takes values from [0,1]:

  1 if the model is perfect, and
  0 when model is badly fitted

Method 2 (stepwise regression)

Stepwise regression is a model-building procedure that attempts to maximize the amount of variance possible to explain in dependent variable while simultaneously minimizing the number of independent variables in the statistical model. The stepwise regression is typically used when a large number of predictor variables are available while the best combination of variables to predict the value of the criterion is wanted. It is designed to give a model that predicts as much variability as possible with the smallest number of parameters. The stepwise regression should be interpreted cautiously or avoided entirely when trying to understand theoretical relation. It makes its selection based purely on the amount of variance that variables can explain without any consideration of causal or logical priority. As the consequence independent variables chosen through a stepwise regression are not guaranteed to be the most important factors affecting the criterion variable. A theoretically meaningful variable that explains a larger amount of variability in the criterion variable could be excluded if it also happens to cause changes in other independent variables, because it would be collinear with those variables. Additionally, stepwise regression attempts to maximize the predictive ability for the predictor variables in the one specific sample that was collected. Its selections will therefore be affected by any relations that happen to appear due to chance alone. If it is impossible to come up with a theoretical explanation for an observed relation between predictor and criterion variables it may just be an artifact only found the particular collected sample. There is one circumstance under which stepwise regression should be used at most: when the most important aim is to determine the best predictive model, without interesting in drawing inferences about the relations in the data.

Stepwise regression consists of the two steps: first start with a simple model and gradually add independent variables to it until any significant improvement is not made i.e. minimize probability that all the factors are equal to zero, but drop variables which become no longer "significant" after introduction of new variables. In other words check the “old” set of independent variables each time a new one is added to the model to make sure that they are still significant. Secondly if it turns out that a predictor variable included in an earlier step is no longer making a significant contribution to the prediction of the dependent variable, then the variable is dropped from the model.

The algorithm of the stepwise regression can be presented in the following way. Suppose that and one criterion variable Y. Notation used in the procedure is following:

Notation

-n is number of independent variables

-Y is criterion variable, ex. corrosion rate [mm/yr] or defect rate [def/km]

- X_i indicates the i’th variable where $i = 1,...,n$

- $R^2_{1,...n}$ - R square coefficient for the model: $Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$

Step wise algorithm

Step 1

- $m = 1$
- define n models of the form: $Y = \beta_0 + \beta_1 X_i + \varepsilon$

Step 2

- take the variable (model) k for which: $\max_k R_k^2$ and p-value for the hypothesis: $H_0 : \beta_k = 0$ against $H_1 : B_k \neq 0$ is less than significance level α

Step 3

-for the remaining n-m variables define n-m models of the form: $Y = \beta_0 + \beta_1 X_k + \beta_2 X_l + \varepsilon$, where $l = 1,..k-1, k+1,..n$, and calculate $R_{k,l}^2$ and check the value for difference between R squares: $R = R_{k,j}^2 - R_k^2$

Step 4

-define a new regression model for which R takes maximum value and is statistically significant[46]

Step 5

-**recalculate the p-values for the t-test for all the variables in the model** and check if all are significant, if any of them is insignificant then should be removed and step 3 should be applied one more time to the model consisting only of significant variable.

Steps 3,4 and 5 have to be repeated sequentially, the new model adding/dropping the variables has to be finished when adding or dropping the variables wont improve the determination coefficient (R square) significantly and all the variables in final model will be significant.

[46] In order to check if difference between R squares for two models is significant hypothesis testing based on F statistics has to be applied- see appendix

Appendix B

B: Tutorial file for CoroGas

Author: Lech A. Grzelak, program created for N. V. Netherlandse Gasunie

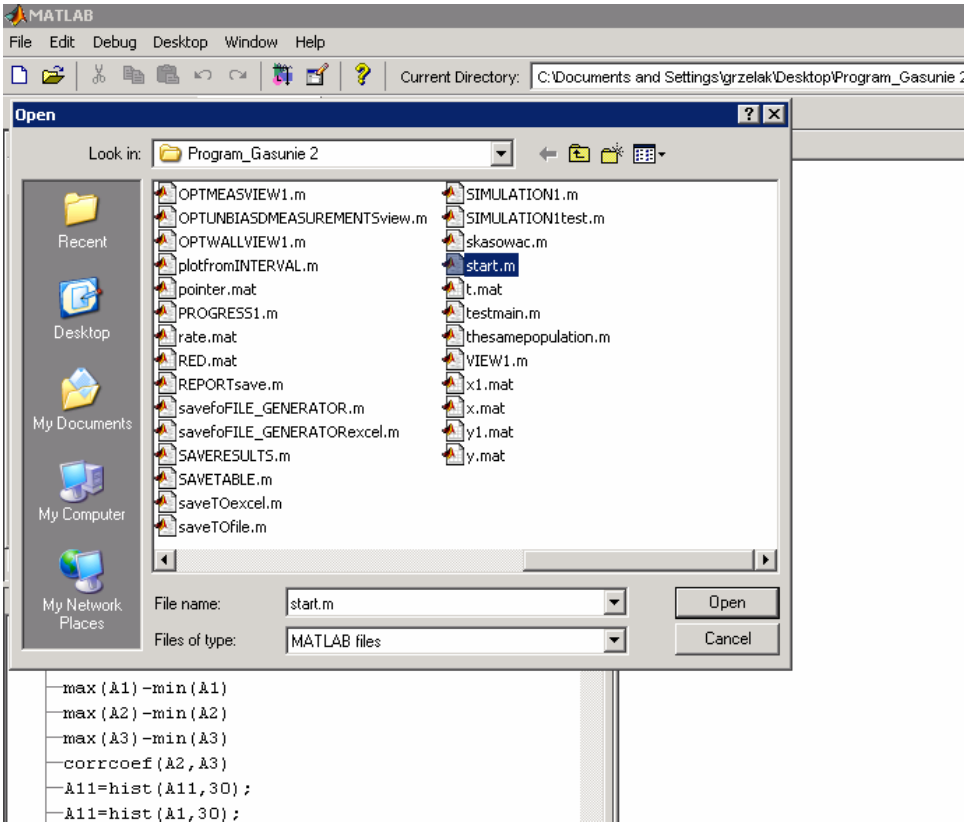
The program does all the theory introduced in part one of the thesis. This tutorial consists of following parts:

a) Data calibration
b) Corrosion rate modeling
c) How to transfer the files into ***.dat** type?
d) Files construction
e) Examples
f) Available datasets descriptions

a) Data Calibration

The first step of using the program "A statistical approach to determine the corrosion rate modeling of the underground gas pipelines" is to open the program in the Matlab environment.

In order to initiate the program the user has to open the matlab file: ***file/open/start.m*** as it is shown on the screenshot below.

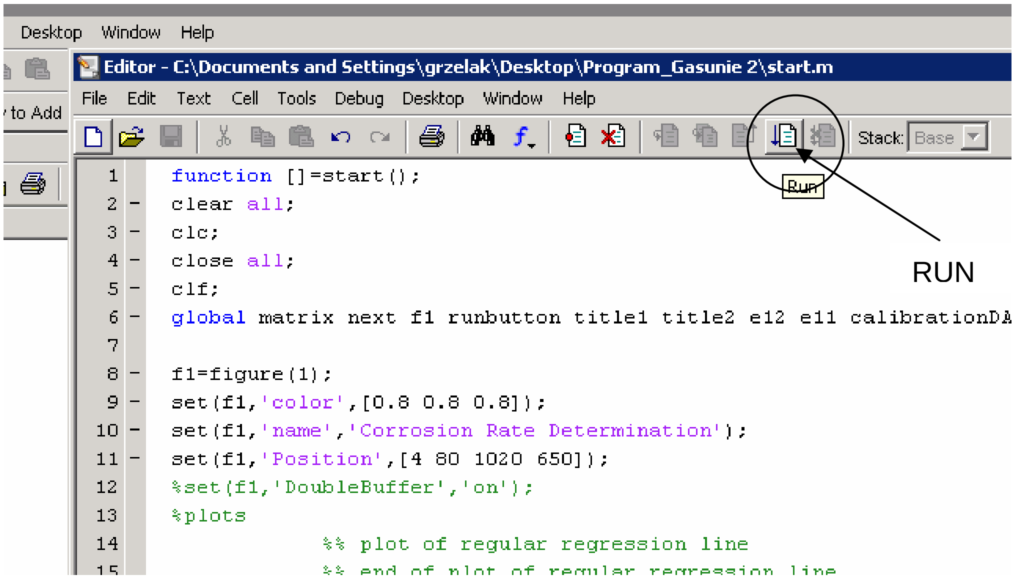

MATLAB
File Edit Debug Desktop Window Help
Current Directory: C:\Documents and Settings\grzelak\Desktop\Program_Gasunie 2
Open
Look in: Program_Gasunie 2
Recent
Desktop
My Documents
My Computer
My Network Places
Open
Cancel

OPTMEASVIEW1.m
OPTUNBIASDMEASUREMENTSview.m
OPTWALLVIEW1.m
plotfromINTERVAL.m
pointer.mat
PROGRESS1.m
rate.mat
RED.mat
REPORTsave.m
savefoFILE_GENERATOR.m
savefoFILE_GENERATORexcel.m
SAVERESULTS.m
SAVETABLE.m
saveTOexcel.m
saveTOfile.m
SIMULATION1.m
SIMULATION1test.m
skasowac.m
start.m
t.mat
testmain.m
thesamepopulation.m
VIEW1.m
x1.mat
x.mat
y1.mat
y.mat

File name: start.m
Files of type: MATLAB files

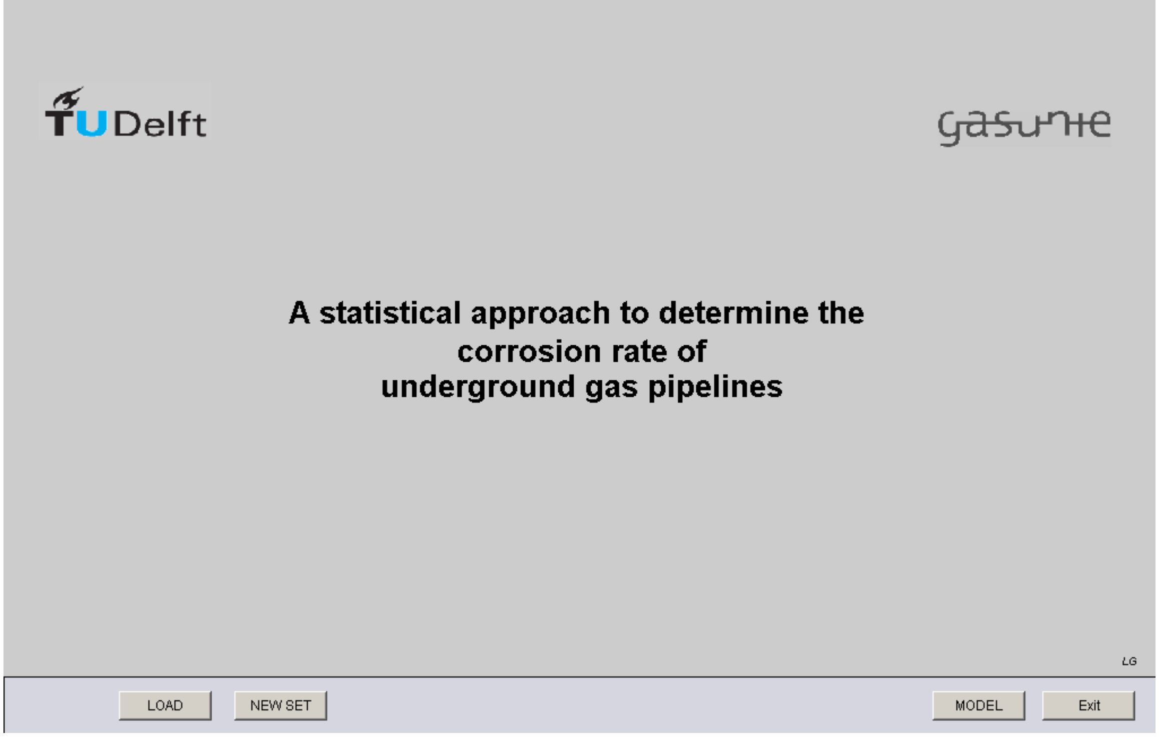
```
max(A1)-min(A1)
max(A2)-min(A2)
max(A3)-min(A3)
corrcoef(A2,A3)
A11=hist(A11,30);
A11=hist(A1,30);
```

When the file ***start.m*** is opened then the initiation of the program can be done in two ways.
The first way is easy - press the run button on the top of the new opened window as it is shown below.


Desktop Window Help
to Add
Editor - C:\Documents and Settings\grzelak\Desktop\Program_Gasunie 2\start.m
File Edit Text Cell Tools Debug Desktop Window Help
Stack: Base
Run
RUN

```
function []=start();
clear all;
clc;
close all;
clf;
global matrix next f1 runbutton title1 title2 e12 e11 calibrationDA

f1=figure(1);
set(f1,'color',[0.8 0.8 0.8]);
set(f1,'name','Corrosion Rate Determination');
set(f1,'Position',[4 80 1020 650]);
%set(f1,'DoubleBuffer','on');
%plots
            %% plot of regular regression line
            %% end of plot of regular regression line
```

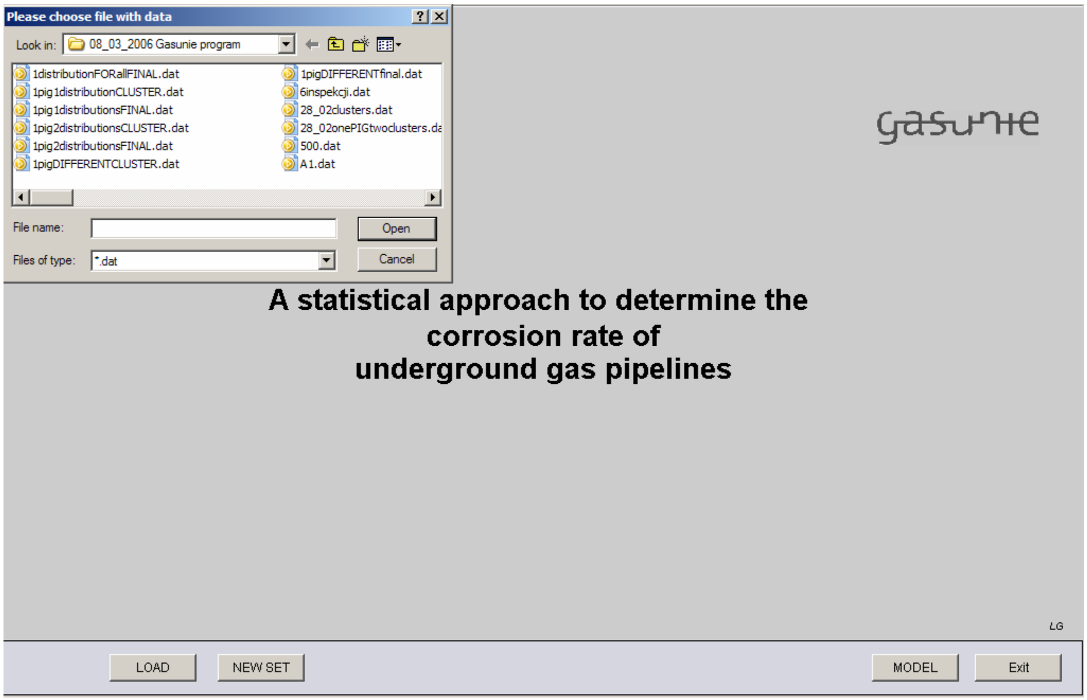
The second way to start the program is to tape the command ***"start.m"*** in the command window.
If the initiation process is successful then the user should see the graphical user interface window.



A statistical approach to determine the
corrosion rate of
underground gas pipelines

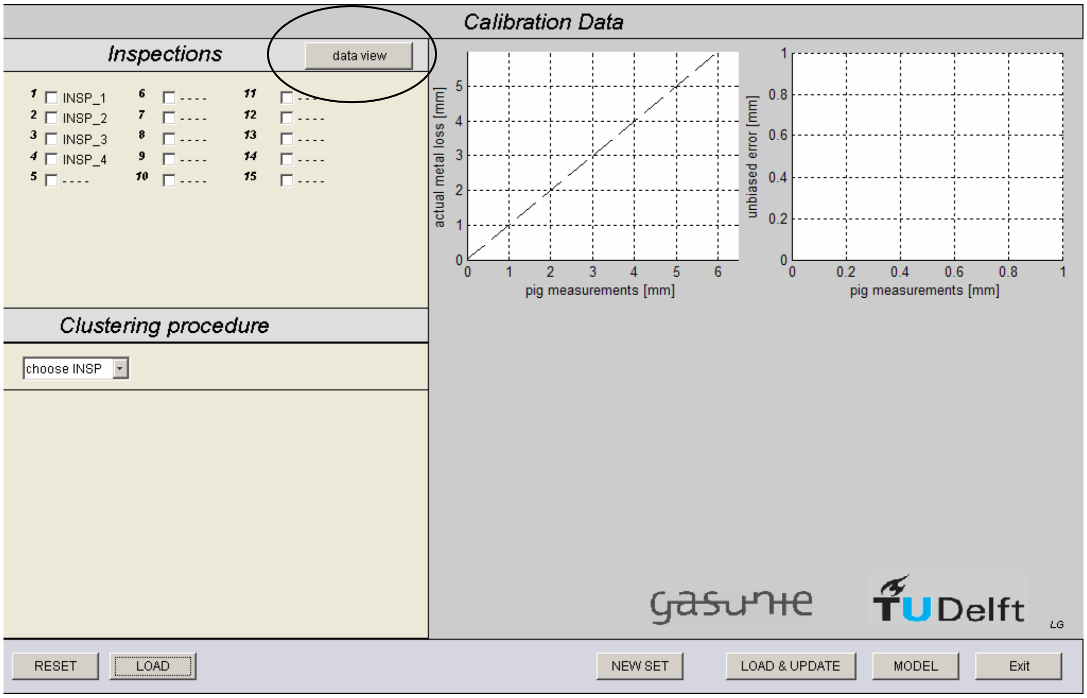
On the bottom of the screen are presented available options. The first one is “***LOAD***” which allows loading previously prepared excavation dataset, and is supposed to consists of calibration data. The detailed description of the file will be presented later on in the text. The Second option “***NEW SET***” gives the opportunity to create user’s calibration dataset (*option under construction*), the third is “MODEL” which goes directly to “corrosion rate modeling” part without calibrating the data, the forth and the last option is “***EXIT***” which terminates the program

Suppose that user wants to ***LOAD*** previously prepared dataset. First a proper path to the calibration data has to be indicated in the load window. The program recognizes the ***“*.dat”*** file (later in text it is shown how to create this type of files).


Please choose file with data
Look in: 08_03_2006 Gasunie program
1distributionFORallFINAL.dat
1pig1distributionCLUSTER.dat
1pig1distributionsFINAL.dat
1pig2distributionsCLUSTER.dat
1pig2distributionsFINAL.dat
1pigDIFFERENTCLUSTER.dat
1pigDIFFERENTfinal.dat
6inspekcji.dat
28_02clusters.dat
28_02onePIGtwoclusters.da
500.dat
A1.dat
File name:
Open
Files of type: *.dat
Cancel
gasunie
LG
LOAD
NEW SET
MODEL
Exit

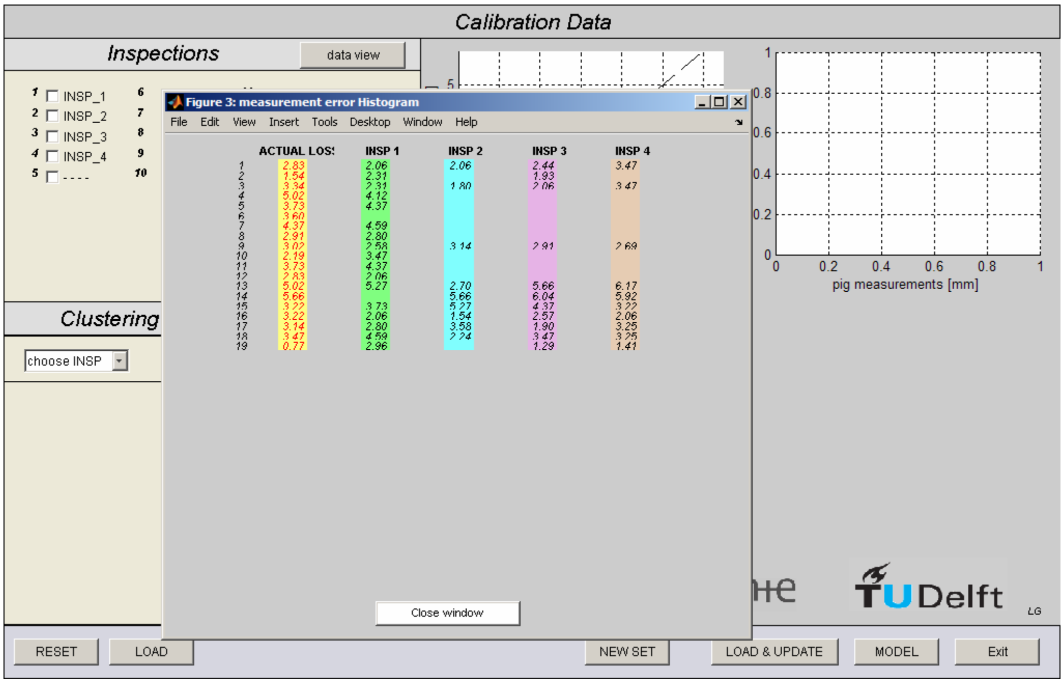
A statistical approach to determine the corrosion rate of underground gas pipelines

When the proper calibration dataset is chosen then the program immediately analyzes the dataset and proceeds to the second stage.



The new window consists of a few parts - the block on the left called “***INSPECTIONS***”, “***CLUSTERING PROCEDURE***”, “**figures of results**”, and “***TOOLBOX***” on the bottom of the window.

- “***INSPECTIONS***”- allows user to have a view on the loaded dataset and also gives the opportunity to look at the graphical representation of the dataset (in the figures on the right). A click on the button “***data view***”, initiates data view.

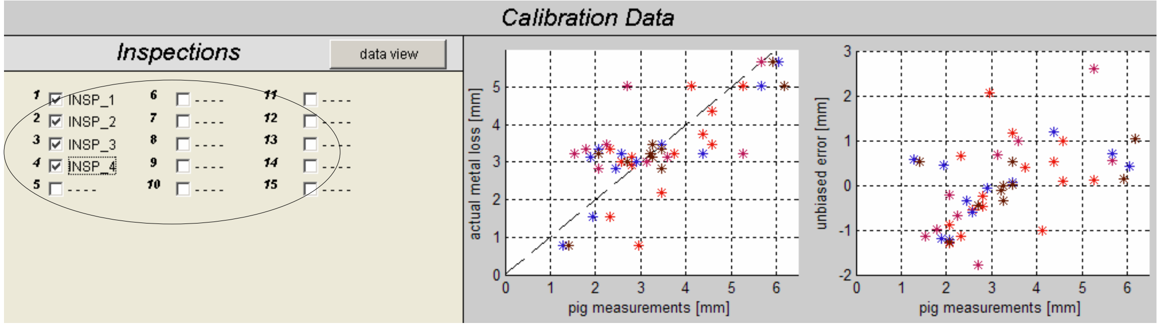


	ACTUAL LOS!	INSP 1	INSP 2	INSP 3	INSP 4
1	2.83	2.06	2.06	2.44	3.47
2	1.54	2.31		1.93	
3	3.34	2.31	1.80	2.06	3.47
4	5.02	4.12			
5	3.73	4.37			
6	3.60				
7	4.37	4.59			
8	2.91	2.80			
9	3.02	2.58	3.14	2.91	2.69
10	2.19	3.47			
11	3.73	4.37			
12	2.83	2.06			
13	5.02	5.27	2.70	5.66	6.17
14	5.66		5.66	6.04	5.92
15	3.22	3.73	5.27	4.37	3.22
16	3.22	2.06	1.54	2.57	2.06
17	3.14	2.80	3.58	1.90	3.25
18	3.47	4.59	2.24	3.47	3.25
19	0.77	2.96		1.29	1.41

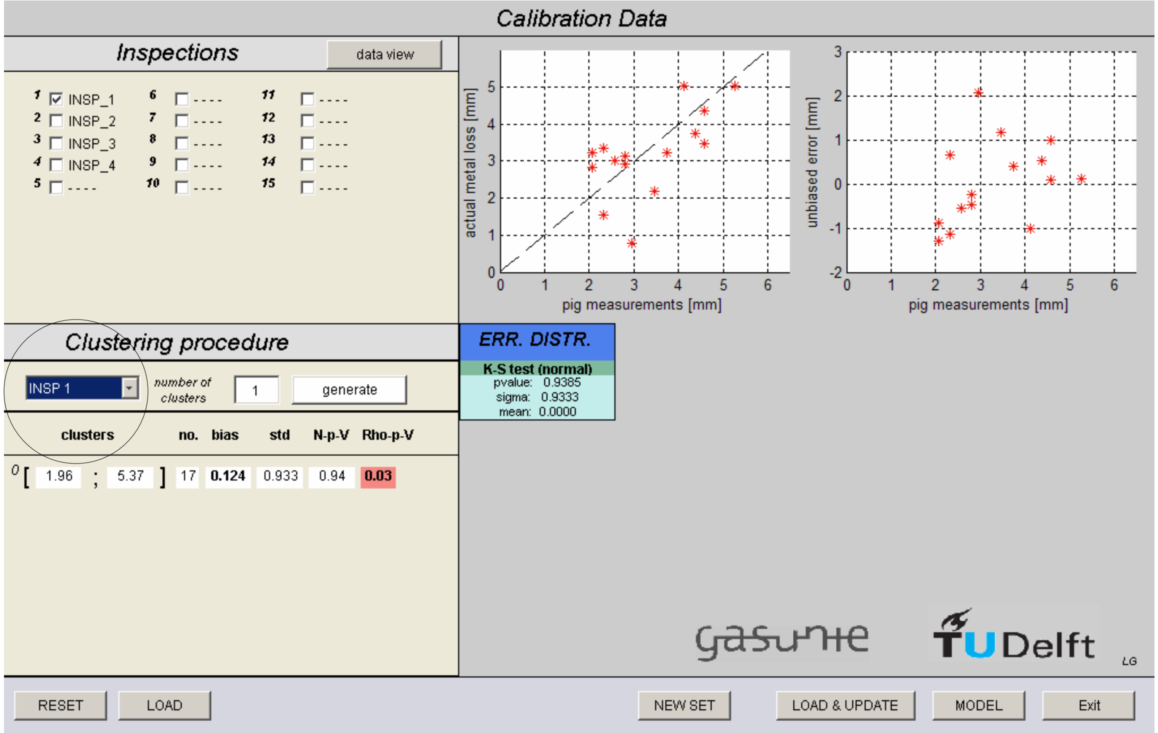
The new window presents dataset loaded for calibration. The first column corresponds to the defects measured at the excavation by an inspector. The other columns are associated with the measurements reported by "intelligent pigs". The empty slots indicate missing data caused either by defect reparation (in current case the defects were repaired after tuboscope inspection) or by clustering procedure. All the measurements are presented in millimeters.

The second available option in this part is to show the calibration graphically.

When user chooses any of the check-boxes corresponding to pigs-inspectors, then the program updates two graphics. The first one shows relation between the "pig measurements" and the excavation data, and the second relation between the pig measurements and the measurement errors (i.e. all data are treated as exactly one cluster).



- ***"CLUSTERING PROCEDURE"***- This part gives an opportunity to calibrate dataset according to a given data (clusters). Firstly user has to use the so called popup-menu and choose a pig for calibration.



1	☑ INSP_1	6	□ ----	11	□ ----
2	□ INSP_2	7	□ ----	12	□ ----
3	□ INSP_3	8	□ ----	13	□ ----
4	□ INSP_4	9	□ ----	14	□ ----
5	□ ----	10	□ ----	15	□ ----

INSP 1 ▾ number of clusters 1

clusters	no.	bias	std	N-p-V	Rho-p-V
0 [1.96 ; 5.37]	17	0.124	0.933	0.94	0.03

ERR. DISTR.

K-S test (normal)
pvalue: 0.9385
sigma: 0.9333
mean: 0.0000

When a certain pig is chosen, then immediately the following results appear:

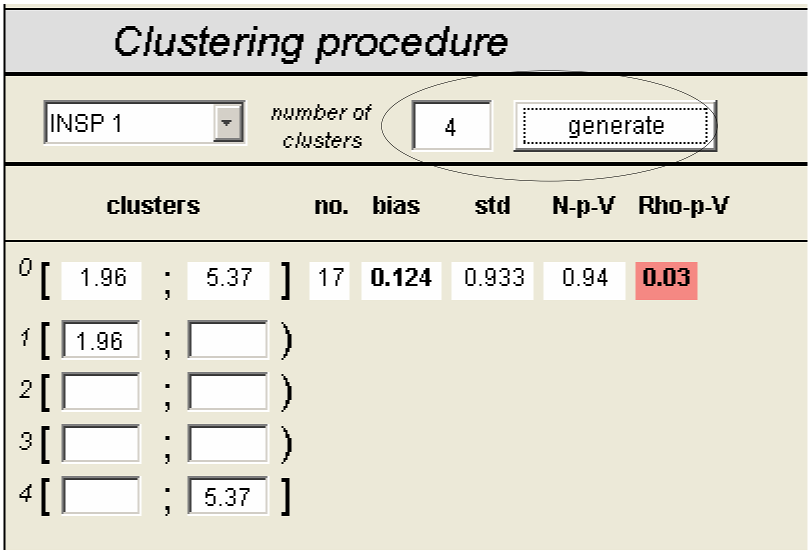
- ***“Number of clusters”***- number of predefined clusters
- ***“Clusters”***- bounds of the interval where pig measurements were reported (interval’s boundaries are extended by 0.1 [mm])
- ***“No.”***- number of defects registered during excavation and associated with pig measurements (within the cluster)
- ***“Bias”***- corresponds to the mean of the difference between the pig measurements and measurements from the excavation (see the “A statistical approach to determine the corrosion rate of underground gas pipelines” report for details)
- ***“Std”***- standard deviation of the measurement error
- ***“N-p-V”*** – is the p-value associated with the following hypothesis:

  H_0 :*Measurement errors are from normal distribution*

  H_1 : *Measurement errors are not from normal distribution*
- ***“Rho-p-V”***- is the p-value associated with the following hypothesis:

  H_0 : *The Spearman’s correlation between measurement errors and pig measurement is 0*

  H_1 : *The Spearman’s correlation between measurement errors and pig measurement is not 0*
- ***“ERR. DISTR”-*** gives the p-value, standard deviation and mean for the measurement error. In the case where the user does not generate additional clusters then the p-value corresponding to Kolmogorov Smirnov test should be the same as given in the column ***“N-p-V”***



Clustering procedure

INSP 1 ▾ number of clusters 4 generate

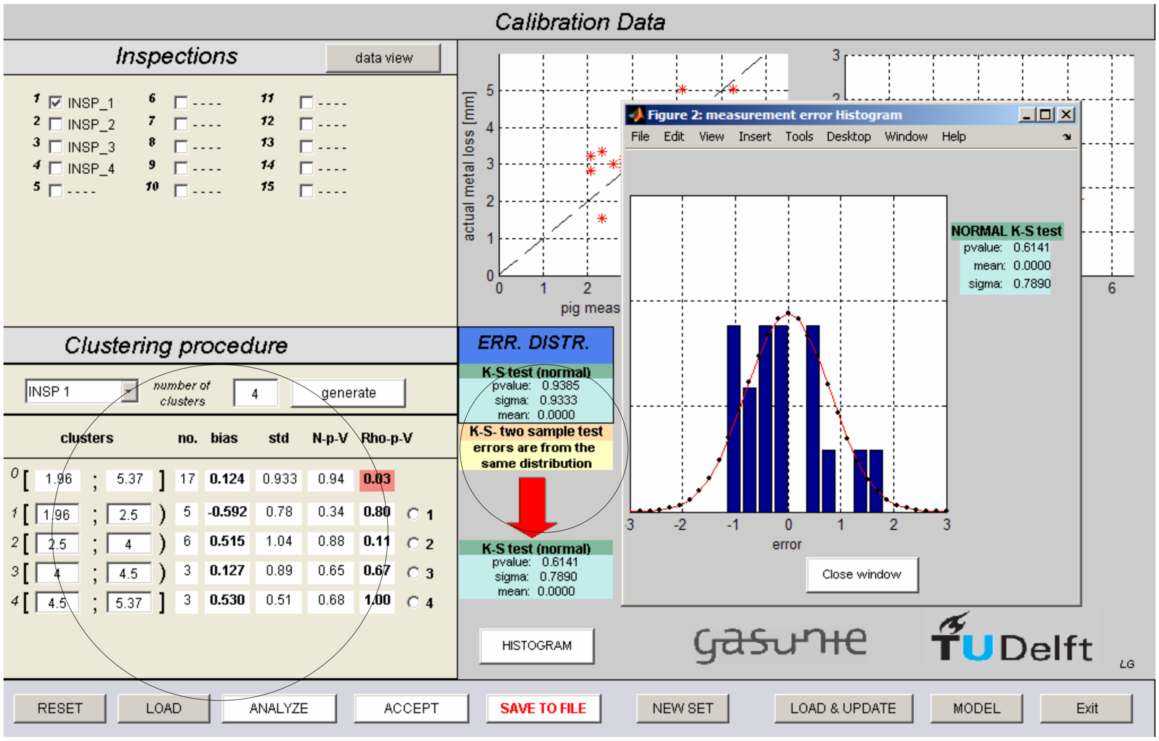
	clusters	no.	bias	std	N-p-V	Rho-p-V
0	[1.96 ; 5.37]	17	0.124	0.933	0.94	0.03
1	[1.96 ;)					
2	[;)					
3	[;)					
4	[; 5.37]					

When user defines the number of clusters ex. according to the techniques introduced in the report “A statistical approach to determine the corrosion rate of underground gas pipelines”[47] then, he is also obligated to define boundaries of these intervals. In order to generate these intervals (clusters) user has to fill in the ***“number of clusters”*** and press the button “***GENERATE***”.
The maximal number of intervals is 6 i.e. the program analyzes up to 6 clusters per one pig. When all the boundaries for clusters are well defined (cover the entire domain, and are not intersecting), then in order to perform the analysis the ***“ANALYZE”*** button has to be pressed. After few seconds a computer should give the answer in the form presented on the picture below. Each row and calculated values correspond to a predefined cluster.

If for the ***“Rho-p-V***” the background color becomes red- then user should have a look closer at the cluster. Red background indicates that p-value for the Spearman’s (rank) correlation hypothesis that measurement error is uncorrelated with the measurements is less than significance level 0.05.

Information available on the screen (yellow rectangle) is the verification of the null hypothesis that errors which come from clusters come from the same population (distribution). The “***HISTOGRAM***” produces the histogram of the unbiased errors (if the hypothesis about the errors coming from the same distribution is not rejected).

[47] Report done by Lech A. Grzelak, for Gasunie Research Department



Clustering procedure

INSP 1 | number of clusters 4 | generate

	clusters		no.	bias	std	N-p-V	Rho-p-V	
0	[1.96	; 5.37]	17	0.124	0.933	0.94	0.03	
1	[1.96	; 2.5)	5	-0.592	0.78	0.34	0.80	○ 1
2	[2.5	; 4)	6	0.515	1.04	0.88	0.11	○ 2
3	[4	; 4.5)	3	0.127	0.89	0.65	0.67	○ 3
4	[4.5	; 5.37]	3	0.530	0.51	0.68	1.00	○ 4

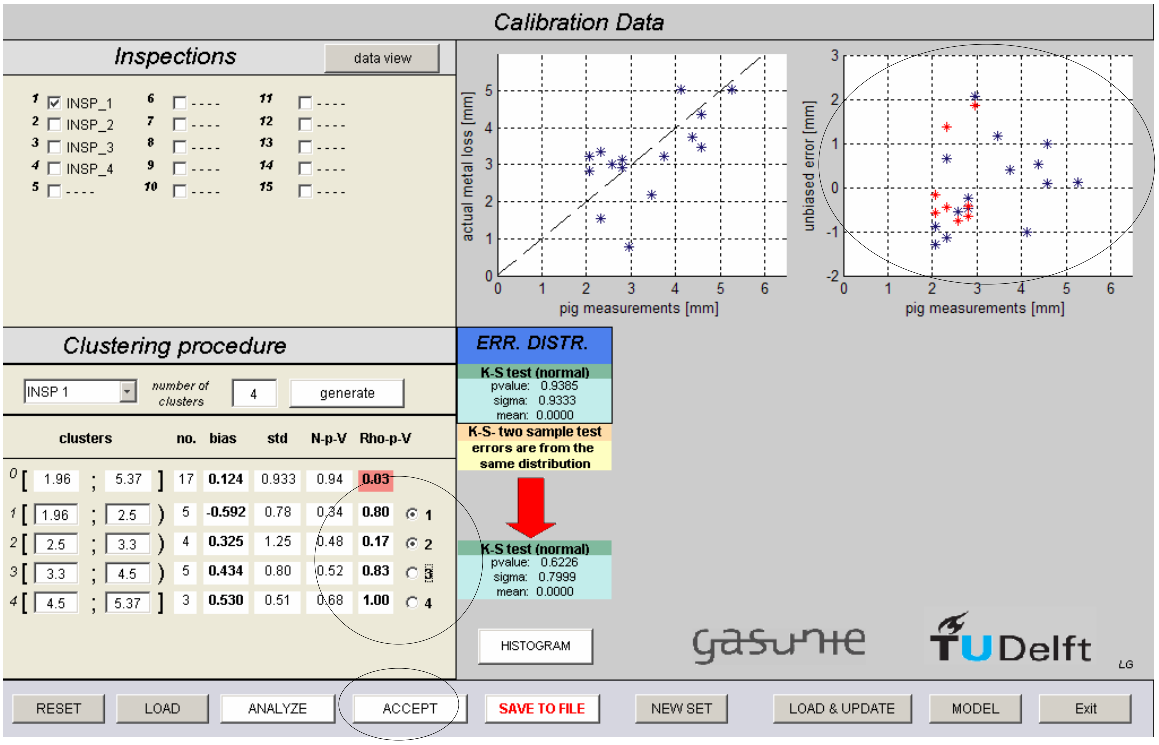
K-S test (normal)
pvalue: 0.9385
sigma: 0.9333
mean: 0.0000

K-S- two sample test
errors are from the same distribution

K-S test (normal)
pvalue: 0.6141
sigma: 0.7890
mean: 0.0000

NORMAL K-S test
pvalue: 0.6141
mean: 0.0000
sigma: 0.7890

Another option available in the clustering procedure is to have a view on the errors associated with the predefined clusters. As it is presented on the picture below; if user clicks on the check box associated with the analyzed cluster, then the error from the cluster is presented like on the picture on the right – the errors without imposed clusters are as a background.



Clustering procedure

INSP 1 | number of clusters 4 | generate

	clusters		no.	bias	std	N-p-V	Rho-p-V	
0	[1.96	; 5.37]	17	0.124	0.933	0.94	0.03	
1	[1.96	; 2.5)	5	-0.592	0.78	0.34	0.80	● 1
2	[2.5	; 3.3)	4	0.325	1.25	0.48	0.17	● 2
3	[3.3	; 4.5)	5	0.434	0.80	0.52	0.83	○ 3
4	[4.5	; 5.37]	3	0.530	0.51	0.68	1.00	○ 4

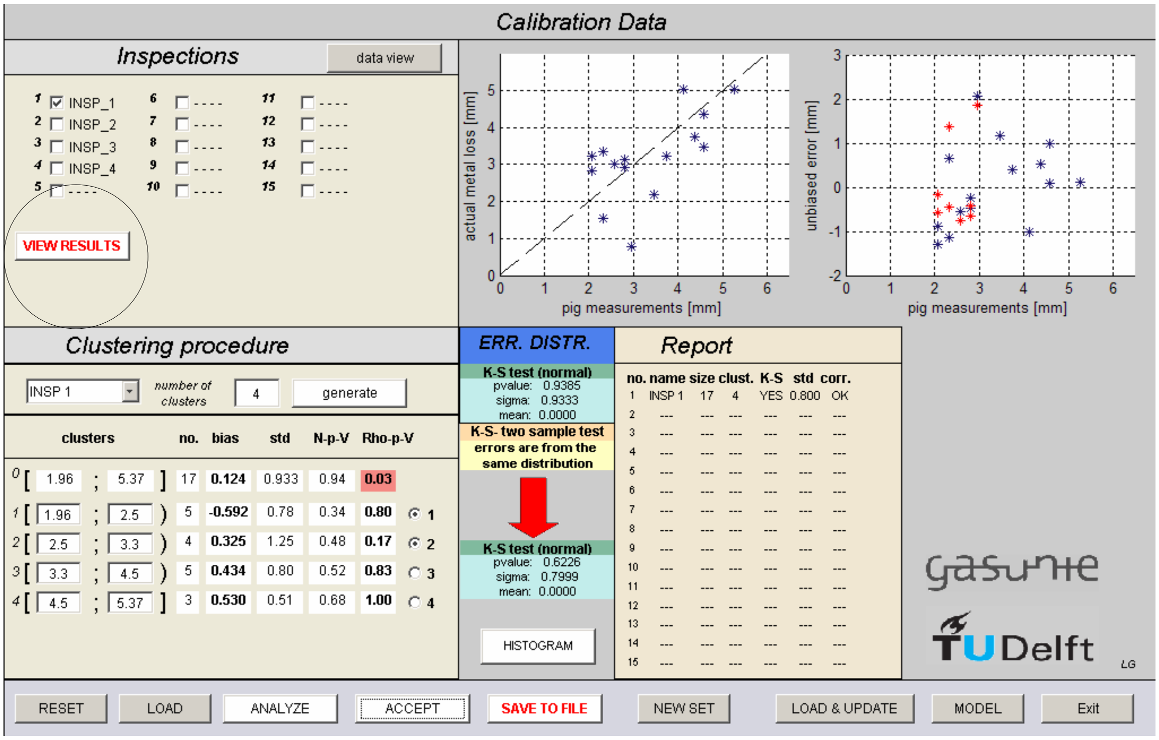
K-S test (normal)
pvalue: 0.9385
sigma: 0.9333
mean: 0.0000

K-S- two sample test
errors are from the same distribution

K-S test (normal)
pvalue: 0.6226
sigma: 0.7999
mean: 0.0000

When user is satisfied with the clusters, then such decision has to be indicated by pressing the button “***ACCEPT***” on the bottom of the window. The button “***ACCEPT***” saves chosen clusters into the memory and initiates the “***REPORT***” in the middle of the window and gives access to the button “***VIEW RESULTS”.*** The columns in the new opened window are:

- “***No***”- index of the inspection
- “***name***”- name of the inspecting pig
- “***clust***.”- number of predefined clusters
- “***K-S***”- takes values ***YES/NO***, which indicate whether measurement errors come from the same population according to Kolmogorov- Smirnov test
- “***std***”- measurement error standard deviation (if the “K-S” is “YES”)
- “***Corr***.”- takes values ***B/OK***- indicates whether measurement errors are correlated with pig measurements, the B indicates “***BAD***”- correlation and “***OK***”- no correlation. The correlation used in the software is the Spearman’s rank correlation



	clusters		no.	bias	std	N-p-V	Rho-p-V	
0	[1.96	; 5.37]	17	0.124	0.933	0.94	0.03	
1	[1.96	; 2.5)	5	-0.592	0.78	0.34	0.80	● 1
2	[2.5	; 3.3)	4	0.325	1.25	0.48	0.17	● 2
3	[3.3	; 4.5)	5	0.434	0.80	0.52	0.83	○ 3
4	[4.5	; 5.37]	3	0.530	0.51	0.68	1.00	○ 4

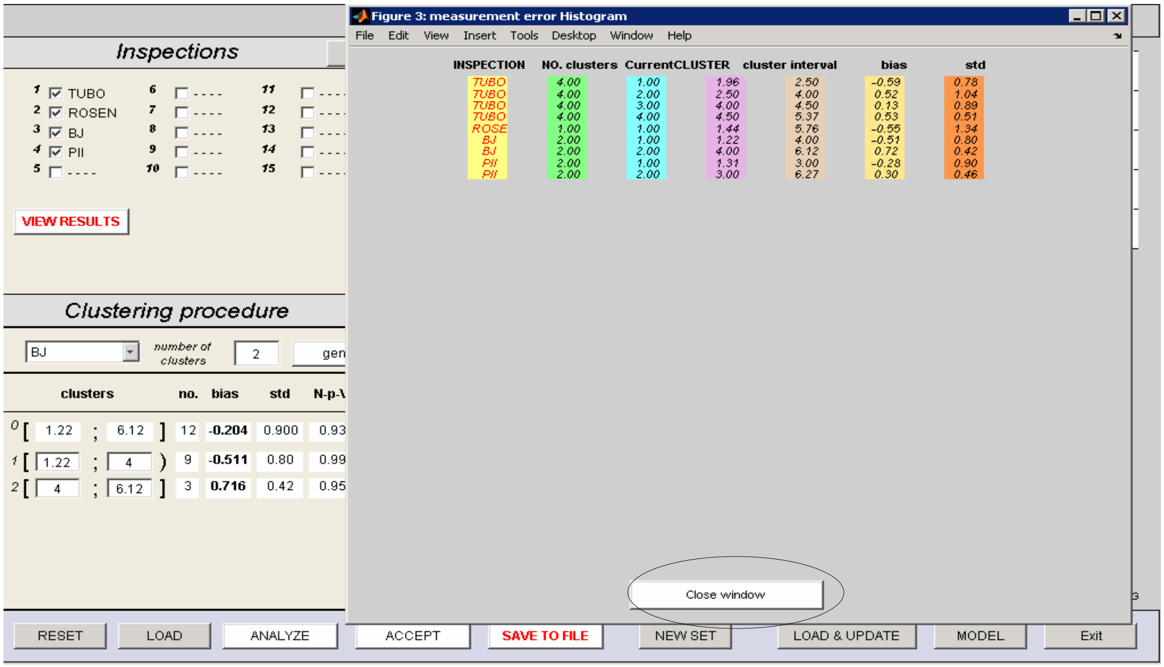
K-S test (normal)
pvalue: 0.9385
sigma: 0.9333
mean: 0.0000

K-S- two sample test
errors are from the same distribution

K-S test (normal)
pvalue: 0.6226
sigma: 0.7999
mean: 0.0000

no.	name	size	clust.	K-S	std	corr.
1	INSP 1	17	4	YES	0.800	OK
2	---	---	---	---	---	---
3	---	---	---	---	---	---
4	---	---	---	---	---	---
5	---	---	---	---	---	---
6	---	---	---	---	---	---
7	---	---	---	---	---	---
8	---	---	---	---	---	---
9	---	---	---	---	---	---
10	---	---	---	---	---	---
11	---	---	---	---	---	---
12	---	---	---	---	---	---
13	---	---	---	---	---	---
14	---	---	---	---	---	---
15	---	---	---	---	---	---

The button “***VIEW RESULTS***” allows to see how the accepted clustered are defined. The first column corresponds to the name of the intelligent pig, the second indicates number of predefined clusters. Column number three shows current cluster, and then respectively are: cluster’s boundaries, bias associated with this cluster and the standard deviation of the errors. This information is collected and later on will be applied in the second stage of the analysis done by the program - namely in the corrosion rate modeling. When the clustering procedure is performed for all the pigs, then in order to save the results user has to press the “***SAVE TO FILE***” button.



INSPECTION	NO. clusters	CurrentCLUSTER	cluster interval		bias	std
TUBO	4.00	1.00	1.96	2.50	-0.59	0.78
TUBO	4.00	2.00	2.50	4.00	0.52	1.04
TUBO	4.00	3.00	4.00	4.50	0.13	0.89
TUBO	4.00	4.00	4.50	5.37	0.53	0.51
ROSE	1.00	1.00	1.44	5.76	-0.55	1.34
BJ	2.00	1.00	1.22	4.00	-0.51	0.80
BJ	2.00	2.00	4.00	6.12	0.72	0.42
PII	2.00	1.00	1.31	3.00	-0.28	0.90
PII	2.00	2.00	3.00	6.27	0.30	0.46

	clusters	no.	bias	std	N-p-\
0	[1.22 ; 6.12]	12	**-0.204**	0.900	0.93
1	[1.22 ; 4)	9	**-0.511**	0.80	0.99
2	[4 ; 6.12]	3	**0.716**	0.42	0.95

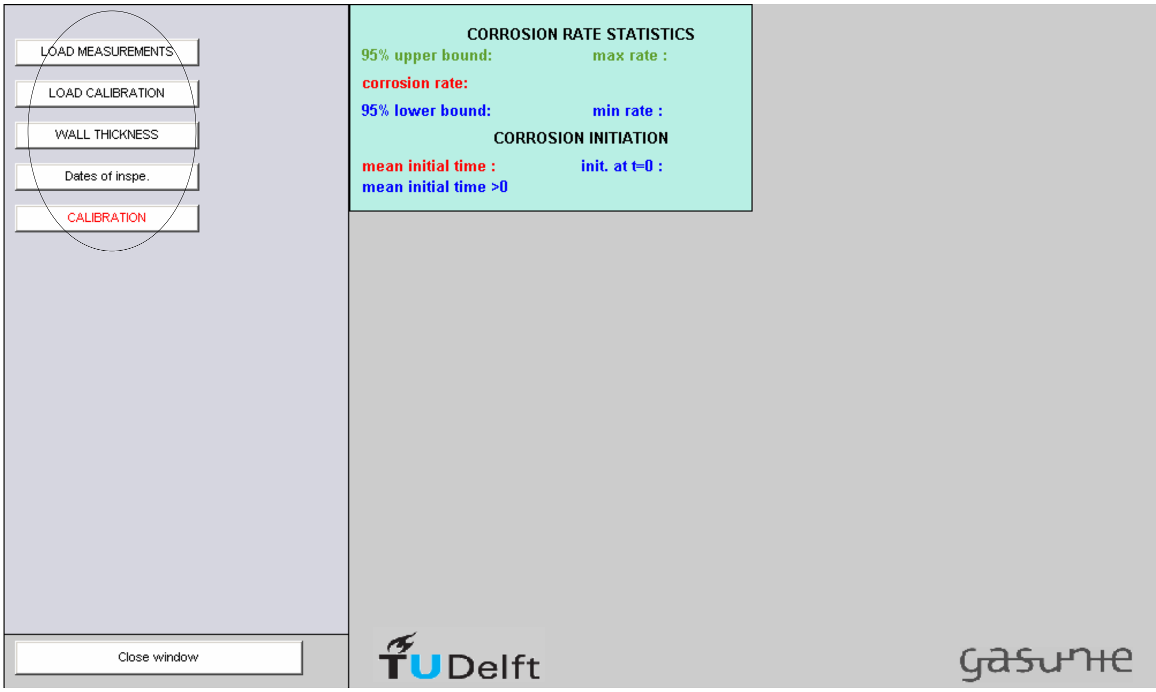
To leave the view window of the defined cluster press the button “***Close window***”.
It is also possible to load the clusters and calibration data from previous session and update them. The option which allows this is the button “***LOAD & UPDATE***”.

b) Corrosion Rate modeling

When the calibration procedure is carried out successfully, then the button ”***MODEL***” gives access to the second part of the program.

The options/buttons available at this stage are:

- “***LOAD MEASUREMENTS***”- user is required to give the path to the file with the measurements collected by intelligent pigs
- “***LOAD CALIBRATION***”- when after the calibration procedure the clusters are defined and saved to the file then user has to indicate according to which file the calibration of the measurements has to be performed
- “***WALL THICKNESS***”- loads the column vector of the nominal wall thicknesses, each row of the vector has to correspond to nominal wall thickness where defects were observed – at this stage program doesn't take into account the uncertainty about wall thickness
- “***DATES OF INSPECTIONS***”- according to the data each inspection was performed at certain time, this option loads the dates of performed inspections- the rows of this vector correspond to the columns of the “***measurements***”
- “***CALIBRATION***”- when all required data are loaded then this button performs the calibration procedure and gives access to the “***SIMULATION***”



CORROSION RATE STATISTICS

95% upper bound: max rate :

corrosion rate:

95% lower bound: min rate :

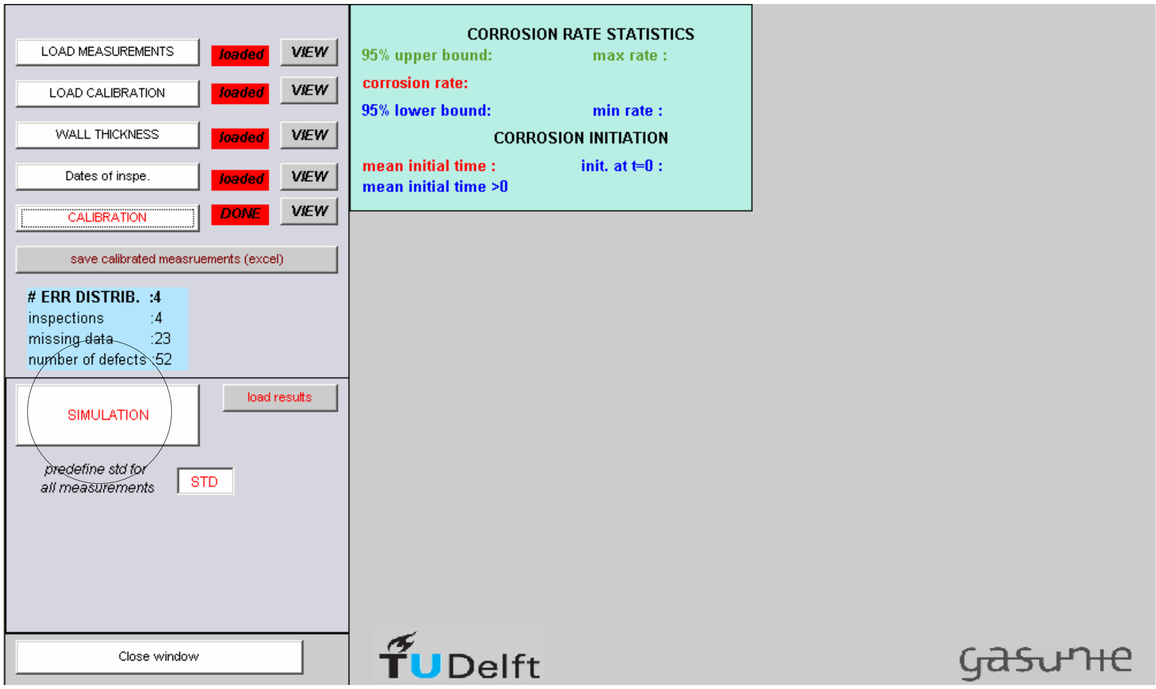
CORROSION INITIATION

mean initial time : init. at t=0 :

mean initial time >0

When the data is loaded then by pressing the button “***VIEW***” user can have a look on the loaded datasets. The button “***CALIBRATION***” calibrates the data and gives some results of the calibration

- ***“# ERR DISTRIB***”- the number of the distributions (sum of distributions associated with all clusters for all pigruns)
- “***INSPECTIONS***”- number of inspections
- “***MISSING DATA***”- number of missing data in the whole measurement dataset
- “***NUMBER OF DEFECTS***”- number of defects observed during inspections
- “***PREDEFINE STD FOR ALL MEASUREMENTS***”- when user types any positive value in the “***STD***” box, then the weights for the inspections are equal (all pigs have the same weight)
- “***LOAD RESULTS***”- gives the possibility to load previously saved Corrosion modeling session



CORROSION RATE STATISTICS

95% upper bound: max rate :

corrosion rate:

95% lower bound: min rate :

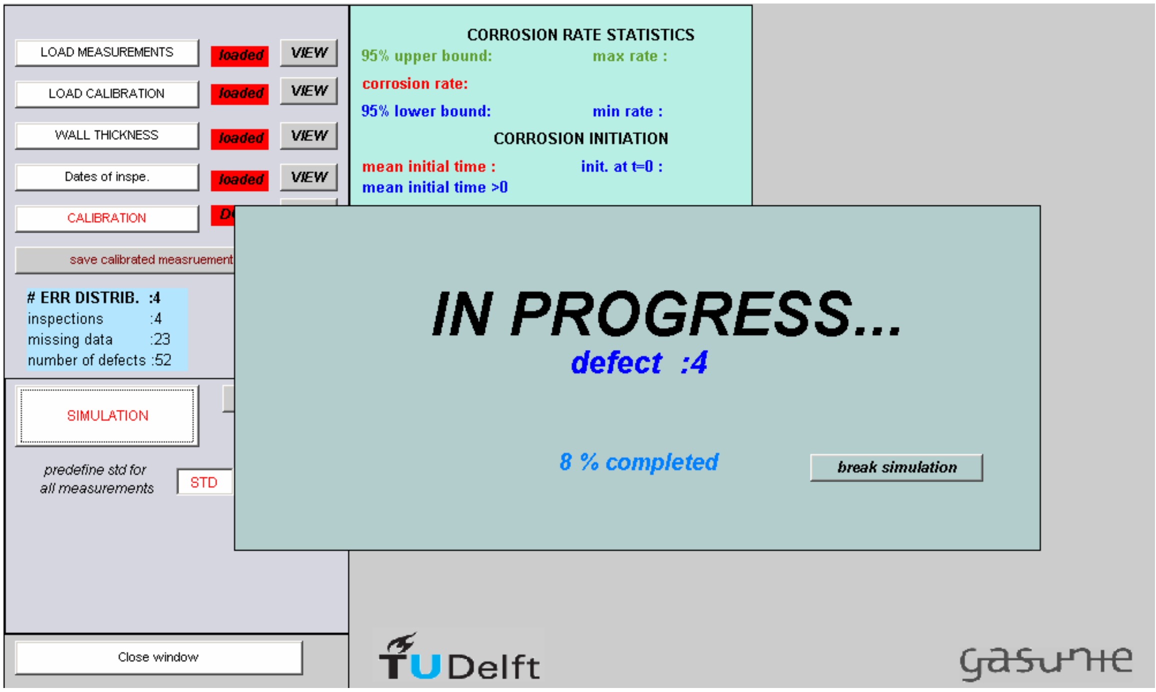
CORROSION INITIATION

mean initial time : init. at t=0 :

mean initial time >0

ERR DISTRIB. :4
inspections :4
missing data :23
number of defects :52

When the datasets are loaded, then the button "***SIMULATION***" initiates the simulation. The simulation produces the optimal solution. Tests show that estimation for one single defect observed at four inspections requires about 15 seconds to obtain the optimal function.



CORROSION RATE STATISTICS

95% upper bound: max rate :

corrosion rate:

95% lower bound: min rate :

CORROSION INITIATION

mean initial time : init. at t=0 :

mean initial time >0

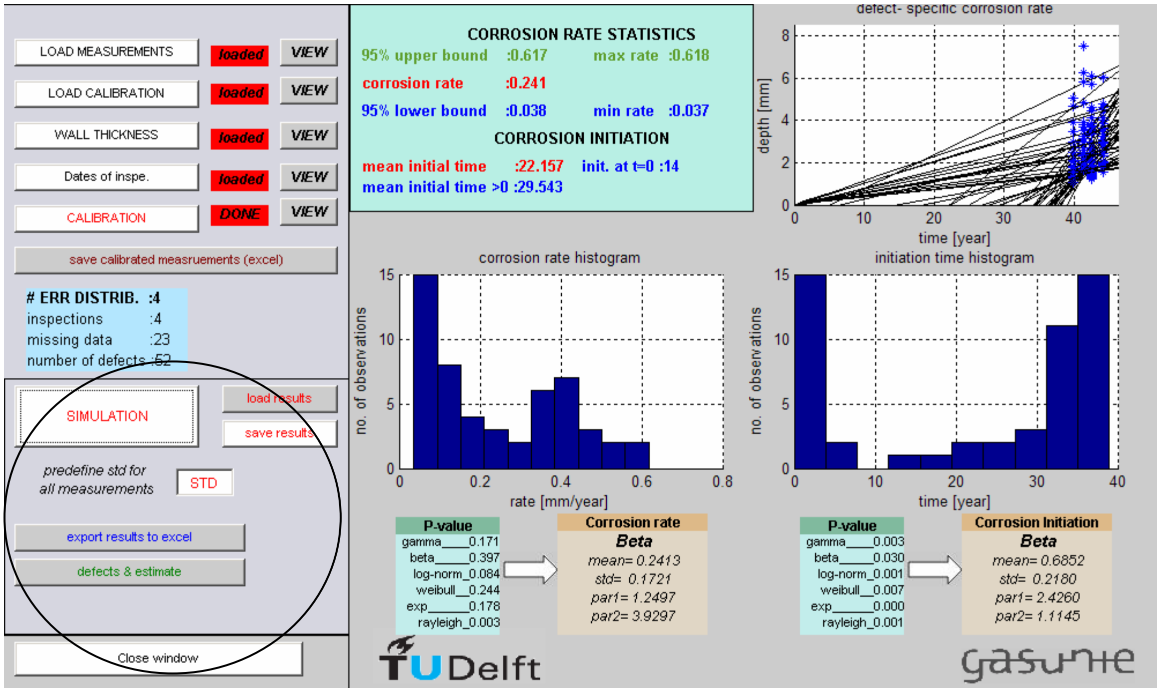
ERR DISTRIB. :4
inspections :4
missing data :23
number of defects :52

IN PROGRESS...

defect :4

8 % completed

When the simulation is complete, after a few seconds a new screen should appear.



ERR DISTRIB. :4
inspections :4
missing data :23
number of defects :52

CORROSION RATE STATISTICS

95% upper bound :0.617 max rate :0.618
corrosion rate :0.241
95% lower bound :0.038 min rate :0.037

CORROSION INITIATION

mean initial time :22.157 init. at t=0 :14
mean initial time >0 :29.543

P-value
gamma____0.171
beta_____0.397
log-norm_0.084
weibull__0.244
exp______0.178
rayleigh_0.003

Corrosion rate
Beta
mean= 0.2413
std= 0.1721
par1= 1.2497
par2= 3.9297

P-value
gamma____0.003
beta_____0.030
log-norm_0.001
weibull__0.007
exp______0.000
rayleigh_0.001

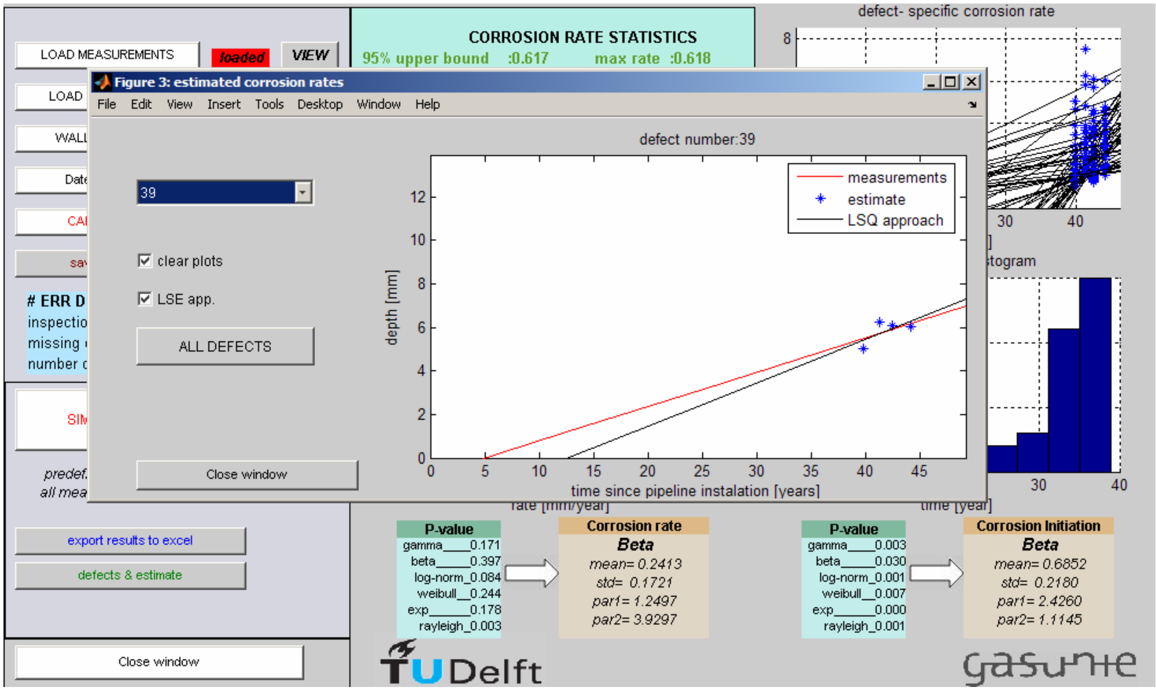
Corrosion Initiation
Beta
mean= 0.6852
std= 0.2180
par1= 2.4260
par2= 1.1145

On the top of the screen is so called "***CORROSION RATE STATISTICS***", which consists of basic statistic measures like

- **corrosion rate** (mean corrosion rate from all rates),
- **95% upper bound of the corrosion rate**
- **95% lower bound of the corrosion rate**
- **Max corrosion rate**
- **Min corrosion rate**
- **mean initial time of corrosion initiation**
- **"init. At t=0"**- indicates how many defects start at time 0 (at pipeline installation time)
- **"mean initial time>0**"- mean initial time of defects initiating after pipeline installation

On the right side, the plot presents the metal loss as a linear function of time for all defects together with the measurements (blue stars). The middle part presents two histograms. First of them corresponds to the distribution of the corrosion rate, and the second one to the initiation time of corrosion per defect. There is a small summary of fitted distributions beneath the histograms -- on the left hand side all available distributions with associated p-values, and on the right distributions with estimated likelihood parameters for which the p-value is the highest.

After the simulation three new options become available. The first of them is "***SAVE RESULTS***" - which allows exporting the results to a ***.dat** file. The next is "**DEFECTS & ESTIMATE",** which allows to have a closer look at the estimated rates. This option allows to check what is the corrosion rate for each defect and to compare estimated curve with the Least Squares approach.



CORROSION RATE STATISTICS

95% upper bound :0.617 max rate :0.618

ERR D
inspectio
missing
number c

39

☑ clear plots

☑ LSE app.

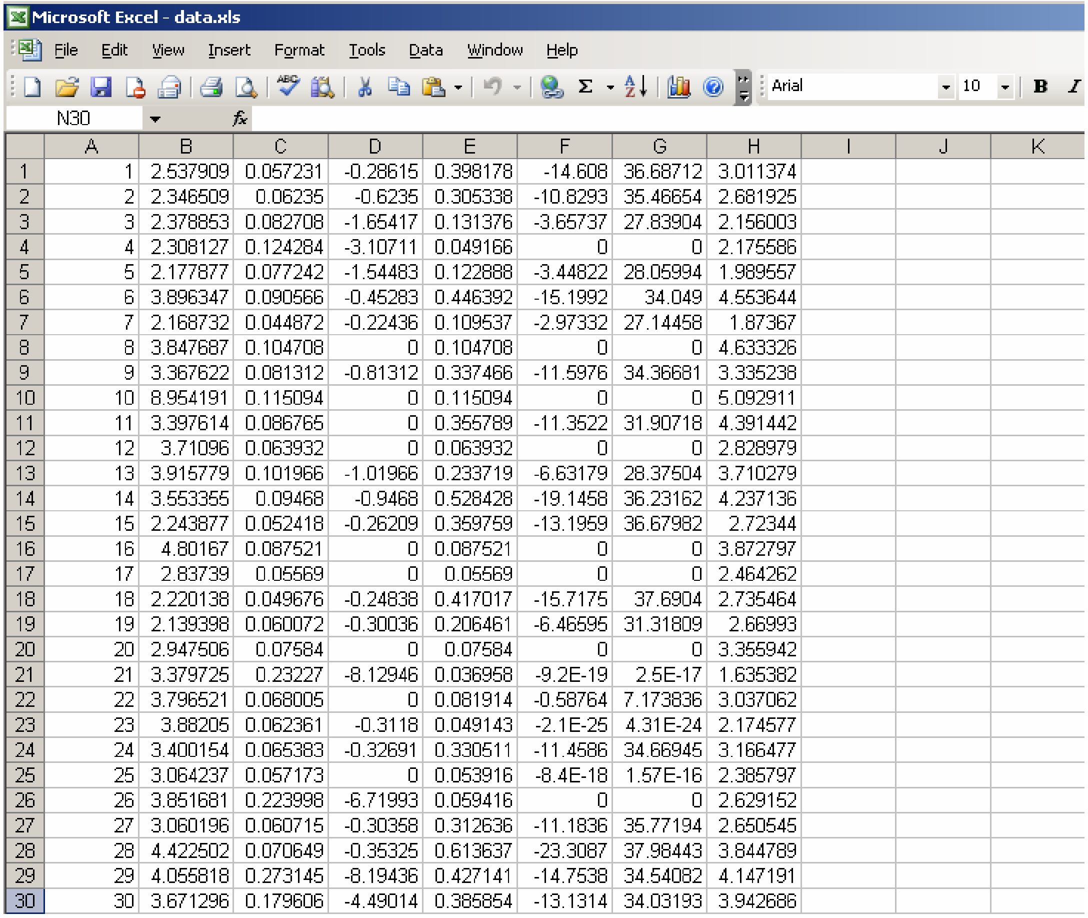
P-value
gamma____0.171
beta_____0.397
log-norm_0.084
weibull__0.244
exp______0.178
rayleigh_0.003

Corrosion rate
Beta
mean= 0.2413
std= 0.1721
par1= 1.2497
par2= 3.9297

P-value
gamma____0.003
beta_____0.030
log-norm_0.001
weibull__0.007
exp______0.000
rayleigh_0.001

Corrosion Initiation
Beta
mean= 0.6852
std= 0.2180
par1= 2.4260
par2= 1.1145

The third option is "***EXPORT RESULTS TO EXCEL***" - exports the results to an Excel file ***.xls**. If the user decides to export results to Excel, then the exported file consists of eight columns where each row corresponds to one defect.


Microsoft Excel - data.xls
File Edit View Insert Format Tools Data Window Help
Arial
10
B
I
N30
fx

	A	B	C	D	E	F	G	H	I	J	K
1	1	2.537909	0.057231	-0.28615	0.398178	-14.608	36.68712	3.011374			
2	2	2.346509	0.06235	-0.6235	0.305338	-10.8293	35.46654	2.681925			
3	3	2.378853	0.082708	-1.65417	0.131376	-3.65737	27.83904	2.156003			
4	4	2.308127	0.124284	-3.10711	0.049166	0	0	2.175586			
5	5	2.177877	0.077242	-1.54483	0.122888	-3.44822	28.05994	1.989557			
6	6	3.896347	0.090566	-0.45283	0.446392	-15.1992	34.049	4.553644			
7	7	2.168732	0.044872	-0.22436	0.109537	-2.97332	27.14458	1.87367			
8	8	3.847687	0.104708	0	0.104708	0	0	4.633326			
9	9	3.367622	0.081312	-0.81312	0.337466	-11.5976	34.36681	3.335238			
10	10	8.954191	0.115094	0	0.115094	0	0	5.092911			
11	11	3.397614	0.086765	0	0.355789	-11.3522	31.90718	4.391442			
12	12	3.71096	0.063932	0	0.063932	0	0	2.828979			
13	13	3.915779	0.101966	-1.01966	0.233719	-6.63179	28.37504	3.710279			
14	14	3.553355	0.09468	-0.9468	0.528428	-19.1458	36.23162	4.237136			
15	15	2.243877	0.052418	-0.26209	0.359759	-13.1959	36.67982	2.72344			
16	16	4.80167	0.087521	0	0.087521	0	0	3.872797			
17	17	2.83739	0.05569	0	0.05569	0	0	2.464262			
18	18	2.220138	0.049676	-0.24838	0.417017	-15.7175	37.6904	2.735464			
19	19	2.139398	0.060072	-0.30036	0.206461	-6.46595	31.31809	2.66993			
20	20	2.947506	0.07584	0	0.07584	0	0	3.355942			
21	21	3.379725	0.23227	-8.12946	0.036958	-9.2E-19	2.5E-17	1.635382			
22	22	3.796521	0.068005	0	0.081914	-0.58764	7.173836	3.037062			
23	23	3.88205	0.062361	-0.3118	0.049143	-2.1E-25	4.31E-24	2.174577			
24	24	3.400154	0.065383	-0.32691	0.330511	-11.4586	34.66945	3.166477			
25	25	3.064237	0.057173	0	0.053916	-8.4E-18	1.57E-16	2.385797			
26	26	3.851681	0.223998	-6.71993	0.059416	0	0	2.629152			
27	27	3.060196	0.060715	-0.30358	0.312636	-11.1836	35.77194	2.650545			
28	28	4.422502	0.070649	-0.35325	0.613637	-23.3087	37.98443	3.844789			
29	29	4.055818	0.273145	-8.19436	0.427141	-14.7538	34.54082	4.147191			
30	30	3.671296	0.179606	-4.49014	0.385854	-13.1314	34.03193	3.942686			

The first column in the exported Excel file is the number of analyzed defects. The second and third column correspond to the initial guess given to the optimizer, the fourth column gives the optimal value of the function “$-\log L$” where the L is the Likelihood function presented in the report. Column indicated as E corresponds to the corrosion rate of estimated defect, the next is the intercept of the linear function. Column G gives the time when the corrosion process has initiated. The last column corresponds to the depth of the defect at the last inspection.

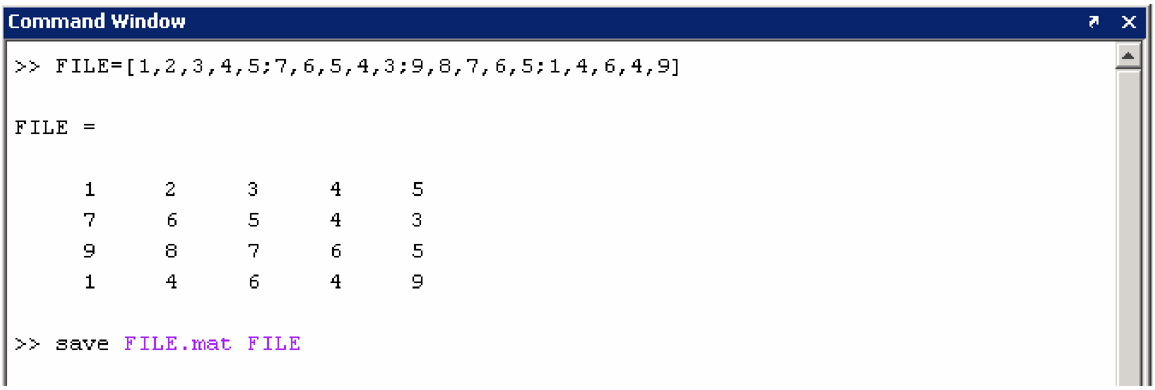
c) How to transform a data to the *.dat file type?

As it was presented previously in the tutorial, the type of files which the program works with is the type with extension ***.dat**. In order to make such a file, user has to apply ***file converter***.

The procedure is following:

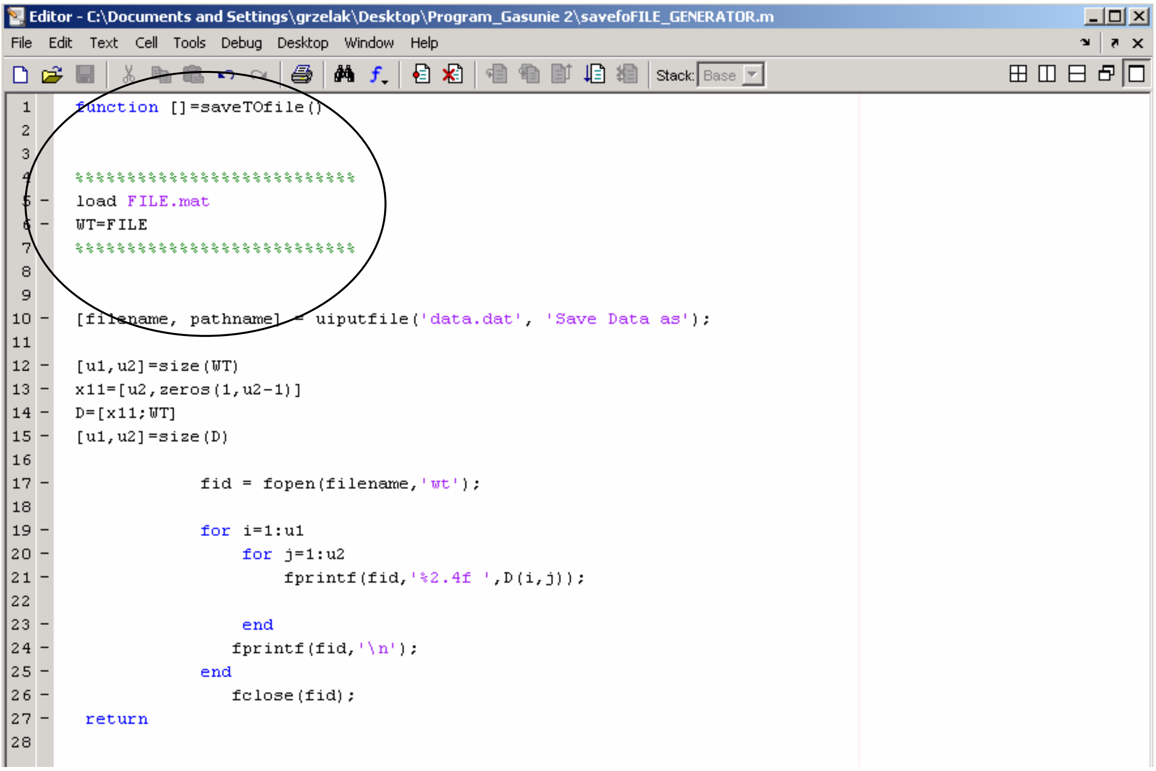
1. Define the dataset in matlab and save it as ***.mat** file- for example create a variable called “**FILE**” and save it using command
   “save FILE.mat FILE”

Remark: Missing data should be indicated as “0” (zero)



```
>> FILE=[1,2,3,4,5;7,6,5,4,3;9,8,7,6,5;1,4,6,4,9]

FILE =

     1     2     3     4     5
     7     6     5     4     3
     9     8     7     6     5
     1     4     6     4     9

>> save FILE.mat FILE
```

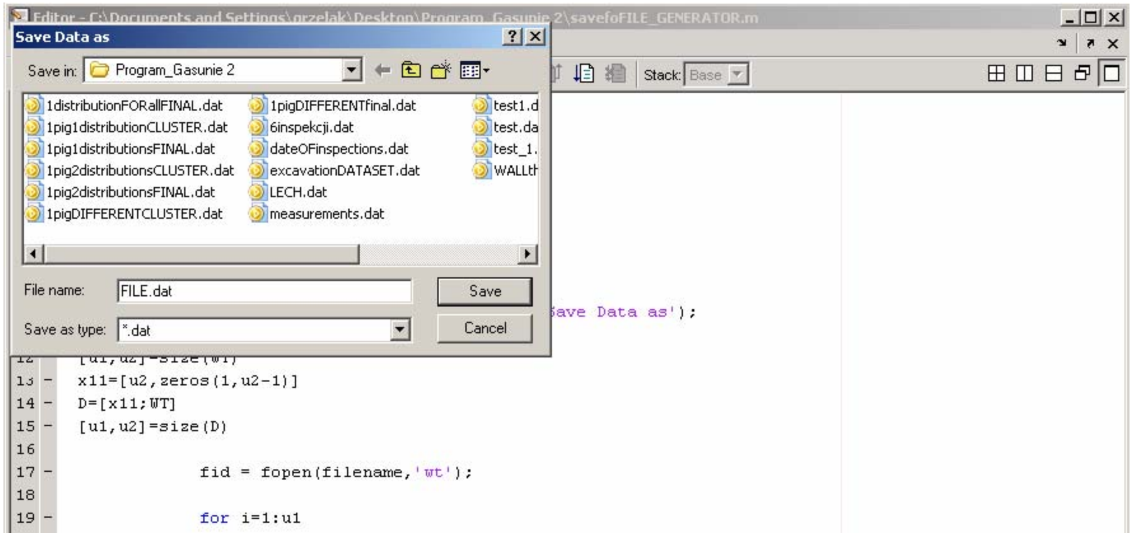
2. open the file “***savetoFILE_GENERATOR.m***”, and in the <u>5 and 6 line</u> write **“load FILE.mat”** and **“WT=FILE”**



Editor - C:\Documents and Settings\grzelak\Desktop\Program_Gasunie 2\savefoFILE_GENERATOR.m
File Edit Text Cell Tools Debug Desktop Window Help
Stack: Base

```
1   function []=saveTOfile()
2
3
4   %%%%%%%%%%%%%%%%%%%%%%%%%%%
5   load FILE.mat
6   WT=FILE
7   %%%%%%%%%%%%%%%%%%%%%%%%%%%
8
9
10  [filename, pathname] = uiputfile('data.dat', 'Save Data as');
11
12  [u1,u2]=size(WT)
13  x11=[u2,zeros(1,u2-1)]
14  D=[x11;WT]
15  [u1,u2]=size(D)
16
17              fid = fopen(filename,'wt');
18
19              for i=1:u1
20                  for j=1:u2
21                      fprintf(fid,'%2.4f ',D(i,j));
22
23                  end
24                 fprintf(fid,'\n');
25              end
26                 fclose(fid);
27   return
28
```

3. run the program and choose the name of <u>***.dat**</u> file, since now the FILE will be transformed to the ****.dat*** file


Editor - C:\Documents and Settings\grzelak\Desktop\Program_Gasunie 2\savefoFILE_GENERATOR.m
Stack: Base

Save Data as

Save in: Program_Gasunie 2

1distributionFORallFINAL.dat
1pig1distributionCLUSTER.dat
1pig1distributionsFINAL.dat
1pig2distributionsCLUSTER.dat
1pig2distributionsFINAL.dat
1pigDIFFERENTCLUSTER.dat
1pigDIFFERENTfinal.dat
6inspekcji.dat
dateOFinspections.dat
excavationDATASET.dat
LECH.dat
measurements.dat
test1.d
test.da
test_1.
WALLth

File name: FILE.dat

Save as type: *.dat

```
ave Data as');
```

```
x11=[u2,zeros(1,u2-1)]
D=[x11;WT]
[u1,u2]=size(D)

            fid = fopen(filename,'wt');

            for i=1:u1
```

d) Files construction

Each file in the program has his own construction, and so:

- “***excavation measurements***” consists of n+1 columns (where *n* is the number of inspections), the first column corresponds to data recorded from the excavation, all the rest correspond to inspections, units are millimeters
- “***measurement* set**”- has exactly n columns and m rows, m is the number of observed defects and n is the number of inspections, units are millimeters
- “***dates of inspections***”- is the column vector consisting of dates of the inspections, all the dates are counted since pipeline installation (pipeline installation is at time t=0) units are years
- “***wall thickness***”- is the column vector, where each row corresponds to each defect, values are presented as nominal wall thickness in millimeters

5 Examples

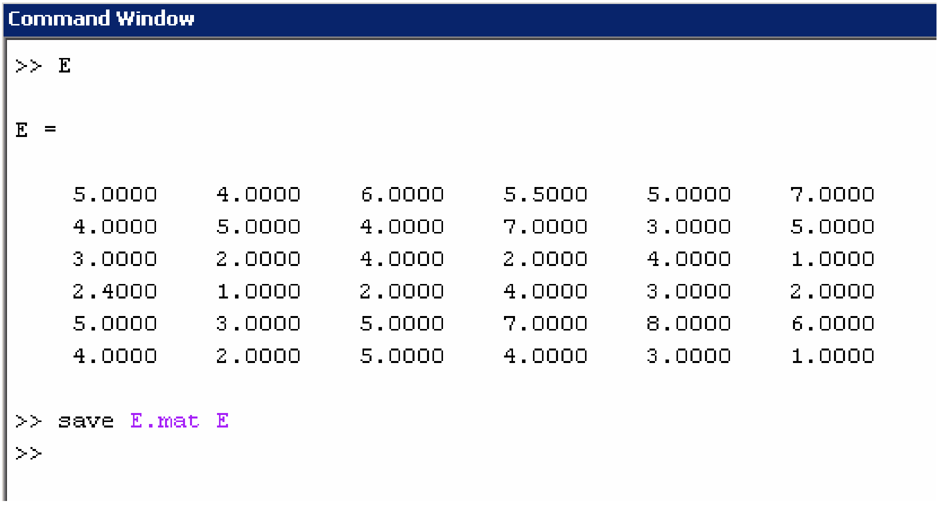
Suppose that we have carried out 5 inspections, the excavation was done after first pigrun. The calibrating dataset is following:

Excavation	Insp 1	Insp 2	Insp 3	Insp 4	Insp 5
5	4	6	5.5	5	7
4	5	4	7	3	5
3	2	4	2	4	1
2.4	1	2	4	3	2
5	3	5	7	8	6
4	2	5	4	3	1

Table 46: excavation dataset [mm]

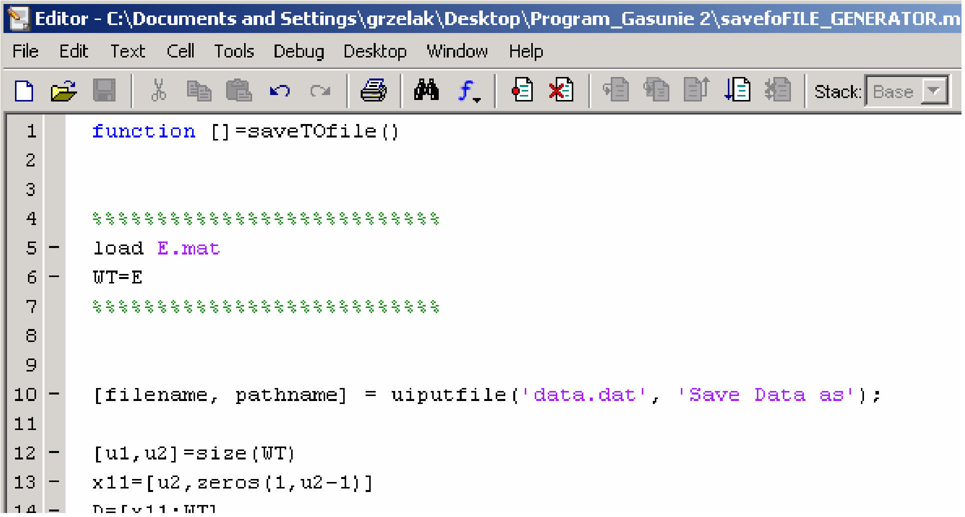
When the dataset is prepared we need to convert it into ***.dat** file, in the manner explained before.
And so:

1. We type the data in matlab and save it as ***E.mat***



```
>> E

E =

    5.0000    4.0000    6.0000    5.5000    5.0000    7.0000
    4.0000    5.0000    4.0000    7.0000    3.0000    5.0000
    3.0000    2.0000    4.0000    2.0000    4.0000    1.0000
    2.4000    1.0000    2.0000    4.0000    3.0000    2.0000
    5.0000    3.0000    5.0000    7.0000    8.0000    6.0000
    4.0000    2.0000    5.0000    4.0000    3.0000    1.0000

>> save E.mat E
>>
```

2. Second step is to convert the dataset to the ***E.dat*** file; hence we change the code in the file "***savetoFILE_GENERATOR.mat***" and turn the program on.


Editor - C:\Documents and Settings\grzelak\Desktop\Program_Gasunie 2\savefoFILE_GENERATOR.m
File
Edit
Text
Cell
Tools
Debug
Desktop
Window
Help
Stack: Base

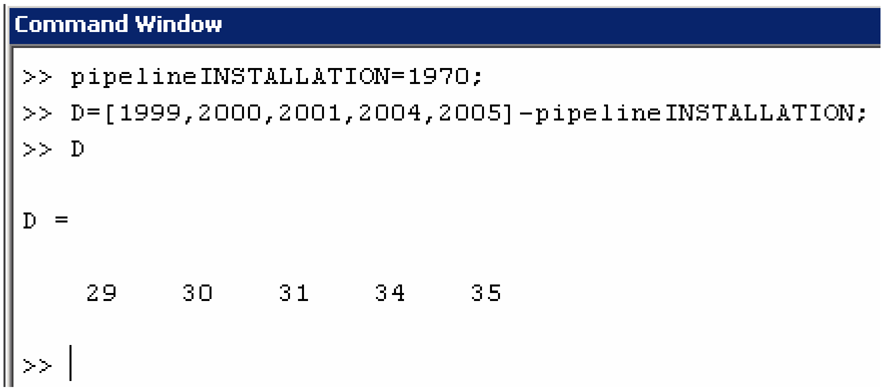
```
function []=saveTOfile()


%%%%%%%%%%%%%%%%%%%%%%%%%%%
load E.mat
WT=E
%%%%%%%%%%%%%%%%%%%%%%%%%%%


[filename, pathname] = uiputfile('data.dat', 'Save Data as');

[u1,u2]=size(WT)
x11=[u2,zeros(1,u2-1)]
D=[x11;WT]
```

The hypothetical measurements reported by the pigs during inspections are following:

Defect no.	**Insp 1 1999**	**Insp 2 2000**	**Insp 3 2001**	**Insp 4 2004**	**Insp 5 2005**
1	4	6	5.5	5	7
2	5	4	7	3	5
3	2	4	2	4	1
4	1	2	4	3	2
5	3	5	7	8	6
6	2	5	4	3	1

Table 47: example of the measurements

The theoretical pipeline was constructed in 1970. So the dates of inspections are following:



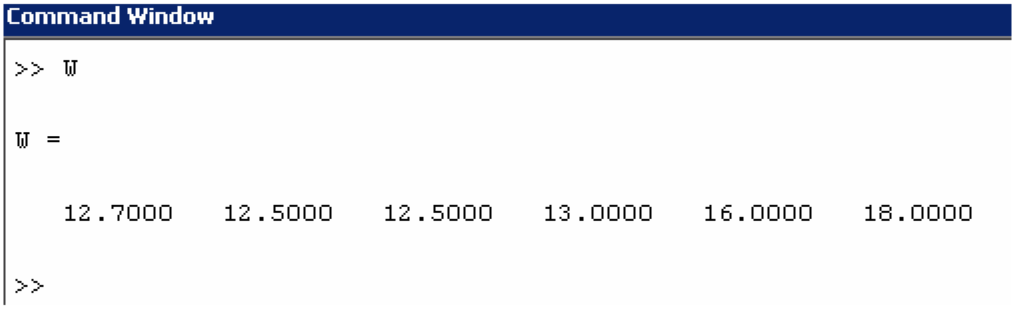
```
>> pipelineINSTALLATION=1970;
>> D=[1999,2000,2001,2004,2005]-pipelineINSTALLATION;
>> D

D =

    29    30    31    34    35

>> 
```

Now, let’s save the variable **D** (dates of inspection) as ***D.dat*** and the measurements **M** as ***M.dat,.***
For the nominal wall thickness we assume:

defect no.	Nominal wall thickness
1	12.7
2	12.5
3	12.5
4	13
5	16
6	18

Table 48: The nominal wall thickness

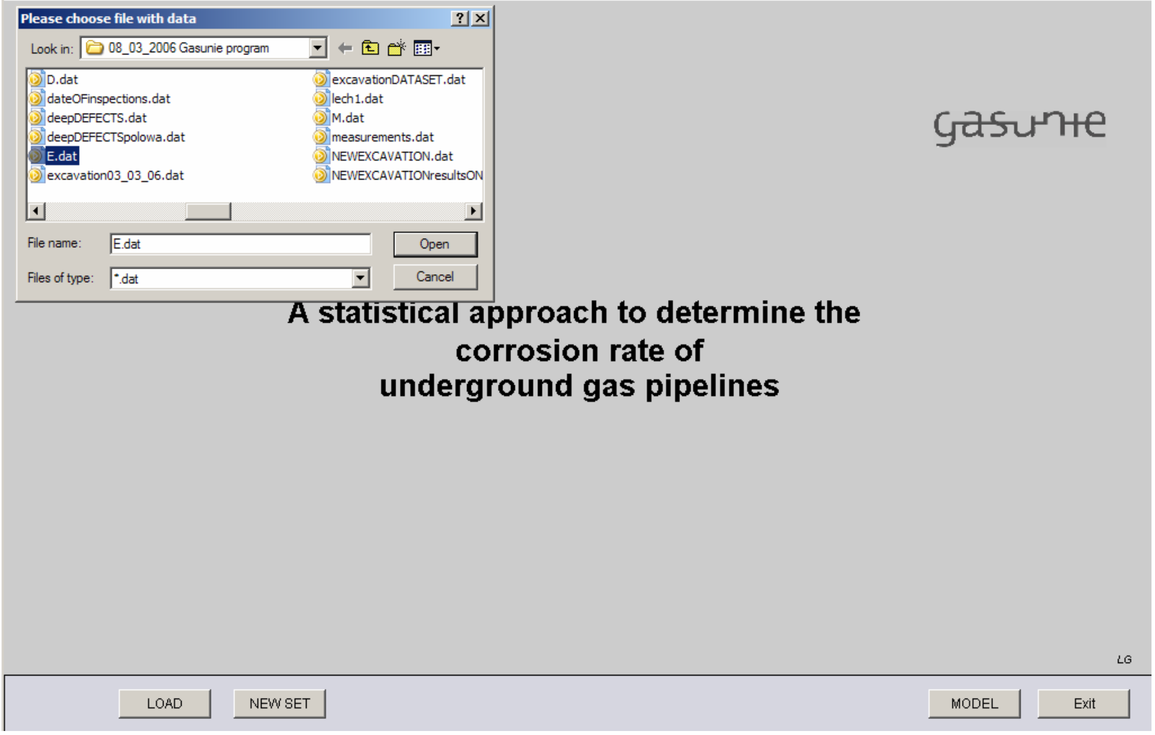
In this case the wall thickness vector we denote as **W** and perform the same transformation procedure as for the others.



```
>> W

W =

   12.7000   12.5000   12.5000   13.0000   16.0000   18.0000

>>
```

Now, when all necessary data are collected, typed, and transformed into ***.dat** type file, so the analysis can be carried out.

1. First: run the program and load the excavation data (here **E.dat** file)



Please choose file with data

Look in: 08_03_2006 Gasunie program

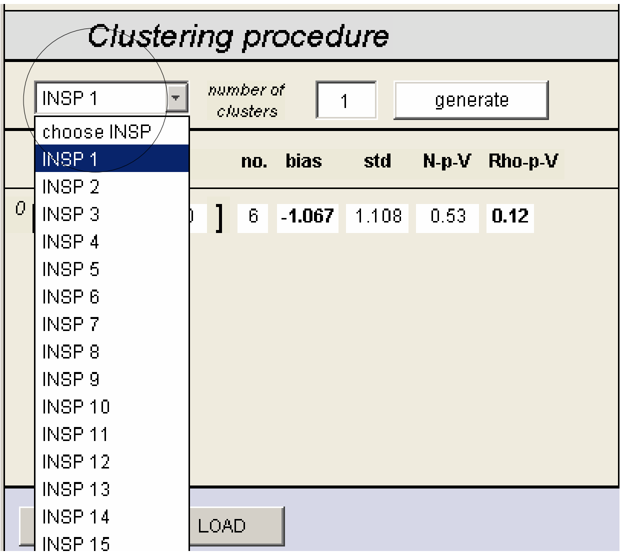
File name: E.dat

Files of type: *.dat

A statistical approach to determine the corrosion rate of underground gas pipelines

2. When excavation data is loaded then program immediately goes to the calibration stage, for simplicity define only one cluster for each inspection (one cluster contains all the measurements). So for each inspection generate exactly one cluster and confirm by pressing ***GENERATE***, ***ANALYZE*** and ***ACCEPT*** buttons.

   a. **Choose inspection**



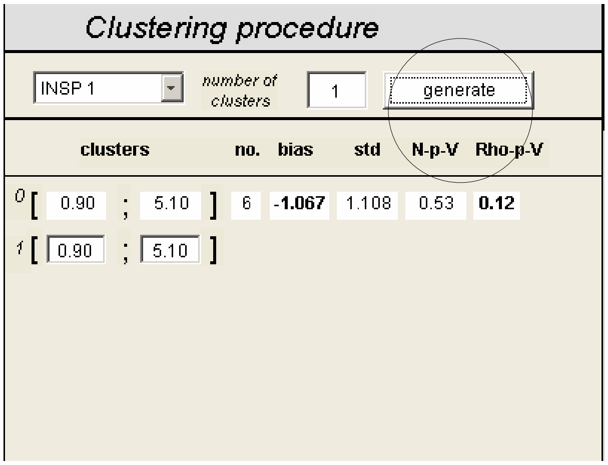
Clustering procedure

INSP 1

number of clusters: 1

no.	bias	std	N-p-V	Rho-p-V
6	-1.067	1.108	0.53	0.12

b. Generate one cluster

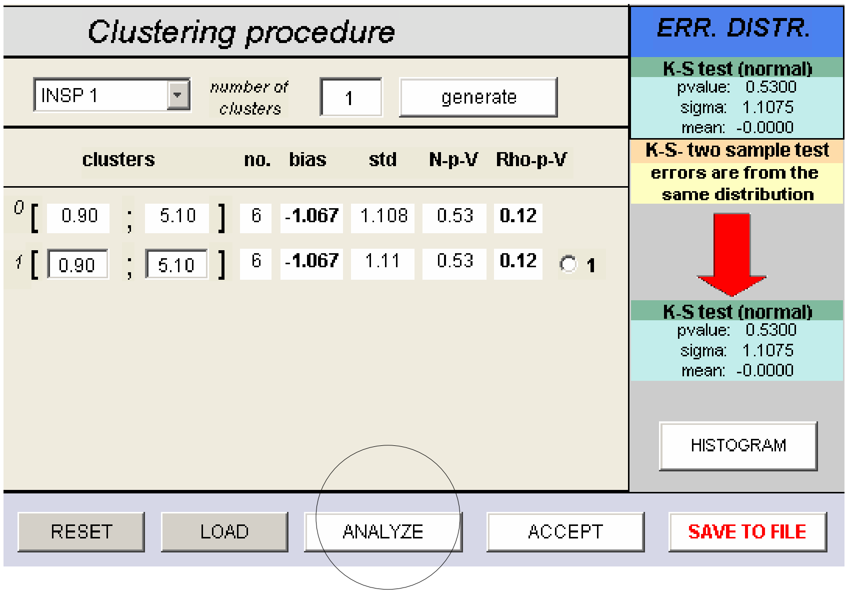


Clustering procedure

INSP 1 number of clusters 1

	clusters			no.	bias	std	N-p-V	Rho-p-V
0	[0.90	;	5.10]	6	-1.067	1.108	0.53	0.12
1	[0.90	;	5.10]					

c. Analyze the clusters



Clustering procedure

INSP 1 number of clusters 1

	clusters			no.	bias	std	N-p-V	Rho-p-V	
0	[0.90	;	5.10]	6	-1.067	1.108	0.53	0.12	
1	[0.90	;	5.10]	6	-1.067	1.11	0.53	0.12	○ 1

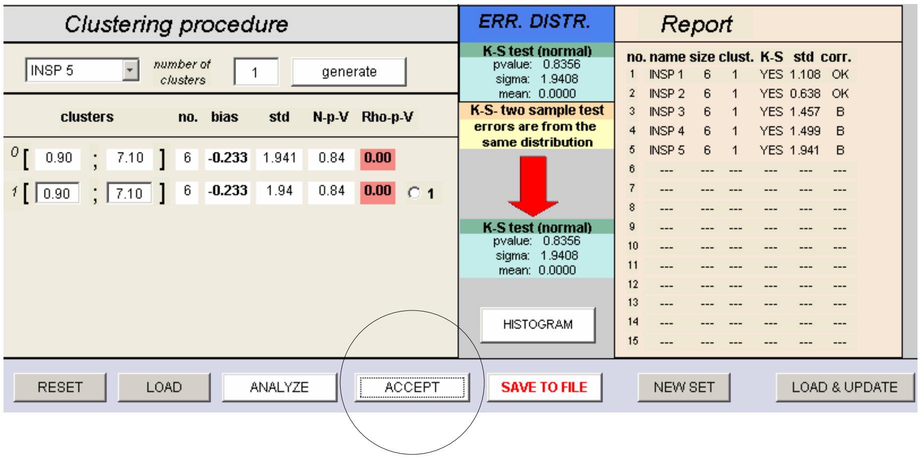
ERR. DISTR.

K-S test (normal)
pvalue: 0.5300
sigma: 1.1075
mean: -0.0000

K-S- two sample test
errors are from the same distribution

K-S test (normal)
pvalue: 0.5300
sigma: 1.1075
mean: -0.0000

d. Accept the results



	clusters		no.	bias	std	N-p-V	Rho-p-V	
0	[0.90	; 7.10]	6	-0.233	1.941	0.84	0.00	
1	[0.90	; 7.10]	6	-0.233	1.94	0.84	0.00	○ 1

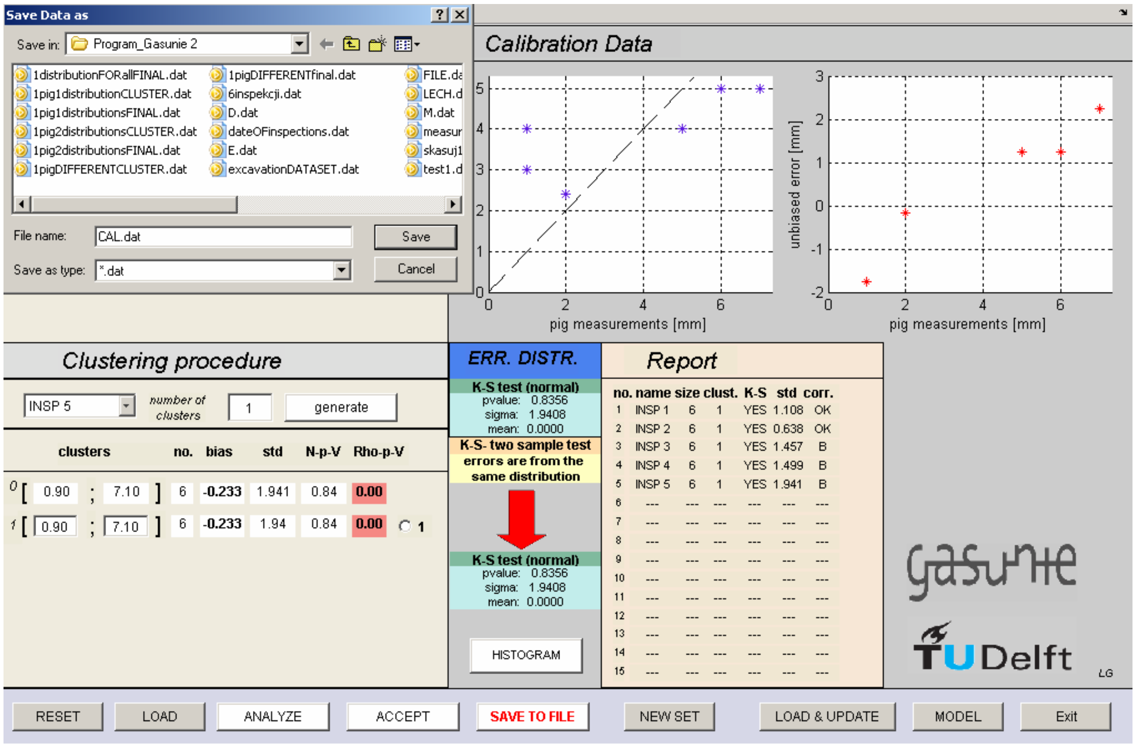
K-S test (normal)
pvalue: 0.8356
sigma: 1.9408
mean: 0.0000

K-S- two sample test
errors are from the same distribution

K-S test (normal)
pvalue: 0.8356
sigma: 1.9408
mean: 0.0000

no.	name	size	clust.	K-S	std	corr.
1	INSP 1	6	1	YES	1.108	OK
2	INSP 2	6	1	YES	0.638	OK
3	INSP 3	6	1	YES	1.457	B
4	INSP 4	6	1	YES	1.499	B
5	INSP 5	6	1	YES	1.941	B
6	---	---	---	---	---	---
7	---	---	---	---	---	---
8	---	---	---	---	---	---
9	---	---	---	---	---	---
10	---	---	---	---	---	---
11	---	---	---	---	---	---
12	---	---	---	---	---	---
13	---	---	---	---	---	---
14	---	---	---	---	---	---
15	---	---	---	---	---	---

e. **Save to file (red button) as *CAL.dat* – in this case no file transformation is required. Matlab immediately saves the results as the **.dat* file.**



File name: CAL.dat
Save as type: *.dat

	clusters		no.	bias	std	N-p-V	Rho-p-V	
0	[0.90	; 7.10]	6	-0.233	1.941	0.84	0.00	
1	[0.90	; 7.10]	6	-0.233	1.94	0.84	0.00	○ 1

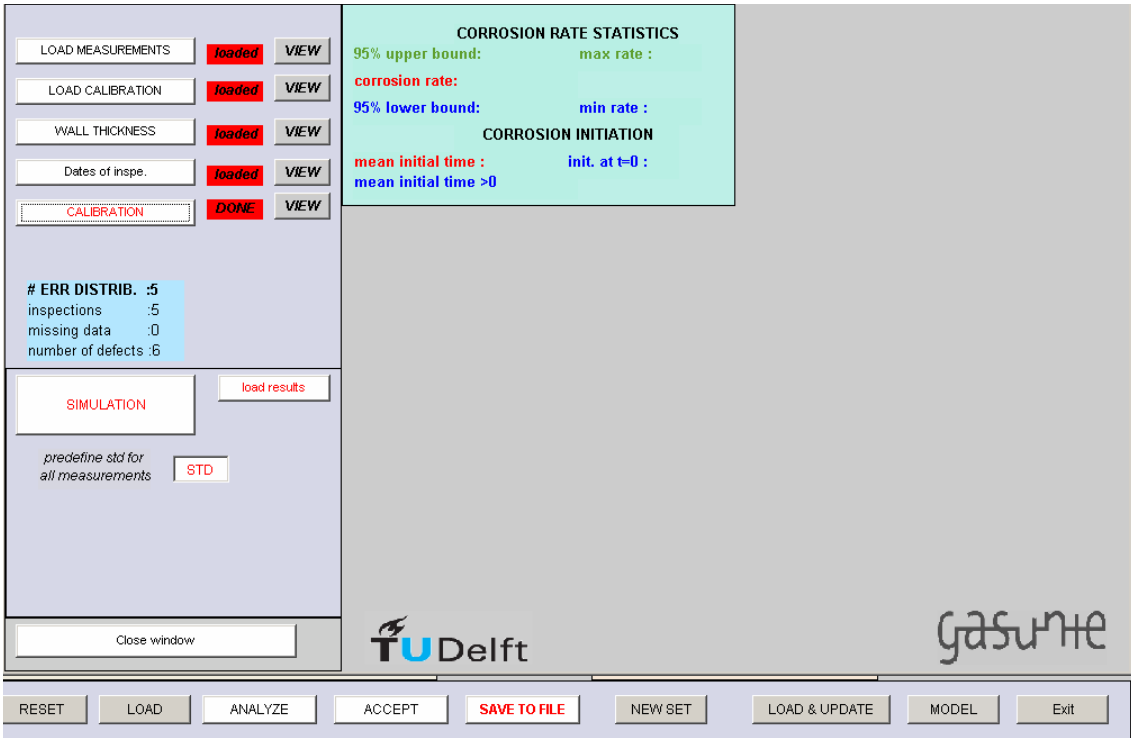
K-S test (normal)
pvalue: 0.8356
sigma: 1.9408
mean: 0.0000

K-S- two sample test
errors are from the same distribution

K-S test (normal)
pvalue: 0.8356
sigma: 1.9408
mean: 0.0000

no.	name	size	clust.	K-S	std	corr.
1	INSP 1	6	1	YES	1.108	OK
2	INSP 2	6	1	YES	0.638	OK
3	INSP 3	6	1	YES	1.457	B
4	INSP 4	6	1	YES	1.499	B
5	INSP 5	6	1	YES	1.941	B
6	---	---	---	---	---	---
7	---	---	---	---	---	---
8	---	---	---	---	---	---
9	---	---	---	---	---	---
10	---	---	---	---	---	---
11	---	---	---	---	---	---
12	---	---	---	---	---	---
13	---	---	---	---	---	---
14	---	---	---	---	---	---
15	---	---	---	---	---	---

3. When the calibration procedure is performed, then go to the second stage-namely to the “**MODEL**”. Firstly, all measurements should be loaded (**wall thickness (W), measurements (M), calibration data (CAL), dates of inspections (D)**) and the “***CALIBRATION***” button should be pressed.



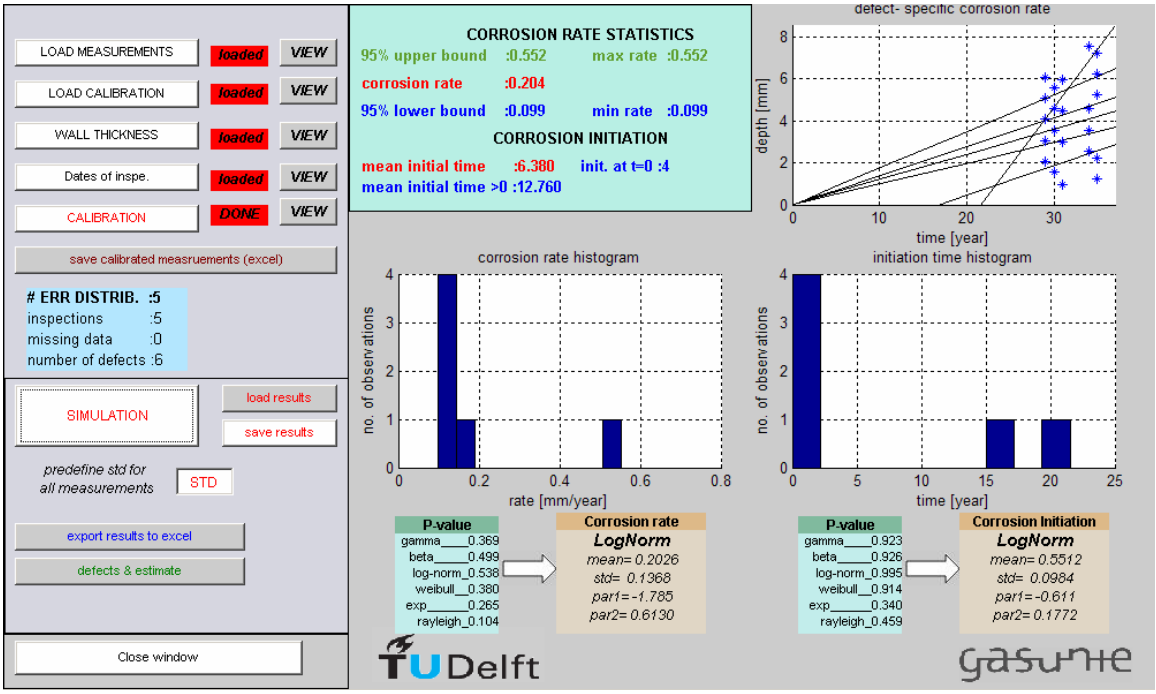
CORROSION RATE STATISTICS
95% upper bound: max rate :
corrosion rate:
95% lower bound: min rate :
CORROSION INITIATION
mean initial time : init. at t=0 :
mean initial time >0

# ERR DISTRIB. :5
inspections :5
missing data :0
number of defects :6

predefine std for
all measurements

A small summary shows that from the calibration procedure we have exactly 5 distributions (one distribution per inspection), the number of defects is 6, the number of missing data is zero.

Now we can proceed and press button "***SIMULATION***". After a few second the results should appear.



CORROSION RATE STATISTICS
95% upper bound :0.552 max rate :0.552
corrosion rate :0.204
95% lower bound :0.099 min rate :0.099
CORROSION INITIATION
mean initial time :6.380 init. at t=0 :4
mean initial time >0 :12.760

# ERR DISTRIB. :5
inspections :5
missing data :0
number of defects :6

predefine std for
all measurements

P-value
gamma____0.369
beta_____0.499
log-norm_0.538
weibull__0.380
exp______0.265
rayleigh_0.104

Corrosion rate
LogNorm
mean= 0.2026
std= 0.1368
par1=-1.785
par2= 0.6130

P-value
gamma____0.923
beta_____0.926
log-norm_0.995
weibull__0.914
exp______0.340
rayleigh_0.459

Corrosion Initiation
LogNorm
mean= 0.5512
std= 0.0984
par1=-0.611
par2= 0.1772

All the files are attached to the delivered program, hence the files conversion is not required.

e) Available DATABASES/ EXAMPLES

1. The database based on the real data collected by Gasunie,
   Inspections done by 4 pigs, one excavation done after the first inspection

Files:

- "***excavationDATASET.dat***"- consists of excavated metal loss depths and data necessary for calibration
- "***measurements.dat***"- measurements done during 4 inspections- 52 reported defects
- "***dateOFinspections.dat***"- consists of data concerning information about carried out dates of inspection
- "***WALLthickness.dat***"- vector of wall thickness for where 52 defects were reported
- "***1pig1distributionCLUSTER.dat***"- result of calibration procedure- each pig has exactly one cluster- measurement error distribution- can be applied in the MODEL
- "***1pig2distributionsCLUSTER.dat***"- idem, but now each pig has 2 clusters- distributions

2. Example presented in the tutorial

- ***"M.dat***"- measurements
- ***"D.dat"***- dates of inspections
- ***"E.dat"***- calibrating dataset- excavation dataset
- ***"W.dat"***-wall thickness file
- ***"CAL*.*dat***" – result of calibration procedure- each pig exactly one distribution