Extreme Value Analysis of Corrosion Data

Master Thesis

Marcin Glegola

Delft University of Technology,
Faculty of Electrical Engineering,
Mathematics and Computer Science,
The Netherlands

July 2007

Abstract

In the oil industry, corrosion is one of the important factors that can cause system failure. This can cost a lot of money or can pose a danger to human lives. Therefore, corrodible structures are subjected to inspections. Usually, it is not possible to inspect 100% of the system. This means that methods for inference about the state of corrosion on the not inspected area, on the basis of inspection data, are needed. Since extreme defect depths influence the reliability of the entire system, the extreme-value methods are sensible to apply.

The thesis covers two main issues. Firstly, the methods for statistical inference about extreme defect depths are presented. They are the generalised extreme-value and the generalised-Pareto distribution. For both models, techniques to extrapolate the results to the not inspected part of the system are described. The second issue is taking into account local dependence of the underlying observations while inferencing about extremes of corrosion. For this purpose, data declustering is performed. In order to find a proper number of data clusters, the extremal index method and agglomerative hierarchical clustering algorithm are used. The methods are applied to simulated and real data sets.

The results show that data declustering improves the consistency of the results given by the generalised extreme-value and the generalised-Pareto distributions.

Key words: extremal index method, data declustering, generalised-extreme value distribution, generalised-Pareto distribution, extrapolation.

Members of the Committee

Chairperson of Graduate Comitee: Prof. Dr. Ir. Jan M. van Noortwijk

Graduate Committee:

Prof. Dr. Ir. Jan M. van Noortwijk	Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science
Prof. Dr. Ir. Geurt Jongbloed	Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science
Dr. Marco Giannitrapani	Shell Global Solutions, Amsterdam, Statistical Consulting Department
Drs. Fred Hoeve	Shell Global Solutions, Amsterdam, Statistical Consulting Department
MSc Ir. Sebastian Kuniewski	Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science

Acknowledgements

I would like to express my sincere gratitude to Jolanta Misiewicz, Roger M. Cooke and Dorota Kurowicka, for giving me this fantastic opportunity to study at TU Delft.

I am very grateful to Jan M. van Noortwijk and Sebastian Kuniewski for help, inspiration, corrections and comments. I thank people from Shell Global Solutions, especially Marco Giannitrapani, Fred Hoeve, Sieger Terpstra and Peter van de Camp, for giving me the opportunity to work on this interesting project and for sharing their knowledge and experience.

Many people have inspired and helped me during my stay at TU Delft. Thanks to them, it was not only hard work but fun also. Thank you all my friends.
I express sincere appreciation to my family for their care and support.
At the end I thank the most Monika for her true love and encouragement.

Contents

List of Figures

List of Tables

Introduction

In the oil industry corrosion is one of the main factors influencing the reliability of the systems. Therefore the information about the corrosion process is an important issue in the maintenance of corrodible structures. If the extreme wall loss becomes greater than the nominal wall thickness of some object (pipe, tank) then we have a system failure. This can cost a lot of money or even human lives. To avoid such situations proper maintenance actions are performed. However, to make right decisions about actions to be taken (e.g. replacement of a component, repair of a component) the information about the deterioration state of the system is needed. This can be obtained through inspection, which covers usually only part of the system. The information about the condition of the remaining, not inspected part of the system is unknown. Therefore methods for inference about the state of these parts of the system are needed. One of the solutions is fitting the statistical models that will allow for inference about the corrosion process and extrapolation of the results. A family of such models form the extreme-value distributions, which are suitable for statistical inference about extremes of the given phenomenon. In their standard form they are applied under the assumptions that the underlying observations are independent. However, in the corrosion context this assumptions is questionable since it is very likely that defects on the surface are locally dependent (see Figure 1).

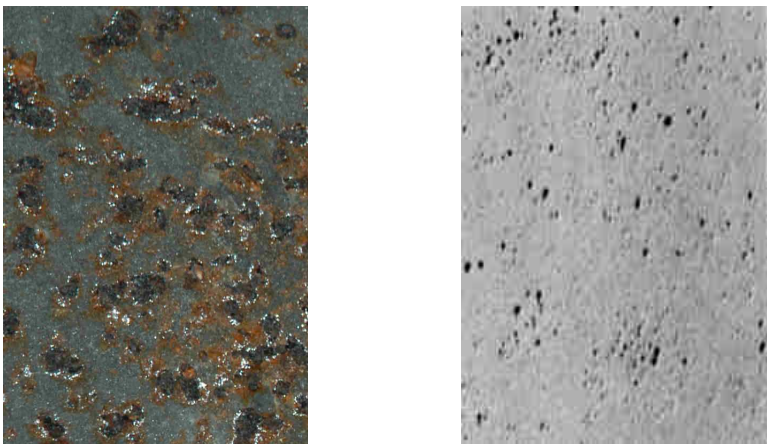

Figure 1: Example of pitting (localised) corrosion

Therefore it is our aim to study how the extreme-value tools can be used without the assumption about the defect depths independence.

Objective of the Thesis

The objective of the thesis is to present the statistical tools to model the extremes of corrosion data taking into account local defect dependence. For this purpose we use two extreme-value distributions, namely the generalised extreme-value and the generalised-Pareto distribution. We show how these models can be applied while not assuming independence of the underlying observations. Through theoretical arguments and examples of application we show that they are very consistent and that both allow for statistical inference about extremes of corrosion.

Outline of the Thesis

The thesis is organised as follows. In Chapter 1, methods for statistical inference about extreme defects caused by corrosion are introduced. They are the generalised extreme-value and the generalised-Pareto distribution. In order to inference about the extent of the corrosion on the not inspected part of the system, methods for spatial results extrapolation are described.

In Chapter 2, it is shown how the above methods can be applied to corrosion data, while not assuming independence of underlying observations. Moreover, the framework for modelling extreme defects caused by corrosion with extreme-value distributions is presented.

The conclusions and recommendations for future research are presented in Chapter 3.

Chapter 1

Extreme-Value Analysis of Corrosion Data

In this chapter we will introduce two probability distributions which we will use to model statistical behaviour of extreme defects caused by corrosion. They are the generalised extreme-value and the generalised-Pareto distribution.

1.1 Generalised Extreme-Value Distribution

This section we will start with a brief description of the generalised extreme-value distribution which is applicable to block maxima data. We will introduce methods for parameter estimation and assessment of goodness-of-fit of the model to data. Because the extent of corrosion on the not inspected area is of primary interest, the extrapolation methods will be described. At the end, an application of the model to real life data will be presented.

The generalised-extreme value distribution is used to model statistical behaviour of **block maxima** data (Coles 2001). If $\{X_1, X_2, X_3, \ldots\}$ is a sequence of independent and identically distributed random variables then the block maxima

are defined as:

$$Z_i = \max(X_1, \ldots, X_n),\ i = 1, \ldots, m. \tag{1.1}$$

It can be shown that for sufficiently large block size n the cumulative distribution function (cdf) of Z_i converges in distribution to the generalised extreme-value distribution given in Definition 1.1.

Definition 1.1 *A random variable Z is said to have a* ***generalised extreme-value distribution*** *(GEV) with scale parameter $\sigma > 0$, location parameter μ and shape parameter ξ, if its cumulative distribution function is given by:*

$$G(z) = \begin{cases} \exp\left\{-\left[1+\xi\left(\dfrac{z-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left[-\left(\dfrac{z-\mu}{\sigma}\right)\right]\right\}, & \xi = 0, \end{cases} \tag{1.2}$$

where $[z]_+ = \max(0, z)$. The corresponding probability density function is given by:

$$g(z) = \begin{cases} \dfrac{1}{\sigma}\left[1+\xi\left(\dfrac{z-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}-1} \exp\left\{-\left[1+\xi\left(\dfrac{z-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ \dfrac{1}{\sigma}\exp\left\{-\left(\dfrac{z-\mu}{\sigma}\right)\right\}\exp\left\{-\exp\left[-\left(\dfrac{z-\mu}{\sigma}\right)\right]\right\}, & \xi = 0. \end{cases} \tag{1.3}$$

The support of the GEV distribution is bounded by $\eta = \mu - \dfrac{\sigma}{\xi}$, what is given in (1.4):

$$\begin{aligned} -\infty < z < \eta \qquad & \text{when } \xi < 0 \\ \eta < z < \infty \qquad & \text{when } \xi > 0 \\ -\infty < z < \infty \qquad & \text{when } \xi \to 0 \end{aligned} \tag{1.4}$$

This fact has a meaning for extrapolation purposes and will be mentioned in the later sections.

The mean and the variance of the GEV distribution are given by (1.5) and (1.6), respectively:

$$E(Z) = \begin{cases} \mu - \dfrac{\sigma}{\xi}(1 - \Gamma(1-\xi)), \ \xi < 1, & \xi \neq 0 \\ \mu + \sigma\gamma, & \xi = 0 \end{cases} \tag{1.5}$$

$$Var(Z) = \begin{cases} \dfrac{\sigma^2}{\xi^2}(\Gamma(1-2\xi) - \Gamma^2(1-\xi)), \ \xi < 1/2, & \xi \neq 0 \\ \dfrac{\pi}{6}\sigma^2, & \xi = 0 \end{cases} \tag{1.6}$$

where γ is the Euler-Mascheroni constant, approximately equal to $0.57721\ldots$ (for details see Appendix A) and $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ is the gamma function.

The single family of the GEV distributions contains all three possible limiting distributions for block maxima, i.e. the Gumbel, Fréchet and the Weibull families (Coles 2001, Beirland, Teugels, Vynckier et al. 1996). Moreover, the Gumbel distribution (for the definition of the Fréchet and Weibull distribution see Appendix A) corresponds to the case when $\xi = 0$, the Fréchet when $\xi > 0$ and the Weibull when $\xi < 0$, respectively. This fact supports the usage of the GEV distribution as a tool for statistical modelling of extreme values. One does not have to make a subjective *a priori* assumptions about the most appropriate type of distribution for data. Through inference on ξ, the data themselves determine the most appropriate type of distribution.

In corrosion application, the suitable data for statistical modelling of extreme defect depths using the GEV distribution can arise as the largest wall loss measurement from distinct specimens, sometimes called coupons (Scarf & Laycock 1996).

1.1.1 Parameter Estimation

One of the most commonly used method to estimate the unknown parameters of the GEV distribution is the **maximum likelihood** method. Under the assumption that $Z_1, ..., Z_m$ are independent variables having the GEV distribution, the log-likelihood of the unknown parameters is given by (1.7):

$$l(z_1, \ldots, z_m|\xi, \sigma, \mu) = \begin{cases} -m\log(\sigma) - (1+1/\xi)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right] \\ -\sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-1/\xi}, & \xi \neq 0 \\ \\ -m\log(\sigma) - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right) \\ -\sum_{i=1}^{m}\exp\left\{-\left(\frac{z_i-\mu}{\sigma}\right)\right\}, & \xi = 0 \end{cases} \tag{1.7}$$

provided $1+\xi\left(\frac{z_i-\mu}{\sigma}\right) > 0$ for $i = 1, \ldots, m$; otherwise $l(z_1, \ldots, z_m|\xi, \sigma, \mu) = -\infty$. The estimates $(\hat{\xi}, \hat{\sigma}, \hat{\mu})$ of the unknown parameters can be found by maximising (1.7) with respect to parameter vector (ξ, σ, μ). The solution is found by numerical methods.

1.1.2 Goodness-of-fit

In this section methods of goodness-of-fit assessment will be introduced. They are the quantile plot, the Anderson-Darling test for the upper tail of distribution and the Kolmogorov-Smirnov test for the entire distribution.

Quantile plot

Quantile plot is a graphical method used to assess if the assumed probability distribution is the appropriate model for the data.

Definition 1.2 *Given an ordered sample of observations* $z_1 \leq, \ldots, \leq z_m$ *from a population with estimated cumulative distribution function* $\hat{G}$*, the* ***quantile plot*** *consists of the points:*

$$\left\{ \left(\hat{G}^{-1} \left(\frac{i}{m+1} \right), z_i \right) : i = 1, \ldots, m \right\} \tag{1.8}$$

If $\hat{G}$ is a reasonable estimate of G, then the quantile plot should consist of points close to the diagonal. Moreover on the basis of the delta method the confidence bounds to this plot can added (for details see Appendix A).

Anderson-Darling test for upper tail

Another approach to assess the goodness-of-fit is the Anderson-Darling goodness-of-fit test for upper tail of the distribution (Chernobai, Rachev & F.Fabozzi 2005), which is based the test statistic given in (1.9):

$$AD^2_{up} = 2 \sum_{i=1}^{m} \log(1 - z_i) + \frac{1}{n} \sum_{i=1}^{m} (1 + 2(m - i)) \frac{1}{1 - z_i}, \tag{1.9}$$

where $z_i = \hat{F}(x(i))$, $\hat{F}$ is the cumulative distribution function of the fitted distribution and $x_{(1)} \leq, \ldots, \leq x_{(n)}$ is and ordered sample.
The test is formulated as:

$$\begin{aligned} H_0 &: \text{The data follow the specified distribution } F. \\ H_a &: \text{The data do not follow the specified distribution } F. \end{aligned} \tag{1.10}$$

The distribution of the AD^2_{up} is approximated by simulation. If the p-value, i.e. the probability of obtaining the value of the test statistic greater than actually calculated, exceeds the assumed significance level α then H_0 is rejected in favour of H_a. Otherwise there are no reasons to reject H_0.
This test puts more weight to the upper tail of the distribution, which is important for extrapolation accuracy.

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test can be used to decide if a sample comes from a population with a specified distribution. It is based on the empirical distribution function given in Definition 1.3.

Definition 1.3 *Let $X_1, \ldots, X_n$ be random variables with realisations $x_i \in \mathbb{R}$, $i = 1, \ldots, n \in \mathbb{N}$. The **empirical distribution function** $F_n(x)$ based on the sample $x_1, \ldots, x_n$ is a step function defined as:*

$$F_n(x) = \frac{\textit{number of elements in the sample} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(x_i \leq x)},$$

where $1_{(A)}$ is an indicator function defined as:

$$1_A(x) = \begin{cases} 1 & \textit{if } x \in A, \\ 0 & \textit{if } x \notin A. \end{cases}$$

The Kolmogorov-Smirnov test statistic is given by (1.11):

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{1.11}$$

where F is the hypothesised distribution. The test is formulated as:

$$\begin{aligned} H_0 &: \text{The data follow the specified distribution } F. \\ H_a &: \text{The data do not follow the specified distribution } F. \end{aligned} \tag{1.12}$$

Under the hypothesis H_0, $\sqrt{n}D_n$ converges in distribution to the Kolmogorov distribution which does not depend on F, provided that F is continuous. Therefore this test is based on the critical values of the Kolmogorov distribution (for more details see Appendix A).

1.1.3 Extrapolation

To extrapolate the results over larger areas we can use two approaches. One is based on the return-level method and the other is based on determining the probability distribution (the GEV distribution) of the maximum corresponding to the area of extrapolation.

The return-level method

For a given value of probability p, $(0 < p < 1)$, we determine a value z_p satisfying (1.13):

$$G(z_p) = 1 - p, \tag{1.13}$$

where G is a cumulative distribution function.

When G belongs to the family of the GEV distributions, solving equation (1.13) by inverting equation (1.2) gives (1.14):

$$z_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma \log\{-\log(1-p)\}, & \xi = 0. \end{cases} \tag{1.14}$$

Note that (1.13) is equivalent to $Pr\{X > z_p\} = p$, where X has a cumulative distribution function G. It means that any particular block maximum exceeds the value z_p with probability p. In common terminology z_p is called the **return-level** associated with **return-period** $1/p$. The level z_p is expected to be exceeded on average once every $1/p$ blocks (Coles 2001).

In corrosion application the return-level method can be used in the following way. If the GEV distribution was fitted to block maxima data $z_1, \ldots, z_m$, where the physical block size is k units, then if we want to extrapolate to the area $M \times k$ we must set $p = \frac{1}{M}$. Then the determined level z_p is expected to be exceeded on average once every $M \times k$ units of the considered area.

Using the **profile likelihood** method (for details see Appendix A) we can determine the confidence interval for return-level z_p, which is usually more accurate than the one obtained by the delta method (Coles 2001).

Extrapolation of the GEV distribution

Another approach for extrapolation is based on the implied distribution of the maximum over the area which is some multiple, say M of the sampled areas. More precisely, suppose that block maxima data $z_1, \ldots, z_m$ come from blocks of the size k units. If the GEV distribution fitted to observed maxima $z_1, \ldots, z_m$ is assumed to be the distribution of the maximum corresponding to the area of k units, then the cumulative distribution function of the random variable X_M, i.e. of the maximum over the area $M \times k$ units, is given by (1.15):

$$G_M(z) = G(z)^M \tag{1.15}$$

where G is the GEV distribution given by (1.2). Then (1.15) is equal to (1.16):

$$G_M(z) = \begin{cases} \exp\left\{-\left[1+\xi\left(\dfrac{z-\mu+\sigma(1-M^{\xi})/\xi}{\sigma M^{\xi}}\right)\right]_+^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left[-\left(\dfrac{z-(\mu+\sigma\log M)}{\sigma}\right)\right]\right\}, & \xi = 0 \end{cases} \tag{1.16}$$

It is worth to mention that this distribution is within the GEV family of distributions and its parameters are related to the original parameters by (1.17):

$$\begin{cases} \mu_M = \mu - \sigma(1-M^{\xi})/\xi,\ \sigma_M = \sigma M^{\xi},\ \xi_M = \xi & if\ \xi \neq 0 \\ \mu_M = \mu + \sigma\log M,\ \sigma_M = \sigma & if\ \xi = 0 \end{cases} \tag{1.17}$$

This gives a direct way to determine the distribution of the maximum wall loss on the extrapolated area.

1.1.4 Example of Application

This example shows the application of the GEV model to real life data. We will start with the data set description. After parameter estimation the goodness-of-fit of the GEV distribution to data will be assessed. At the end the extrapolation results will be presented.

The data consists of corrosion wall loss measurements from a 36" diameter buried pipe of 19 [mm] wall thickness, and of 300 [m] total length. The inspected area was about 13.51 [m] (for further analysis this region is denoted as $C1$). On the excavated area, scans were made with the screening technique. The automated scanner first takes the measurements each $dy = 5$ $[mm]$ in circumferential direction, and then moves with step size $dx = 58$ $[mm]$ to the next axial position. Because corrosion is present mainly at the bottom of the pipe, the measurements were made along the pipe around the "6 o'clock" position with width 0.8 $[m]$. This gave in total 37280 measurements which are stored

in a 160×233 matrix. In each column of this matrix there is data gathered in one step of the screening machine in axial direction. All measurements are presented in Figures 1.1 and 1.2. The first step in our analysis is the block

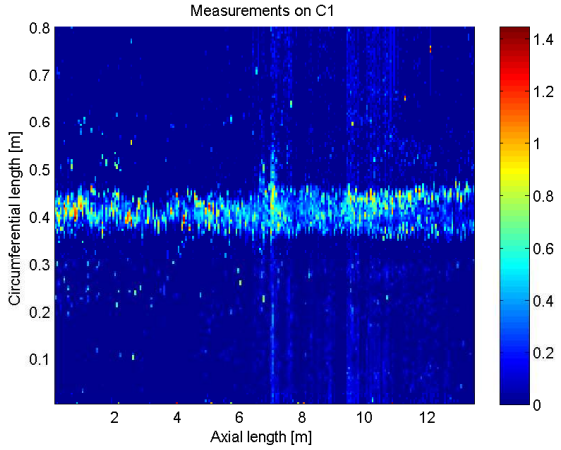

Figure 1.1: Scan results on the excavated part of pipe-C1 in [mm]

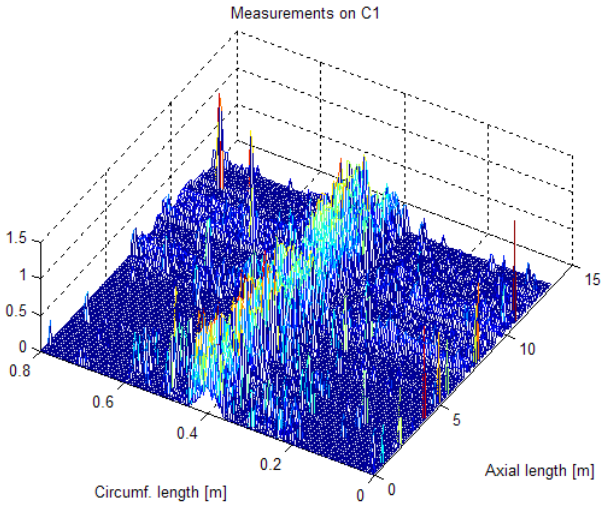

Figure 1.2: Scan results on the excavated part of pipe-C1 in [mm]

definition. We have to specify how we are going to read the maxima from the presented data set. As can be seen on Figure 1.1, corrosion is present mainly on the bottom of the pipe. This suggests that we can define one block as number B_s of columns in the matrix where data is stored. Since it is hard to judge which number to choose we will perform the goodness-of-fit tests for the GEV distribution versus different values of B_s. The results are presented in Figure 1.3 from which we can see that $B_s = 2$ is a good choice.

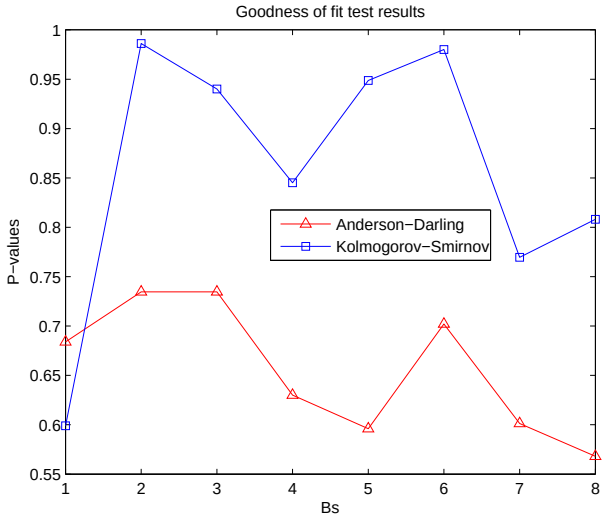

Figure 1.3: Goodness-of-fit test results for the GEV distribution for different block sizes. B_s-number of columns corresponding to one block

Then the block maxima seems to be stationary, what indeed is confirmed by Figure 1.4.

The histogram and the results of fitting the GEV distribution to block maxima data are presented in Figure 1.5 and Table 1.1.

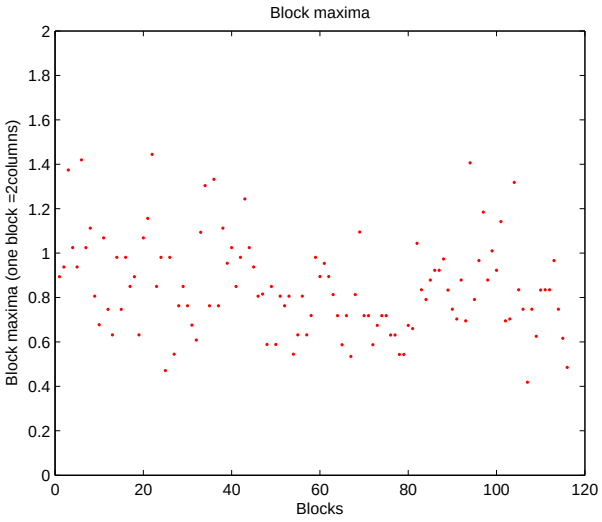

Figure 1.4: Measured block maxima

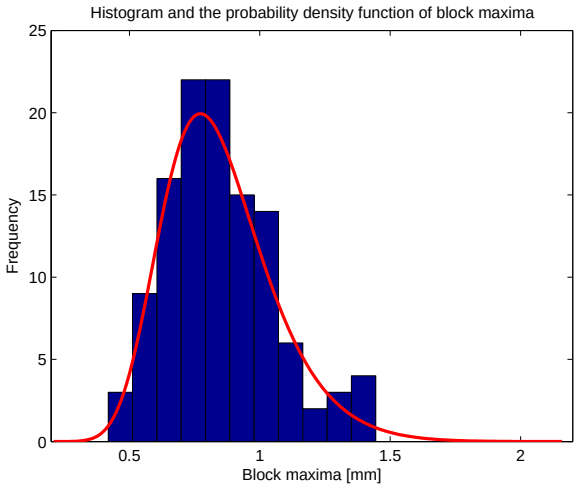

Figure 1.5: Histogram and the fitted probability density function of the GEV dsitribution to block maxima data

$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.082 (-0.211; 0.046)	0.182 (0.158; 0.210)	0.757 (0.719; 0.794)	0.713	0.986

Table 1.1: GEV fit to block maxima data

The fact that $\hat{\xi} < 0$ implies that the support of the GEV distribution is bounded from above by $\hat{\eta} = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} = 2.970\ [mm]$. This means that regardless of the size of the area to which we want to extrapolate the results, the predicted wall loss should never be greater than 2.970 $[mm]$. However, the 95% confidence bounds for $\hat{\eta}$, equal to $(0; 6.312)$[1], are very wide. Figure 1.6 and the high p-value of goodness-of-fit tests (see Table 1.1) both confirm that the GEV is well fitted to data.

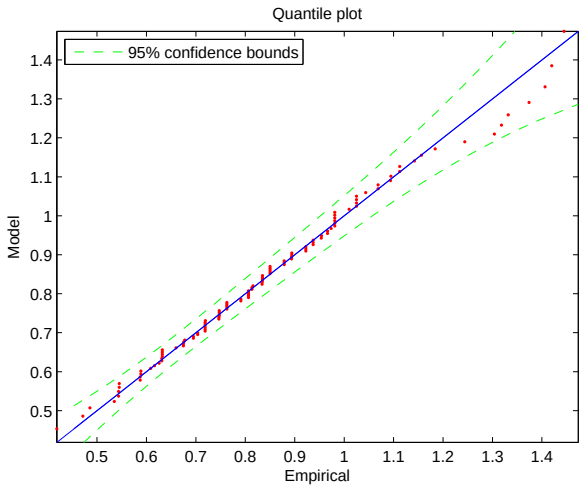

Figure 1.6: Quantile plot for the GEV distribution

In order to extrapolate the results to the not inspected area the two introduced methods in Section 1.1.3 will be used. The return level and the implied distribution of the extreme wall loss on the not inspected part of the pipe will be determined.

To calculate the return level, first we have to determine the so-called return period M (see Section 1.1.3) which is just a size multiple of one block. Since data was gathered from 13.51 $[m]$ of the pipe of 300 $[m]$ total length then by block

[1] in fact the exact interval is $(-0.373; 6.312)$, the left point of the interval is negative due to the delta method used

definition $M = \frac{300\ [m] - 13.51\ [m]}{2 \times dx} \approx 2470$, where $dx = 0.058\ [m]$ (here we can also interpret M as the number of not inspected blocks on the remaining part of the pipe). Then $p \approx 0.0004$ and the corresponding return level $\hat{z}_p$ along with 95% confidence bounds based on the profile likelihood method are presented in Table 1.2.

$\hat{z}_p$	95% *confidence bounds*
1.806	(1.553; 2.528)

Table 1.2: Estimated return level and profile likelihood based confidence interval [mm]-GEV distribution

Hence, the wall loss that is expected to be exceeded on average once on the not inspected part of the pipe is 1.806 $[mm]$ $(1.553; 2.528)$. It is worth to mention that the observed maximum wall loss during inspection was 1.447 $[mm]$.
The implied cumulative distribution function of the extreme pit depth on the not inspected part of the pipe, together with the calculated return level and 95% confidence bounds, are presented in Figure 1.7.

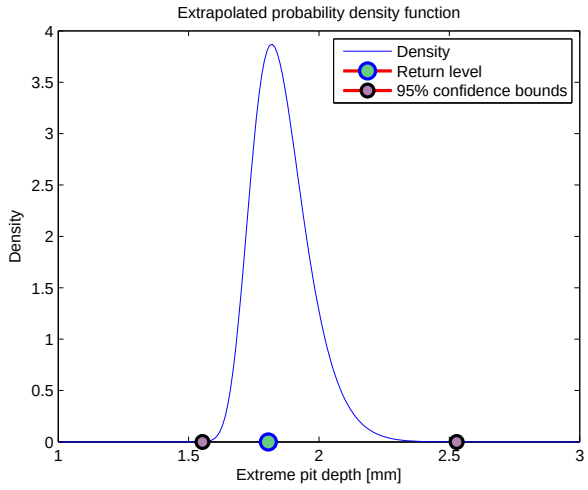

Figure 1.7: Extrapolated GEV distribution

As we can see both methods gave very consistent results. The calculated return level equal to 1.806 $[mm]$ is very close to the mode (for definition see Appendix A) of the implied (extrapolated) probability density function.

In the next section we will present another approach to statistical modelling of extreme values, namely Peaks Over Threshold method and the generalised-Pareto distribution.

1.2 Generalised Pareto Distribution

Another approach to modelling the statistical behaviour of extreme events is based on the generalised-Pareto distribution. This model is applicable to **excess over threshold data**. Similarly as for the GEV distribution, we will describe the methods for parameter estimation and extrapolation. However, some additional and important topics will be mentioned as threshold selection or the so-called Poisson frequency of threshold exceedances. We will end this section showing the application of the described methods, to the data used in Example 1.1.4.

When we use the generalised-Pareto distribution, as extreme we regard those events X_i from a sequence $\{X_1, X_2, \ldots\}$ of independent and identically distributed random variables with common distribution function F, which exceed some high threshold u (see Figure 1.8).

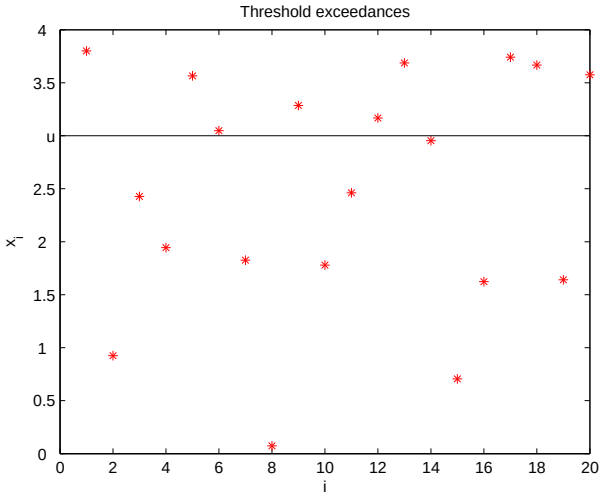

Figure 1.8: Exceedances over threshold

Let X denote an arbitrary term in a sequence $\{X_1, X_2, \ldots\}$. Then we define the excess over threshold u as $X - u$, for $X > u$. If the distribution function F of X would be known, then the distribution of the excess over threshold u could be easily calculated from:

$$Pr\{X - u > y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}$$

However in practise, F is not known and approximation methods are needed. It can be shown (Coles 2001) that under appropriate conditions, for large enough value of u, the distribution function of $X - u$ conditioned on $X > u$ is approximately within the Generalised Pareto family of distributions, given in Definition 1.4.

Definition 1.4 *A random variable Y is said to have the* ***generalised-Pareto distribution*** *(GP) with shape parameter ξ and scale parameter $\bar{\sigma}$, if its cumulative distribution function is given by:*

$$H(y) = \begin{cases} 1 - \left[1 + \dfrac{\xi y}{\bar{\sigma}}\right]_+^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left(-\dfrac{y}{\bar{\sigma}}\right), & \xi = 0 \end{cases} \tag{1.18}$$

where $[z]_+ = \max(0, z)$, and $y > 0$.

The corresponding probability density function is given by:

$$h(y) = \begin{cases} \dfrac{1}{\bar{\sigma}}\left[1 + \dfrac{\xi y}{\bar{\sigma}}\right]_+^{-1/\xi - 1}, & \xi \neq 0 \\ \dfrac{1}{\bar{\sigma}} \exp\left(-\dfrac{y}{\bar{\sigma}}\right), & \xi = 0 \end{cases} \tag{1.19}$$

The support of the GP distribution is bounded by $\gamma = -\dfrac{\sigma}{\xi}$, what is given in (1.20):

$$\begin{aligned} 0 < y < \gamma \qquad & \text{when } \xi < 0 \\ 0 < y < \infty \qquad & \text{when } \xi \geq 0 \end{aligned} \tag{1.20}$$

The mean and variance are given by:

$$E(Y) = \begin{cases} \dfrac{\bar{\sigma}}{1-\xi} & ,\xi < 1, \quad \xi \neq 0 \\ \bar{\sigma}, & \xi = 0 \end{cases} \tag{1.21}$$

$$Var(Y) = \begin{cases} \dfrac{\bar{\sigma}^2}{(1-\xi)^2(1-2\xi)} & ,\xi < 1/2, \quad \xi \neq 0 \\ \bar{\sigma}^2, & \xi = 0 \end{cases} \tag{1.22}$$

To model the extreme events using the GP distribution we proceed as follows. From the measurements $x_1, x_2, \ldots, x_n$, extreme events are identified as those exceeding a high threshold u, i.e. $\{x_i : x_i > u\},\ i = 1, \ldots, n_u$. Next we label these exceedances by $x_1, x_2, \ldots, x_{n_u}$ and define the threshold excess by $y_i = x_i - u$ for $i = 1, \ldots, n_u$, to which the GP distribution is fitted.

1.2.1 Poisson Frequency of Threshold Exceedances

The Poisson frequency of threshold exceedances leads to the so-called Poisson-GP model (Smith 2003). This model is closely related to the block maxima data and the GEV distribution. Indeed, if $X_1, \ldots, X_n$ is a sequence of independent and identically distributed random variables corresponding to one block then it can be shown that:

- the number N of exceedances of the threshold u in any block has a Poisson distribution with rate λ_{GEV};

- conditionally on $N \geq 1$, the excess values $Y_i = X_i - u$, for $i = 1, \ldots, N$, are independent and identically distributed random variables from the GP distribution.

If M denotes a maximum of one block ($M = \max_{1 \leq i \leq N}(Y_i) + u$) then it can be shown (for details see Appendix A) that for $x > u$

$$Pr\{M \leq x\} = \exp\left\{-\lambda_{GEV}\left(1 + \xi\frac{x-u}{\bar{\sigma}}\right)_+^{-1/\xi}\right\} \tag{1.23}$$

where $x_+ = \max(x, 0)$.

If we substitute (1.24):

$$\bar{\sigma} = \sigma + \xi(u - \mu),\ \lambda_{GEV} = \left(1 + \xi\frac{u-\mu}{\sigma}\right)^{-1/\xi} \tag{1.24}$$

to (1.23) then it reduces to the GEV distribution given in (1.2). It is worth to underline that $\bar{\sigma}$ and λ_{GEV} in (1.24) are the scale parameter of the GP distribution and the rate of exceedances above threshold u corresponding to the area of one block, respectively. The parameters of the GP distribution are uniquely determined by those of the GEV distribution. In particular, the shape parameter in (1.18) is equal to that of the corresponding GEV

This relationship implies that once we fit the GEV distribution to block maxima data and determine the threshold u, we get extra information about the rate of threshold exceedances on the area of one block and about the parameters of the GP distribution.

If for some reason the block maxima data is not available, but instead we have excess over threshold data $y_i = x_i - u$ for $x_i > u$ and $i = 1, \ldots, n_u$, gathered from area of size $|S|$, then the maximum likelihood estimate of the rate of exceedances

above threshold u is given by:

$$\hat{\lambda}_{GP} = \frac{n}{|S|} \tag{1.25}$$

It is worth to stress the difference between λ_{GEV} and λ_{GP}. The first one, i.e. λ_{GEV} denotes the rate of exceedances above threshold u in the area corresponding to one block (average number of exceedances per block), whereas λ_{GP} is the rate of threshold exceedances per area unit of S, where S is the part of the system that was inspected.

The Poisson frequency of the threshold exceedances can be determined by means of different models depending on the data format. In case of block maxima data, the GEV-based model is used, whereas for excess over threshold data, the GP-based approach is preferred. It is worth to stress that the distribution of $X - u$ conditioned on $X > u$ is approximately within the Generalised Pareto family of distributions for a large enough value of threshold u. Therefore, methods of threshold selection are needed, which we introduce in next section.

1.2.2 Threshold Selection Methods

The issue of the threshold selection implies a balance between a bias and variance. A too low threshold is likely to violate the asymptotic basis of the model, leading to bias; a too high threshold will result in few excesses with which the model could be estimated, leading to a high variance (Coles 2001). Thus we want to determine a threshold as low as possible, preserving the asymptotic properties of the model. To do so, we can use two approaches. The first is based on threshold selection prior to model estimation and uses a **mean residual life plot**. The other one is an assessment of the **parameter stability** based on the fitting of the model across the range of thresholds. It is worth to stress that if the GP distribution is a valid approximation of the distribution of

the excess over threshold u_0, then it is valid also for all $u > u_0$.

The first method of threshold selection is based on the mean of the GP distribution (1.21). It can be shown (see Appendix A) that if the GP distribution is valid for some threshold u_0 then the estimate of the mean should change approximately linearly with u for all $u > u_0$. In other words if we plot the points:

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} x_i - u \right) : u < \max_{1 \le j \le n} (x_j) \right\} \tag{1.26}$$

where $x_1, \ldots, x_{n_u}$ consists of n_u points from measurements $x_1, \ldots, x_n$, that exceed the threshold u, the resulting plot, called the mean residual life plot should be approximately linear above threshold u_0 for which the GP is a valid approximation. The confidence intervals can be added to the plot based on approximate normality of the sample mean (for details see Appendix A).

The second method is based on the estimation of the model at a range of thresholds. Above the level u_0 for which the asymptotic motivation for the GP distribution is valid, estimates of the shape parameter should be approximately constant, while the estimates of $\bar{\sigma}$ should be linear in u. If we denote $\bar{\sigma}_u$ as a scale parameter corresponding to threshold u, then by (1.24) for $u > u_0$ we get:

$$\bar{\sigma}_u = \bar{\sigma}_{u_0} - \xi u_0 + \xi u$$

Let us reparameterise $\bar{\sigma}$ according to:

$$\sigma^* = \bar{\sigma}_u - \xi u \tag{1.27}$$

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

the confidence bounds to the plot can be added (for details see Appendix A).

1.2.3 Parameter Estimation

Similarly as for the GEV distribution we introduce the maximum likelihood method of parameter estimation.

If $y_1, \ldots, y_k$ are excesses over threshold u then the log-likelihood is given by:

$$l(\xi, \bar{\sigma}) = \begin{cases} -k\log(\bar{\sigma}) - (1 + 1/\xi)\sum_{i=1}^{k}\log(1 + \xi y_i/\bar{\sigma}), & \xi \neq 0 \\ -k\log(\bar{\sigma}) - \frac{1}{\bar{\sigma}}\sum_{i=1}^{k} y_i, & \xi = 0 \end{cases} \tag{1.28}$$

provided $1 + \xi y_i/\bar{\sigma} > 0$ for $i = 1, \ldots, k$; otherwise $l(\xi, \bar{\sigma}) = -\infty$.

To find the estimates of the unknown parameters we maximise (1.28) with respect to the parameter vector $(\xi, \bar{\sigma})$. The solution is found numerically.

The goodness-of-fit will be assessed using the same tools as for the GEV (see Section 1.1.2).

1.2.4 Extrapolation

Like for the GEV distribution the two approaches for the results extrapolation will be used.

Return-level method.

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

$$y_p = \begin{cases} \frac{\bar{\sigma}}{\xi}\left[p^{-\xi} - 1\right], & \xi \neq 0 \\ -\bar{\sigma}\log(p), & \xi = 0 \end{cases} \tag{1.29}$$

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

be rewritten as $y_p + u$.

In terms of the rate of threshold exceedances the return level (1.29) is given in (1.30):

$$y_p = \begin{cases} \frac{\bar{\sigma}}{\xi}\left[\left(\frac{1}{\lambda_{GP} \times |E|}\right)^{-\xi} - 1\right], & \xi \neq 0, \\ -\bar{\sigma}\log\left(\frac{1}{\lambda_{GP} \times |E|}\right), & \xi = 0, \end{cases} \tag{1.30}$$

where $|E|$ is the size of not inspected part of the system and $\lambda_{GP} \times |E|$ is the expected number of threshold exceedances on the area E.

Analogously as for the GEV return-level extrapolation, the confidence bounds for the determined return-level y_p, can be determined using the profile likelihood method (for more details see Appendix A).

Distribution function of the maximum corresponding to the not inspected area

Using the relation (1.23), the cumulative distribution function of the random variable X_M, i.e. of the maximum corresponding to the area which is M size multiple of one inspected block is given by:

$$\begin{aligned} Pr\{X_M \leq x\} &= \exp\left\{-\lambda_{GEV}\left(1 + \xi\frac{x-u}{\bar{\sigma}}\right)_+^{-1/\xi}\right\}^M \\ &= \exp\left\{-M \times \lambda_{GEV}\left(1 + \xi\frac{x-u}{\bar{\sigma}}\right)_+^{-1/\xi}\right\}, \end{aligned} \tag{1.31}$$

where $x > u$.

1.2.5 Example of Application

In this section, we will show the application of the GP distribution to data set used in Example 1.1.4.

First, we will try to determine a proper value of threshold u. After the parameter estimation, we will assess the goodness-of-fit of the GP distribution to excess data. At the end, the extrapolation results will be presented.

To determine a proper value of threshold u, we apply the two methods introduced in Section 1.2.2. The plots of the mean residual life and of the estimated ξ and σ^* for a range of thresholds are presented in Figures 1.9, 1.10 and 1.11, respectively. Additionally, we will perform the goodness-of-fit test (Section 1.1.2) for different threshold values.

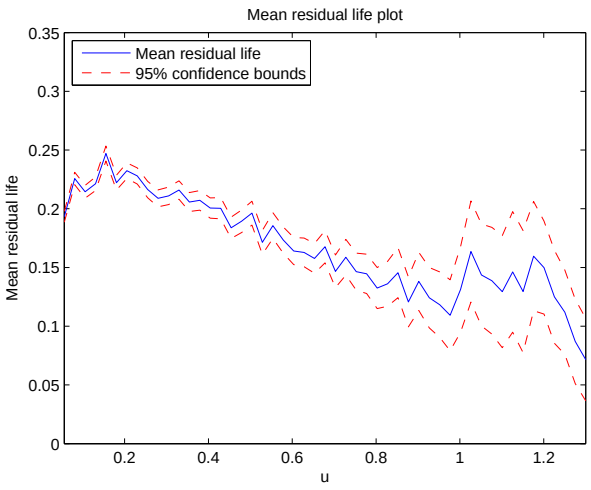

Figure 1.9: Mean residual life plot

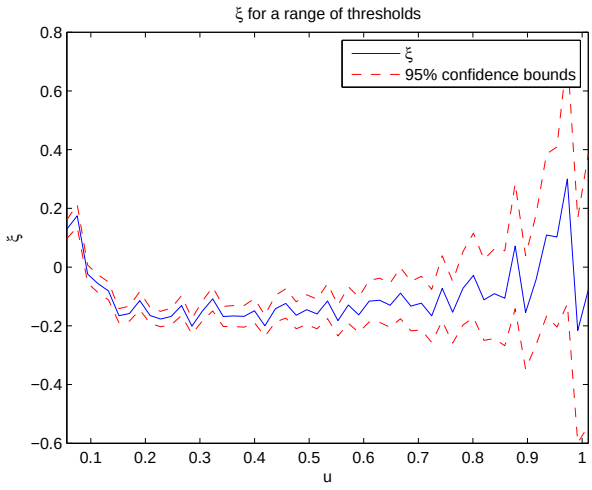

Figure 1.10: Estimate of ξ at a range of thresholds

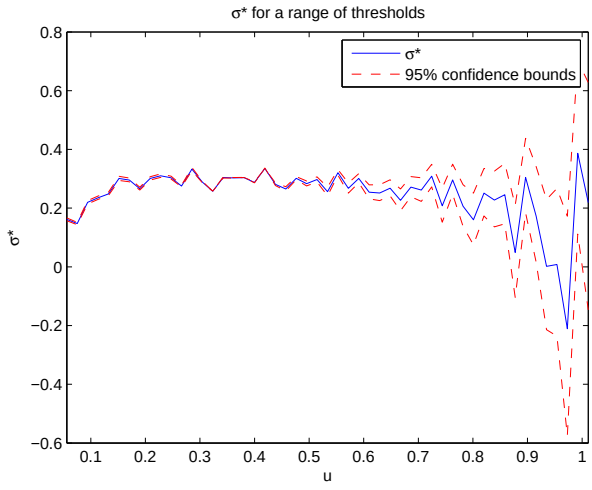

Figure 1.11: Estimate of σ^* at a range of thresholds

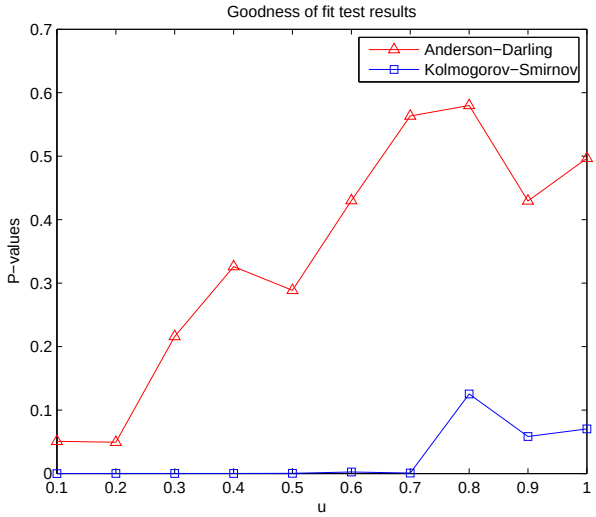

Figure 1.12: Goodness-of-fit test results for the GP distribution for different threshold values u

Remind that if u_0 is a proper threshold (i.e. is such a threshold that the cumulative distribution function of the excesses $Y_i = X_i - u_0$ given $X_i > u_0$ for $i = 1, \ldots, k$ can be approximated by the GP distribution) then for all $u > u_0$ the mean residual life plot should be linear in u and the estimates of ξ and σ^* should be approximately constant. From Figures 1.9, 1.10, 1.11 and 1.12, it follows that the threshold equal to $u_0 = 0.8$ $[mm]$ is a reasonable choice. Then, on the inspected part of the pipe there are 216 exceedances above u_0. The histogram and the results of fitting the GP distribution to excess above $u_0 = 0.8$ $[mm]$ data are presented in Figure 1.13 and Table 1.3.

$\hat{\xi}$	$\hat{\tilde{\sigma}}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.037 (-0.178; 0.104)	0.1400 (0.115; 0.170)	0.597	0.125

Table 1.3: GP fit to excess data

Figure 1.14 and high p-value of Anderson-Darling goodness-of-fit test (see Table 1.3),

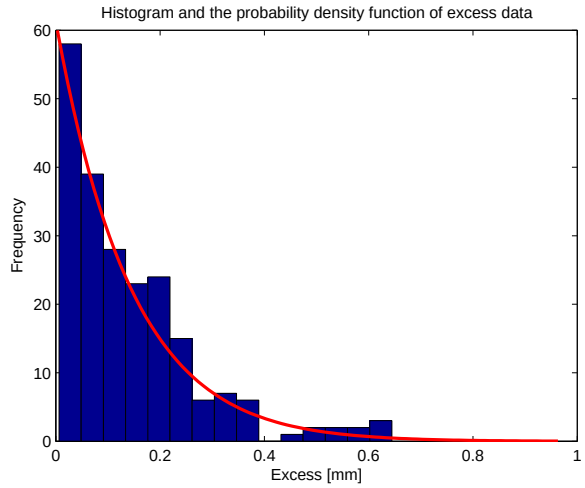

Figure 1.13: Histogram and the probability density function of the GP distribution fitted to excess data

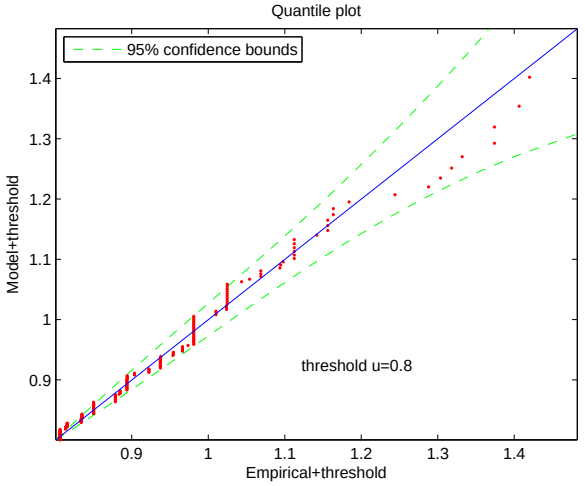

Figure 1.14: Quantile plot for the GP distribution

both confirm that the GP is reasonably fitted to data.

To extrapolate to the not inspected part of the pipe, we have to determine the corresponding expected number of threshold exceedances. If N_S denotes the

number of threshold exceedances on the surface S then the expected value of N_S is given by:

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

where λ is the rate of exceedances per area unit of S and $|S|$ is the size of S. Using (1.25), we get:

$$\hat{\lambda}_{GP} = \frac{216}{13.51\ [m] \times 0.8\ [m]} \approx 19.98/\ m^2$$

Then, the expected number of defects (wall loss) on the not inspected part of the pipe whose depths exceed 0.8 $[mm]$ is given by:

$$E(N_{not\ insp.}) = 19.98 \times 0.8 \times (300 - 13.51) \approx 4579$$

Using (1.29) with $p = \dfrac{1}{E(N_{not\ insp.})} = 0.00022$, we get the estimate of the return-level given in Table 1.4.

$\hat{z}_p$	95% $confidence\ bounds$
1.812	(1.549; 2.622)

Table 1.4: Estimated return level [mm]-GP distribution

This means that on the not inspected part of the pipe, the wall loss that is expected to be exceeded on average once is 1.812 $[mm](1.549; 2.622)$. Moreover looking at Figure 1.15, we can see that the estimated return level is quite robust against threshold selection.

Additionally, we compare the probability density function of the block maxima, obtained through the GP distribution with the probability density function of the block maxima corresponding to the GEV. From Figure 1.16 we can see

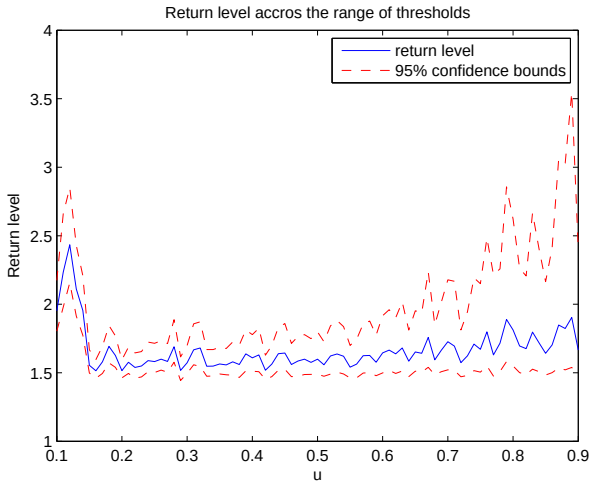

Figure 1.15: Estimate of return level at a range of thresholds along with 95% confidence bounds derived by means of the profile likelihood method

that they are slightly different. It is also worth to stress that the calculated

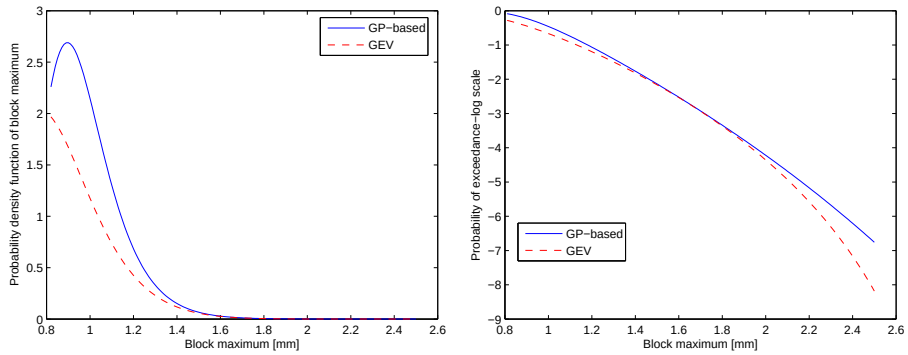

Figure 1.16: Comparison of the probability density functions of the block maximum

return level is not very far from the one in Example 1.1.4. The difference in the estimate of confidence bounds for the return level $\hat{z}_p$ is caused by the wider confidence bounds obtained for shape parameter $\hat{\xi}$ in the GP distribution than for the same parameter in the GEV distribution.

1.2.6 Summary

In this chapter, we introduced and showed a real life application of two methods for extreme-value analysis of corrosion data. Depending on the data format, it is possible to use either the GEV distribution (block maxima data) or the GP distribution (threshold exceedances data). These two distributions are closely related and give consistent results. However, both models rely on the assumption that the measurements are independent. In real life corrosion defects on the surface seem to be locally dependent (one pit/defect can influence the growth of neighbouring pits/defects). Fortunately, these models can be applied also in the case of local defect dependence. The extension of applicability of the GP and GEV models to locally dependent data will be introduced in Chapter 2.

Chapter 2

Extreme-Value Analysis of Corrosion Data with Locally Dependent Defect Depths

While applying the GEV or GP models to corrosion data we had to assume that the underlying observations, i.e. defect depths, are independent. However, this is not the case in practise because it is likely that growth of one defect can influence growth of the others, the neighbouring defects. Thus, it is said that the defects are locally dependent. It can be shown (Leadbetter, Lindgren & Rootzen 1983, Coles 2001) that if the underlying observations are stationary and locally dependent, then the statistical behaviour of extreme events can be modelled by extreme-value distributions (for definition of stationarity see Appendix B). The dependence condition (called $D(u_n)$[1] condition, for definition see Appendix B) requires that the extent of long-range dependence at extreme levels is limited. It means that extreme events are approximately independent when they are far enough apart. In application to corrosion this means that if

[1] For the purpose of illustration, we introduce the dependence condition for one dimensional data. However, it can be extended to higher dimensions. For more details see Turkman (2006)

two extreme depth defects are far away from each other they can be considered as approximately independent. This is a reasonable assumption with respect to corrosion. This motivates of the usage of extreme-value distributions to model corrosion data.

The extent of local dependence at extreme levels is captured by the parameter called *extremal index*. It measures the degree of clustering, i.e. the tendency of extreme events to occur in groups. Therefore, the extremal index is often interpreted as the propensity of the limiting mean cluster size.

We will start the chapter introducing the notion of extremal index. Then, we will define the extremal index for stationary random fields. In the next section, the approaches to model data extremes by means of the GEV and GP distributions will be described. It will be shown that for locally dependent data prior to fitting the GP distribution, the data declustering has to be done. For this purpose the two methods will be introduced. One is based on the extremal index parameter the other uses clustering algorithm with validation criteria with respect to the number of determined data clusters. For completeness, we will introduce also the clustering tendency test. At the end, some illustrative examples, based on real and simulated data, will be presented. We will end the chapter with brief a summary of the results.

2.1 The notion of extremal index

The **extremal index**, $\theta \in (0, 1]$, is a measure of the propensity of the stationary process to cluster at extreme levels (see Figure 2.1). It is often written as (Coles 2001):

$$\theta := \frac{1}{limiting\ mean\ cluster\ size} = \frac{n_c}{n_e} \tag{2.1}$$

where n_c is the number of clusters and n_e is the number of exceedances above some high threshold u.

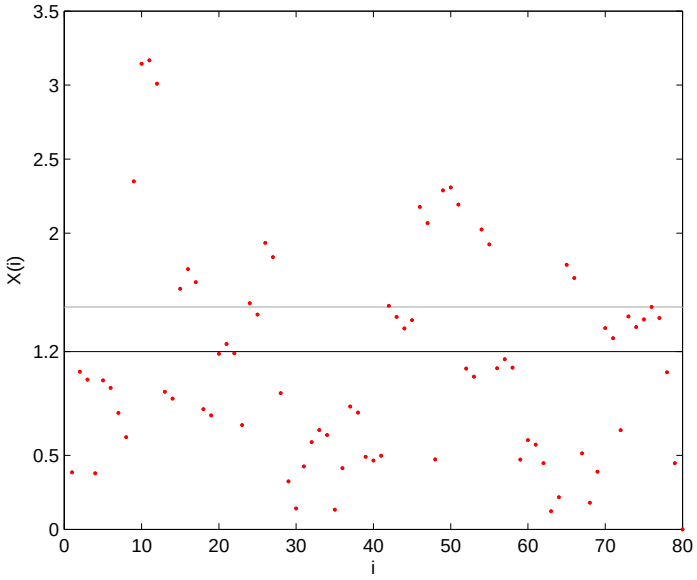

Figure 2.1: Example of stationary sequence $\{X_i\}$ with tendency to cluster at extreme levels. $X_0 = Y_0$, $X_i = \max\{aY_{i-1}, Y_i\}$ for $i = 1, \ldots, n$, where a=0.95, n=160 and $Y_i \sim gamma(1.2, 0.8)$

If $\{X_1, X_2, \ldots\}$ is a sequence of independent random variables then $\theta = 1$, but the converse is not true. For stationary sequences, the lower the value of θ, the higher the local dependence, and the higher the tendency to clustering at extreme levels.

The notion of extremal index can be extended to higher dimensions (Turkman 2006, Leadbetter & Rootzen 1998), For a two-dimensional stationary random field $X(i, j)$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ (for definition see Appendix B) the extremal index θ is defined as:

$$\theta = \theta_1 \times \theta_2 \tag{2.2}$$

where $\theta_1,\ \theta_2$ correspond to the x-coordinate and y-coordinate clustering propensity, respectively. To be more precise, θ_1 indicates how the largest values of the y-columns, i.e.:

$$V_i = \max_{1\leq j\leq n_2}\{X(i,j)\},\ i=1,\ldots,n_1$$

cluster along the x-direction. On the other hand, θ_2 indicates how the large values of the process

$$W_{x_0} = X(x_0,1), X(x_0,2),\ldots,X(x_0,n_2)$$

cluster along the y-direction at a fixed x-coordinate.

Theoretically speaking, since the random field is assumed to be stationary the value of θ_2 should not depend upon the choice of x_0. However, in real life applications the variation in the estimate of θ_2 due to the choice of x_0 is possible. Therefore we propose to calculate θ_2 not for a chosen column process but for a process composed of column processes. We define this process as:

$$\begin{aligned} W &= X(1,n_2), X(1,n_2-1),\ldots,X(1,1), X(2,1), X(2,2),\ldots,X(2,n_2) \\ &\ldots\ X(3,n_2), X(3,n_2-1),\ldots,X(3,1),\ldots,X(n_1,1), X(n_1,2),\ldots,X(n_1,n_2) \end{aligned} \tag{2.3}$$

The way we read the values of the random field on the lattice grid to define the process W is schematically presented in Figure 2.2.

In order to estimate the extremal index for spatial corrosion data (in the format as presented in Example 1.1.4) we use the following approach:

- store data in $n\times m$ matrix $A=[a_{ij}],\ i=1,\ldots,n,\ j=1,\ldots,m$;

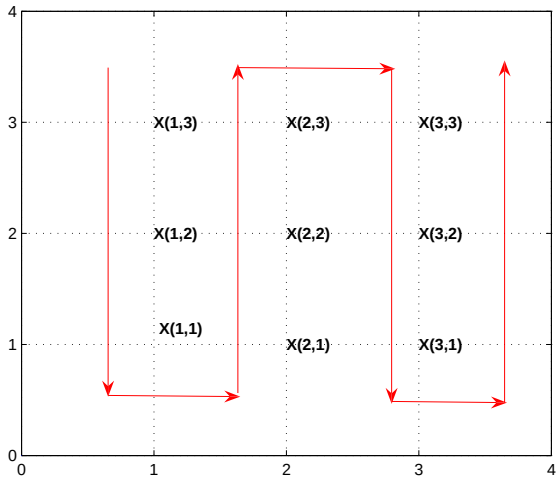

Figure 2.2: Definition of column process W

- estimate the extremal index θ_1 for the sequence of column maxima

$$v_j = \max_{1 \le i \le n} \{a_{ij}\}, \ j = 1, \ldots, m$$

- estimate extremal index θ_2 for the series composed from columns merged in the following order:

$$a_{11}, a_{21}, \ldots, a_{n1}, a_{n2}, a_{n-12}, \ldots, a_{12}, a_{13}, a_{23} \ldots, a_{n3}, a_{n4}, a_{n-1,4}, \ldots, a_{14}, \ldots, a_{nm}$$

- the estimate of the extremal index for the whole data set is given by $\hat{\theta} = \hat{\theta}_1 \times \hat{\theta}_2$

The method of extremal index estimation is the issue of next section.

2.1.1 Extremal Index Estimation

There are many extremal index estimators (Galbraith & Zernovi 2006). Some of them require an *a priori* choice of the parameters whose values can significantly influence the estimated value of θ. In this section we will present the **intervals estimator** that does not require secification of any parameters (Ferro 2002, Ferro 2003).

Let $x_1, \ldots, x_n$ be a realisation from the stationary process and let u be a high threshold and N be the number observations exceeding u. Let

$$1 \leq S_1 <, \ldots, < S_N \leq n$$

be the exceedance times. Then inter-arrival times are defined as

$$T_i = S_{i+1} - S_i,\ i = 1, \ldots, N-1$$

and the intervals estimator for the extremal index is given by:

$$\hat{\theta} = \begin{cases} \min\left\{1, \dfrac{2\left(\sum_{i=1}^{N-1} T_i\right)^2}{(N-1)\sum_{i=1}^{N-1} T_i^2}\right\} & if\ \max\{T_i : 1 \leq i \leq N-1\} \leq 2 \\ \min\left\{1, \dfrac{2\left(\sum_{i=1}^{N-1} (T_i - 1)\right)^2}{(N-1)\sum_{i=1}^{N-1} (T_i-1)(T_i-2)}\right\} & if\ \max\{T_i : 1 \leq i \leq N-1\} > 2 \end{cases} \tag{2.4}$$

To estimate the extremal index θ, we have to choose a high enough threshold value u. By definition of θ for spatial data and by (2.3), we can base the choice of u on the methods presented in Section 1.2.2. Thus, we can apply the methods of mean residual life and parameter stability.

It is worth to mention that the higher the threshold value, the higher the value of θ and the lower the number of exceedances.

2.2 Modelling Extremes of Dependent Data

In this section, we will present two approaches to modelling extremes of stationary and locally dependent data.

2.2.1 Modelling block maxima data

Modelling block maxima of stationary and locally dependent data does not differ from modelling block maxima of stationary but independent data (Coles 2001). This follows from the fact that if the long-range dependence at extreme levels is very weak (what is true if the $D(u_n)$ condition holds; for more details, see Appendix B), block maxima can be considered as approximately independent. Therefore to estimate the unknown parameters of the GEV, we still can use the maximum likelihood method introduced in Section 1.1.1. The estimated parameters will include information about data local dependence (for more details, see Appendix B).

We can reasonably assume that corrosion data is characterised by a limited extent of long-range dependence. Therefore we can still apply the GEV method introduced in Section 1.1.

2.2.2 Modelling excess over threshold data

For modelling the statistical behaviour of excesses over threshold of stationary data satisfying assumptions about the long-range dependence, we can still use the generalised-Pareto distribution (Coles 2001). However, since neighbouring exceedances may be dependent we cannot use maximum likelihood estimation

to this data directly. The most widely used method to solve this problem is **data declustering** which is based on filtering out dependent observations such that the remaining exceedances are approximately independent. This suggests the following framework for modelling excesses over threshold of stationary data with the generalised-Pareto distribution:

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

The key issue of data declustering is cluster identification. There are many algorithms that can be used to find a given number of clusters in data (Everitt, Landau & Leese 2001). However, if we will wrongly specify this number, the algorithms can group data into artificial clusters. This can create a different data structure than the true one. Therefore, the estimation of the number of clusters is of great importance. We propose to do it in two ways:

- Estimate the number of clusters prior to finding any clusters in data. Use the fact that the extremal index θ (in the limit sense) can be interpreted as the reciprocal of the mean cluster size. Hence, by (2.1):

$$n_c = \theta \times n_e \tag{2.5}$$

  where n_c is the number of clusters and n_e is the number of exceedances above some high threshold u.

- Run cluster algorithms for a range of n_c. As proper value of n_c, choose the one that optimises given validity criteria.

Thus, we have two approaches to answer the question about the proper number of clusters in data. The first one uses the extremal index method and is applied

prior to cluster identification; the other one is based on the clustering algorithm used to find the given number of clusters in data. As proper value of n_c we choose the one that optimises given validity criteria. The above implies two important remarks. Firstly, we can compare the results given by the two methods to see if they are consistent. Secondly, if for some reason the estimation of the extremal index θ is difficult or not possible (this will be shown later in Section 2.4.2), we can still use the other method to find data clusters.
The clustering method that we want to use is the issue of next section.

2.3 Clustering method

There are many methods that can be used to find clusters in data. In this section however, we are going to shortly present only one of them, namely **agglomerative hierarchical clustering** method.
The agglomerative hierarchical clustering method produces a series of partitions of data into clusters (Everitt et al. 2001). Starting with single points as individual clusters, and at each step the closest pair of clusters is merged. The last partition consists of the one cluster including all data points. In this way we create multilevel hierarchy where clusters at one level are joined into new clusters at the next higher level.
To decide which clusters to merge we need the definition of cluster **proximity**. We have to define the rule that will tell us which clusters are close to each other and which are far apart. There are many definitions of proximity between objects. Some of the standard are single link, complete link and group average (Tan, Steinbach & Kumar 2006). The single link rule defines proximity between clusters as the proximity between the two closest points that are in different clusters. For the complete link, we look at the proximity of the two furthest points in different clusters. The group average rule defines cluster proximity as

the average of pairwise proximities of all pairs of points from different clusters. The above are illustrated in Figures 2.3, 2.4, 2.5.

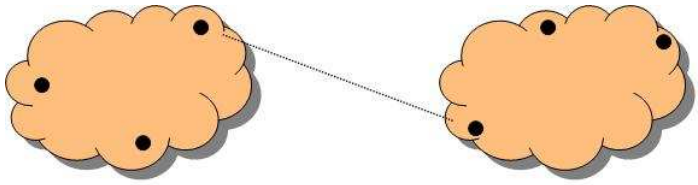

Figure 2.3: Single link

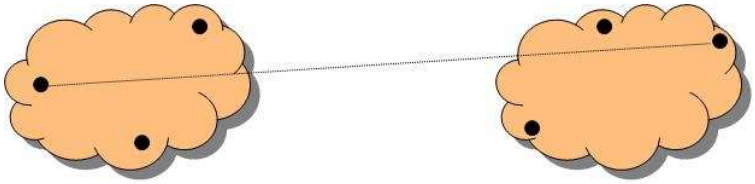

Figure 2.4: Complete link

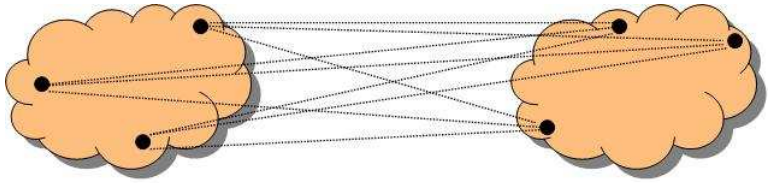

Figure 2.5: Average link

In later sections we will try the following proximity measures between objects: **single linkage** - also called nearest neighbour, uses the smallest distance between objects in the two clusters

$$d(r,s) = \min_{i,j}(dist(x_{ri}, x_{sj})),\ i = 1,\ldots,n_r,\ j = 1,\ldots,n_s \tag{2.6}$$

where r and s are clusters and $x_{ri} \in r$, $x_{sj} \in s$

complete linkage - also called furthest neighbour, uses the largest distance between objects in the two clusters

$$d(r,s) = \max_{i,j}(dist(x_{ri}, x_{sj})),\ i = 1,\ldots,n_r,\ j = 1,\ldots,n_s \tag{2.7}$$

average linkage - based on the average distance between all pairs of objects in two clusters

$$d(r,s) = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \tag{2.8}$$

centroid linkage - based on the Euclidean distance between centroids of the two clusters

$$d(r,s) = dist(\bar{x}_r, \bar{x}_s) \tag{2.9}$$

where $\bar{x}_r = \dfrac{1}{n_r}\sum_{i=1}^{n_r} x_{ri}$

median linkage - uses the Euclidean distance between the weighted centroids of the two clusters

$$d(r,s) = dist(\tilde{x}_r, \tilde{x}_s) \tag{2.10}$$

where $\tilde{x}_r$ and $\tilde{x}_s$ are weighted centroids of clusters r and s. If the cluster r was created by combining clusters p and q, then $\tilde{x}_r$ is defined recursively as $\tilde{x}_r = \dfrac{1}{2}(\tilde{x}_p + \tilde{x}_q)$.

Note that some of the above proximity measures are defined in terms of the Euclidean distance (centroid linkage and median linkage). Since we deal with spatial data, we will use the Euclidean distance measure as *dist*.

The choice of a particular linkage determines the hierarchical clustering algorithm. Therefore, in we will test which one performs the best for the type of data we have.

To better understand the idea of hierarchical clustering let us consider a simple example. Suppose there are given five points (objects) on the lattice grid presented in Figure 2.6.

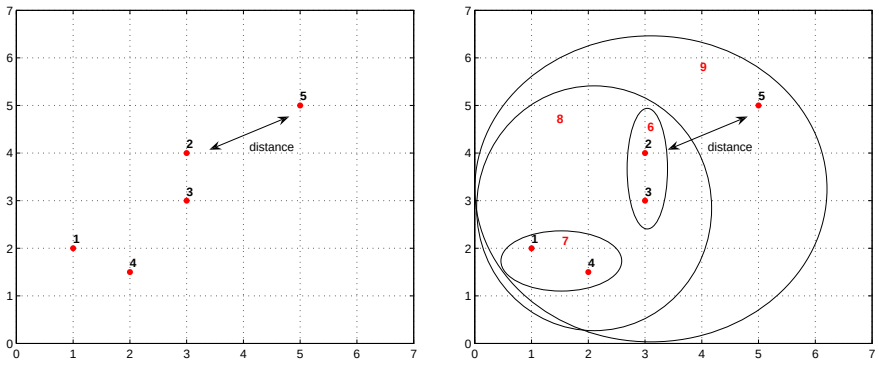

Figure 2.6: Example of hierarchical grouping

The algorithm works as follows. At the first step, all five object are treated as clusters. At this level, the objects number 2 and 3 form the closest pair of clusters; thus, they will be merged into one cluster, creating object number 6. In the second step, the objects 1 and 4 are merged creating object 7. At the last step, objects 8 and 5 are joined what leads to one data cluster, denoted in Figure 2.6 by number 9. The process of merging objects is visualised in the diagram called **dendrogram** (Figure 2.7).

The numbers along the horizontal axis are labels of the original objects (data points) and the vertical represents the distance between clusters. The links between objects are represented as upside-down U-shaped lines. The agglomerative hierarchical clustering method produces a tree with the top cluster containing all the data points. Therefore, finding the right number of clusters means

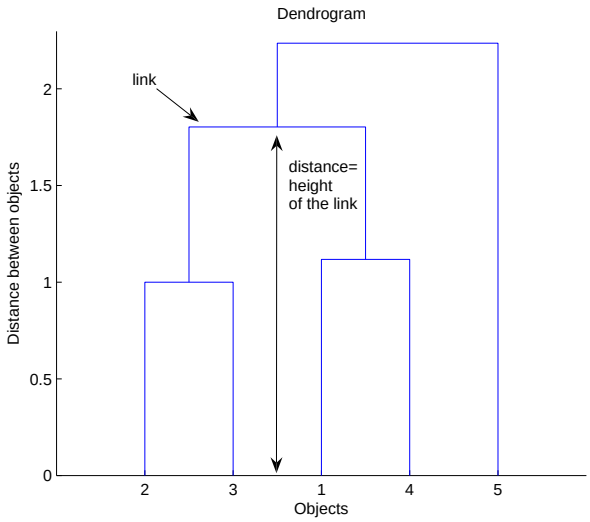

Figure 2.7: Example of dendrogram

cutting the tree at some level. This is directly connected with cluster validation methods which will be described in the next section.
Summing up the cluster identification for spatial corrosion data by means of the agglomerative hierarchical clustering method will be performed in the following steps:

- calculate the distance between every pair of points with wall loss above threshold u;
- use linkage rule (e.g. single link or group average) to build up the hierarchical tree;
- use cluster validation methods to cut the tree at the proper level.

To decide which proximity measure to choose for identifying clusters in data and assess the overall goodness of clustering, we can use the so-called **cophenetic correlation coefficient** (Tan et al. 2006). In a hierarchical cluster tree, any two

objects in the original data are eventually linked together at some level. The height of the link reflects the distance between the two clusters that contain those two objects (see Figure 2.7). This height is called the cophenetic distance between the two objects. The cophenetic correlation coefficient measures the correlation between the entries of the distance matrix of the original data set and the entries of the cophenetic distance matrix produced by the hierarchical clustering algorithm. The entries of the distance matrix corresponding to the original data set are pairwise distances between data points. If the clustering is valid then there should be strong correlation between the cophenetic distances and the distances corresponding to the original data points. Therefore the closer the value of the cophenetic correlation coefficient to 1, the better the found clusters reflect the natural clusters present in data.
In further examples we will use the group average link between clusters because for this measure and data type the calculated cophenetic correlation coefficient was the highest.

2.3.1 Cluster Validation Methods

There are many cluster validation methods that can be helpful to determine the proper number of clusters in data. Two commonly used are the **silhouette plot** and the **Davies-Bouldin index**. Our choice of these two particular methods is motivated by the ease of calculations and good experimental results.
The silhouette plot shows the relation between the number of clusters and the average silhouette coefficient. The silhouette coefficient for each point can be computed in the following steps (Tan et al. 2006):

- For the i^{th} point calculate the average distance to all the other points in its clusters, call this value a_i;
- For the i^{th} point and any cluster not containing the point, calculate its

average distance to all the points in the given cluster. Let b_i be the minimum of such values with respect to all clusters;

- For the i^{th} point, the silhouette coefficient is given by:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{2.11}$$

The silhouette coefficient takes values between -1 and 1. The negative values are undesirable because in this case $a_i > b_i$ what corresponds to the situation where the average distance to points in a cluster is greater to the minimum average distance to point in another cluster. We want a_i to be as close to zero as possible because the coefficient attains the value of 1 when $a_i = 0$. The overall measure of the goodness of clustering can be computed as the average silhouette coefficient of all points. Then the best (according to this measure) number of clusters for a given data set is the one that maximises this value.

Similarly as the average silhouette coefficient, the Davies-Bouldin index aims at identifying the sets of clusters that are compact and well isolated. The value of the index is calculated according to the formula (Bolshakova & Azuaje 2003):

$$DB(U) = \frac{1}{n_c} \sum_{i=1}^{n_c} \max_{i \neq j} \left\{ \frac{\triangle(X_i) + \triangle(X_j)}{\delta(X_i, X_j)} \right\} \tag{2.12}$$

where U is a partition of the data set into n_c clusters $X_1, \ldots, X_{nc}$ and $\triangle(X_i)$, $\delta(X_i, X_j)$ are the measures of intracluster and intercluster distance, respectively.
For further analysis, as the intercluster distance measure, we take the centroid

distance which is given by:

$$\delta(X_i, X_j) = d(C_{X_i}, C_{X_j}) \tag{2.13}$$

where $C_{X_i} = \frac{1}{|X_i|} \sum_{z \in X_i} z$, $C_{X_j} = \frac{1}{|X_j|} \sum_{z \in X_j} z$ are the centroid locations of clusters X_i and X_j respectively and $d(\cdot\,,\cdot\,)$ is the Euclidean distance measure.

As intracluster distance measure, we will use the so-called centroid diameter:

$$\triangle(X_i) = 2 \left(\frac{\sum_{z \in X_i} d(z, C_{X_i})}{|X_i|} \right) \tag{2.14}$$

There are other possible intercluster and intracluster distance measures (Bolshakova & Azuaje 2003). Our choice is supported by the computational ease and very similar results (obtained for data we have) to the ones given by the other measures (e.g. average distance measure).

Small values of the Davies-Bouldin index correspond to the case when clusters are compact and their centres are far from each other. Therefore, small values of the $DB(U)$ index are desired and the optimal number of clusters is the one for which $DB(U)$ attains its minimum.

2.3.2 Clustering Tendency Test

In the previous sections we introduced methods to estimate the number of clusters in data. One method is based on the extremal index θ, the other one uses cluster algorithms together with some validation criteria. If the corrosion data

is spatially stationary then we can use the two methods and compare their results. If the estimate of θ is less than 1, it means that there is a clustering tendency in data. This motivates further cluster identification and data declustering. However, when data is not spatially stationary (even if it is stationary at extreme levels as in Figure 1.1) the estimation of the extremal index may not be possible and the clustering algorithm with validation criteria are used. The cluster algorithm we use will always find clusters given data. Of course we can validate which number of clusters is the best but then still there is a possibility to cluster data that does not possess the natural clustering tendency. To solve this problem we can try to evaluate whether the data set has clusters, without running clustering algorithm. The most common approach, especially for data in Euclidean space, is to use the statistical test for spatial randomness (Tan et al. 2006). One of such a test, simple but powerful (Benerjee & Davae 2004, Tan et al. 2006) is based on the **Hopkins statistics**. The test is based on the idea of choosing at random M, so-called sampling origins (i.e. points randomly distributed across the data space, where $M << N$ and N is the number of points in space). Then actual M data points, called marked points are sampled. For both sets, the distance to the nearest neighbour in the original data set is calculated. If u_i is denoted as the nearest neighbour distance of the sampling origins and w_i as the nearest neighbour distance of the marked points from original data set, the Hopkins statistic is defined as:

$$H = \frac{\sum_{i=1}^{M} u_i^2}{\sum_{i=1}^{M} w_i^2 + \sum_{i=1}^{M} u_i^2} \tag{2.15}$$

It has been suggested (Benerjee & Davae 2004) that when $M < 0.1 \times N$, then all the $2M$ nearest neighbour distances are statistically independent and H has

a beta distribution with parameters (M, M). Values of H close to 1 indicate that the data is highly clustered, values around 0.5 indicate randomness and values close to 0 suggest regular data spacing.

2.4 Examples of Application

In this section, we are going to show the application of the above methods to two kinds of data. First, we will apply the GEV and GP tools to dependent simulated data. The next application will be based on a real data set, the same as used in Example 1.1.4.

What we want to show is the comparison of the results given by the GEV and GP applied to dependent data and declustered data. Remind, that if the data is locally dependent then we can fit the GEV distribution directly and for the GP distribution prior to model fitting the data declustering has to be done. Then, theoretically speaking, the estimates of the shape parameters of the GEV distribution fitted to the original data and the GP distribution fitted to the declustered data should be the same. Moreover, we expect that fit of the GP distribution to original data will give overestimated results with respect to extrapolation.

2.4.1 Simulated Corroded Surface

In this example, we want to apply the introduced methods to a simulated corroded surface. The data is a matrix (200×500) of locally dependent defect depths. They were generated by the gamma-process model, described in detail by Ostrowska (2006). The basic idea behind the model is to generate dependent defect depths from the gamma distribution with prescribed dependence

structure between them. The dependence is expressed in terms of the product moment correlation coefficient defined as:

$$\rho(X_k, X_l) = \exp\left\{-d\left(\sum_{i=1}^{2}|dist_i|^p\right)^{q/p}\right\} \tag{2.16}$$

where $X_k = (x_k, y_k)$ and $X_l = (x_l, y_l)$ are points on the Cartesian grid, $dist_1 = |x_k - x_l|$, $dist_2 = |y_k - y_l|$, d is the parameter regulating the strength of dependence and p and q are the parameters associated with the vector norm. For the purpose of illustration, we used $d = 0.3$, $p = 2$, $q = 1$ (in this case (2.16) is based on the Euclidean norm) and a gamma distribution with shape parameter $a = 0.1$ and scale parameter $b = 0.5$. The corresponding strength of correlation is presented in Figure 2.8. We can see that the defect dependence will be local.

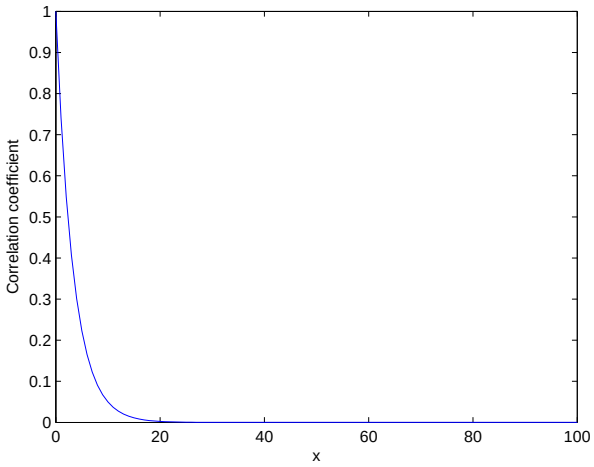

Figure 2.8: Strength of correlation (in one direction) for $d = 0.3,\ p = 2,\ q = 1$

It is worth to mention that the data will be generated in one step and the

simulated defect depths will correspond to a accumulated in time deterioration process.

The simulation of the corroded surface with prescribed dependence structure is computationally expensive and was restricted to the dimension 50×50 which is not enough for the purpose of our example. To create a larger data set, we merged 40 smaller matrices 50×50 where each matrix has the desired local dependence structure. This is shown schematically in Figure 2.9.

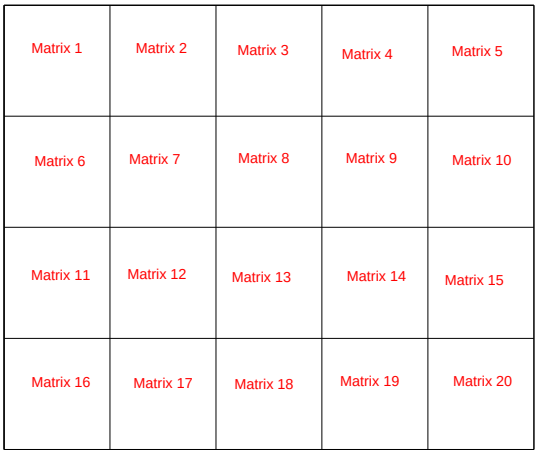

Figure 2.9: Data set generation

One can easily notice that merging matrices in this way does not preserve the dependence structure along the neighbouring matrix boundaries. This however does not add any extra dependence that would violate the assumption that defect depths are locally dependent (in fact defect depths along the neighbouring boundaries are independent). The resulting data set is presented in Figures 2.10[2] and 2.11.

[2]surface generated by courtesy of Mr MSc Ir. Sebastian Kuniewski, Delft Institute of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, The Netherlands

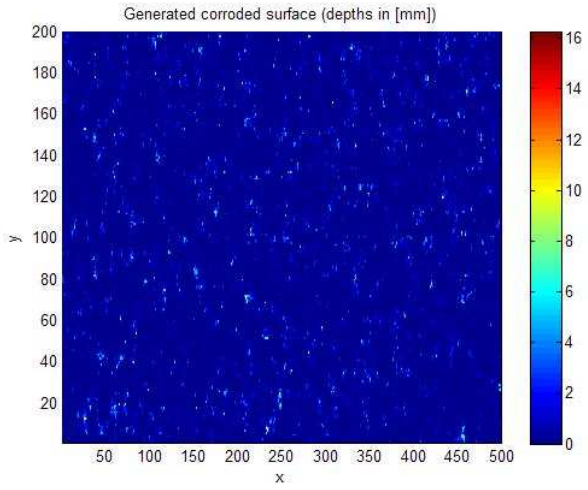

Figure 2.10: Generated corroded surface

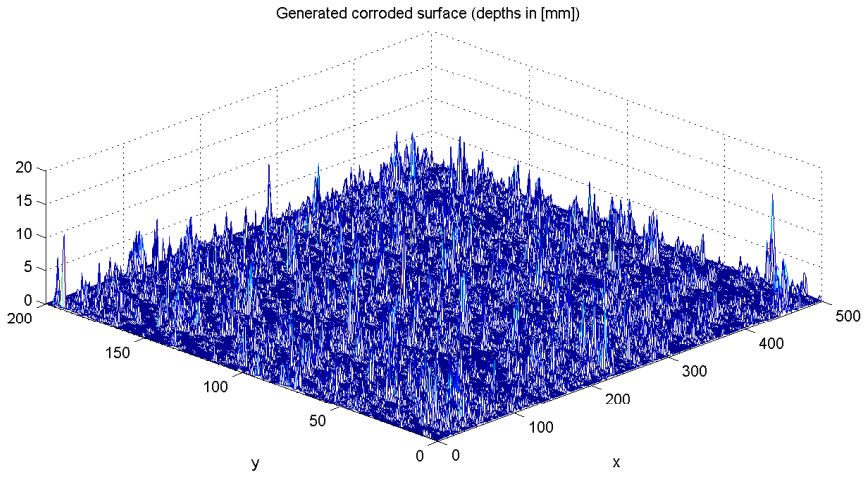

Figure 2.11: Generated corroded surface

In order to fit the GEV distribution to the data shown in Figure 2.10 we have to define blocks. Similarly as in Example 1.1.4, we can define one block as the number B_s of columns from the data matrix. To choose the proper number of columns per block, we perform the goodness-of-fit tests for different block

sizes. In Figure 2.12, we can see that there is no significant difference between p-values for block size 4 and 5, therefore for further analysis we will use $B_s = 4$ because then there are more observations to fit the GEV distribution.

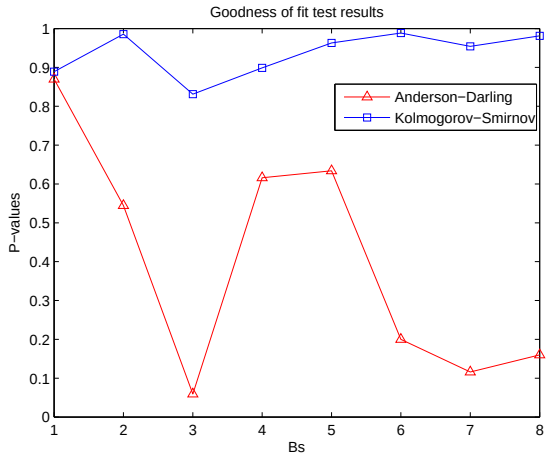

Figure 2.12: Goodness-of-fit test results for the GEV distribution for different block sizes. B_s-number of columns corresponding to one block

The fit of the GEV distribution to 125 block maxima obtained in this way, is presented in Table 2.1 and Figure 2.13.

$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.007	1.627	5.637	0.621	0.899
(-0.122; 0.107)	(1.414; 1.872)	(5.320; 5.955)		

Table 2.1: GEV fit to block maxima data

To extrapolate the results in space, we proceed as data would result from the screening technique as in Example 1.1.4 with $dy = 5$ $[mm]$, $dx = 58$ $[mm]$. This gives 1 $[m] \times 29$ $[m]$ of scanned area. Further, we assume that the total pipe area is 1 $[m] \times 300$ $[m]$. We want to extrapolate to the remaining part of the pipe.

The estimated maximal defect depth that is expected to be exceeded on the not

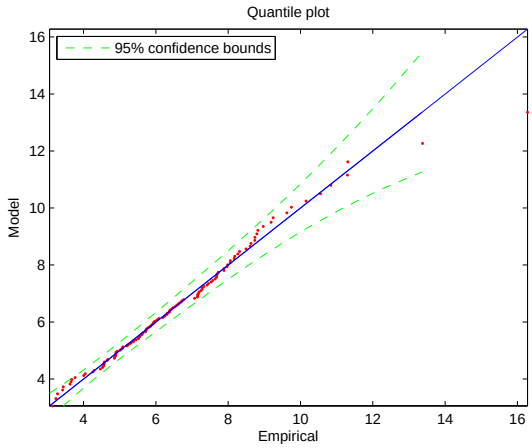

Figure 2.13: Quantile plot for the GEV distribution

inspected part of the system on average once, along with the 95% confidence interval is presented in Table 2.2.

$\hat{z}_p$	95% $confidence\ bounds$
16.834	(14.140; 23.904)

Table 2.2: Estimated return level and profile likelihood based confidence interval[mm]-GEV distribution

Further, we fit the GP distribution to original and declustered data. On the basis of Figures 2.14, 2.15 and 2.16 we choose the threshold value to be 5.6 $[mm]$. This gives 197 exceedances which are presented in Figure 2.17.

From Figure 2.17 we can see that there are some data clusters. The clustering tendency is indeed confirmed by statistical test, whose results are presented in Figure 2.18. On the significance level $\alpha = 0.05$, we reject the hypothesis that there is no data clustering (the calculated p-value is smaller than significance level $\alpha = 0.05$).

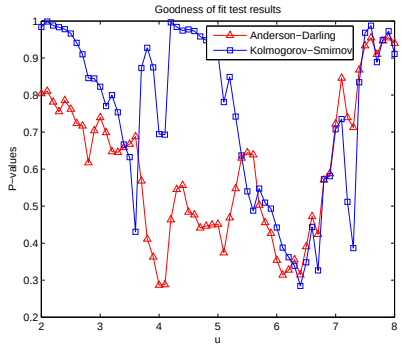

Figure 2.14: Goodness-of-fit test results for the GP distribution for different threshold values u

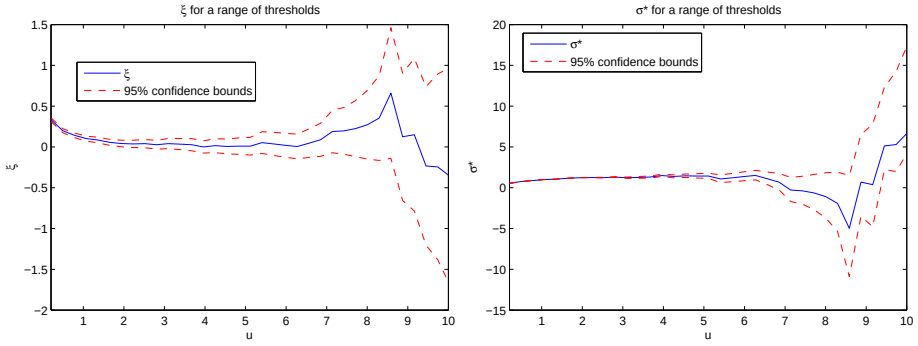

Figure 2.15: Estimate of ξ and σ^* for the range of thresholds

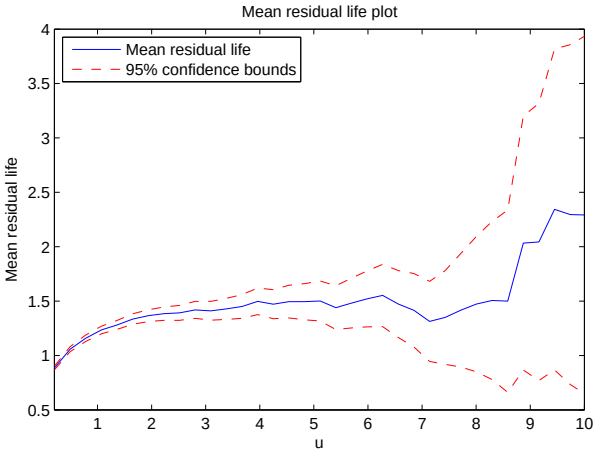

Figure 2.16: Mean residual life plot

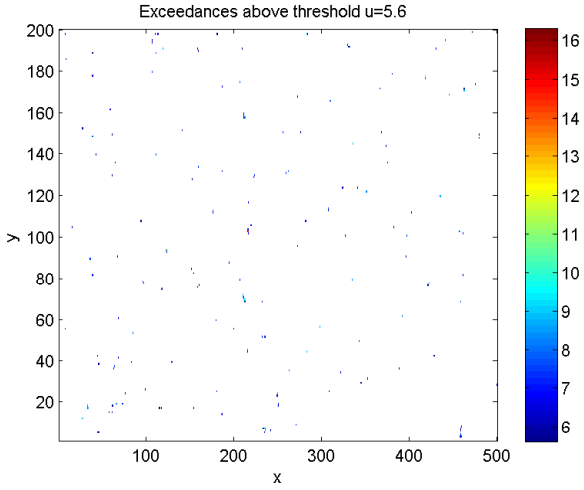

Figure 2.17: Exceedances over threshold $u = 5.6$ mm

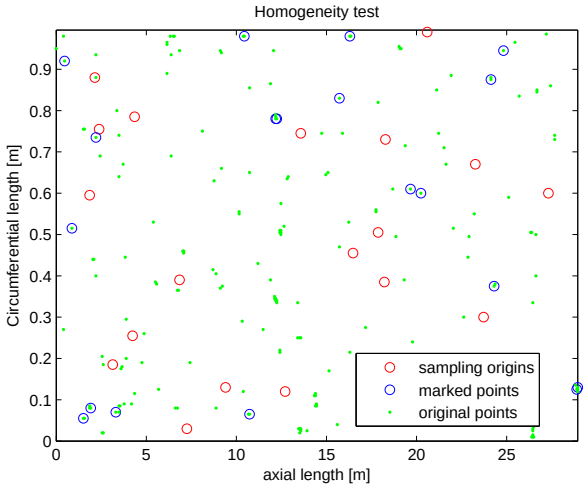

Figure 2.18: Clustering tendency test results. P-value =0.006, H=0.697

In the next step, we estimate the extremal index and the corresponding number of clusters for the range of thresholds (see Figures 2.19 and 2.20). We can see that in this example these results are quite robust against threshold choice.

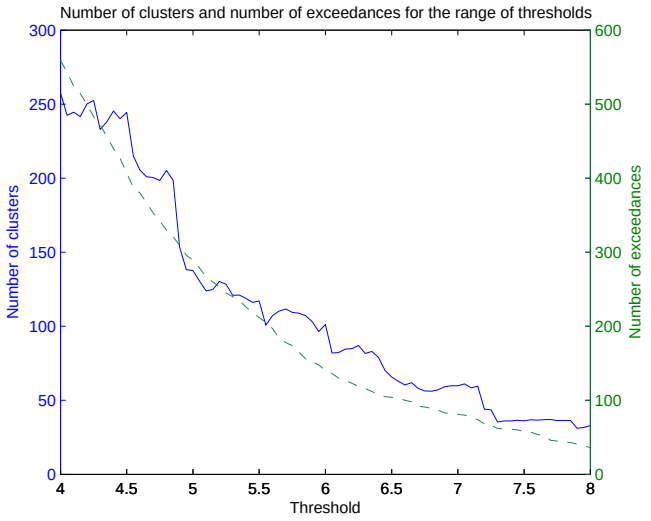

Figure 2.19: Estimate of the number of clusters and determined number of exceedances for the range of thresholds

Further using the clustering algorithm and the introduced validation criteria (see Section 2.3.1), we determine the number of data clusters. Diagnostic plots and the results are presented in Figure 2.21 and Table 2.3.

$\hat{\theta}$	n_c	$n_{c_{alg}}$
0.544	107	121

Table 2.3: The estimate of extremal index and determined number of clusters. n_c-determined number of clusters using extremal index method, $n_{c_{alg}}$-determined number of clusters using clustering algorithm

Since the estimated number of clusters given by the two methods differs we want to check which one is more correct. For this purpose, we decluster data assuming 107 and 121 clusters and for both cases perform goodness-of-fit tests.

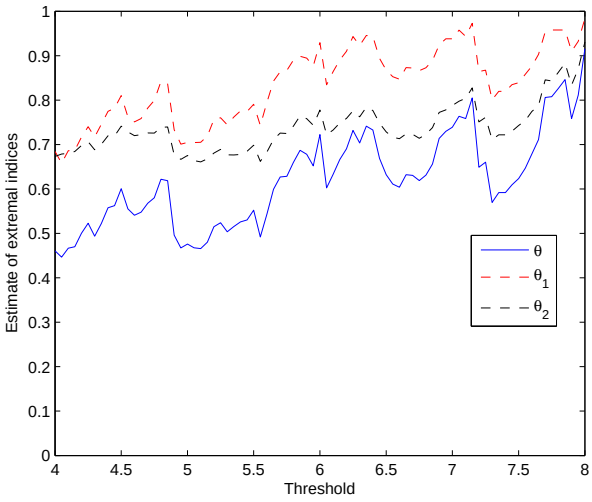

Figure 2.20: Estimate of the extremal index for the range of threshold

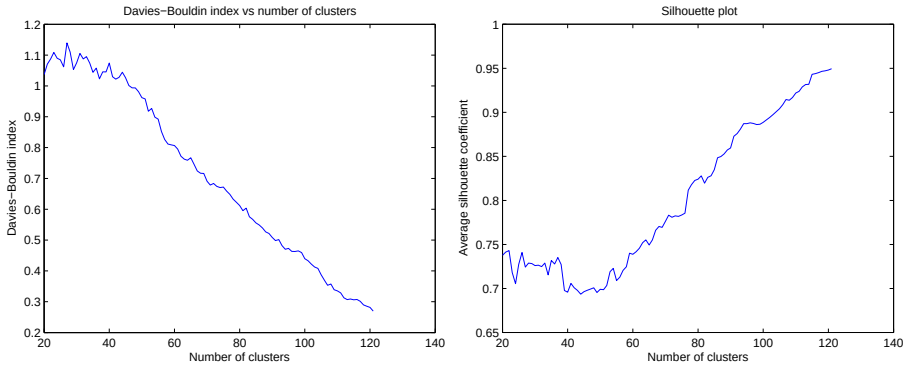

Figure 2.21: Number of cluster validation

The results are presented in Table 2.4.

Number of clusters	AD^2_{up} $p-v.$	KS $p-v.$
107	0.376	0.204
121	0.549	0.493

Table 2.4: Goodness-of-fit test results for different number of clusters

We can see that working with 121 instead 107 data clusters gives better model

fit results. Therefore for further analysis we assume that there are 121 data clusters.

The fit results of the GP distribution to dependent and declustered data are presented in Tables 2.5, 2.6 and Figure 2.22.

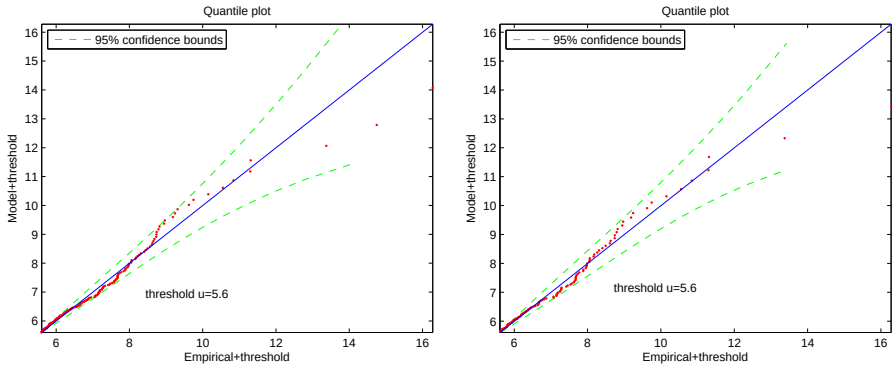

Figure 2.22: Quantile plot for the GP distribution. Left dependent data, right declustered data.

$\hat{\xi}$	$\hat{\tilde{\sigma}}$	AD^2_{up} $p-v.$	KS $p-v.$
0.065 (-0.082; 0.213)	1.341 (1.095; 1.643)	0.647	0.488

Table 2.5: GP fit to excess dependent data

$\hat{\xi}$	$\hat{\tilde{\sigma}}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.0137 (-0.177; 0.146)	1.6839 (1.325; 2.139)	0.555	0.493

Table 2.6: GP fit to excess of declustered data

We can see that data declustering influenced the results. The p-value of the goodness-of-fit test is slightly lower, but when we compare the estimate of the shape parameter we can see that the one corresponding to declustered data is very close (theoretically, they should be the same) to the estimate of the shape parameter for the fitted GEV distribution (Table 2.1). The effect of

data declustering is also visible in the extrapolation results (Table 2.7 and 2.8). In the case of declustering, the results are closer the ones given by the GEV model (Table 2.2). What is even more interesting is the fact that for the return level based on the GP distribution and declusterd data, we obtained narrower confidence bounds than for the return level determined on the basis of the original data. Even if declustering caused a reduction of the number of data points used to fit the model (what should imply wider confidence bounds), the change of the sign of the fitted shape parameter resulted in more accurate confidence bounds than the ones corresponding to original data.

$\hat{z}_p$	95% $confidence\ bounds$
18.623	(14.977; 30.865)

Table 2.7: Estimated return level and profile likelihood based confidence interval[mm]-GP distribution, dependent data

$\hat{z}_p$	95% $confidence\ bounds$
16.886	(14.177; 27.043)

Table 2.8: Estimated return level and profile likelihood based confidence interval[mm]-GP distribution, declustered data

2.4.2 Real Data

In this section we will use the same data set as presented in Example 1.1.4. This data is not spatially stationary (although it seems to be stationary a extreme levels). This is visible when we analyse the plot (Figure 2.23) of the part of column process W defined in equation (2.3). Therefore, in this case we cannot apply the extremal index method and we will use the introduced clustering algorithm to determine the number of clusters in data.

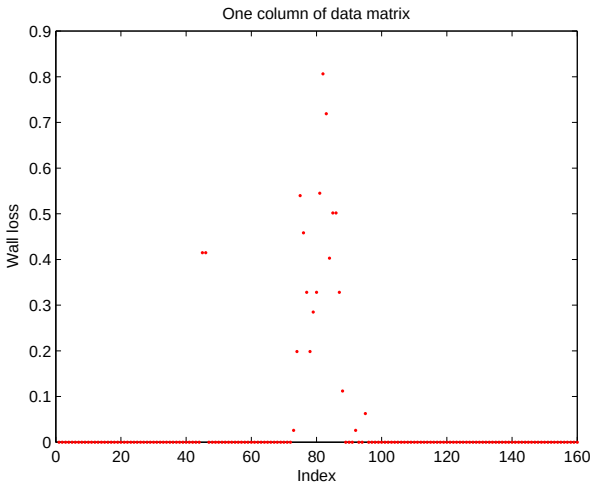

Figure 2.23: One column of data matrix

In Example 1.1.4 we chose the threshold $u_0 = 0.8\ [mm]$. This choice was mainly based on the goodness-of-fit test results. However, we have to remember that while fitting the GP distribution for different threshold values we did not take into account local data dependence and the results could be biased. Therefore, in this step we treat $u_0 = 0.8\ [mm]$ only as good threshold candidate. We are going to check goodness-of-fit test results for different threshold values (close to $0.8\ [mm]$) for declustered data. The results are presented in Table 2.9.

Threshold	$KS\ p-value$	$AD^2_{up}\ p-value$
0.78	0.803	0.628
0.79	**0.654**	**0.717**
0.8	0.767	0.681
0.81	0.692	0.646
0.82	0.625	0.618

Table 2.9: Goodness-of-fit test results for different threshold values and declustered data

Because we put more importance to the goodness-of-fit of the tail of distribution (which is more important for extrapolation) for further analysis as threshold value we take $u_0 = 0.79\ [mm]$. However, before fitting the GP distribution, we have to identify data clusters which are visible in Figure 2.24. The clustering tendency is additionally confirmed by Figure 2.25. On the significance level $\alpha = 0.05$, we reject the hypothesis that there is no data clustering.

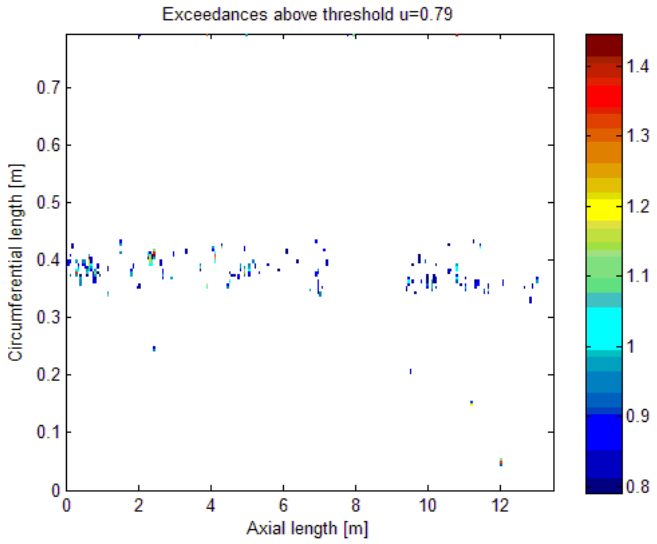

Figure 2.24: Exceedances above threshold $u = 0.79\ [mm]$

Using the clustering algorithm and the validation criteria, we find (see Figure 2.26) that there are 105 data clusters.

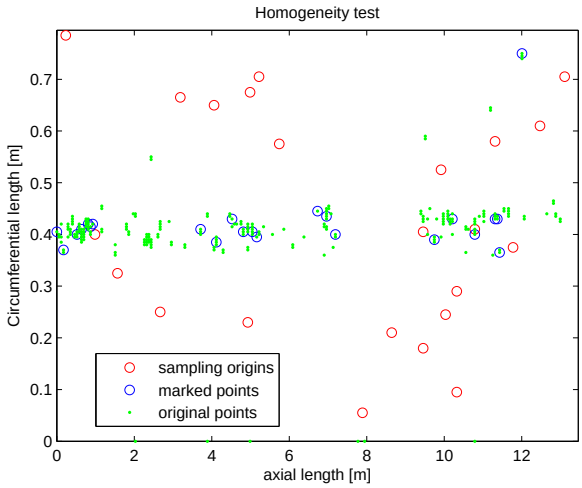

Figure 2.25: Clustering tendency test results. P-value =0, H=0.973

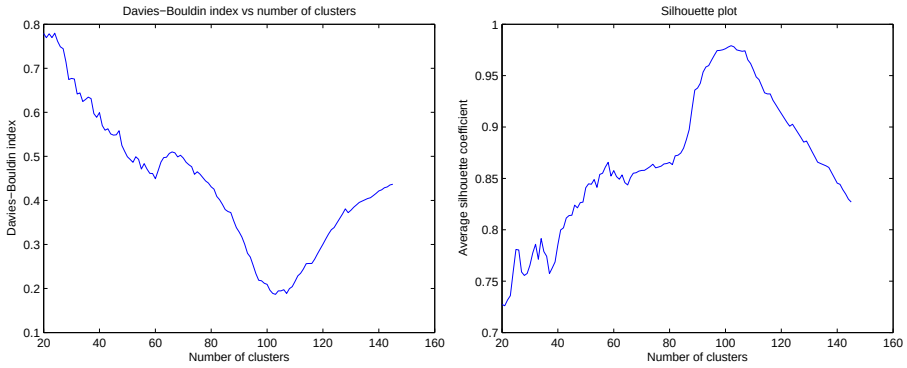

Figure 2.26: Number of cluster validation

The fit results of the GEV to block maxima and GP to dependent and declustered data are presented in Tables 2.10, 2.11 and 2.12, respectively.

$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.082	0.182	0.757	0.713	0.986
(-0.211; 0.046)	(0.158; 0.210)	(0.719; 0.794)		

Table 2.10: GEV fit to block maxima data

$\hat{\xi}$	$\hat{\tilde{\sigma}}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.008 (-0.150; 0.135)	0.133 (0.110; 0.161)	0.616	0.074

Table 2.11: GP fit to excess of dependent data

$\hat{\xi}$	$\hat{\tilde{\sigma}}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.069 (-0.286; 0.148)	0.172 (0.129; 0.229)	0.717	0.654

Table 2.12: GP fit to excess of declustered data

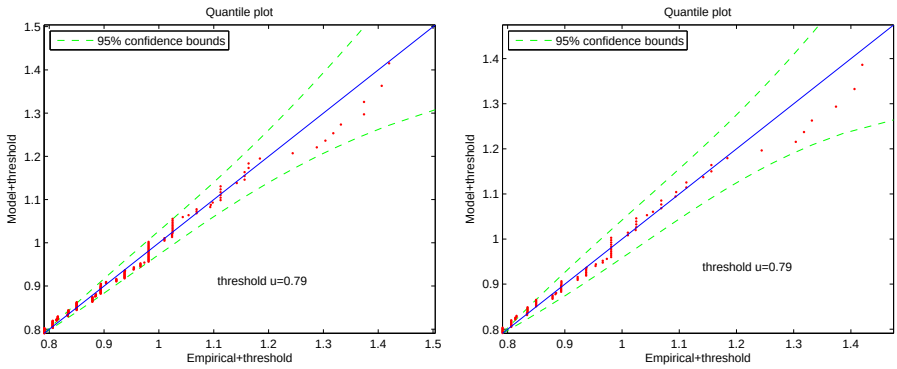

Figure 2.27: Quantile plot for the GP distribution. Left dependent data, right declustered data.

We can see that similarly to the previous example the data declustering improved the model fit. The estimate of the shape parameter of the GP fitted to declustered data is closer to the estimate of the shape parameter of the GEV distribution. Moreover, for declustered data the estimate of return level for the GEV and GP distribution are closer. However, data declustering reduces the number of observations used to fit the model. This results in wider confidence bounds for the return level determined (Table 2.15).

$\hat{z}_p$	95% $confidence\ bounds$
1.806	(1.553; 2.528)

Table 2.13: Estimated return level and profile likelihood based confidence interval [mm]-GEV distribution

$\hat{z}_p$	95% *confidence bounds*
1.889	(1.581; 2.856)

Table 2.14: Estimated return level and profile likelihood based confidence interval [mm]-GP distribution, dependent data

$\hat{z}_p$	95% *confidence bounds*
1.816	(1.508; 3.253)

Table 2.15: Estimated return level and profile likelihood based confidence interval [mm]-GP distribution, declustered data

In order to better understand the effect of data declustering, the probability density functions of the block maximum obtained through the GP distribution for dependent and declustered data, will be compared. Moreover we will plot the implied distribution functions corresponding to the maximum wall loss on the not inspected area. From Figures 2.28, 2.29, 2.30 and 2.31 we can see that in the case of declustered data the results given by the GP model are closer to the ones of the GEV.

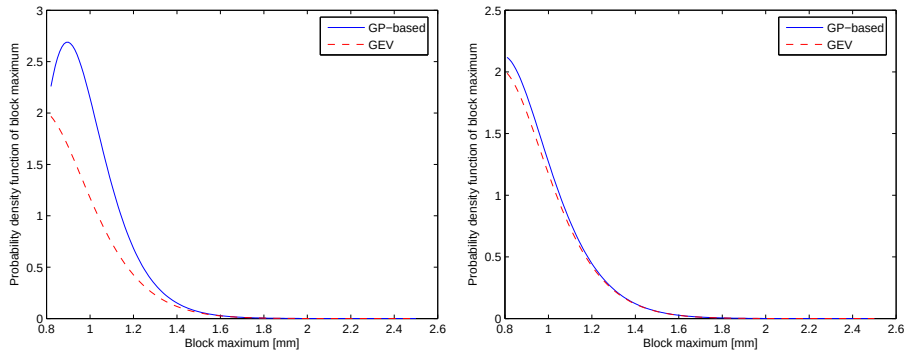

Figure 2.28: Comparison of the probability density functions of the block maximum. Left dependent data, right declustered data.

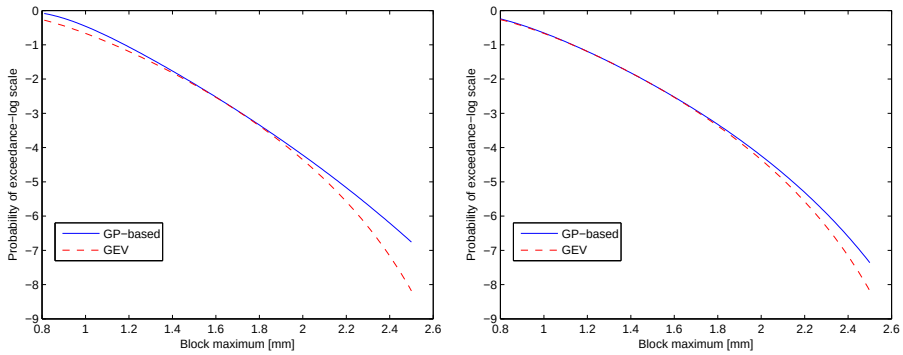

Figure 2.29: Comparison of the probabilities of exceedance for the block maximum. Left dependent data, right declustered data.

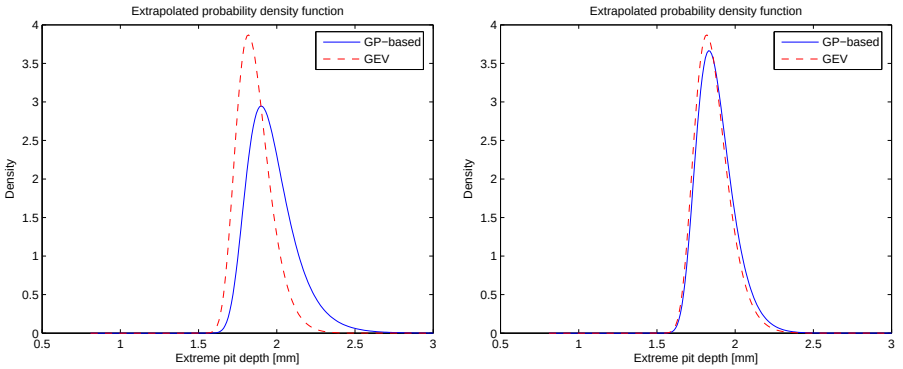

Figure 2.30: Comparison of the probability density functions of the maximum on the not inspected area. Left dependent data, right declustered data.

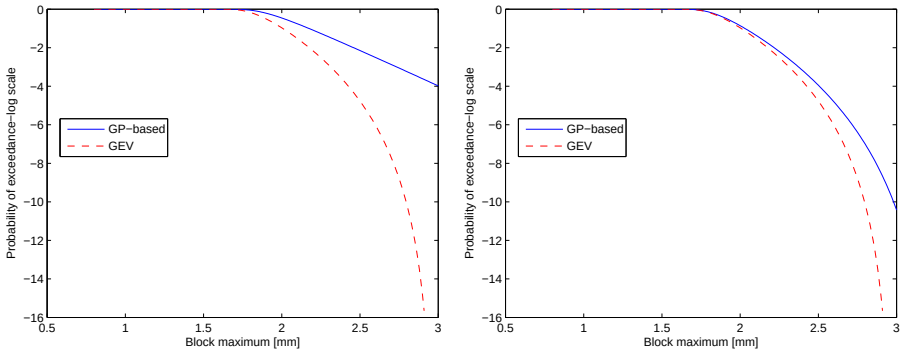

Figure 2.31: Comparison of the probabilties of exceedance for the maximum on the not inspected area. Left dependent data, right declustered data.

One of the questions we might have after going through the thesis is "Why should we bother about the GP distribution and data declustering, since we can have proper results fitting the GEV ?".
First of all, by applying the two methods to a given data set we can validate results since theoretically they should be consistent. The other reason is that it is not always possible or easy to apply the GEV distribution due to the need of the block definition. This is illustrated in the following example.

To show that it is not always straightforward to apply the GEV distribution we will present the data set resulting from a simulation of the Poisson cluster process (Diggle 1983). In general settings, the simulation of such a process can be summarised in steps as follows:

- sample the so-called parent events from a Poisson process with intensity λ_P;
- each parent produces a random number S of offspring, realised independently and identically for each parent according to a probability distribution $\{p_s : s = 0, 1, \ldots\}$;
- the positions of the offspring relative to their parents are independent and identically distributed according to a bivariate distribution.

The described procedure is a one-step simulation. However, to simulate data we performed the above simulation a number of times, i.e. 18. Moreover, we slightly modified the above procedure. In each step we sample parents and the corresponding offspring. Additionally sampled offspring in step t_{i-1} is treated as parents in step t_i and in this step can produce new offspring. The sampling space is the two dimensional lattice grid.
The parameters used are:

- $\lambda_P = \lambda t^q$, where $\lambda = 150,\ q = .2,\ t = 1, 2, \ldots, 18$;
- the random number S of the offspring is generated according to

$$S = \min(N,\ NC)$$

where N has a Poisson distribution with $\lambda_e = 0.18 \times NC$ and NC is the number of cells for possible extension. This is visualised in Figures 2.32, 2.33. Hence, for instance if some parent does not have any neighbours then $NC = 8$ and $S = \min(N,\ 8)$. Otherwise, as shown in Figure 2.33, NC is 8 diminished with the number of already existing neighbours (for the parent, denoted as green square $NC = 6$);

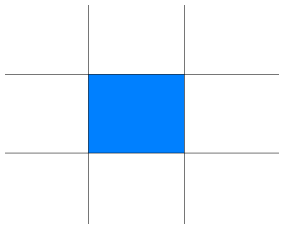

Figure 2.32: Explanation of offspring generation

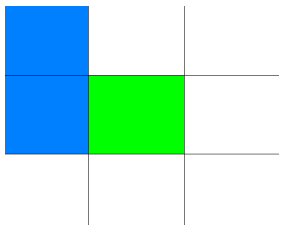

Figure 2.33: Explanation of offspring generation

- offspring positions are chosen randomly from a possible set of positions NC

The defect depths are sampled from a gamma distribution with shape parameter $a = 2$ and scale parameter $b = 3.33$. The depths generated in each step of the simulation are added. Moreover to obtain a non-homogeneous surface, we use some spatial function that gives a set of possible locations for parents and their offspring. The results are presented in Figure 2.34[3].

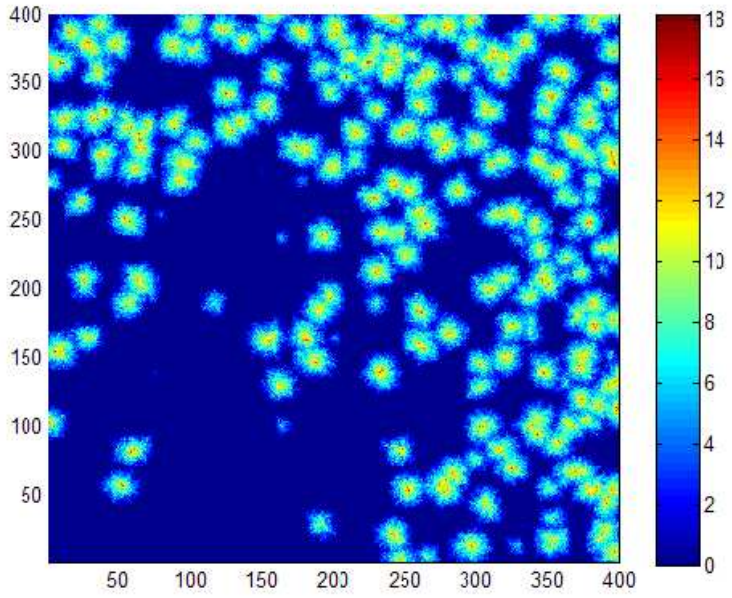

Figure 2.34: Example of spatially non-homogenous surface

For the data set as presented in Figure 2.34, it is difficult to define blocks such that the number of observations per block does not vary significantly. Therefore in this case it is easier to apply the GP distribution to declustered data.

[3]surface generated by courtesy of Mr MSc Ir. Sebastian Kuniewski, Delft Institute of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, The Netherlands

2.5 Proposed framework to model the extremes of corrosion data

In this section, we want to summarise our findings and propose some kind of framework that could lead to proper usage of extreme-value tools for corrosion data. We restrict our attention to the kind of data as introduced in Examples 1.1.4 or 2.4.1, i.e. to the case when data results from the full scanning of the system. We assume that the input data is stationary and locally dependent. When we model extremes of stationary and locally dependent data with the GEV distribution we proceed as for the independent data by the argument given in Section 2.2. We put attention to the block definition because this can influence the results. For the data type considered, we simply took a number of matrix columns as one block. Clearly there are other choices possible. It is important however, to check which choice gives a good model fit. Therefore the approach applied can be summarised in Figure 2.35.

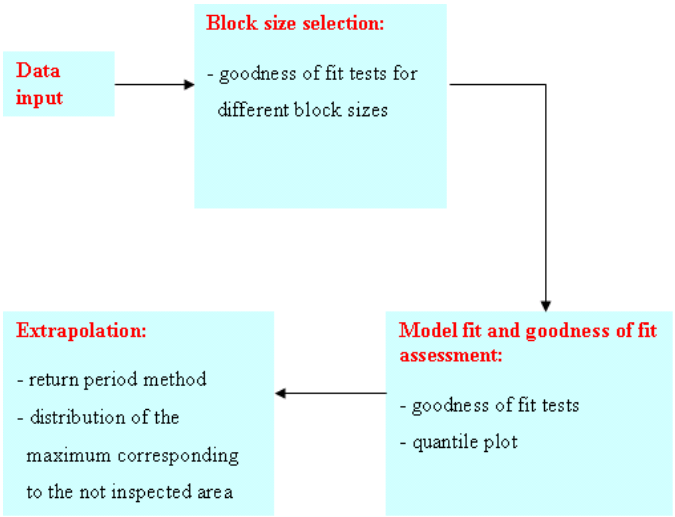

Figure 2.35: GEV framework to model corrosion data

When we want to apply the GP distribution we have to determine the proper threshold value. If there is clustering tendency confirmed then, theoretically speaking, the best approach would be to calculate introduced goodness of threshold choice measures (p-values of statistical tests, mean residual life plot, parameter stability) for declustered data. However, this is computationally expensive. Therefore, we propose to compute the measures mentioned not for declustered but for original data. This should give some estimate u_0 of the good threshold value. In the next step, we decluster excess over threshold u_0 data and repeat the same for a number of thresholds (say 5) close to u_0. For each threshold we check the goodness-of-fit and choose the best one. To determine the number of data clusters, we can use if possible two methods, namely the extremal index method and clustering algorithm together with validation criteria. If the determined number of clusters is different then we decluster data for the number of clusters given by the two methods, fit the GP distribution and compare the goodness-of-fit by statistical tests. Next, as the proper number of data clusters we chose the one for which the p-values are higher. Finally, when we find the proper threshold value and decluster data, we fit the model and extrapolate the results. This is schematically presented in Figure 2.36.

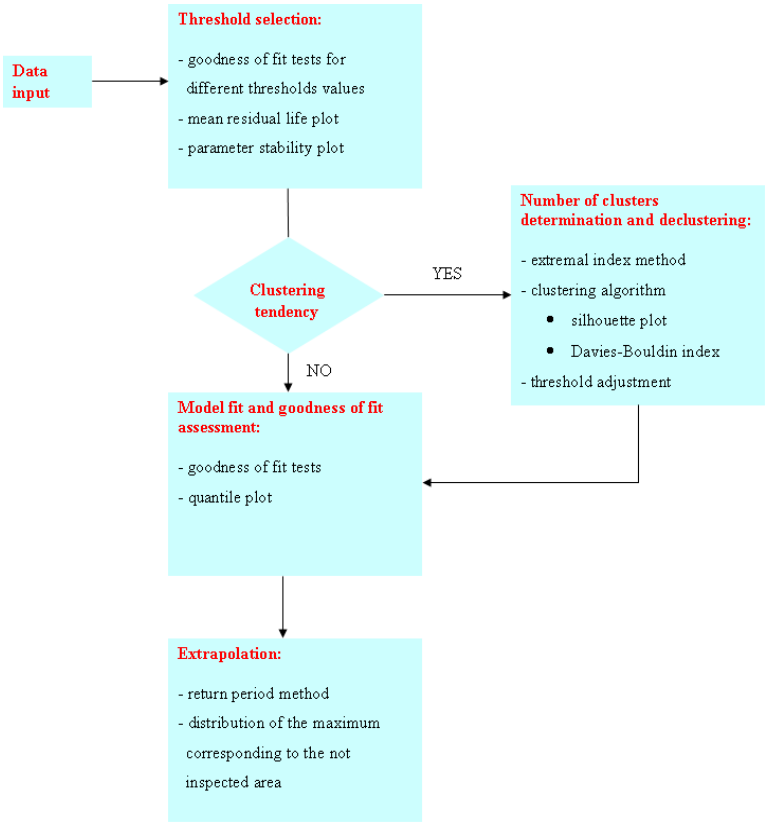

Figure 2.36: GP framework to model corrosion data

2.6 Summary

In this chapter we showed the application of the extreme-value methods to stationary corrosion data when taking into account local defect dependence. When we apply the GEV distribution we proceed as data would be independent. For the GP distribution prior to model fitting data declustering has to be done. The number of data clusters can be estimated by two methods: the extremal index method and clustering algorithm with introduced validation criteria. When possible, both methods can be used and the results compared and validated. Moreover, a framework within which the results given by the GEV and GP models should be consistent was presented. This helps to assess the general validity of the results.

Chapter 3

Conclusions and recommendations

The two approaches to model statistical behaviour of extreme defect depths in corrosion were introduced in the thesis. For block maxima data, the generalised extreme-value distribution is used, whereas for excess over threshold data the generalised-Pareto is applied. These two distributions are closely related and should lead to the same inference about extreme values.

We showed how the above methods can be applied when the underlying observations are stationary and locally dependent. When the GEV distribution is used, we proceed as data was independent because the information about data dependence will be incorporated into fitted parameters during the parameter estimation process. However, for the GP distribution a change of practise is needed. Prior to model fitting, the data declustering has to be done, which is based on filtering out the dependent observations such that remaining exceedances are approximately independent.

The key issue in data declustering is cluster identification. This can be done

by the clustering algorithm. To estimate the number of data clusters we can use two methods. The first approach is based on the extremal index parameter, which is the measure of the degree of clustering of the process at extreme levels. The other approach is based on the clustering algorithm used together with validation criteria. For this purpose, we introduced the agglomerative hierarchical algorithm, silhouette plot and Davies-Bouldin index. The application of the above methods was shown on the simulated and real data sets.
The benefit of data declustering is higher consistency of the results given by the GEV and GP models.

In order to realise the goal of the thesis we had to touch several topics like model fitting, results extrapolation or cluster analysis. Clearly there is still much more that could be done. It is often the case that corrosion data, due to changing environmental conditions or physical features of the equipment used (like a pipe placed under a certain slope) is not spatially stationary. It means that certain locations are influenced by more severe corrosion than the others. Then in order to be able to analyse such data together or extrapolate the results to the areas with space varying environmental conditions, covariate-dependent extreme-value models with trends could be used.

Bibliography

(n.d.).

Azis, P. M. (1956), 'Application of the statistical theory of extreme values to the analysis of maximum pit depth for aluminium', *Corrosion* **12**.

Beirland, J., Teugels, J., Vynckier, P. et al. (1996), *Practical Analysis of Extreme Values*, Leuven University Press.

Benerjee, A. & Davae, R. N. (2004), 'Validating clusters using the hopkins statistics'.

Blischke, W. R. & Murthy, D. N. P., eds (2003), *Case Studies in Reliability and Maintenance*, Wiley & Sons, New Jersey.

Bolshakova, N. & Azuaje, F. (2003), 'Cluster validation techniques for genome expression data', *Signal Processing* **83**(4).

Chernobai, A., Rachev, S. & F.Fabozzi (2005), 'Composite goodness of fit tests for left-truncated loss samples'.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer, London.

Coles, S. (2004), *The Use and Misuse of Extreme Value Models in Practise*, CRC Press LLC.

de Haan, L. & Ferreira, A. (2006), *Extreme Value Theory, An Introduction*, Springer, New York.

Diggle, P. J. (1983), *Statistical Analysis of Spatial Point Patterns. Mathematics in Biology*, Academic Press Inc., London.

Dubes, R. (1987), 'A test for spatial homogeneity in cluster analysis', *Journal of Classification* **4**(1).

Elderidge, G. G. (1957), 'Analysis of corrosion pitting by extreme value statistics and its application to oil well tubing caliper surveys', *Corrosion* **13**.

Everitt, B., Landau, S. & Leese, M. (2001), *Cluster Analysis*, Oxford University press Inc., New York.

Ferro, C. (2004), *Extremes of Stationary Time Series*, John Wiley & Sons, Ltd.

Ferro, C. A. (2002), 'Automatic declustering of extreme values via an estimator for the extremal index'.

Ferro, C. A. (2003), 'Inference for clusters of extreme values', *Journal of the Royal Statistical Society* **65**.

Galbraith, J. W. & Zernovi, S. (2006), 'Extreme dependence in the nasdaq and s&p 500 composite indexes'.

Jakel, J. & Nollenburg, M. (n.d.), 'Validation in the cluster analysis of gene expression data'.

Karatzas, I., Rajput, B. S. & Taqqu, M. S., eds (1998), *Stochastic Processes and Related Topics. In Memory of Stamatis Cambanis 1943-1995*, Birkhauser Boston.

Kotz, S. & Nadarajah, S. (1999), *Extreme Value Distributions.Theory and Applications*, Imperial College Press.

Laycock, P. J. & Scarf, P. A. (1993), 'Exceedances, extremes, extrapolation and order statistics for pits, pitting and other localized corrosion phenomena', *Corrosion Science* **35**(1-4).

Laycock, P. J., Scarf, P. A. & R.A.Cottis (1990), 'Extrapolation of extreme pit depths in space and time', *Journal of Electrochemical Society* **137**(1).

Leadbetter, M. R., Lindgren, G. & Rootzen, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, Springer Verlag, New York.

Leadbetter, M. R. & Rootzen, H. (1998), On extreme values in stationary random fields, *in* 'Stochastic Processes and Related Topics. In Memory of Stamatis Cambanis 1943-1995', Birkhauser, Boston.

Ostrowska, A. (2006), Simulating inspections on corroded surfaces, Master's thesis, Delft University of Technology.

Reiss, R. D. & Thomas, M. (1997), *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and other Fields*, Birkhauser Verlag, Basel Switzerland.

Scarf, P. A. & Laycock, P. (1992), 'Extrapolation of extreme pit depths in space and time using the r deepest pit depths', *Journal of Electrochemical Society* **139**(9).

Scarf, P. A. & Laycock, P. J. (1996), 'Estimation of extremes in corrosion engineering', *Journal of Applied Statistics* **23**(6).

Shibata, T. (1990), 'Evaluation of corrosion failure by extreme value statistics', *ISIJ International* **31**(2).

Smith, R. L. (2003), 'The statistics of extremes'.

Tan, P.-N., Steinbach, M. & Kumar, V. (2006), *Introduction to Data Mining*, Addison-Wesley.

Tsuge, H., Akashi, M. et al. (1994), *Introduction to Life Prediction of Industrial Plant Materials: Application of Extreme Value Statistical Methods for Corrosion Analysis*, Allerton Press,New York.

Turkman, K. (2006), 'A note on extremal index for space-time processes', *Journal of Applied Probability* **43**.

Webb, A. (2002), *Statistical Pattern Recognition*, Chichester Wiley .

Zempleni, A. (2004), 'Goodness of fit tests in extreme value applications'.

Appendix A

Definition 1 *A random variable Z is said to have a* ***Fréchet distribution*** *with scale parameter $\sigma > 0$, location parameter μ and shape parameter $\alpha > 0$, if its cumulative distribution function is given by:*

$$G(z) = \begin{cases} 0 & z \leq \mu, \\ exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{-\alpha}\right\} & z > \mu. \end{cases}$$

Definition 2 *A random variable Z is said to have a* ***Weibull distribution*** *with scale parameter $\sigma > 0$, location parameter μ and shape parameter $\alpha > 0$, if its cumulative distribution function is given by:*

$$G(z) = \begin{cases} \exp\left\{-\left(-\frac{z-\mu}{\sigma}\right)^{\alpha}\right\} & z < \mu, \\ 1 & z \geq \mu. \end{cases}$$

Confidence intervals for quantile plot of the GEV - delta method

To calculate confidence intervals for the quantile plot of the GEV distribution using the delta method, we have to first determine the gradient vector corresponding to the quantile z_p, i.e

$$\nabla z_p^T = \left[\frac{\partial z_p}{\partial \xi}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \mu}\right] \tag{3.1}$$

where

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(p)\}^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma \log\{-\log(p)\}, & \xi = 0. \end{cases} \tag{3.2}$$

Then

$$Var(z_p) = \nabla z_p^T V \nabla z_p$$

where

$$V = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix}$$

is the variance-covariance matrix corresponding to the parameter vector (ξ, σ, μ). The approximate confidence interval $(1 - \alpha)$ for z_p is given by:

$$z_p \pm x_{\alpha/2}\sqrt{Var(z_p)}$$

where $x_{\alpha/2}$ is $(1 - \alpha/2)$ quantile of the standard normal distribution.

Definition 3 *The **Kolmogorov distribution** is the distribution of the random variable*

$$Y = \sup_x |W(x)|,$$

where $W(x)$ is the Wiener process. The cumulative distribution function of the random variable Y is given by:

$$\Pr\{Y \leq y\} = 1 - s\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2y^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty}e^{-(2i-1)^2)\pi^2/(8y^2)}$$

It is known that $Z_n = \sqrt{n}D_n$, where D_n is the Kolmogorov-Smirnov test statistic, **converges in distribution** to the Kolmogorov distribution. This means that if F_1, F_2, ... is a sequence of cumulative distribution functions corresponding to the random variables Z_1, Z_2, ..., and that F is a distribution function corresponding to a random variable Y, then

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every real number x at which F is continuous.

Confidence intervals based on profile likelihood

Construction of confidence intervals using profile likelihood is based on Theorem 1 given below (Coles 2001).

Theorem 1 *Let $x_1, \ldots, x_m$ be independent retaliations from a distribution within a parametric family F, and let $\hat{\theta}_0$ denote the maximum likelihood estimator of the d-dimensional model parameter $\theta_0 = (\theta^{(1)}, \theta^{(2)})$, where $\theta^{(1)}$ is a k-dimensional subset of θ_0. Then, under suitable regularity conditions, for large m*

$$D_p\left(\theta^{(1)}\right) = 2\left\{l\left(\hat{\theta}_0\right) - l_p\left(\theta^{(1)}\right)\right\} \sim \chi^2_k, \tag{3.3}$$

where χ^2_k is a Chi-square distribution with k-degrees of freedom.

Then for a single component θ_i, $C_i = \{\theta_i : D_p(\theta_i) \leq c_\alpha\}$ is a $(1-\alpha)$ confidence interval, where c_α is a $(1-\alpha)$ quantile of the χ^2_1 distribution.

Confidence interval for the GEV-return-level

To obtain confidence interval for the return-level z_p we have to reparametrise the GEV distribution. More precisely we want to incorporate z_p to GEV as a parameter. Using equation (1.14) we get:

$$\begin{cases} \mu = z_p + \dfrac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \xi \neq 0 \\ \mu = z_p + \sigma \log\{-\log(1-p)\}, & \xi = 0. \end{cases} \tag{3.4}$$

Then replacement of μ in (1.2) with (3.4) gives desired effect of expressing the GEV in terms of the parameters (ξ, σ, z_p). To obtain the profile likelihood for return-level z_p we fix $z_p = z_{p0}$ and maximise the log-likelihood of the GEV with respect to the remaining parameters. This is repeated for a range of values of z_{p0}. The corresponding maximised values of the log-likelihood constitute the profile log-likelihood for z_p, from which Theorem 1 leads to obtain approximate intervals.

Confidence interval for the GP-return-level

For the GP distribution we proceed similarly as for the GEV. Using the relation (following from equation (1.29)) we get:

$$\bar{\sigma} = \begin{cases} \dfrac{y_p \xi}{(p^{-\xi} - 1)}, & \xi \neq 0 \\ -\dfrac{y_p}{\log(p)}, & \xi = 0 \end{cases} \tag{3.5}$$

Then replacement of $\bar{\sigma}$ in (1.18) results in expressing the GP in terms of (ξ, y_p). To obtain the profile likelihood for return-level y_p we proceed analogously as for the GEV distribution.

Definition 4 *The **mode** is the most frequent value assumed by a random variable, or occurring in a sampling of a random variable. Hence, it is the value of the random variable for which (if it exists) the probability density function is maximal.*

Euler-Mascheroni constant

The Euler-Mascheroni constant is a mathematical constant defined as a limiting difference between harmonic series and the natural logarithm:

$$\gamma = \lim_{n \to \infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \log(n) \right)$$

Its approximate value is 0.57721566490153286060651209008240243104215933 5.

Mean residual life plot

We know that if Y has the GP distribution then:

$$E(Y) = \begin{cases} \frac{\bar{\sigma}}{1-\xi} & ,\xi < 1, \quad \xi \neq 0 \\ \bar{\sigma}, & \xi = 0 \end{cases} \tag{3.6}$$

Suppose that the GP distribution is a valid model for the excesses over a threshold u generated by the series $X_1, \ldots, X_n$. If we denote an arbitrary term of this series by X then:

$$E(X - u_0 | X > u_0) = \begin{cases} \frac{\bar{\sigma}_{u_0}}{1-\xi} & ,\xi < 1, \quad \xi \neq 0 \\ \bar{\sigma}_{u_0}, & \xi = 0 \end{cases} \tag{3.7}$$

where σ_{u_0} is the scale parameter corresponding to threshold u_0.

But if the GP model is valid for threshold u_0, it must be valid for all $u > u_0$. By (1.24) it follows that:

$$\bar{\sigma}_u = \bar{\sigma}_{u_0} - \xi u_0 + \xi u$$

Then we have:

$$E(X - u | X > u) = \begin{cases} \frac{\bar{\sigma}_u}{1-\xi} = \frac{\bar{\sigma}_{u_0} - \xi u_0 + \xi u}{1-\xi} & ,\xi < 1, \quad \xi \neq 0 \\ \bar{\sigma}_u = \bar{\sigma}_{u_0} - \xi u_0 + \xi u, & \xi = 0 \end{cases} \tag{3.8}$$

Therefore for $u > u_0$, the mean residual life plot should be linear in u.

Confidence intervals for mean residual life plot

To calculate the confidence intervals for the mean residual life plot, we use the fact (following from the Central Limit Theorem) that the distribution of the sample mean can be approximated by the normal distribution. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables and let $Y_i, \ldots, Y_{n_u}$ be defined as $Y_i = X_i - u$ for a threshold u and $\{X_i : X_i > u\}$, $i = 1, \ldots, n_u$. Let us denote an arbitrary term in $Y_i, \ldots, Y_{n_u}$ by Y. If both the expected value μ and the standard deviation σ of Y exist and are finite, then for large n_u the approximate distribution of the random variable $\bar{Y} = \frac{1}{n_u}\sum_{i=1}^{n_u} Y_i$ is $N(\mu, \frac{\sigma^2}{n_u})$. We estimate σ^2 by $\frac{1}{n_u - 1}\sum_{i=1}^{n_u}(y_i - \bar{y})^2$ and the $(1-\alpha)$ confidence interval for mean residual life is calculated from:

$$\frac{1}{n_u}\sum_{i=1}^{n_u} y_i \pm \frac{\hat{\sigma}}{\sqrt{n_u}} x_{\alpha/2} \tag{3.9}$$

where $x_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution and $y_j = x_j - u$ for $j = 1, \ldots, n_u$ are realisations of the random variable Y.

Confidence intervals for parameters of GP distribution obtained through delta Method - parameter-stability threshold selection

The confidence intervals for $\hat{\xi}$ are immediately obtained from the variance-covariance matrix V of $(\hat{\xi}, \hat{\tilde{\sigma}})$, where

$$V = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

Then the approximate confidence interval $(1-\alpha)$ for $\hat{\xi}$ is given by:

$$\hat{\xi} \pm x_{\alpha/2}\sqrt{v_{11}}$$

where $x_{\alpha/2}$ is $(1-\alpha/2)$ quantile of the standard normal distribution. For $\hat{\sigma}^*$ we use the delta method and get:

$$Var(\sigma^*) = \nabla\sigma^{*T} V \nabla\sigma^*$$

where

$$\nabla\sigma^* = \left[\frac{\partial\sigma^*}{\partial\bar{\sigma}}, \frac{\partial\sigma^*}{\partial\xi}\right]$$

is a gradient vector. Using equation (1.27) we get $\nabla\sigma^* = [1, -u]$. Then the approximate confidence interval $(1-\alpha)$ for $\hat{\sigma}^*$ is given by:

$$\hat{\sigma}^* \pm x_{\alpha/2}\sqrt{Var(\hat{\sigma}^*)}$$

GEV distribution and the Poisson frequency of threshold exceedances

Let $M = \max_{i\leq i\leq N}(Y_i) + u$, then

$$\begin{aligned}
Pr\{M \leq x\} &= Pr\{\max_{i\leq i\leq N}(Y_i) \leq x-u\} \\
&= Pr\{N=0\} + \sum_{n=1}^{\infty} Pr\{N=n, Y_1 \leq x-u, \ldots, Y_n \leq x-u\} \\
&= e^{-\lambda} + \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \left\{1 - \left(1+\xi\frac{x-u}{\bar{\sigma}}\right)_+^{-1/\xi}\right\}^n \\
&= \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \left\{1 - \left(1+\xi\frac{x-u}{\bar{\sigma}}\right)_+^{-1/\xi}\right\}^n \\
&= \exp\left\{-\lambda\left(1+\xi\frac{x-u}{\bar{\sigma}}\right)_+^{-1/\xi}\right\} \qquad (3.10)
\end{aligned}$$

Appendix B

Definition 5 *A sequence of random variables $X_1, X_2, \ldots$ is **stationary** if the joint distribution of $(X_{i_1}, \ldots, X_{i_n})$ is identical to the joint distribution of $(X_{i_1+m}, \ldots, X_{i_n+m})$ for any choice of $n,\ i_1, \ldots, i_n,\ m$.*

Definition 6 *A stationary series $X_1, \ldots, X_n$ is said to satisfy the $\mathbf{D(u_n)}$ **condition** if, for all $i_1 < \ldots < i_p < j_1 < \ldots < j_q$ with $j_1 - i_p > l$,*

$$\begin{aligned} & |Pr\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\} \\ - \quad & Pr\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\} \times Pr\{X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\}| \leq \alpha(n, l), \end{aligned} \tag{3.11}$$

where $\alpha(n, l_n) \to 0$ for some sequence l_n such that $l_n/n \to 0$ as $n \to \infty$.

$D(u_n)$ condition ensures that (Coles 2001), for sets of variables that are far enough apart, the difference of probabilities in (3.11) while not zero, is sufficiently close to zero to have no effect on the limit lows for extremes. This is stated more formally in the Theorem 2.

Theorem 2 *Let $X_1, X_2, \ldots$ be a stationary process and define $M_n = \max(X_1, \ldots, X_n)$. Then if $\{a_n > 0\}$ and $\{b_n\}$ are sequences of constants such that*

$$Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \tag{3.12}$$

where G is non-degenerate distribution function, and the $D(u_n)$ condition is satisfied with $u_n = a_n z + b_n$ for every real z, G is a member of the generalised extreme-value family of distributions.

*A **degenerate distribution** is the probability distribution of a discrete random variable that assigns all of the probability, i.e. probability 1, to a single number.*

Theorem 3 *Let $X_1, X_2, \ldots$ be a stationary process and $X_1^*, X_2^*, \ldots$ be a sequence of independent variables with the same marginal distribution. Define $M_n = \max(X_1, \ldots, X_n)$ and $M_n = \max(X_1^*, \ldots, X_n^*)$. Under suitable regularity conditions,*

$$Pr\{(M_n^* - b_n)/a_n \leq z\} \to G_1(z) \tag{3.13}$$

as $n \to \infty$ for normalising sequences $\{a_n > 0\}$ and $\{b_n\}$, where G_1 is a non-degenerate distribution function, if and only if

$$Pr\{(M_n - b_n)/a_n \leq z\} \to G_2(z) \tag{3.14}$$

where

$$G_2(z) = G_1^{\theta}(z) \tag{3.15}$$

for a constant θ such that $0 < \theta \leq 1$.

Theorem 3 says that if the cumulative distribution function of maxima of stationary sequence can be approximated by the GEV distribution (what is valid if the $D(u_n)$ condition is satisfied , see Theorem 2) then this distribution is related to the limiting distribution of maxima of independent sequence according to equation (3). This means that the effect of short-range dependence in a stationary sequence is captured by the parameter θ, called *extremal index* (Coles 2001).

Note that if G_1 is the limiting distribution of maxima of independent sequence, that corresponds to the GEV distribution with parameters (ξ, σ, μ), then,

$$G_1^{\theta}(z) = \begin{cases} \exp\left\{-\left[1 + \xi\left(\dfrac{z - \mu + \sigma(1-\theta^{\xi})/\xi}{\sigma\theta^{\xi}}\right)\right]_+^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left[-\left(\dfrac{z - (\mu + \sigma\log\theta)}{\sigma}\right)\right]\right\}, & \xi = 0 \end{cases} \tag{3.16}$$

where

$$\begin{cases} \mu_\theta = \mu - \sigma(1-\theta^\xi)/\xi,\ \sigma_\theta = \sigma\theta^\xi,\ \xi_\theta = \xi & if\ \xi \neq 0 \\ \mu_\theta = \mu + \sigma\log\theta,\ \sigma_\theta = \sigma & if\ \xi = 0 \end{cases} \tag{3.17}$$

This means that the parameters of the GEV distribution are affected by dependence in the stationary series. Since the $D(u_n)$ condition is assumed to hold, the block maxima can be considered as approximately independent. This justifies the usage of the maximum likelihood estimation of the unknown parameters (equation (1.7)). The estimated parameters will be different than the ones that would have been obtained if the series had been independent (it follows from (3.17)). Thus the information about local dependence will be included in the estimated parameters.

Random field

In the simplest language a **random field** is a set of random variables whose values are mapped onto the n-dimensional space.

In the case of data we have, random values of corrosion wall loss, are mapped onto the 2-dimensional space, namely Cartesian grid. Each coordinate has assigned random wall loss.