Extreme value analysis (EVA) of inspection data and its uncertainties

Daniel Benstock, Frederic Cegla*

Mechanical Engineering Department, Imperial College London, Exhibition Road, South Kensington, London SW7 2AZ, United Kingdom

A R T I C L E   I N F O

Keywords:
Extreme value analysis
Statistical evaluation of inspection data
Ultrasound

A B S T R A C T

Extreme value analysis (EVA) is a statistical tool to estimate the likelihood of the occurrence of extreme values based on a few basic assumptions and observed/measured data. While output of this type of analysis cannot ever rival a full inspection, it can be a useful tool for partial coverage inspection (PCI), where access, cost or other limitations result in an incomplete dataset. In PCI, EVA can be used to estimate the largest defect that can be expected. Commonly the return level method is used to do this. However, the uncertainties associated with the return level are less commonly reported on. This paper presents an overview of how the return level and its 95% confidence intervals can be determined and how they vary based on different analysis parameters, such as the block size and extrapolation ratio. The analysis is then tested on simulated wall thickness data that has Gaussian and Exponential distributions. A curve that presents the confidence interval width as a percentage of the actual return level and as a function of the extrapolation ratio is presented. This is valid for the particular scale parameter (σ) that was associated with the simulated data. And for this data it was concluded that, in general, extrapolations to an area the size of 500–1000 times the inspected area result in acceptable return level uncertainties (<20% at 95% confidence). When extrapolating to areas that are larger than 1000 times the inspected area the width of the confidence intervals can become larger than 30–50% of the actual return level. This was deemed unacceptable: for example of wall thickness mapping that is used throughout this paper, these uncertainties can represent critical defects of nearly through wall extent. The curve that links the confidence interval width to the return value as a function of extrapolation ratio is valid only for a particular scale parameter value of the EVA model. However, it is imaginable that a few of such relations for different scale parameters σ could be simulated. By picking the relation with the closest σ value (based on observation or estimation) for the inspection dataset, the presented approach can then be used to quickly estimate the uncertainty associated with an EVA extrapolation.

1. Introduction

Statistical modelling is an important tool in many areas of science and engineering. It has been used in a wide range of applications from quantifying the uncertainty of the output of a measurement tool [1] to predicting the lifetime of engineering components [2,3]. In non-destructive testing (NDT), specifically, statistical modelling has been used to study the probability of missing critically sized defects in an inspection [4], studying the reliability of inspections [5,6] and the probability of component failure [7]. A statistical model describes the behaviour of a random variable, providing a probability of a given value of the random variable occurring. While the statistical tools described in this paper are not limited to a particular problem, this paper illustrates their use by application to a particular problem: assessment of wall loss due to corrosion from ultrasonic inspection data. For this particular application area coverage is very important. Often full coverage of a plant/component cannot be achieved due to cost, access or other limitations. Therefore partial coverage inspection (PCI) is required. PCI is based on the construction of a statistical model from inspection data collected across a sample area (as illustrated in Fig. 1). The findings from the sample area are then used to form a picture of the condition of a larger area or the whole component. In the particular case that we focus on in this paper the random variable is the measured thickness (or the wall loss) across the inspection area and the operator is particularly interested in the largest defect (the thinnest wall thickness that is to be expected/measured).

Fig. 1 shows that the key contribution of the statistical model is the extrapolation of information from the known domain (inspected area) to the full region of interest (area to which one extrapolates). There are different ways of carrying out this extrapolation, it can be done based on the cumulative distribution function (CDF) of measured thickness values or based on a sample of minima within sub-populations of the
available data population, extreme value analysis (EVA) is an example of the latter. The key difference between the two approaches is that the CDF is a model that focuses on describing the whole thickness distribution, whereas EVA focuses more tightly onto the extremes (in this case the largest defects/thinnest part of the wall).

The study of extreme values is a well-developed topic, finding applications to topics as varied as finance [8,9], structural design [10–12], environmental modelling [13,14] and even the assessment of the risk of terrorist attacks [15]. Its use for analysis of corrosion data is discussed in [3]. Kowaka provides a number of examples of the use of extreme value analysis to extrapolate from C-scans of reduced areas of a plant to larger areas of components. Further examples are provided in [16–29]. EVA gained some traction in Japan in the 1970s. However, it then fell out of favour. The authors believe that this might have been due to the lack of available computational power which made processing large amounts of C-scan data infeasible.

In recent years, the use of an extreme value approach has regained popularity. A report prepared for the Health and Safety Executive (in the UK) assessed current available methods and the barriers to their adoption. It concluded that there are readily available statistical methods for the analysis of corrosion data. However, these methods are not used due to poor dissemination to engineers and the lack of any readily available computational tools [30]. There is also a lack of knowledge about the uncertainties of EVA predictions and this paper aims to address this.

This paper is structured as follows: first the theoretical framework behind the cumulative distribution function and EVA are outlined and the process of extrapolation from the known data is described. Following this a method of determining the confidence intervals of the return level is described. Then 3 different numerical studies are presented: The first study demonstrates that EVA and the return level are an effective way of estimating the minimum thickness when extrapolating from an inspected area to a larger area over which an assessment is made. Simulated inspection data is used to show to compare the minimum thickness values in actual thickness distributions to those estimated by the return level method. The second study investigates the variation of the confidence interval width as a function of the block size (number of thickness minima) and the extrapolation ratio that are used in the EVA. The third study highlights the trade off between precision (confidence interval width) of an EVA prediction and the accuracy of the prediction, specifically with regards to the number of minima that are used to construct the EVA model. In order to do so groups of actual minimum thicknesses from very large surfaces and their EVA estimates are compared. Then results are summarised and conclusions are drawn.

2. Background theory

2.1. The cumulative distribution function

Ultrasonic inspection data of corroded components usually comes in the form of a C-scan thickness map (Fig. 2(a)). The thickness at each position in the map is represented by a coloured pixel, providing a qualitative overview of the degradation in the inspection area. The thickness map can be converted into a more quantitative presentation of the data by calculating an estimate of the cumulative probability distribution of the thickness measurements. This is calculated by sorting the thickness measurements in ascending order, assigning each thickness measurement a rank and using this rank to calculate the empirical cumulative distribution (ECDF) function:

\[
F(x) = \frac{i}{N+1}
\]

where \(x\) is a measurement of thickness, \(i\) is its rank and \(N\) is the total number of thickness measurements. \(F(x)\) is the probability of measuring a thickness of less than \(x\). An example of a cumulative distribution
function generated from the thickness map in Fig. 2(a) is shown in Fig. 2(b). The ECDF is an estimate of the probability of measuring a thickness less than a given value. A number of examples of ECDFs generated from inspection data collected from in service engineering components can be found in [31].

In addition to the empirical estimate of the cumulative probability distribution, it is common practice to fit a probability distribution to the thickness measurements. Thickness measurements can be distributed in many different ways. Their distribution is determined by the conditions the component is subjected to. Different temperatures, pH, and surface conditions can all produce different distributions. There are usually many different degradation mechanisms occurring across the inspection area, consequently, by the central limit theorem, the overall thickness distribution often tends to be Gaussian or Gaussian correlated [32]. However, localised corrosion mechanisms can often produce exponential [33] or more exotic distributions [31].

The Gaussian distribution is defined by two parameters: the mean \( \mu \) and the standard deviation \( \sigma \). The probability of obtaining a measurement of less than \( x \) from a Gaussian distribution is given by:

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

where \( F(x) \) is the probability of obtaining a measurement of less than \( x \). In Fig. 2(b) this distribution has been fitted to thickness measurements using maximum likelihood estimation (MLE) [34, p. 824] and plotted alongside the empirical cumulative distribution function.

MLE is a method of estimating the parameters of a distribution from a given set of data. The method is based on the joint distribution of the set of measurements, conditioned on the parameters of the distribution. For example, for the normal distribution the joint distribution is given by:

\[
f(x_1, x_2, x_3, \ldots, x_n \mid \mu, \sigma) = f(x_1 \mid \mu, \sigma) f(x_2 \mid \mu, \sigma) \cdots f(x_n \mid \mu, \sigma)
\]

where \( x \) and \( s \) are the mean and standard deviation of the distribution and \( x \) are a set of observations. Alternatively, this can be thought of in terms of the likelihood function, which is defined as:

\[
L(\mu, \sigma \mid x_1, x_2, x_3, \ldots, x_n) = f(x_1, x_2, x_3, \ldots, x_n \mid \mu, \sigma)
\]

where \( L \) is the likelihood of the parameters \( \mu \) and \( \sigma \) given \( x \). The larger the likelihood of a pair of parameters \( (\mu, \sigma) \), the better the fit of model to the data.

Maximisation of the likelihood will provide a best estimate for the parameters given the observed data. In practice the likelihood is normally expressed as the log-likelihood as it is often easier to work with:

\[
L(\mu, \sigma \mid x_1, x_2, x_3, \ldots, x_n) = \log(L(\mu, \sigma \mid x_1, x_2, x_3, \ldots, x_n)) = \sum_{i=1}^{n} f(x_i \mid \mu, \sigma)
\]

as the logarithm is a monotonic function, maximisation of Eq. (5) is equivalent to maximisation of Eq. (4). For the normal distribution the log-likelihood function is given by:

\[
L(\mu, \sigma \mid x_1, x_2, x_3, \ldots, x_n) = \frac{-1}{\sigma \sqrt{2\pi}} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}
\]

Maximisation of this function will provide estimates for the parameters of the normal distribution given the set of data. There are alternative methods to maximum likelihood estimation, such as the method of moments and least squares regression. However, in this work, MLE estimation is used exclusively because it is very flexible and the framework can be used to estimate distributional parameters for a large variety of problems.

The Gaussian distribution provides a good fit to the set of data, as shown by the red dashed line in Fig. 2(b). Both the ECDF and the fitted distribution are models for the thickness measurements in the inspection area (blue area in Fig. 1). Both models provide information on the average behaviour of the thickness measurements. However, condition assessment is most concerned with assessing the largest extent of the wall loss in a component. From a constructed model, the probability that \( x \) is the smallest thickness in an area the size of the inspection area is \( \Psi(x) = 1 - F(x) \). Consequently, the probability that \( x \) is the smallest thickness in an area \( M \) times the initial inspection area is given by:

\[
\Psi(x) = \Psi(x)^M = (1 - F(x))^M
\]

where \( \Psi(x) \) is an empirical distribution of the smallest thickness in an area \( M \) times the inspection area. Hypothetically, a fitted distribution such as Eq. (2) could be used with Eq. (7) to compute the exact analytical form of the distribution of the minimum thickness in a given area. However, any small inaccuracy in \( \Psi(x) \) will be magnified in \( \Psi(x)^M \).

An alternative approach is to accept that \( \Psi(x) \) is unknown and to look for a limiting form of the distribution \( \Psi(x)^M \). Extreme value analysis describes such an approach. It shows that, under certain assumptions, the distribution of \( \Psi(x)^M \) will be a generalized extreme value distribution (GEVD).

2.2. Extreme value analysis

EVA is based on the Fisher-Tippet-Gnedenko theorem [35] which provides a limiting form for Eq. (7) (the probability of obtaining a minimum thickness \( x \) in an area \( N \) times the inspection area). The theorem shows that in the limit \( N \to \infty \) Eq. (7) is a generalized extreme value distribution (GEVD) with CDF:

\[
\Phi(x; \mu, \sigma, k) = \exp \left\{ -1 + k \left( \frac{x - \mu}{\sigma} \right)^{1/k} \right\}
\]

where \( \mu \) is the location parameter, which determines the size of the minima; \( \sigma \) is the scale parameter, which determines the spread of the minima; and \( k \) is the shape parameter, which determines the shape of the distribution. The shape parameter can loosely be thought of as controlling the skewness of the distribution, that is the difference between the mean and the median of the distribution. For the particular case of \( k=0 \) a different equation might need to be used [36]. \( \Phi(x; \mu, \sigma, k) \) is the probability of measuring a thickness minimum of less than \( x \). A full discussion of a suitable size for \( N \) for Eq. (8) to be a valid model can be found in [37]. Examples of the pdfs of GEVDs are shown in Fig. 3.

Eq. (8) can be used as a model for the smallest thickness measurement in a prescribed area. However, for this to be possible, a data analyst must calculate values of \( \mu, \sigma \) and \( k \). With a standard statistical model this can be accomplished by fitting the model to the collected data. From the authors’ experience thickness minima are usually modelled using distributions with \( k \leq 0 \), real examples of this can be found in Hawn [17,33,16]. The derivation of Eq. (8) is based on the convergence of the sequence of renormalised thickness measurements. The shape of the distribution is determined by the way this sequence converges, which depends on the family of distributions the thickness measurements belong to. The thickness distributions studied in this work are restricted to be Gaussian and exponential, which both lead to extreme value distributions with \( k \leq 0 \) [11]. In practice, it would be unusual for a GEVD with \( k > 0 \) to occur in components undergoing only wall loss as this would imply wall thickness growth.

An extreme value (EV) model is a model for the minimum thickness in an area as opposed to the thickness. The data analyst requires a sample of thickness minima to calculate values of \( \mu, \sigma \) and \( k \). A sample of minima can be extracted from a number of different thickness maps from different areas [17], or from a single thickness map [3]. Performing multiple inspections to obtain different thickness maps can often be infeasible for time, cost or access reasons. As a consequence, selecting a sample of minima from a single thickness map is often the best option for PCI applications. A sample of minima
is the probability of measuring a thickness measurement. An example of a distribution with
parameters \( \mu \) and \( \sigma \) can be selected from a thickness map by dividing the map into a
number of equally sized blocks (shown in Fig. 4). The minimum thickness is selected from each of these blocks. These measurements form a set of minima which can be used to calculate the parameters in
Eq. (8). An example of a histogram of a set of 100 thickness minima extracted from a correlated Gaussian surface with RMS height 0.1 mm and correlation length 2.4 mm is shown in Fig. 5. Each thickness measurement is represented by a bar, the height of which is proportional to
the frequency of the measurement occurring. The GEVD provides a good description of the data.

2.3. Extrapolation using EVA

Suppose ultrasonic thickness C-scan data of a fraction of a component has been acquired. The resulting thickness map can be used to construct an extreme value model for the smallest thickness
measurements by partitioning the thickness map into \( N_{\text{blocks}} \) equally sized blocks. The blocks are used to construct a sample of thickness minima by selecting the smallest thickness measurement in each block.

If the thickness minima are selected from sets of thickness measurements that are independently and identically distributed, the generalized extreme value distribution (GEVD) is the limiting form of the thickness minima distribution. Estimates of \( \mu \), \( \sigma \), and \( k \) can be extracted using maximum likelihood estimation (MLE).

Once an extreme value model has been constructed from the thickness minima, a data analyst can use this model to perform PCI. This will require extrapolations to areas larger than the initial inspection using the extreme value model. There are two methods available for extrapolation of an extreme value model: the return level method and the distributional method. The return level method is used to draw conclusions about areas larger than the initial inspection by mapping them to quantiles of the EV model, whilst the distributional method attempts to directly construct a model for the minimum in the extrapolated area.

The \( M_{\text{th}} \) return level, \( r_{M_{\text{th}}} \) is defined as the thickness value that will not be exceeded only once in \( M \) blocks of an inspection. The expected number of thickness measurements greater than the return level in a sample of \( M \) thickness minima is given by:

\[
E(N(x > r_M)) = MP(x > r_M) = M(1 - \Phi(x\mu, \sigma, k))
\]

where \( P(x > r_M) \) is the probability of measuring a thickness measurement greater than the return level. From the definition of return level (\( E(N(x > r_M)) = 1 \)),

\[
MP(x > r_M) \Phi(x\mu, \sigma, k) = 1 - \frac{1}{M}
\]

Eq. (10) is visualised in Fig. 6 as the area bound to the right of the red dashed line and the probability density function. It is equivalent to the \( M_{\text{th}} \) quantile of the GEVD. The position of the red dashed line can be calculated by rearranging the GEV distribution (a derivation can be found in Coles [38]):

\[
r_M = \mu - \frac{k}{\sigma} \left[ 1 - \log\left(1 - \frac{1}{M}\right) \right]^{-1/k}
\]

This is the value of thickness that the model predicts will be exceeded at least once in \( M \) blocks. It can be interpreted as an estimation of the smallest thickness that is expected to be found in an area the size of \( M \) blocks.

Extrapolations to areas larger than the inspection region can be
performed by calculating the return level corresponding to a number of blocks greater than the initial sample of minima \( N_{\text{blocks}} \). For example, a data analyst could estimate the minimum thickness in an area twice the size of the inspection area by calculating the return level corresponding to \( M = 2N_{\text{blocks}} \). This return level is the reference value that the model expects would not be exceeded in an area two times the initial inspection area.

It is impossible to directly validate an extrapolation, it would require data from outside of the inspection area, which is unavailable to the data analyst. However, one can show that the model constructed is reasonable given the available data and the assumptions made by EVA, which can be achieved using a method previously presented by the authors [37]. Once the assumptions made by the model have been shown to be reasonable, the uncertainty associated with the extrapolations can be quantified by calculating a 95% confidence interval around the return level.

### 2.4. Confidence intervals for the return levels

The uncertainty around the return level that arises from statistical variations can be quantified with a confidence interval. A 95% confidence interval for the return level is the bounds which contains 95% of the possible estimates of the return level. For example, if 100 models were generated from different samples from inspections of the component (with the same thickness distribution), 95 of the estimates will lie within these bounds. Confidence intervals are a reflection of the data analyst’s belief in the return level estimate.

Confidence intervals around the return level can be calculated using the profile likelihood method. The profile likelihood is, for a given parameter, using \( \mu \) as an example, following the method described by Coles [38, p. 57]:

\[
L(\mu) = \max_{\sigma, k} \{ L(\mu, \sigma, k|x) \}
\]  

In other words, the profile likelihood function for \( \mu \), \( L(\mu) \), is the maximised likelihood function with respect to \( \sigma \) and \( k \). For the return level, the profile likelihood is:

\[
L(r_M) = \max_{\sigma, k} \{ L(r_M, \sigma, k|x) \} = N \ln(\sigma) - \left( 1 + \frac{1}{k} \right)
\]

\[
\sum_{i=1}^{N} \ln \left( 1 + k \left( x_i - (r_M - \frac{\sigma}{k} \{ 1 - \{ -\log(\frac{1}{\alpha}) \}^k \} \right) \right)
- \sum_{i=1}^{N} \left( 1 + k \left( x_i - (r_M - \frac{\sigma}{k} \{ 1 - \{ -\log(\frac{1}{\alpha}) \}^k \} \right) \right)
\]

Eq. (13) can be used to calculate a confidence interval for the return level with the deviance function. The deviance function is defined as:

\[
D(r_M) = 2(L(r_M) - L(\hat{r}_M))
\]

where \( L(r_M) \) and \( L(\hat{r}_M) \) are the profile likelihoods for an estimate of \( r_M \), \( \hat{r}_M \), and the true value of \( r_M \). It can be shown that the statistic \( D(r_M) \sim \chi^2_d \), where \( \chi^2_d \) is the chi-squared function [39] with \( d \) degrees of freedom. \( d \) describes the number of factors affecting the likelihood function.

As the deviance statistic follows a \( \chi^2 \) distribution, we can calculate the bounds in which 95% of its estimates lie. This is defined as the set \( \{ \theta : D(\theta) \geq \chi^2_\alpha \} \) where \( \chi^2_\alpha \) is a \( (1 - \alpha) \) quantile of the \( \chi^2_d \) distribution. This set is described graphically by Fig. 7.

In all EVA examples in the literature, extrapolated return levels are reported as a single value. This is an incomplete representation of the data. No information about the uncertainty associated with the extrapolation is revealed. Confidence intervals allow a data analyst to quantify some of the uncertainty associated with the model, which is key to understanding the conclusions drawn from extreme value models.

### 3. Behaviour of EVA extrapolations

One of the aims of this paper is to make EVA theory more accessible to data analysts and NDE professionals. Simulation studies that show how EVA predictions of the return level behave for Gaussian and Exponential surfaces are shown here. One of the most important aspects of EVA is that the return value itself has an associated pdf. Commonly only the modal value is quoted. However, based on the EVA model parameters that are obtained from a dataset, a range of return levels is likely. The range of the likely return levels is best expressed by confidence intervals. In this section the size of the confidence intervals as a function of the block size and extrapolation ratio will be investigated. Finally, the accuracy of predictions will be discussed.
4. The return level as a function of extrapolation ratio

The aim of these simulations is to show that the return level and the calculated confidence intervals are useful metrics for condition assessment purposes. With this in mind, 1000 samples of 48 by 48 mm Gaussian height distributed Gaussian correlated surfaces and 1000 samples of 48 by 48 mm Exponential height distributed Gaussian correlated surfaces with mean thickness 10 mm, RMS=0.1 mm and $\lambda = 2.4$ mm were generated using the rough surface algorithm described in [40]. These samples are equivalent to inspections of 2, 304, 000 mm$^2$ (2.3 m$^2$) of a component. Extreme value models were constructed from a subset of these surfaces corresponding to one 240 by 240 mm Gaussian and a 240 by 240 mm exponential surface, i.e. 25 of the 48 by 48 mm surfaces or $N_{\text{blocks}} = 25$. The model provides a description of the minimum thickness in an area the size of a single block and this and extrapolation results to larger areas can be compared to the distribution of minima in the actual population of the 1000 simulated surfaces.

Extrapolations using these models can, conveniently, be considered as extrapolations to multiples of this block size. The model was used to calculate return levels and their corresponding confidence intervals to areas ranging from 2 to 1000 blocks. These areas are rescaled in terms of the initial inspection area, which consisted of 25 blocks, to define the extrapolation ratio:

$$ER = \frac{EA}{IA}$$ (15)

where $ER$ is the extrapolation ratio, $EA$ is the area to which one extrapolates and $IA$ is the inspected area from which data is available. If $ER < 1$ one effectively is interpolating, because the area for which a prediction from the model is made is smaller than the area for which data is available. The model is only truly used for extrapolation when $ER > 1$. In terms of the extrapolation ratio, return levels corresponding to extrapolation ratios ranging from 0.08 to 40 were calculated. The full population of 1000 surfaces corresponds to $ER=40$.

Fig. 8 (a) and (b) show box plots of the actual thickness minima from the 1000 48 by 48 mm correlated Gaussian surfaces and 1000 48 by 48 mm correlated exponential surfaces respectively. The length of the box is the inter-quartile range of the sample of thickness minima, which is a measure of spread of the measurements. The solid line in the middle of the box is the median thickness minimum, whilst the whiskers contain 99% of the thickness minima. Return levels corresponding to the extrapolation ratios on the x-axis were calculated from the model generated from both the Gaussian and the exponential surfaces using Eq. (11). These are shown as black crosses. 95% confidence intervals around these return levels are shown as blue crosses. With an increasing extrapolation ratio, the return levels decrease, indicating that in a larger area a smaller minimum thickness is expected.

For extrapolation ratios less than 1, the return level is modelling the minimum thickness in an area less than the inspection area (an area from which the inspection data is taken). Subsequently, the confidence intervals for this region are narrow and the return level provides a good description of the data. For example, the return level for the extrapolation ratio of 0.08 is very close to the median of the sample of thickness minima. This is expected as the return level for this extrapolation ratio will be exceeded once every two blocks, so around 50% of the thickness measurements should be less than this value. This trend is continued with the return levels for the extrapolation ratios ranging from 0.16 to 1, with the return levels matching up with the appropriate quantiles in the sample.

Once the extrapolation ratio exceeds 1, the exact value of the return level predicted by the model does not necessarily match up with the correct quantile of the thickness sample. This is expected as extrapolations, by their very nature, will not provide a perfect description of the data. In these situations, the confidence interval around the return level is key to interpreting the results of the model. For example, as the extrapolation ratio is increased to the point where it corresponds to the size of the total population ($ER=40$), the return level gets closer and closer to providing a set of bounds for the smallest thickness measurement in the population. For the data used in this study, there are no thickness measurements less than the return level with an extrapolation ratio of 40. However, for different realisations, there could be thickness measurements less than this value because extrapolation can lead to uncertainties in the prediction of the return level.

However, the confidence interval around the model’s estimate of return level will contain the true value of the return level 95% of the time. The confidence interval can be interpreted as the bounds inside which the smallest thickness measurement in an extrapolated area will lie. Rather than reporting just a single value for the minimum thickness, the return level, a data analyst can state a range of values and a measure of his confidence that the minimum thickness lies in this set of bounds. This will allow plant operators to make a decision about the condition of a component with knowledge of the uncertainty.

---

1 To clarify, by exceed the return level, we mean that there will be at least one thickness measurement less than the return level.
around the data analyst’s estimate.

The width of the confidence intervals increases with the extrapolation ratio. Ideally a data analyst will try to minimise the extrapolation ratio for the extrapolation they are performing as it will minimise the width of the confidence bounds around the return level. Knowledge of how large a confidence interval will be on average, as a function of extrapolation ratio, will allow a data analyst to make a decision about the amount of inspection area required to obtain bounds on the minimum thickness of a given width. This is addressed in the next section.

5. Return level confidence intervals as a function of extrapolation ratio

There are a number of factors which determine the width of the confidence intervals. For example, if the data analyst has a large sample of minima, the confidence intervals will be narrower and they will be more confident in the predictions made by the model. However, the sample size of minima is determined by the ratio of the block size to the inspection area. If the minima have been selected using a block size which is a small fraction of the inspection area, some of the samples will not be representative of extremes of the population. Then the model will not provide a good description of the condition of the component.

Knowledge of how the confidence intervals behave with different inspection designs will allow for improvements in the design of partial coverage inspections. With this goal in mind many confidence intervals were calculated from simulated inspections partitioned using different block sizes. The inspection data was simulated by generating 50 correlated Gaussian surfaces of size 240 by 240 mm and 50 correlated exponential surfaces of the same size. The surfaces all had an RMS height 0.1 mm and a correlation length of 2.4 mm.

From each surface an extreme value model was generated using block sizes corresponding to different numbers of minima, summarised in Table 1. The return levels corresponding to extrapolation ratios ranging from 0.08 to 400 were calculated, along with the corresponding confidence intervals. For each set of simulations the average width of the confidence intervals is calculated and expressed as a percentage of the return levels. This percentage is plotted as a function of the extrapolation ratio.

Figs. 9(a) and (b) show the average size of the confidence intervals (expressed as percentage of mean return level) as a function of extrapolation ratio. Although Figs. 9(a) and (b) were produced using data collected from surfaces with different distributions, they show very similar behaviour. As the extrapolation ratio is increased the width of the confidence intervals is nearly the same up to an extrapolation ratio of 1. For extrapolation ratios less than one, the return levels correspond to areas smaller than the inspection area. Once the extrapolation ratio increases past 1, the return levels correspond to areas greater than the inspection area. Consequently, the average size of the confidence intervals begins to increase, as there is more uncertainty due to the extrapolation to an area larger than the inspection area. Beyond ER=1 the width of the confidence interval increases rapidly with increasing extrapolation ratio.

Past a certain point, the confidence intervals indicate that estimates of the return level are no longer a useful tool. The confidence interval indicates the range in which the true value of return level lies, if it is too large, the model cannot be used for useful extrapolations. For example, in both Fig. 9(a) and (b), at an extrapolation ratio of 400 the average width of the confidence intervals is at least 30% of the size of the return level, e.g. for a return level of 5 mm the confidence interval has a range of 3.5–6.5 mm. For extrapolations of these sizes, more data is required and a data analyst should conclude that the inspection area should be increased. (As illustrative example remember that the mean wall thickness for these simulations is 10 mm, the return level is 5 mm and the confidence interval width indicates that 3.5 mm is possible, i.e 35% of nominal thickness. This would most likely be considered a critical defect).

The presented data seems to suggest that using more minima is the correct approach to narrow the width of the confidence intervals. However, this should be considered carefully. To obtain a larger sample, the minima are selected using a smaller block size. The smaller the block size, the lower the number of thickness measurements in each block. Each block is a smaller sample of the underlying thickness distribution. As the size of the sample of thickness measurements in the block decreases, the expected number of extremes in the sample will decrease. Consequently, it is likely that the minimum thickness measurement in this block may not be representative of the extremes of the distribution. A model constructed from a set of minima from

![Fig. 9](image.png)

Fig. 9. The average width of the return level confidence intervals, expressed as a percentage of the return level and as a function of extrapolation ratio, calculated from 50 extreme value models constructed from (a) 240 mm by 240 mm Gaussian surfaces and (b) 240 mm by 240 mm exponential surfaces. Up to an extrapolation ratio of 1 (the size of the inspection area), the width of the confidence intervals is constant. Past this point the model is being used to extrapolate and the confidence intervals grow. The rate of growth is determined by the number of minima used to construct the model.

Table 1

The different block sizes and the corresponding number of minima used to generate the extreme value models.

| Block Size/mm | Number of minima |
|---------------|------------------|
| 24            | 100              |
| 30            | 64               |
| 40            | 36               |
| 48            | 25               |
| 60            | 16               |
these blocks may be inaccurate. In the next section, this problem is investigated using a series of simulated inspections taken from across very large surfaces.

6. Testing the accuracy of an extrapolation

In order to determine the quality of extrapolations from models generated using different block sizes, Gaussian and exponential surfaces (with RMS height of 0.1 mm and correlation length of 2.4 mm) of size 2400 by 2400 mm, 12,000 by 12,000 mm and 24,000 by 24,000 mm were generated. From each surface 50 different inspection areas of 240 by 240 mm (such that the total size of the surfaces corresponded to extrapolation ratios of 10, 50 and 100) were chosen and EV models were generated from each inspection area using block sizes ranging from 24 to 60 mm (corresponding to minima sample sizes ranging from 100 down to 16). The EV models generated from these inspection areas were used to calculate return levels and confidence intervals for extrapolation ratios corresponding to the size of the surfaces. The return levels and confidence intervals were averaged over the 50 inspection areas and compared to the smallest thickness measurement across the surface.

The average return levels and average confidence intervals were plotted as a function of the number of minima in the top of Figs. 10 [(a), (b) and (c)] for the Gaussian surfaces corresponding to extrapolation ratios of 10, 50 and 100 respectively. The data collected from the exponential surfaces is shown at the bottom of Figs. 10 [(a), (b) and (c)]. The red crosses are the average confidence bounds on the return level, the blue cross is the average return level and the dashed line is the smallest thickness across the entire surface. The average return level in each figure overestimates the minimum thickness across the surface. This is expected as the return level is expected to overestimate the minimum thickness as, by definition, it is larger than the minimum thickness.

Fig. 10. The average return level (blue crosses) and confidence intervals (red crosses) calculated from 50 different extreme value models constructed using 240 mm by 240 mm inspections of a Gaussian surface (top) and Exponential surface (bottom) corresponding to (a) an extrapolation ratio of 10 (b) an extrapolation ratio of 50 and (c) an extrapolation ratio of 100. Each model was constructed using a different number of minima (x-axis) and the average confidence intervals were compared to the smallest thickness across the surface (black dashed line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Fig. 11. The return level (blue cross) and confidence intervals (red cross) corresponding to an extrapolation ratio of 50 calculated using an extreme value model constructed from a single 240 mm by 240 mm inspection of a Gaussian surface corresponding to an extrapolation ratio of 50. The smallest thickness across the surface is shown as the black dashed line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

The return level is larger than the minimum thickness.

2 The return level is larger than the minimum thickness.
a threshold which will be exceeded at least once in the extrapolated area. The average return level decreases slightly as the number of minima in the sample is decreased. This is reflective of the fact that a smaller number of minima is collected using a larger block size, the average minimum thickness in the sample will be smaller, resulting in smaller return levels.

The size of the confidence intervals increases as the number of minima is decreased. This is in agreement with Figs. 9(a) and (b). Encouragingly, and as speculated, for extrapolation ratios of 10 and 50, the minimum thickness measurement lies within the average confidence interval. That is, on average, the confidence interval provides a set of bounds in which the minimum thickness lies. However, this is the average behaviour over 50 different examples. Models from individual inspections can sometimes predict bounds which do not contain the minimum thickness. An example of this is shown in Fig. 11, which is an example of the return levels and confidence intervals for an extrapolation ratio of 50 predicted by a model from a single inspection area from the 12,000 by 12,000 mm Gaussian surface. The minimum thickness is contained by both the confidence bounds calculated from the EV models constructed using 16 and 36 minima, but it lies outside the confidence bounds calculated using 64 and 100 minima.

As the size of the surface increases, the minimum thickness decreases. Simultaneously, the size of the average return level confidence intervals increases. If the extreme value models are providing an adequate description of the minimum thickness of the surface, on average the return level prediction of the minimum thickness should remain within the confidence interval. However, the minimum thickness decreases at a faster rate than the growth of the confidence intervals. Consequently, there is a point where the minimum thickness is no longer contained by the confidence bounds. At that point the model is not providing an adequate description of the damage in the extrapolation area. The number of minima (block size) used to generate the model is key to ensuring that this is unlikely to occur.

In general, the confidence intervals from models generated using larger block sizes (less minima) more consistently provide a bounds for the minimum thickness. For the Gaussian surface, the return level confidence intervals calculated from the models with 16 and 36 minima contained the minimum thickness in 82% and 76% of cases respectively for an extrapolation ratio of 50, compared to 68% and 60% for sample sizes of 64 and 100. For an extrapolation ratio of 100, the number of models which predict bounds which contain the minimum thickness reduces further. The confidence intervals only contain the minimum thickness in 44%, 58%, 72% and 68% of cases for models generated with 100, 64, 36 and 16 minima respectively. Arguably the models generated using the larger block sizes could still be used for extrapolations of this size, the minimum thickness will lie outside the bounds 30% of the time, but for most applications this error rate will be too high.

Both of these cases can be compared to the results for an extrapolation ratio of 10. For this case, the return level confidence intervals contained the minimum thickness for 76%, 82%, 86% and 90% of the models constructed using 100, 64, 36 and 16 minima respectively. The confidence intervals contain the minimum the majority of the time. However, it is clear that, even at a small extrapolation ratio a model generated with fewer minima is able to more accurately model the thickness measurements in the extrapolation area.

The confidence bounds calculated from models constructed using the exponential behaviour show a similar trend. For an extrapolation ratio of 10, the confidence interval contained the minimum thickness across the surface 92%, 86%, 66% and 54% of the time for sample sizes of 16, 36, 64 and 100 minima. With an extrapolation ratio of 50, the minimum thickness lay in the bounds 92%, 74%, 80% and 70% of the time for sample sizes of 16, 36, 64 and 100. Finally, with an extrapolation of 100, the bounds contained the minimum thickness in 92%, 80%, 74% and 74% of cases.

The increased error rate that occurs with the combination of a model generated using a large sample of minima (small block size) and a large extrapolation ratio originates in the bias introduced into the EV model by the sample of minima. Sample minima collected using smaller block sizes are on average larger than those collected with larger block sizes. Therefore, the predictions made by an extreme value model will overestimate the size of the minimum thickness (predict it to be thicker than it actually is) in the extrapolated area. This leads to an overestimation of the return level. Consequently, the model will be biased, it will underestimate the probabilities of the smallest measurements of thickness occurring. In addition, as the sample is larger, there is more evidence the model is ‘correct’ so the width of the confidence intervals increases at a slower rate than for smaller sets of minima (as shown in Fig. 9), which leads to the minimum thickness lying outside the confidence bounds of the return level.

7. Practical use of the information in this paper

A key purpose of this paper is to draw attention to the uncertainties that are associated with EVA models. The presented work in particular focused on uncertainties that arise as a function of the data analysis process, i.e. the construction of the model of thickness minima, the partitioning of the measurement data into several sub samples and the degree to which the model is extrapolated. This information can be used as a simple practical guide to look up the expected uncertainty that an EVA model will result in.

It can be used in several ways: (1) Planning an EVA inspection: one has a rough idea of the wall thickness and minimum thickness of the component that is to be inspected and the overall component surface area is known. A confidence interval for the minimum thickness is chosen, i.e. 10% of the return level. Then the curves in Fig. 9 are used to determine the corresponding extrapolation ratio (ER) that is allowed. In the case of 10% confidence interval width, ER=100, with 16 minima this means that the inspection will need to be carried out over an area of at least 1% of the overall surface area of the vessel. (2) Having inspected a particular area and having calculated the return level using EVA (which is very simple to do using Eq. (11)), it is possible to assess what the confidence interval of the return level estimate is if the ratio of the overall surface area to the inspected area (ER) is known. If the confidence interval is too wide for comfort (i.e. in excess of 30–50% of the return level), further inspections have to be made. This avoids having to compute the confidence interval of the return level using the profile likelihood method which is very time consuming. In both of these cases it is expected that data will have been collected by a standard inspection technique, e.g. an ultrasonic C-scan. The data will need to be appropriately partitioned into blocks and minimum thickness are extracted to construct the EVA model. It is recommended that during the partitioning the right block size is selected and it is checked whether the data is suitable for EVA (i.e. data compatibility with the assumptions of EVA are checked). This can be achieved by using a simple blocking algorithm as described in [37].

It is important to note however that the data that was simulated in this study was limited. It does not take into account all sources of uncertainty that will be encountered in practice; e.g. the investigated surfaces only had a moderate variation in wall thicknesses [RMS 0.1 mm] and thickness measurement errors, which can also be of the order of 0.1 mm, were ignored. It is expected that these additional uncertainty factors would mainly influence the scale parameter \( \sigma \) of EVA models and hence the width of the confidence intervals. Therefore, the data of Fig. 9, which has a particular ratio of confidence interval width to return level at ER=1 cannot be generally used for estimation of uncertainties of any EVA analysis. However, it could be envisaged that a few such curves are produced for characteristic underlying scale parameter \( \sigma \) values. The curve with the most appropriate scale parameter is then used for a quick assessment of the uncertainty.
8. Conclusions

Understanding uncertainty associated with extrapolation is a key part of using extreme value analysis for partial coverage inspection. In the existing literature, there have not been any studies in relation to NDE of the errors associated with extrapolation from extreme value models. This has been a barrier to the use of extreme value models in the NDE and inspection community. In this paper the theory behind EVA is reviewed and studies showing the uncertainties associated with extrapolation using the return level method are presented. Both

Gaussian and exponential surfaces were studied and at every stage of this study it was found that models constructed from them performed similarly. This is consistent with theory of EVA which says that the methodology is independent of the thickness distribution of the component that is being studied. It was found that the precision of the return level predicted by EVA mainly depends on the number of minima that is used in creating the model and the size of the area to which one extrapolates. The behaviour is roughly the same for Gaussian and exponential surfaces and below extrapolation ratios of 500 the confidence intervals of the return levels are less than 30% of the actual return level value which should result in acceptable levels of uncertainties. The curves of confidence interval width as a function of extrapolation ratio that are provided in this paper are a quick and easy way to look up the uncertainty that is to be expected when performing an EVA on inspection data. The data that has been used in the particular simulations of this paper has a particular inherent spread or scale parameter ($\sigma$) and does not contain measurement errors it therefore cannot be generally applied to any EVA model but only those with a similar scale parameter. However, it is easily imaginable that the concept can be used generally and a few master curves with different underlying $\sigma$ values could be produced so that a very quick assessment of uncertainties of EVA models is possible. The accuracy of predictions can be influenced by the number of minima that are used in the EVA model. Using larger numbers of minima will tighten confidence levels but it does not necessarily improve accuracy as the minima that enter the analysis might not be extremes and they therefore bias the EVA model to give less conservative/missing results. For the presented work using block sizes that resulted in a minima population size of 16–25 minima resulted in the best results and it is believed that this will generally be the case for all EVA models.

9. Data accessibility

Readers who are interested in accessing data associated with this paper are referred to www.imperial.ac.uk/non-destructive-evaluation where either the data or details of how to obtain the data can be found.

Acknowledgements

F. Cegla would like to acknowledge funding from EPSRC, grant reference EP/K033565/1.

References

[1] Hughes I, Hase T. Measurements and their uncertainties: a practical guide to modern error analysis. Oxford University Press; 2010.
[2] Balkema A, de Haan L. Residual life time at great age. Ann Probab 1974;2:792–804.
[3] Kowaka M. Introduction to life prediction of industrial plant materials: application of extreme value statistical method for corrosion analysis. New York, USA: Allerton Press, Inc.; 1994.
[4] Ali A, Ballint D, Temple A, Leever P. The reliability of defect sentencing in manual ultrasonic inspection. NDT E Int 2012;51:101–10.
[5] N. Brierley, T. Tippett, and P. Cavley, Improving the reliability of automated non-destructive inspection. In: AIP Conference Proceedings, American Institute of Physics, May 2012, vol. 1430, pp. 1824–31.
[6] Burbata SF, Stow BA, Will M. Computer modelling for the prediction of the probability of detection of ultrasonic corrosion mapping. In: The Non-Destruct Test Cond Monit 2005;47(12):761–4.
[7] L. Shibay, Statistical methods for extremes and degradation analysis (Ph.D. thesis), Iowa State University, 2013.
[8] Embrechts P, Kluppelberg C, Mikosch T. Modelling extremal events for insurance and finance. Berlin, Heidelberg: Springer-Verlag; 1997.
[9] Wiggins JB. Estimating the volatility of S & P 500 futures prices using the extreme value method. J Futures Mark 1992;12(3):265–73.
[10] Gumbel E. Statistics of extremes. New York: Dover Publications, Inc.; 2004.
[11] Coles S, Tawn J. Statistical methods for multivariate extremes - an application to structural design. J R Stat Soc Ser C (Appl Stat) 1994;43(1):48–62.
[12] Coles SG, Tawn JA. Statistics of coastal flood prevention. Philos Trans R Soc A Math Phys Eng Sci 1990;332(1627):457–76.
[13] Tawn JA. An extreme-value theory model for dependent observations. J Hydrod 1988;101:227–50.
[14] Katz RW. Statistics of extremes in climate change. Clim Change 2010;100(1):71–6.
[15] H. Mohr, Assessing the risk of terrorism using extreme value statistics. In: Institute of Food Technologists’ Proceedings of the first annual food protection and defense research conference, Atlanta, Georgia, 2005.
[16] Azir P. Application of the Statistical theory of extreme values to the analysis of maximum pit depth data for Aluminum. Corrosion 1956;12(10):35–46.
[17] Hawn DE. Extreme value prediction of maximum pits on pipelines. Mater Perform 1977;16:29.
[18] Shibata T. Application of extreme-value statistics to corrosion. J Res Natl Inst Stand Technol 1994;99(4):327.
[19] Shibata T. Evaluation of corrosion failure by extreme value statistics. ISIJ Int 1991;31(2):115–21.
[20] Rivas D, Caley F, Valor A, Hallen JM. Extreme value analysis applied to pitting corrosion experiments in low carbon steel: comparison of block maxima and peak over threshold approaches. Corros Sci 2008;50(11):3193–204.
[21] Scarf P, Laycock P. Applications of extreme-value theory in corrosion engineering. J Res Natl Inst Stand Technol 1994;99(4):313.
[22] Vajo JJ, Wei R, Phelps AC, Reiner L, Herrera GA, Cervantes O, Gidanian D, Bavarian B, Kappes CM. Application of extreme value analysis to crevice corrosion. Corros Sci 2003;45(3):497–508.
[23] Caley F, Velázquez JC, Valor A, Hallen JM. Probability distribution of pitting corrosion depth and rate in underground pipelines: A Monte Carlo study. Corros Sci 2009;51(9):1925–32.
[24] C. Schneider, Application of extreme value analysis to corrosion mapping data. In: Proceedings of the 4th European-American Workshop on Reliability of NDE, 2009, pp. 1–8.
[25] Melchers RE. Extreme value statistics and long-term marine pitting corrosion of steel. Prod Eng Mech 2008;23(4):882–8.
[26] Laycock PJ, Cottis RA, Scarf PA. Extrapolation of extreme pit depths in space and time. J Electrochem Soc 1990;137(1):64–9.
[27] J. Wallace Jon, R. Reddy, D. Pugh, and J.L. Pacheco, Sour service pit growth predictions of carbon steel using extreme value statistics, January 2007.
[28] Najjar D, Bigerele M, LeFebvre C, Iost A. A new approach to predict the pit depth extreme value of a localized corrosion process. ISIJ Int 2003;43(5):720–5.
[29] M. Iaccasone, M. Strom, H. Rootstein, and O. Lunder, Galvanically induced atmospheric corrosion on magnesium alloys: a designed experiment evaluated by extreme value statistics and conventional techniques. Technical report, SAE International, February 1997.
[30] TWI, Guidelines for use of statistics for analysis of sample inspection of corrosion. Technical report, The Welding Institute, 2002.
[31] M. Stone, Wall thickness distributions for steels in corrosive environments and determination of suitable statistical analysis methods. In: Proceedings of the 4th European-American Workshop on Reliability of NDE, 2009.
[32] Thomas TR. Rough Surfaces. London: Longman; 1982.
[33] Strutt JE, Nicholls JR, Barbier B. The prediction of corrosion by statistical analysis of corrosion profiles. Corros Sci 1985;25(5):305–15.
[34] Harris J, Stecker H. Handbook of mathematics and computational science. Burlington, USA: Elsevier Science; 2014.
[35] Fisher RA, Tippett LHC. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceed Cambridge Philos Soc 1928.
[36] Burch SF, Stow BA, Wall M. Computer modelling for the prediction of the probability of detection of ultrasonic corrosion mapping. In: The Non-Destruct Test Cond Monit 2005;47(12):761–4.
[37] Shibata T. Probability and statistical evaluation of corrosion data.Ulrich's Corrosion HandbookThird Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2011. p. 365–84.
[38] Benstock D, Cegla F. Sample selection for extreme value analysis of inspection data collected from corroded surfaces. Corros Sci 2015;103:206–14.
[39] Coles S. An introduction to statistical modeling of extreme values. London, UK: Springer-Verlag: London; 2001.
[40] Boas M. Mathematical methods in the physical sciences. New York, USA: Wiley-Interscience; 2006.
[41] Benstock D, Cegla F, Stone M. The influence of surface roughness on ultrasonic thickness measurements. J Acoust Soc Am 2014;136(6):3028–39.