

# Pitting corrosion: Comparison of treatments with extreme value distributed responses

A.-L. FOUGÈRES<sup>1,3</sup>, S. HOLM<sup>2,3</sup> and H. ROOTZÉN<sup>2,3</sup>

## SUMMARY

In this paper we develop Statistical Extreme Value Theory as a method to validate and improve experiments with extremal responses, and to extrapolate and compare results. Our main motivation is corrosion tests performed at Volvo Car Company: Localized, or "pitting", corrosion can limit the usefulness of aluminum, magnesium and other new lightweight materials. It makes judicious choice of alloys and surface treatments necessary. Standard methods to evaluate corrosion test are based on weight loss due to corrosion and ANOVA. These methods fail in two ways. The first is that it usually is not weight loss but the risk of perforation, i.e. the depth of the deepest pit which is of interest. The second is that the standard ANOVA assumption of homogeneity of variances typically is not satisfied by pit depth measurements, and that normality doesn't give credible extrapolation into extreme tails.

*Key words:* Pitting corrosion; Extreme values; Comparisons of treatments; designed experiments.

*AMS 2000 Subject classifications:* Primary 62P30, secondary 62G32.

---

<sup>1</sup>CNRS, Département GMM, Institut National des Sciences Appliquées, 135 Avenue de Rangueil, 31077 Toulouse Cedex 04, France. Email: fougères@insa-toulouse.fr

<sup>2</sup>Department of Mathematics, Chalmers University of Technology, S-412 96 Gothenburg, Sweden. Email: rootzen@math.chalmers.se

<sup>3</sup>Research supported by VCC/Ford, by the Knut and Alice Wallerberg foundation, and by TMR Spatial and Computational Statistics Network

## 1. INTRODUCTION

In this paper we develop methods to validate and improve experiments with extremal responses, and to extrapolate and compare treatments. Our main application is to corrosion experiments at Volvo Car Company, and the methods were developed for this purpose as an engineering tool ready for routine use.

Making cars lighter is important for reducing fuel consumption, and is a central challenge for the automotive industry. Localized corrosion (also called “pitting” or “galvanic” or “bimetallic” corrosion) limits the use of new lightweight materials, such as magnesium or aluminum alloys. A key issue hence is to reduce pitting corrosion, via coating, surface treatment, choice of alloy or use of isolating washers.

Standard methods to evaluate corrosion test are based on weight loss due to corrosion and ANOVA. These methods fail in two ways. The first is that it usually is not weight loss but the risk of perforation, i.e. the depth of the deepest pit which is of interest (see e.g. Isacson et al. (1997)). The second is that the standard ANOVA assumption of homogeneity of variances typically is not satisfied by pit depth measurements, and that normality doesn't give credible extrapolation into extreme tails. Extreme Value Statistics has appeared as a theoretically well founded alternative way to analyse data on pitting corrosion, see e.g. Shibata (1996). This was suggested already by Aziz (1956) and Gumbel (1958), and in many subsequent papers. Much of this was developed and promoted in a book, Kowaka (1994). Likelihood, generalized Pareto and Extreme Value distribution methods for analysis and extrapolation were proposed in a series of papers by Scarf, Laycock, Cottis (see e.g. Scarf and Laycock (1994) and the references in it). However, the literature does not seem to include advice on how to check experimental conditions or on comparison of corrosion reducing treatments.

Another application is to the fatigue limit i.e. the “threshold stress for nonpropagation of the cracks” (Murakami and Beretta (1999)). For metallic materials this

threshold stress is determined by the size of the largest nonmetallic inclusion or defect (Murakami and Usuki (1989), Takahashi and Sibuya (1996), Murakami and Beretta (1999)). A further application could be to experimentation with synthetic portfolios of financial instruments, where risks are evaluated from historical extreme price fluctuations over a number of time intervals.

This paper reports on the first part of a continuing effort, where the distant goal is a full theory of design and analysis of experiments with extreme value distributed responses. Presumably such a theory would also be likelihood based, and incorporate results from this paper, but would in particular add covariate models for the entire experiment, see e.g. Alec Stephenson's very recent R program on covariate models for extreme values, <http://cran.us.r-project.org/>; package 'evd'. We in fact already tried this approach in a pitting corrosion setting in Isacson et al. (1997). However, we now believe that substantial further development is needed before such a theory can be widely useable. This development should include improvement and better understanding of numerical routines and extensive experience with and analysis of properties of estimators and tests. It should also include much better understanding of the effects of different choices of parametrisation, and the ability of models to respect the stochastic monotonicity implied by the non-reversibility of the corrosion process.

In this paper analysis is based on block maxima. In a complementary method, the Peaks over Threshold method, analysis uses not only maxima, but all values exceeding a large threshold (or a predetermined number of the largest values), see e.g. Coles (2001). It is reasonably straightforward to translate our methods to the Peaks over Thresholds setting. The main changes would be to replace Gumbel distributions with exponential ones and Extreme Value distributions with Generalized Pareto distributions. However, in corrosion testing, to measure the pits is a major part of the experimental effort, and the part the experimenters like the least. Hence the choice of methods is determined by measurement convenience and not by statistical consid-

eration. At Volvo, engineers find it easier to quickly locate the deepest pit and then measure it carefully, rather than to make careful measurements of several pits, some of which subsequently turn out to be too shallow to be included in the analysis. Hence the block maxima method is the Volvo standard, and we have chosen to present our methods in this setting.

The description of the method is given in the context of pit corrosion. Section 2 summarizes some basic tools for extreme value modeling and Section 3 discusses pit corrosion on Magnesium and the Volvo experiment. Section 4 describes the method and analyzes the Magnesium corrosion dataset. Section 5 deals with some statistical and modeling issues which arise. Section 6 contains our conclusions. Some technical issues are relegated to appendices.

## 2. STATISTICAL EXTREME VALUE THEORY

Statistical extreme value theory models and analyzes data which is obtained as the maxima of many (approximately) independent and identically distributed (i.i.d.) underlying variables. Useful recent introductions to the area are Coles (2001) and Embrechts et al. (1997). Coles gives an up-to-date account of statistical methods while Embrechts et al. contains the basic theory, from an econometric perspective.

The central result of extreme value theory is that the natural model for maxima is the Extreme Value (EV) distribution (sometimes also called the “Generalized Extreme Value distribution”) with distribution function

$$G(x) = \exp \left[ - \left\{ 1 + \xi (x - \mu) / \sigma \right\}^{-1/\xi} \right],$$

where  $\sigma > 0$ ,  $\mu, \xi \in \mathbb{R}$ , and the formula is valid for  $1 + \xi(x - \mu)/\sigma > 0$ . The parameters  $\mu$ ,  $\sigma$  and  $\xi$  are the location, scale and shape parameters, respectively. For  $\xi = 0$  the formula should be interpreted as the limiting (as  $\xi \rightarrow 0$ ) Gumbel distribution  $G(x) = \exp \left[ - \exp \left\{ - (x - \mu) / \sigma \right\} \right]$  and for  $\xi$  negative the distribution has a finite upper bound. This model is supported by two related basic properties:

- The EV distribution is obtained as the only possible limit (under linear normalization) of the distribution of the maximum of  $n$  i.i.d. random variables as  $n \rightarrow \infty$ , and
- the EV distribution is the only one which is stable under change of blocksize, ie such that if maxima over smaller i.i.d. blocks have this distribution, then maxima over bigger blocks have the same distribution.

Several methods for estimation of the EV parameters have been proposed, see for example Hosking et al. (1985), and the review by Johnson et al. (1994, Volume 2, Chapter 22). Maximum likelihood estimation in particular gives good results as soon as the sample size is not too small (see Section 5 and Appendix 2 for further discussion) and is much more general and flexible than the competitors. In the present paper we use maximum likelihood estimation and the delta method for confidence intervals throughout (see e.g. Coles (2001), p.33).

We consistently use suitably adapted and modified versions of so called Gumbel plots. These plots illustrate the adequacy of the EV fit and provide easy graphical interpretation and extrapolation of results. If  $X_1, \dots, X_n$  are i.i.d. observations, the Gumbel plot shows the graph

$$\left\{ X_{(i)}, -\log \left( -\log \frac{i}{n+1} \right) \right\}, \quad i = 1, \dots, n,$$

where  $X_{(1)} \leq \dots \leq X_{(n)}$  denote the observations ordered in ascending order. The values are scattered around a straight line if they come from a Gumbel distribution. The distribution function of the fitted EV distribution is also shown in the plots, and appears as a convex curve if the estimated shape parameter  $\xi$  is negative, a straight line for the Gumbel case  $\xi = 0$ , and a concave curve if  $\xi > 0$ .

### 3. PIT CORROSION

In this section we give a rapid sketch of galvanic corrosion, and then describe the Volvo magnesium corrosion experiment.

Galvanic corrosion is the consequence of an oxidation-reduction reaction. The reaction is caused by the potential difference which is created when two different metals are in electrical contact and in contact with an electrolyte to form a “galvanic cell”. The rate and amount of corrosion depends strongly on environmental factors, such as temperature and the precise composition of the solution, and also on the geometry of the galvanic cell and on surface structure and treatments. Important gaps still remain in the basic chemical knowledge of the corrosion mechanism. Hence the automotive industry has to resort to experimentation and experience to be able to manufacture sufficiently corrosion resistant cars.

In particular, sophisticated experimentation systems, such as climate chambers have been developed. These chambers make possible laboratory tests with carefully controlled conditions of humidity, salinity and temperature, and complement field tests in an important way.

The following laboratory experiment performed at Volvo Car Company is typical of many similar data sets. Circular plates of the magnesium alloy Mg AZ91D were combined with three different types of bolt, untreated steel bolts (denoted “Fe”), black-chromated zinc-steel bolts (denoted “Fe/Zn C4”), and JS500 zinc coated steel bolts (denoted “Zn JS500”) to form an experimental assembly (Figure 1, left). The plates were covered with synthetic dirt (89 % washed sea sand, 9 % kaolin, 1 % active carbon, 1 % sodium chloride) and the assemblies were placed in a climate chamber. Then they were exposed to climate cycling according to the “Volvo Indoor Corrosion Test” protocol, without acid rain, i.e. the temperature was kept constant at 35 degrees and the humidity was cycled between 50 % and 95 % twice a day (Isacson et al, 1997). These conditions are aimed at accelerating the corrosion process, from years to a matter of weeks. A basic and very difficult problem is to make this acceleration uniform for different surface treatments and alloys, and to make the translation from laboratory experiments to reality.

The experiment was performed with 9 plates per type of bolt. Of these,  $n = 3$  assemblies (“replicates”) with each type of bolt were taken out of the climate chamber after 2 weeks of exposure, after 4 weeks of exposure, and after 6 weeks of exposure, respectively. Thus in the terminology of design of experiments (which is somewhat incompatible with corrosion terminology), the treatments are the three types of bolt, and the three timepoints. During the experiment the plate rested on an inclined surface, and the orientation of each plate was recorded. After the end of the exposure corrosion products were dissolved from the plates, and each plate was divided into  $k = 8$  sectors. The maximum pit depth in each sector was measured by direct radiography, using a technology provided by AGFA (trademark: DirectRay). The dataset obtained thus consists of an  $8 \times 27$  matrix of observations of the maximum pit depth in a sector, with the sectors ordered according to their position relative to the incline of the plate.

The more general framework is thus the following: For each specific treatment (e.g. choice of alloy, surface coating of the bolt, duration of corrosion exposure ...), a given number  $n$  of experimental assemblies are used. After the experiment is concluded, the “measurement unit” (e.g. the plate) is divided into  $k$  blocks ( e.g sectors), and the maximum pit depth in each block is measured. A typical dataset hence consists of  $nk$  measurements of block maxima for each treatment.

Now, how should one compare the efficiency of the treatments? In the next section



Figure 1: *Left: Specimens of magnesium plates. Right: Numbering of the sectors in terms of their location.*

we propose a method for analysis of such datasets.

#### 4. METHOD AND DATA ANALYSIS

Recall that the experimental assemblies consisted of circular magnesium plates (the units) joined to steel bolts which were treated in different ways. Each treatment was applied to 3 assemblies, and each unit was divided into 8 sectors (or blocks). The data set consists of measurements of the deepest pit in each such sector.

The method divides the analysis into 3 parts: a preliminary study of the data; a separate analysis of each treatment; and pairwise comparisons of the treatments. Each part consists of several steps. For each step, similar elements are provided: a graph; a parametric likelihood ratio test based on a Gumbel or an EV model; and randomization tests. The latter are suggested as a way to corroborate the results in cases where there is doubt if sample sizes are large enough to make the likelihood ratio tests sufficiently accurate.

In each step we first describe the method for a general situation, and then apply it to the Volvo corrosion experiment.

##### **Steps 1 and 2: preliminary study of the data**

The first two steps check that units are replicates and homogeneous, or in statistical terms that the  $nk$  observations for a specific treatment are i.i.d. The experiments are designed to achieve this, and we expect the measurements to pass the test. However, if they don't, this may indicate a need for improving the experimental setup. It also would mean that one cannot proceed with the following steps in the way outlined below – modifications would be needed.

*1. Are units homogeneous?* For each treatment, observations from sectors at similar locations are combined into groups, and the values in the different groups are plotted on separate lines in a dot-diagram. If the groups are well chosen these graphs make it possible to see systematic differences (“inhomogeneities”) between

sectors with different locations. Next, a Gumbel distribution is fitted to each group of sectors. Inhomogeneities then correspond to different parameter values in the different groups. This is checked by a likelihood ratio test. Sample sizes for this are often small, and if they are, say, below 20, it is prudent to corroborate the likelihood ratio test by randomization tests. We use three such tests. The first one is based on the Gumbel likelihood ratio statistic. The other two are completely nonparametric, and are based on statistics measuring location heterogeneity and dispersion heterogeneity, respectively. See Fougères et al. (2002) for more details.

*Data analysis:* To check if the pit depths were influenced by the position of the sector relative to the incline of the plate, the sectors were divided into four groups as shown in Figure 1, right. Thus, for each set of three replicate plates there are four groups of two sectors, with  $3 \times 2 = 6$  pit depths measured for each group. The Fe bolts (top row in Figure 2) seem to have slightly deeper pits at the top of the plate and the 2 weeks Fe/Zn C4 measurements include two high values in sectorgroup 3. However, no consistent pattern that would indicate a serious influence of the position of the sectors on the incline emerges from Figure 2.

This conclusion mainly agrees with the results of the formal statistical tests reported in Table 1, Columns 1 to 4. The p-values for the more specific parametric LR-tests are smaller than for the location and dispersion tests. The location and dispersion tests measure different kinds of deviations from the null hypotheses, and the p-values also differ.

2. *Are units replicates ?* The graphical test for systematic differences between experimental units (i.e. if units are “replicates”) is a Gumbel plot (see Section 2 above) where each unit has its own symbol. This is complemented by a likelihood ratio test of the hypothesis that the separate parameters for the different experimental units in fact are the same. Since sample sizes were small (less than 20) we in addition performed randomization tests in the same way as for Step 1.

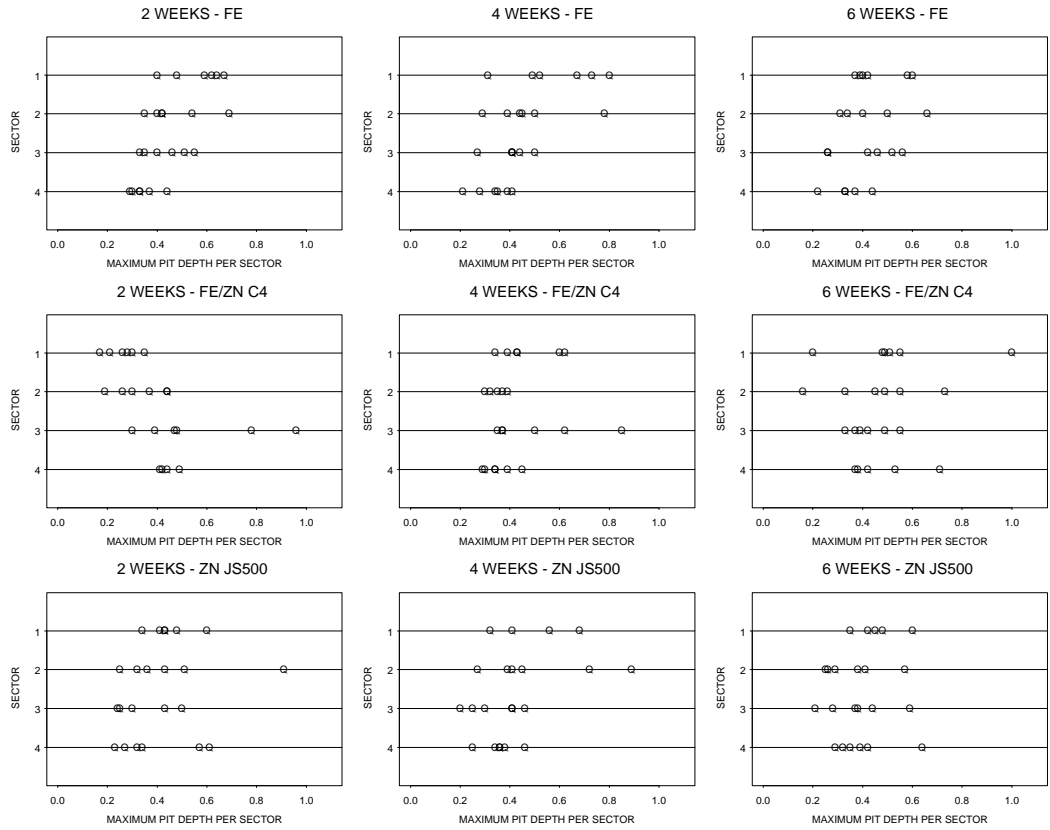


Figure 2: *Pit depth maxima by sectors. The first column provides the 2 weeks data, the second one the 4 weeks data, and the third one the 6 weeks data. Rows show from top to bottom “Fe”, “Fe/Zn C4” and “Zn JS500”.*

*Data analysis:* Figure 3 contains the Gumbel plots for each set of three replicate units. The results of likelihood ratio and randomization tests are reported in Table 1, columns 5 and 6. No consistent pattern which would indicate that units are not replicates is seen.

*Comments:* The presumption is that the experiment has been carried out so that sectors are homogeneous and units are replicates. The analysis thus is only aimed at safeguarding against gross deviations. In particular, in the subsequent analysis we use the more flexible EV distribution, and don't restrict ourself to the Gumbel model.

Table 1: *P-values of tests in Steps 1 and 2: Columns 1 to 4 test if sectors are homogeneous and Columns 5 and 6 if plates with the same treatment are replicates.*

		homogeneous?				replicates?	
		LR	Rand-disp	Rand-loc	Rand-LR	LR	Rand-LR
2 weeks	Fe	0.01	0.24	0.6	0.01	0.13	0.12
	Fe/Zn C4	0.00	0.27	0.26	0.00	0.40	0.55
	Zn JS500	0.37	0.66	0.38	0.39	0.16	0.17
4 weeks	Fe	0.08	0.52	0.54	0.17	0.24	0.38
	Fe/Zn C4	0.01	0.40	0.71	0.11	0.41	0.59
	Zn JS500	0.13	0.42	0.09	0.23	0.01	0.01
6 weeks	Fe	0.31	0.81	0.8	0.50	0.02	0.03
	Fe/Zn C4	0.13	0.24	0.47	0.52	0.12	0.44
	Zn JS500	0.32	0.63	0.41	0.40	0.30	0.34

However, sample sizes in Steps 1 and 2 typically are too small (less than, say, 20, see Appendix 2) for successful likelihood estimation of EV parameters, and we believe that for the present data with values of the shape parameter close to zero (see Table 2), LR tests based on approximation by a Gumbel distribution will detect gross deviations.

Corroboration by graphical and randomisation tests may be prudent. In particular, it can be noted that the randomized version of the LR-test gives correct p-values regardless of whether the data actually comes from a Gumbel distribution or not. As expected the LR test and the randomized LR test led to the same conclusion in most cases. Thus, even for such small sample sizes the simpler asymptotic test worked well. (The exception was the 4 weeks Fe/Zn C4 data. For this case the Gumbel plot was non-linear, and the randomized LR-test should be preferred.) As a further precaution

we also made randomized dispersion and location tests. Generally, these were not as sensitive as the LR tests. This is probably explained by the role of the model and indicates that the model based LR test quantities should be used.

Formally the test of homogeneity uses the assumption that experimental units are replicates, and correspondingly the test of whether units are replicates uses the assumption of homogeneity. However, because of the symmetry of the design it seems very unlikely that this “circularity” could hide the gross deviations we are interested in guarding against.

Of course, even if deviations are not expected, if they occur they would invalidate the subsequent analysis, and indicate a need for improvement of the experimental setup.

To proceed we assume that the analysis in Steps 1 and 2 has not given reason to doubt homogeneity and that units are replicates. We then for the rest of the analysis pool all the observations which stem from the same treatment, and assume that those observations are i.i.d.

### **Steps 3 to 5: analysis of one treatment at a time**

Step 3 is to standardize to meaningful units. This means that results should be presented and discussed in terms of quantities which are of central interest to the problem at hand, rather than in terms of, say, the distribution of the deepest pit in a sector which has no practical meaning outside of the experiment. An example of such a quantity could be the distribution of the deepest corrosion pit on an entire car.

Step 4 is to test if a Gumbel distribution is enough to describe the data from the separate treatments. It is only sometimes relevant. A reason to perform it could be that experience from similar situations indicate that the Gumbel model is likely to be suitable. There is also some theoretical justification for the Gumbel distribution: the lack of memory property of the (approximately) exponential tails of individual variables which are linked to the Gumbel limit distribution for maxima.

Step 5 makes Gumbel plots and fits EV distribution with confidence intervals for each treatment.

3. *Standardization to meaningful units.* The raw data are the maximum pit depths in the sectors. However, as noted above, sectors are introduced only for the purpose of analysis and have no intrinsic interest. Thus it is useful to transform observations and the fitted EV distribution to meaningful units. This could be the experimental units. Or, as an example, in the automotive context, the interest is centered on a car as a unit, and a car may contain several assemblies like the experimental unit, and the standardization should then be made accordingly. For the subsequent analysis, all plots and presentations of results should be made after standardization to meaningful units whenever possible. It is straightforward to do this standardization: see Appendix 1.

*Data analysis:* The maximum pit depth per plate is the interesting quantity rather than the maximum pit depth per sector. We hence standardized to plate as unit wherever possible in the following steps.

4. *EV fit versus Gumbel fit.* As discussed above, in some situations it may be reasonable to check the fit of the Gumbel distribution. We use graphics and a likelihood ratio test of a Gumbel distribution against a general EV distribution for this. Of course, for the latter, lack of evidence against the null hypothesis is not in itself positive proof of good fit of the Gumbel distribution, it just shows that the fit of the EV distribution is no better.

*Data analysis:* Figure 4 contains Gumbel plots with a fitted Gumbel distribution in addition to the EV distribution, for two treatments chosen to illustrate good and less good fit of the Gumbel distribution. Standardization to units is not done in this plot since standardization is model-dependent and would yield different scales on the  $x$ -axes for the Gumbel and EV distributions. Now, consider e.g. the 2 weeks Fe data (Figure 4, top). For these, the Gumbel model gives very good fit, in fact with p-value

equal to 0.99 in the likelihood ratio test of the Gumbel distribution. However, the same conclusion does not apply in all the cases, as for example for the 4 weeks Fe/Zn C4 data (Figure 4, bottom). We hence preferred to use the EV distribution for all the main Gumbel plots in Step 5 below.

*5. Gumbel plots.* Next, EV (or if preferred, Gumbel) distributions with separate parameters for each treatment are fitted, this distribution and the observations are transformed to meaningful units, and the result is presented as a Gumbel plot, with confidence intervals obtained by the delta method. Additional information is inserted by providing the plot with two different y- scales: the left one shows the probability of the pit depth exceeding the level (in percent), (this is  $100(1 - i/(n + 1))$  for the  $i$ -th largest observation), and the other one gives its inverse,  $1/(\text{Probability of exceeding})$ . This is the expected number of units needed to achieve a given depth, and is often termed the “return period”.

*Data analysis:* Figure 5 shows one example of the Gumbel plots with fitted EV distribution and confidence interval for one of the data sets. The plot is standardized to units (=plates) using the fitted EV distribution.

*Comments:* The Gumbel plots from Step 5 contain all the information obtained from statistical analysis performed separately for each treatment. In particular, the answers to many basic questions may be read directly from the plots.

E.g the answer to “What is the expected number of perforated units if one has 1000 units with 1 mm thick plates”, is obtained from Figure 5 by reading that the probability of a pit depth exceeding 1 mm is .0716 and hence the answer is  $1000 \times .0716 = 71.6$ . Preferably this point estimate should be complemented by a confidence interval, which in the same way can be read from the graph to be (21, 230). However, for such extreme quantiles, the likelihood function is rather skew, and profile likelihood intervals (see e.g. Coles (2001), p. 34) give a better representation of the real uncertainty than the delta method, but of course are more computationally demanding.

Similarly, to find out “How thick should the plates be if one wants the expected number of perforated units out of 1,000 to be at most 40”, one reads the x-value corresponding to the probability  $40/1,000 = .04$  from the graph and gets the answer 1.12 mm. Again a delta method or, preferably, a profile likelihood, confidence interval can be constructed to quantify the uncertainty of this estimate. We leave this to the reader.

In Step 4, we have had difficulties fitting EV distributions for sample sizes around 10, (e.g. estimation failed 20 % of the time for sample size 8) while non-convergence was rare (less than 1 %) for sample size 20 or larger, see Appendix 2. The Gumbel distribution may hence be the only viable alternative for small sample sizes. However it of course should only be used if it fits reasonably well.

### **Steps 6 – 7: pairwise comparisons of treatments**

In Step 6 we check if pairs of treatments “lead to the same corrosion mechanism”. Step 7 outlines how pairwise comparisons of treatments can be made, both graphically and formally by computing confidence intervals. A basic property of the present model is that one of a pair of treatments may be preferable in one region, while the other one may be best in another region. Because of this it is possible for the model to discern between situation with many shallow pits, and other, potentially more dangerous, situations with few but deep pits.

*6. Are the corrosion mechanisms the same?* Given two treatments '1' and '2', the observations for Treatment  $i$  are supposed to follow an EV distribution with parameter  $(\xi_i, \sigma_i, \mu_i)$ ,  $i = 1, 2$ , cf. Step 4 above. Let  $G_1$  and  $G_2$  be the distribution functions for the two treatments and write  $\bar{G}_i(x) = 1 - G_i(x)$ ,  $i = 1, 2$ , for the corresponding tail functions. We interpret “different mechanisms” in statistical terms, to mean that differences are not just in location and scale, but in the shape of the distribution. On a more qualitative and from an engineering viewpoint important scale, if the shape parameter  $\xi$  of the EV distribution is negative, then there is an upper bound for the

possible pit depths, while a zero or positive  $\xi$  means that such a bound does not exist. (Note that distributions with infinite upper endpoints often give the best description in the range of interest and should not be ruled out by appealing to finite thickness of the plate. Doing this would be similar to ruling out normal distributions for weights or heights on the ground that any normal distribution gives positive probability to negative values).

Equality of shape is investigated graphically by Gumbel plots with fitted EV distributions, where fits are shown both with the shape parameters assumed to be equal and with free shape parameters. It is also checked by likelihood ratio tests of the hypothesis  $\xi_1 = \xi_2 = \xi$ .

*Data analysis:* The estimates of the shape parameter  $\xi$  were positive for the 2 and

Table 2: *Maximum likelihood estimates of  $\xi$  for the magnesium data. Standard deviations in parentheses.*

	Fe	Fe/Zn C4	Zn JS500
2 weeks	0.084 (0.260)	0.088 (0.148)	0.13 (0.204)
4 weeks	0.027 (0.161)	0.384 (0.205)	0.091 (0.163)
6 weeks	-0.12 (0.167)	-0.079 (0.105)	-0.098 (0.203)

4 weeks data, corresponding to an unbounded distribution, for all types of bolt, while the 6 weeks estimates of the shape parameter are negative and indicate an upper bound for pit depths (Table 2). This could mean different mechanisms for the different time periods, perhaps with a “transition period” at 4 weeks. This, however, is of course quite speculative. The tests of equality of shape parameters are illustrated in Figure 6. The first row in Figure 6 shows an example where both treatments had the same exposure time and where the assumption of equality of the shape parameters does not change the fit. In the second pair one treatment had four weeks of exposure

and the other had six weeks of exposure, and the fit obtained with free parameters looks somewhat different than when the shape parameters are equal. However, the likelihood ratio test did not reject the hypothesis of equality of shape parameters (Table 3). Nevertheless, in this paper we confine attention to comparisons for 2 weeks

Table 3: *P-values for likelihood ratio tests of equality of the shape parameters  $\xi_1$  and  $\xi_2$  for pairs of treatments. “w” stands for “weeks”.*

		Fe		Fe/Zn C4			Zn JS500		
		4w	6w	2w	4w	6w	2w	4w	6w
Fe	2w	0.85	0.50	0.99	0.36	0.55	0.89	0.98	0.58
	4w		0.52	0.78	0.16	0.58	0.69	0.78	0.63
	6w			0.35	0.05	0.82	0.32	0.36	0.92
Fe/Zn C4	2w				0.24	0.35	0.86	0.99	0.46
	4w					0.03	0.40	0.26	0.09
	6w						0.32	0.38	0.93
Zn JS500	2w						0.88	0.41	
	4w							0.47	

and 6 weeks of exposure.

7. *Which treatment is best?* This question is here answered via pairwise comparisons of treatments. Typically several, or all, pairs are compared. This sometimes may lead to considerations of “multiple inference”, see the end of Section 5.

Now, consider a pair of treatments and assume that the previous analysis has not contradicted that the corrosion mechanisms for the two treatments in the pair are same. It is hence assumed that  $\xi_1 = \xi_2 = \xi$  and the EV distributions with parameters  $(\xi, \sigma_1, \mu_1, \sigma_2, \mu_2)$  fitted by maximum likelihood in Step 6 are used. ‘Treatment 1’ is better than ‘Treatment 2’ for a given pit depth  $x_0$ , if the tail functions satisfy

$\bar{G}_1(x_0) \leq \bar{G}_2(x_0)$ , or equivalently if the ratio of the return periods for Treatment 1 and for Treatment 2, i.e.  $\bar{G}_2(x_0)/\bar{G}_1(x_0)$ , is greater than 1. (A stronger statement would be that the ratio is greater than 1 for all  $x$  – however neither the present data nor scientific knowledge of the corrosion process seemed enough for such strong conclusions from the bolt comparisons). To present the comparisons graphically we first recalculate to relevant units (cf. Step 3) and then plot the ratio on a nonlinear scale obtained from a linear scale for  $\bar{G}_2(x_0)/\{\bar{G}_1(x_0) + \bar{G}_2(x_0)\}$ . Finally, two confidence intervals for the ratio are included in the plot. One is obtained by the delta method, and the other from a standard parametric bootstrap. If these intervals do not include 1 at  $x = x_0$ , then there is a statistically significant difference between the treatments for the pit depth  $x_0$ .

*Data analysis:* Figure 7 shows estimates of  $\bar{G}_2(x_0)/\bar{G}_1(x_0)$  as a function of the maximum pit depth  $x_0$ , together with 90% confidence intervals calculated with the delta method and by the parametric bootstrap. In two cases the delta method and the bootstrap confidence intervals differ markedly. For the 2 weeks data, the confidence bounds throughout included 1. For two 6 weeks cases, the confidence bounds did not include 1 for large pit depths, indicating that for plate thicknesses above a certain value the magnesium alloy AZ91D was better in combination with 'Fe' bolts than with 'Fe/Zn C4' bolts. In the same way, 'Zn JS500' bolts were found to be better than 'Fe/Zn C4' bolts. The sizes of these effects can be read off the diagrams, and depend on which thickness one is interested in.

The comparisons were also made using the Gumbel model instead of the EV's. However this led to very similar results, and is not presented here.

*Comment:* There were large differences between the delta method and bootstrap confidence bounds in two cases. These probably were a result of a change of estimated shape parameter from negative to positive in some of the bootstrap samples. This is an indication that the model is not completely stable and that moderate changes of data

can cause large changes in inferences. One should be cautious in the interpretation of such cases.

The confidence bounds in Figure 7 show pointwise intervals, one for each  $x$ -value, and are intended to be used as such: one is interested in a particular material thickness and wants to read off the confidence interval for this  $x$ -value. Bounds which apply to all thicknesses simultaneously would be wider, and this difference could well matter in other applications.

The engineers who performed the experiment had noticed the presence of increasing quantities of corroded materials which could act as a hinderance to further development of the pits. Step 6 provides some additional indication of this possibility. Such speculation should however be corroborated by further chemical and physical knowledge before taken seriously.

## 5. SOME STATISTICAL AND MODELING ISSUES.

In the literature on localized corrosion, except for work by Laycock, Scarf, Cottis, see e.g. Scarf and Laycock (1994) and Cottis et al. (1990), attention has mainly been focused on Gumbel rather than general EV modeling. In this paper, we prefer (as the authors mentioned above) the more flexible EV family. In particular, it can indicate if pits do not continue to grow indefinitely (for  $\xi < 0$ ). This can of course be synonymous with an important reduction of costs. A price for this increased flexibility is that the EV fits require slightly larger sample sizes than Gumbel fits. We performed some simulations to compare the numerical convergence of the maximum likelihood estimations in the two models for small sample sizes. One result was that the numerical maximum likelihood routines we used (the Splus routine “nlminb”, and the R routine “optim”) did not converge for the EV distribution in one fifth of the cases for sample size 8, and the convergence problems were even worse for smaller sample sizes, see Appendix 2. This practical problem is known for small sample sizes, see e.g. Tables 2 and 4 in Drees, de Haan & Li (2005) in a slightly different context.

The first two steps in our method presented are tests of homogeneity and replication. Sample sizes in the Volvo experiment at those steps are very small ( $n = 6$  or  $8$ ). That's the reason for using the Gumbel rather than the EV distribution in these steps, even in cases where the Gumbel distribution may not fit perfectly. Further, as discussed above, in our experience Gumbel based LR-tests are more sensitive than non-specific randomization tests. The randomized version of the LR test, if available, should be preferred to the simpler asymptotic LR test, in particular in case of imperfect Gumbel fit. The aim of the steps is rough safeguards to detect if the experimental conditions have turned out to be different than intended.

The confidence intervals in this paper use the delta method and, for Figure 7, a parametric bootstrap method. Sometimes, in particular when extreme quantiles are estimated, the likelihood function can be quite asymmetric. Profile likelihood methods are then preferable, but require much heavier calculations.

Depending on the setup, prior knowledge, perhaps physical arguments or statistical evidence from similar situations, can speak for more specific models with fewer parameters. Possible candidates are an additive model when the effect of 'Treatment 2' is gotten from 'Treatment 1' by translation, and a multiplicative model when the effects of 'Treatments 1 and 2' are related by a multiplicative change of scale. Specifically, additivity means that the parameters of the underlying EV distributions satisfy the restrictions  $\xi_1 = \xi_2$  and  $\sigma_1 = \sigma_2$ . In the multiplicative model instead  $\xi_1 = \xi_2$ ,  $\sigma_2 = \lambda \sigma_1$  and  $\mu_2 = \lambda \mu_1$ , for some  $\lambda > 0$ . The parameters of the model are straightforwardly estimated by maximum likelihood, and the goodness of fit assessed by looking at Gumbel plots with estimated distribution lines and through likelihood ratio test. In our example comparisons using the additive and multiplicative did not lead to more significant results than the full model.

In the presentation we so far have assumed that the purpose of the experiment was explorative/hypotheses generating. However, if many tests or many confidence

intervals are used, the overall significance level (which controls the risk that at least one of the intervals or tests leads to the wrong conclusion) can be much less than the one for the individual comparisons. Accordingly, if one wants to make formal inference with a controlled overall significance level, multiple inference methods have to be used. In our analysis based on asymptotic normality, Tukey's method with an infinite number of degrees of freedom is appropriate (see Hsu 1996, p. 119). According to this method, if one e.g. has 6 different treatments one obtains an overall 5 % confidence level for all  $6 \times 5$  possible confidence intervals for pairwise differences by just making all intervals 45 % wider than for a single comparison. As further examples, for 12 treatments the intervals have to be 67 % wider and for 16 treatments 75 % wider. Similarly, for testing, treatment as multiple tests with a predetermined level of significance gives the corresponding scale changes in the power function, again see Hsu (1996).

## 6. SUMMARY AND CONCLUSIONS.

In this paper we have developed a strategy for comparing treatments with extreme value distributed responses. The method was successfully applied to an experiment on pit corrosion for magnesium alloys. It was motivated by needs of the automotive industry. We believe it is a useful tool for many kinds of corrosion problems and also in other contexts, such as in material fatigue and some medical and financial settings.

The approach throughout uses graphical methods, and is based on fitting EV distributions and on maximum likelihood estimation and testing. Different observation schemes, in the corrosion context measuring all pits deeper than some specified threshold, would instead lead to using the Peaks over Thresholds method and the Generalized Pareto (GP) models, see e.g. Coles (2001). It would be straightforward to translate our method to such situations.

**Acknowledgement:** We want to thank Malte Isacson for initiating this work and him, Gunnar Ström and Zheng Tan for many stimulating discussions and ideas.

The magnesium corrosion experiment was performed and first analyzed by Jenny Andersson, and we would also like to acknowledge her help. Thanks to Nader Tajvidi for constructive discussions and comments. We also thank two referees and an associate editor for very useful and thoughtprovoking comments which led to substantial improvement of the paper. •

#### APPENDIX 1: STANDARDIZATION TO UNITS

The EV distribution is preserved after taking the maximum of i.i.d. variables, as already mentioned in Section 2. More precisely, let us assume that the maximum per block,  $X$ , follows a Gumbel distribution with parameters  $(\mu, \sigma)$ , and suppose that the unit of interest consists of  $k$  independent blocks. Then the maximum per unit,  $X_k$ , follows a Gumbel distribution with parameters  $(\mu + \sigma \log k, \sigma)$ . Analogously, if  $X$  has an EV distribution with parameters  $(\xi, \sigma, \mu)$ , then  $X_k$  follows an EV distribution with parameters  $(\xi, \sigma k^\xi, \mu + \sigma/\xi [k^\xi - 1])$ . As a consequence, the results become expressed *per unit* if the  $x$ -axis of the Gumbel plot is transformed via  $x \mapsto x + \sigma \log k$  ( or via  $x \mapsto k^\xi(x - c) + c$ , where  $c = \mu - \sigma/\xi$ , for the EV distribution). •

#### APPENDIX 2: PERFORMANCE OF MLE FOR EV PARAMETERS

Simulations were performed (with Splus optimization routine "nlminb") to investigate the small sample behavior of the maximum likelihood estimators in the EV and Gumbel models (see Table 4). No numerical convergence problems occurred for the Gumbel distribution. For small sample sizes, maximum likelihood estimation of the parameters of the EV distribution sometimes failed. For more results on the estimation errors, see Fougères et al. (2002). Simulations were also made in R. The results were very similar to those in Table 4. In these simulations we used the R routine "fgev" (in the package 'evd', at <http://cran.us.r-project.org/>) to call the R optimization routine 'optim'. We asked "fgev" to compute standard deviation. If one does not ask

for standard deviation, the percentage of convergence becomes higher, but the results then are not reliable for small sample sizes such as 5 or 10.

Table 4: *Proportion of convergent maximum likelihood estimation in the EV model. For each sample size  $n$  and shape parameter  $\xi$ , 500 samples were simulated. The parameters  $\sigma$  and  $\mu$  were equal to 1 and 0, respectively.*

$n \setminus \xi$	-0.25	-0.1	0	0.1	0.25
5	0.452	0.520	0.536	0.536	0.506
8	0.722	0.784	0.838	0.870	0.892
10	0.818	0.896	0.932	0.952	0.964
15	0.942	0.980	0.988	0.992	1
20	0.996	0.988	1	0.996	1

## REFERENCES

- Aziz, P. M. (1956). Corrosion **13**, 495.
- Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.
- Cottis, R. A., Laycock, P. J. & Scarf, P. A. (1990). Extrapolation of extreme pit depths in space and time. *J. Electrochem. Soc.* **137**, 64-69.
- Drees, H., de Haan, L. & Li, D. (2005). Approximation to the tail empirical distribution function with application to testing extreme value conditions. To appear in *J. Statist. Plan. Inf.*.
- Embrechts, P. Klüppelberg, C. & Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Springer, Berlin.
- Fougères, A.-L., Holm, S., & Rootzén, H. (2002). Pitting corrosion: Comparison of two treatments with extreme value distributed responses, extended version. *Technical report*,

*Mathematical Statistics, Chalmers* **2002:33**. (<http://www.math.chalmers.se/Stat/Research/Preprints/index.cgi>)

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia Univ. Press, New York.

Hosking, J. R. M., Wallis, J. R., & Wood, E. F. (1985). Estimation of the generalized extreme value distribution by the method of probability-weighted moments. *Technometrics* **27**, 251-261.

Hsu, J. C. (1996). *Multiple comparisons, theory and methods*. Chapman & Hall, London. •

Isacsson, M., Ström, M., Rootzén, H. & Lunder, O. (1997). Galvanically induced atmospheric corrosion on magnesium alloys: a designed experiment evaluated by extreme value statistics and conventional techniques. *The Engineering Society for Advancing Mobility Land Sea Air and Space, Technical Paper 970328*.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*. 2nd edition. Wiley, New York.

Kowaka, M. (1994). *An Introduction to Life Prediction of Plant Materials. Application of Extreme Value Statistical Methods for Corrosion Analysis*. Allerton Press, New York.

Murakami, Y. & Beretta, S. (1999). Small defects and inhomogeneities in fatigue strength: Experiments, models and statistical implications. *Extremes* **2** (2), 123-147.

Murakami, Y. & Usuki, H. (1989). Quantitative evaluation of effects of non-metallic inclusions on fatigue strength of high strength steels II: Fatigue limit evaluation based on statistics for extreme value of inclusion size. *Int. J. Fatigue* **11** (5), 299-307.

Scarf, P. A. & Laycock, P. J. (1994). Applications of extreme value theory in corrosion engineering. *J. Res. Natl. Inst. Stand. Technol.* **99**, 313-320.

Shibata, T. (1996). Statistical and stochastic approaches to localized corrosion. *Corrosion* **52** (11), 813-830.

Takahashi, R. & Sibuya, M. (1996). The maximum size of the planar sections of random spheres and its application to metallurgy. *Ann. Inst. Statist. Math.* **48** (1), 127-144.

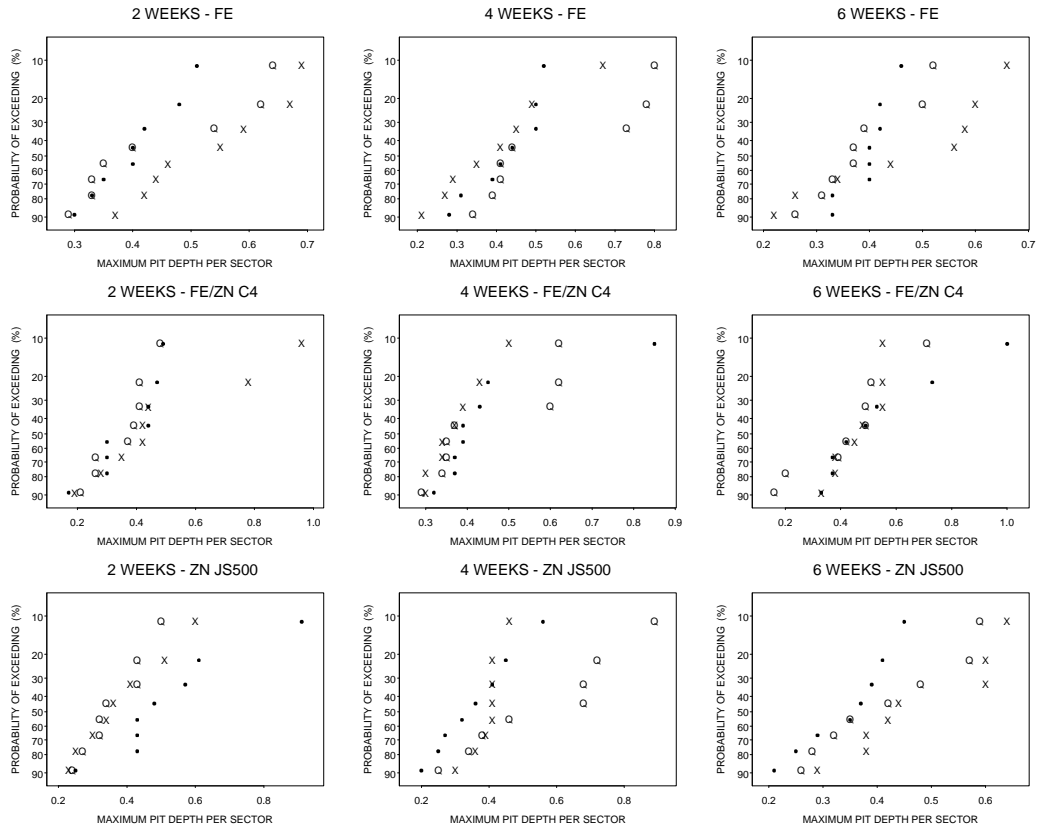


Figure 3: Gumbel plots of maximum pit depths per sector, for the three replicates and for the nine treatments. Symbol '•' is the first sample, symbol 'Q' is the second sample, and symbol 'X' is the third one. The first column provides the 2 weeks data, the second one the 4 weeks data, and the third one the 6 weeks data. Rows show from top to bottom "Fe", "Fe/Zn C4" and "Zn JS500".

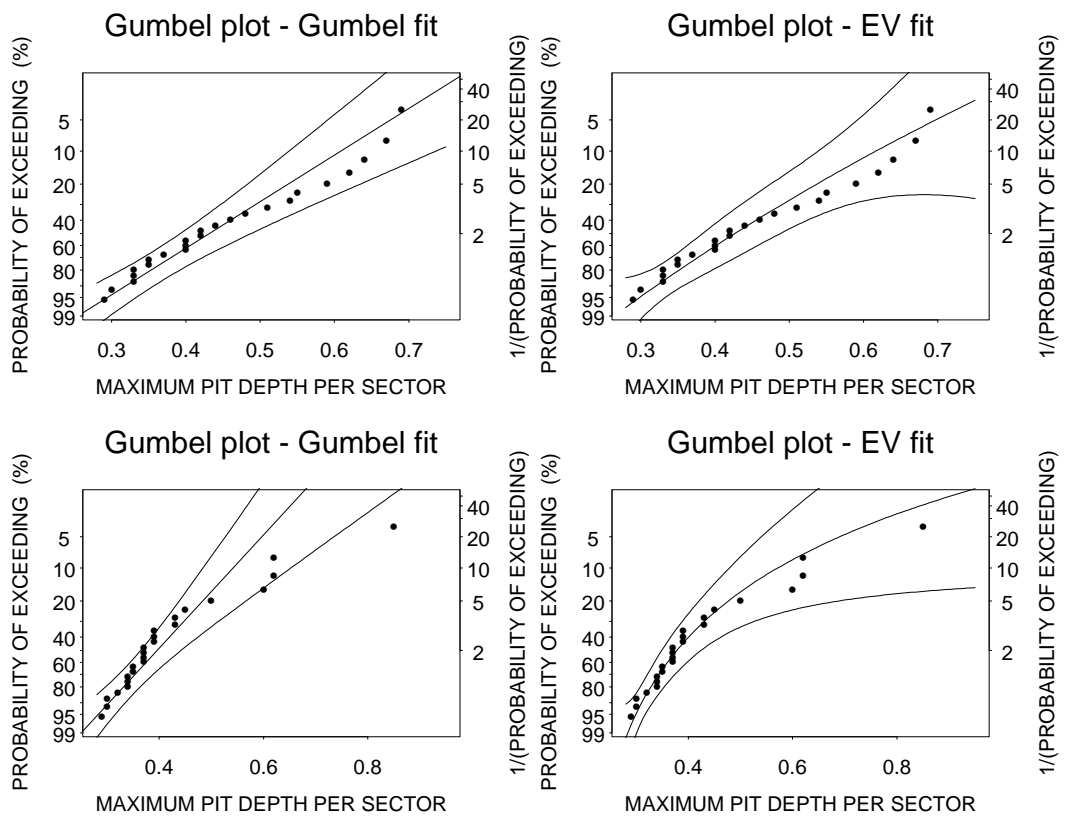


Figure 4: Gumbel plots with Gumbel and EV fits of maximum pit depths per sector, for magnesium data, with associated 95% confidence intervals. First row shows the 2 weeks-Fe data, and second row the 4 weeks-Fe/Zn C4 data.

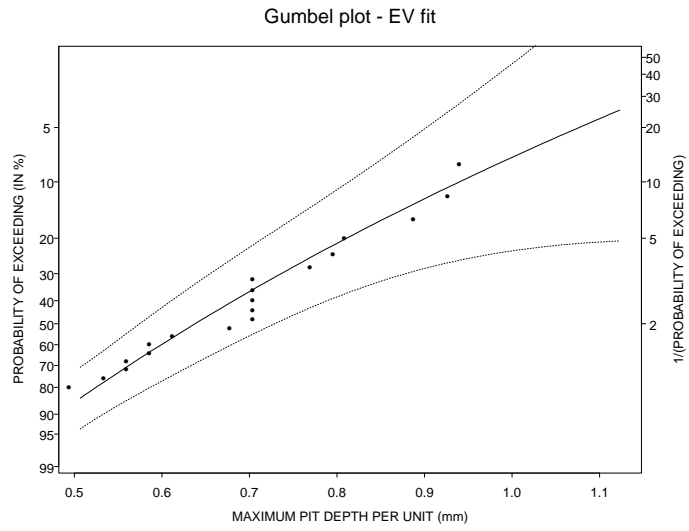


Figure 5: *Gumbel plot with EV fit corresponding to the maximum pit depth per unit, for magnesium data with JS500 coated zinc bolts, after 2 weeks.*

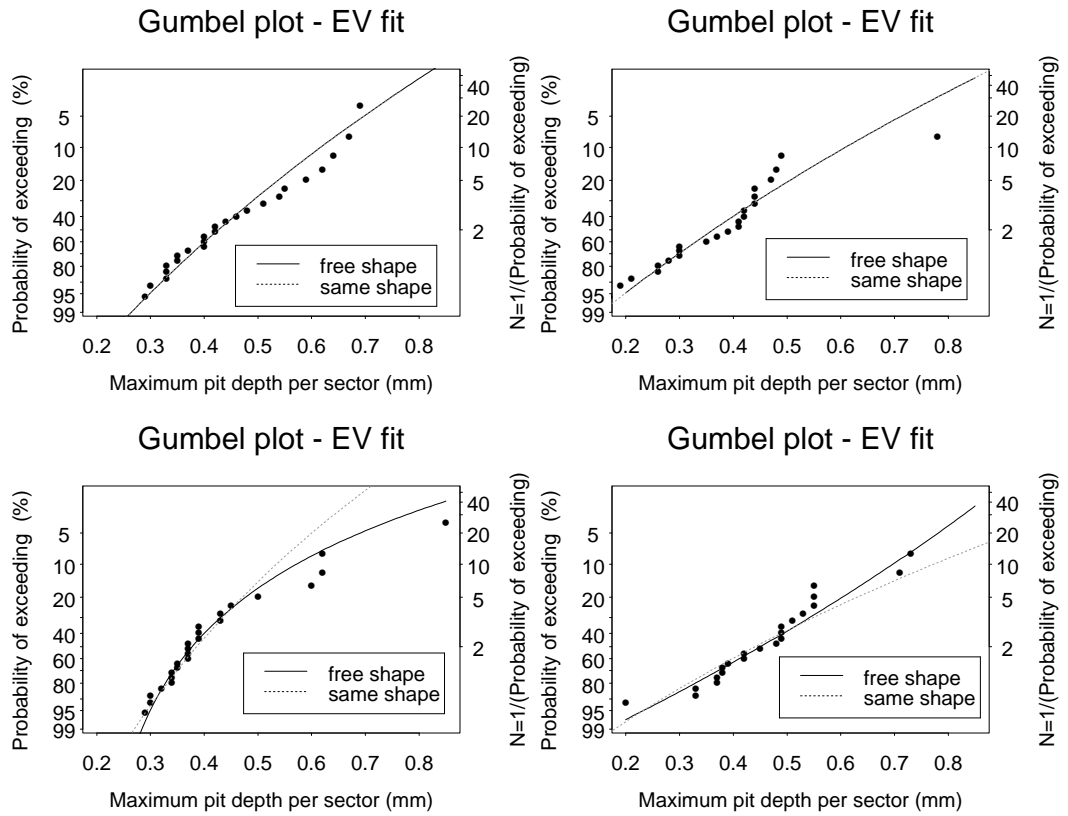


Figure 6: *Gumbel plots with EV fits of maximum pit depths per sector, for pairs of treatments (called Treatment 1, Treatment 2), with shape parameters assumed to be equal or free. Each row shows one pair. The first column shows Treatment 1, where the fitted lines are indistinguishable and the second column Treatment 2. The top pair is 2 weeks-Fe, 2 weeks-Fe/Zn C4 and the bottom pair is 4 weeks-Fe/Zn C4, 6 weeks-Fe/Zn C4.*

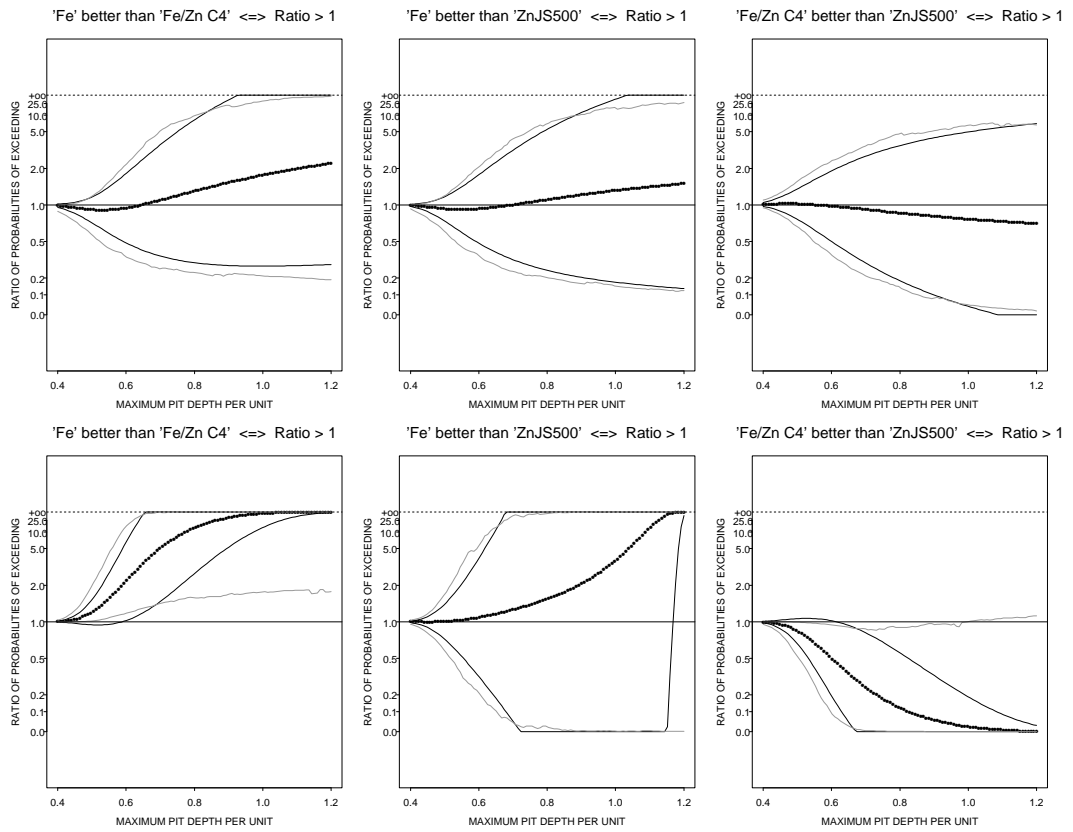


Figure 7: Estimation of the ratio of the return periods in terms of pit depth per unit (.....) with associated 90% confidence interval (—). The irregular lines are the bootstrap confidence intervals. The first row provides the 2 weeks data, and the second one the 6 weeks data.