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

A.-L. FOUGÈRES [1,3], S. HOLM [2,3] and H. ROOTZÉN[2,3]

Summary

Localized, or "pitting", corrosion can limit the usefulness of aluminum, magnesium and other new lightweight materials. Use of such materials reduces weight and hence fuel consumption in automobiles, and is attracting major interest in the car industry. Judicious choice of alloys and surface treatments is necessary to avoid corrosion problems. 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 assumptions of normality and homogeneity of variances typically are not satisfied by pit depth measurements. ANOVA hence will give wrong predictions of corrosion life lengths. In this paper we instead use Statistical Extreme Value Theory to develop methods to validate and improve experiments with extremal responses, and to compare and extrapolate the results. Our approach is intended as an engineering tool ready for routine use. It is presented as a flowchart where each step involves both suitable statistical tests and confidence intervals, and where the results are presented in carefully chosen graphs. The method is applied to corrosion tests performed at Volvo Car Company, and was developed for this purpose. However we believe that it is widely useful both inside and outside the field of corrosion testing.

Some key words: Pitting corrosion; Extreme value modeling; Comparisons of treatments.

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

1. Introduction

Making cars lighter is a main way to reduce fuel consumption, and is a central challenge for the automotive industry. Localized corrosion (also called "pitting" or "galvanic" or

[1]Département de Génie Mathématique et Modélisation, Institut National des Sciences Appliquées 135 Avenue de Rangueil, 31077 Toulouse Cedex 04, France. Email: fougeres@gmm.insa-tlse.fr

[2]Department of Mathematics, Chalmers University of Technology, S-412 96 Gothenburg, Sweden. Email: rootzen@math.chalmers.se

[3]Research supported by VCC/Ford, by the Knut and Alice Wallerberg foundation, and by TMR Spatial and Computational Statistics Network

“bimetallic” corrosion) is, in particular, important for car designers since it puts limits on 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. Isacsson et al. (1997)). The second is that the standard ANOVA assumptions of normality and homogeneity of variances typically are not satisfied by pit depth measurements. Hence, using ANOVA in this context will give misleading results. Instead, Extreme Value Statistics has appeared as a successful alternative way to analyse data on pitting corrosion, see e.g. Shibata (1996). In fact, such methods were suggested in this context already by Aziz (1956) and Gumbel (1958), and in many subsequent papers and is developed and promoted in a book, Kowaka (1994). However, this literature does not seem to include any statistical advise on how to compare different corrosion-reducing treatments. In the present paper we develop new analysis methods for this. The purpose is to make it easily feasible to validate and improve experiments with extremal responses, and to use the results to compare treatments and to extrapolate the experimental results.

Our approach is intended as an engineering tool ready for routine use. It is presented as a flowchart where each step involves both suitable statistical tests and confidence intervals, and where the results are presented in carefully chosen graphs. The method has been applied to corrosion tests for lightweight materials performed at Volvo Car Company, and was developed for this purpose. However we believe that these methods to compare datasets with extreme value responses in fact are much more widely useful, both inside and outside the field of corrosion testing.

As an example of application of the proposed methodology in a different context than corrosion testing, one could consider the fatigue limit of metallic materials. The fatigue limit is the “threshold stress for nonpropagation of the crack which emanated from original cracks, defects or inhomogeneities” (Murakami and Beretta (1999)). This threshold stress is closely related to the presence of nonmetallic inclusions in the material, or, more precisely, to the size of the largest inclusion in the specimen of interest. Extreme value models are therefore widely used in this context, as pointed out by Murakami and Usuki (1989), or Takahashi and Sibuya (1996) and many others (see Murakami and Beretta (1999) for an overview of the litterature in this area). Comparisons between two materials in terms of their fatigue limit hence typically require statistical methods for extreme value distributed data.

Comparison of portfolios of financial assets in terms of the risks they incur, and choice of reinsurance contracts, are examples outside engineering where some of the same questions arise.

The aim of this paper is to provide statistical methodology for comparisons of extreme value data in a form which is easily accessed and used, also for non-experts. Hence its main contribution is the careful selection and packaging of methods. However contributions also include information on the performance of the methods, modification of available graphical methods and introduction of new graphs.

The description of the methodology is given in the context of pit corrosion. Section 2 contains the basic tools of the extreme value modeling and a very brief "picture" of the area of pit corrosion. A flowchart which summarizes the methodology is presented in Section 3, which also contains comments on the individual steps in the flowchart. An illustrating dataset is analysed in Section 4. Section 5 deals with some of the new statistical and modeling issues which arise, and Section 6 contains our conclusions. Some technical issues are relegated to appendices.

2. Extreme value theory and pit corrosion

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. It also deals with suitably chosen subsets of the largest variables, as in the so-called Peaks over Threshold models. The underlying variables may be actually observed or hypothetical. Useful recent introductions to the area are Coles (2001) and Embrechts et al. (1997). Coles gives an uptodate account of statistical methods while Embrechts et al. contains the basic theory, but from an economic perspective. An earlier account of the theory with an engineering point of view is Leadbetter et al. (1983). The theory is semi-parametric, only maxima and large values are modelled. Even if one has a model, such as the normal distribution, for central values of the underlying variables, the models is not trusted to be accurate enough in the tails for reliable use.

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 μ, σ and ξ are the location, scale and shape parameters, respectively. For $\xi = 0$ the formula should be interpreted as the limiting (as $\xi \to 0$) Gumbel distribution

$G(x) = \exp[-\exp\{-(x-\mu)/\sigma\}]$. 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 $n \to \infty$, 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.

The EV distribution in particular is the limiting distribution for maxima of most standard continuous distributions, such as the normal and Weibull ones, and it has a suitable amount of flexibility built in.

Several methods for estimation of the EV parameters have been proposed, see for example Hosking, Wallis and Wood (1985), and the review by Johnson, Kotz and Balakrishnan (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 3 for further discussion) and is much more general and flexible than the competitors. The maximum likelihood method also has the advantage of producing simple asymptotic confidence intervals via the delta method (see e.g. Coles (2001), p. 33), and the possibility of using profile likelihood confidence intervals if additional accuracy is needed (again see Section 5 for details). In the present paper we use maximum likelihood estimation and the delta method for confidence intervals throughout.

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. In Gumbel plots the ordered observations are plotted on a scale which is constructed so that the values are scattered around a straight line if they come from a Gumbel distribution. More precisely, if $X_1, \ldots, X_n$ are i.i.d. observations, the Gumbel plot shows the graph

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

where $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the observations ordered in ascending order. The distribution function of the fitted EV distribution is also shown, and appears as a convex curve with a finite upperbound if the estimated shape parameter ξ is negative, a straight line for Gumbel case $\xi = 0$, and a concave curve if $\xi > 0$.

Our main, and motivating, example is corrosion, and we hence give a very superficial sketch of this area. Numerous examples of applications of extreme value modeling to localized corrosion exist in the literature, see for instance in Kowaka (1994), Table 4.4 p. 32. We believe it is essential to the understanding of new general methodology to have a concrete example in the back of the head, and hence also give a detailed description of a specific experiment with pitting corrosion. This example is used throughout.

In chemical terms galvanic corrosion is the consequence of an oxidation-reduction reaction. Such a reaction is the result of 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 extent 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. At present quantitative practical prediction of corrosion behavior from chemical principles is not possible. 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.

We next describe a typical experiment setup for a galvanic corrosion test, in the specific setting of the dataset analyzed in Section 4 below. It should be re-emphasised that our methodology is general, and applies also to field testing and to other metals, as well as to problems outside the corrosion area.

In an experimental assembly, an experimental unit, in our example a circular plate of magnesium alloy (Figure 1), is joined to another metal, which in our case is a steel bolt.

Please insert Figure 1 about here.

The assembly is then exposed to cycles of exposure to humidity for specified amounts of time and at a specified temperature. The particular 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.

For each specific treatment (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 corrosion products are dissolved from the units. Each unit is divided in k blocks, and the maximum pit depth in each block is measured. For the dataset in Section 4, the measurements were made with direct radiography, using a technology provided by AGFA (trademark: DirectRay). The treatments consisted of different coatings for the bolt and different lengths of exposure. Each circular magnesium plate was divided into $k = 8$ sectors (or blocks) and usually $n = 3$ assemblies were used for each treatment. A typical dataset hence consisted of $nk = 24$ measurements of block maxima for each treatment.

Now, how should one compare the efficiency of the treatments? A methodology for analysis of such datasets is proposed in the following section.

3. The flowchart

Figure 2 shows a flowchart which illustrates the steps of the method.

Please insert Figure 2 about here.

To fix ideas, the figure is in terms of the dataset introduced in the previous section. Recall that it concerned experimental assemblies which consisted of circular magnesium plates (the units) joined to steel bolts which were treated in different ways. Each treatment was applied to three assemblies, and each unit was divided into 8 sectors (or blocks). The dataset consists of measurements of the deepest pit in each such sector. Of course, in general the blocks may be rectangles or other kind of shapes, but for ease of description we still call them "sectors".

A step-by-step description of the flowchart follows. Each step consists of similar elements: First, a graph is provided: this is often the most eloquent part of the method. Then a parametric likelihood ratio test based on a Gumbel or an EV modeling is performed. Finally, randomization tests 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. Explicit formulas for the test statistics are postponed until Appendix 1.

The description in this section is for a general situation, but we still use the graphs from the specific dataset in Section 4 below to illustrate what we mean. For more explanation of the graphs we refer to Section 4.

Steps 1 and 2: preliminary study of the data

The first two steps in the flowchart check that units are replicates and homogeneous, or in statistical terms that the 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 of the flowchart in the way outlined below – modifications would be needed. Corrosion experiments are difficult to perform, and we believe that the possibility of checking and improving experimental methods provided by Steps 1 and 2 is an important part of our analysis.

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 (Figure 4). 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 would then correspond to different parameter values in the different groups and it is checked by a likelihood ratio test if the parameters are significantly different or not. Since sample sizes for this comparison often will be small it might be prudent to corroborate the test by randomization tests. Three such tests are proposed in Appendix 1. 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.

2. Are units replicates ? The graphical test to see if there are 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 (Figure 5). This is then complemented by a likelihood ratio test of the hypothesis that the separate parameters for the different experimental units in fact are the same. If there is doubt about the accuracy of the test it is possible to make randomization tests instead, in the same way as for Step 1.

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 substantial deviations. In particular, in the subsequent analysis we use the general EV distribution, and don't restrict ourself to the Gumbel model. However, sample sizes in Steps 1 and 2 typically are too small (less than, say, 20, see Appendix 3) for successful likelihood estimation of EV parameters, and approximation by a Gumbel distribution should be good enough to detect deviation of interest. Of course, even if deviations are not expected, they would invalidate the subsequent analysis, and indicate a need for improvement of the experimental setup, and it is hence important to look for them.

As with all statistical graphs it requires experience to be able to separate systematic deviations from the random variation inherent in the model. A good way to get experience is to simulate samples with the same structure and distribution and then look at the same plots for the simulated data as for the real data (see Figure 6). Similar experimentation with the tests can also be useful.

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

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 are to standardize to meaningful units, and to make Gumbel plots and fit EV distribution with confidence intervals for each treatment. Step 4 is to test if a Gumbel distribution is enough to describe the data from the separate treatments. This is only sometimes relevant.

3. Standardization to meaningful units. The raw data are the maximum pit depths in the sectors. However, usually sectors are introduced only for the purpose of analysis, to increase the number of observations, 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: the exact procedure is given in Appendix 2.

4. EV fit versus Gumbel fit. Surprisingly little seems to be lost by using the more general EV distribution even if the Gumbel distribution is the true one. However, in some situations it may still be reasonable to use graphics (Figure 7) and a likelihood ratio test to check if the Gumbel distribution is a good model. In particular this might be reasonable if further data with the same structure but sample sizes too small for EV likelihood estimation is to be analysed later.

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 (see Figure 8), 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/(Probability of exceeding). This is the expected number of units needed to achieve a given depth, and is usually termed the "return period".

Comments: The Gumbel plots from Step 5 contains 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 8 by reading that the probability of a pit depth exceeding 1 mm is .00716 and hence the answer is 1000 x .00716 = 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 delta method produces too wide confidence intervals, and profile likelihood intervals (see e.g. Coles (2001), p. 34) may be preferable.

Similarly, to find the answer to "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.

Regarding Step 4, we have experienced difficulties with EV fit 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 3 for further details. The Gumbel distribution may hence be the only viable alternative for small sample sizes, provided one knows that the Gumbel model fits reasonably well.

Steps 6 – 8: pairwise comparisons of treatments

As before, we assume that Steps 1 and 2 indicate that the data for each treatment can be pooled and considered as i.i.d. observations of a same phenomenon. Given two treatments '1' and '2', the observations for Treatment i are supposed to follow an EV distribution with parameter (ξ_i, σ_i, μ_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. In Step 6 we check if the two treatments "lead to the same corrosion mechanism". In Step 7 the two treatments are compared, both graphically and formally by computing confidence intervals. A basic property of the present model is that one of the treatments may be preferable in one region, while the other one may be best in the another region. The final step, Step 8, is only relevant in special situations where one may have reason to believe in specific simpler ("additive" or "multiplicative") structures.

6. Are the corrosion mechanisms the same? We interpret this question in statistical terms to mean "Do the distributions of pit depths for the two treatments have the same shape parameters ξ?". On a more qualitative scale, if the shape parameter ξ of the EV distribution is negative, then there is an upper bound for the possible pit depths, while a zero or positive ξ means that such a bound does not exist. Equality of shape parameters also makes the statistical comparisons of treatments easier. Equality is investigated by Gumbel plots with EV fits, where fits are provided both with the shape parameters supposed to be equal and with free shape parameter (Figure 9), and by likelihood ratio tests of the hypothesis $\xi_1 = \xi_2 = \xi$.

7. Which treatment is best? We assume that the previous analysis has not contradicted that the corrosion mechanisms for the two treatments 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. To present this graphically we first recalculate the survival functions to relevant units (cf. Step 3) and then plot the ratio on a non-linear 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 by a parametric bootstrap (see Appendix 4 for details). If these intervals do not include 1 at $x = x_0$, then there is a significant difference between the treatments for the pit depth x_0. Figure 10 shows examples of these plots.

8. Simpler models? Depending on the setup, a priori knowledge, perhaps physical arguments or statistical evidence from similar situations, can speak for more specific models with fewer parameters. Here we consider an additive model (when the effects for 'Treatment 1' can be deduced from results for 'Treatment 2' via a translation), and a multiplicative model (when the effects for 'Treatment 1' and 'Treatment 2' are by a multiplicative change of scale).

Specifically, the additive model means that the parameters of the underlying EV distributions satisfy $\xi_1 = \xi_2$ and $\sigma_1 = \sigma_2$. In the multiplicative model it is instead assumed that $\xi_1 = \xi_2$, $\sigma_2 = \lambda\,\sigma_1$ and $\mu_2 = \lambda\,\mu_1$, for some $\lambda > 0$. The parameters of the model can then be 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.

Comment: It seems to us that the plot in Step 7 is a simple and very useful way of summarizing the comparisons of two treatments with extreme value distributed responses.

4. Corrosion of magnesium plates with steel bolts.

In this section, the flowchart from Section 3 is used to analyse a laboratory experiment performed at Volvo Car Company. The experiment concerned galvanic corrosion of steel on magnesium. Circular plates of the magnesium alloy Mg AZ91D (Figure 1) 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"). The units were covered with synthetic dirt (89 % washed sea sand, 9 % kaolin, 1 % active carbon, 1 % sodium chloride) and 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 (Isacsson et al, 1997). The experiment was performed with 9 plates per type of bolt. Three plates ("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. 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 8 sectors. The maximum pit depth in each sector was measured using direct radiography. The dataset obtained thus consists of an 8x27 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.

Preliminary study of the data

Step 1: To check whether 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 3.

Please insert Figure 3 about here.

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. These pit depths are shown in the dot-diagram of Figure 4.

Please insert Figure 4 about here.

The Fe bolts (top row in Figure 4) 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 4.

This conclusion mainly agrees with the results of the formal statistical tests (see Appendix 1 for details on the test procedures) reported in Table 1, columns 1 to 4.

The p-values for the more specific parametric LR-tests are smaller than for the randomization tests based on the Gumbel distribution and for the location and dispersion tests, as could be expected. The difference is particularly large for the 4 weeks Zn JS500 measurements, perhaps because these were taken at the 4 weeks "transition period", see Step 6 below. The location and dispersion tests measure different kinds of deviations from the null hypotheses, and the p-values also differ.

Table 1: *P-values of tests performed in Steps 1 and 2 of the preliminary study: 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
	Fe	0.01	0.24	0.11	0.01	0.13	0.12
2 weeks	Fe/Zn C4	0.08	0.52	0.07	0.17	0.24	0.38
	Zn JS500	0.31	0.81	0.65	0.50	0.02	0.03
	Fe	0.00	0.27	0.00	0.00	0.40	0.55
4 weeks	Fe/Zn C4	0.01	0.40	0.12	0.11	0.41	0.59
	Zn JS500	0.13	0.24	0.70	0.52	0.12	0.44
	Fe	0.37	0.66	0.59	0.39	0.16	0.17
6 weeks	Fe/Zn C4	0.13	0.42	0.28	0.23	0.01	0.01
	Zn JS500	0.32	0.63	0.39	0.40	0.30	0.34

Step 2: Figure 5 contains Gumbel plots for each set of three replicate units, with different symbols for each replicate, and provides a graphical test of whether the units were replicates or not. As a basis for judging if deviations are serious, Figure 6 shows the corresponding plot for a set of i.i.d. simulated data from an EV distribution with parameters $(\xi, \sigma, \mu) = (0.02, 0.1, 0.4)$. The results of likelihood ratio and randomization tests of the hypothesis that units are replicates are reported in Table 1, columns 5 and 6. No consistent pattern which would indicate that units are not replicates is seen.

Please insert Figures 5 and 6 about here.

Analysis of one treatment at a time.

Since Step 1 and 2 did not indicate otherwise, all the data coming from a given treatment (i.e. one type of bolt, one period of exposure) was pooled. This yielded 9 datasets of size 24, which were analysed separately in Steps 3 – 5 below.

Step 3: 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.

Step 4: Figure 7 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.

Please insert Figure 7 about here.

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 7, first line of graphics). 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. It therefore would seem preferable, as it provides narrower confidence bands. However, the same conclusion does not apply in all the cases, as for example for the 4 weeks Fe/Zn C4 data. We hence preferred to use the EV distribution for the main Gumbel plots in Step 5 below.

Step 5: Figure 8 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. This is the final result of the analysis of the treatments one by one.

Please insert Figure 8 about here.

Pairwise comparisons of treatments

The final steps are geared at the end goal of comparing the treatments.

Step 6: The estimates of the shape parameter ξ were positive for the 2 and 4 weeks

Table 2: *Maximum likelihood estimates of ξ for the magnesium data. Standard deviations are given between 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)

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. The tests of equality of shape parameters are illustrated in Figure 9. The figure contains Gumbel plots with the EV fits which are obtained when the shape parameters are assumed to be equal and when the shape parameters are "free", i.e. with separate shape parameters for the two treatments. The first row in Figure 9 shows an example where both treatments had 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.

Please insert Figure 9 about here.

However, the likelihood ratio test of equality of shape parameters (Table 3) shows that the differences, if they exist, are not large enough to be statistically significant. Nevertheless, we confine attention to comparisons for 2 weeks and 6 weeks of exposure.

Table 3: *P-values for the likelihood ratio test of the equality of the shape parameters ξ_1 and ξ_2 of the EV fits for two treatments. Abbreviation "w" states 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

Step 7: For a given level x_0, as discussed in Section 3, we say 'Treatment 1' is better than 'Treatment 2' if one has $\bar{G}_2(x_0)/\bar{G}_1(x_0) \geq 1$, and otherwise 'Treatment 2' is better. In Figure 10 the estimates of this ratio as a function of the maximum pit depth, for some combinations of two treatments, are plotted, together with 90% confidence intervals calculated with the delta method and using a parametric bootstrap (see

Appendix 4 for details). In two cases the delta method and the bootstrap confidence intervals differ markedly. The reason may be non-linearities which are too strong for the delta method to work well.

Please insert Figure 10 about here.

For the 2 weeks data, the different types of bolt do not have a statistically significant effect. The 6 weeks data lead to significant differences in two cases, for pits deeper than a certain value: The corrosion behavior of the magnesium AZ91D together with the 'Fe' type of bolt appears significantly better than together with the 'Fe/Zn C4' bolt. Similarly, the 'Zn JS500' bolts showed significantly better corrosion behavior than the 'Fe/Zn C4' bolts.

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

Step 8: In this step we fitted two models with fewer parameters, the additive and the multiplicative one, as discussed in Section 3. The paired comparisons resulting from these models are summarized in Table 4. The simpler models did not lead to more significant results than the general one.

Table 4: *Paired comparisons based on the ratio of the return periods for the* 2 *and* 6 *weeks magnesium data. The sign '* $\equiv$*' means that there is no significant difference, '*$\prec$ *' (or '*$\succ$*') means that the first treatment is significantly worse (or better) than the second treatment for some range of pit depths, at the confidence level 0.95 obtained by the delta method. P-values for the likelihood ratio tests for the validity of the submodels in parentheses.*

		General model	Multiplicative model	Additive model
2 weeks	Fe / Fe/Zn C4	$\equiv$	$\prec$ (0.04)	$\prec$ (0.30)
	Fe / Zn JS500	$\equiv$	$\prec$ (0.21)	$\prec$ (0.55)
	Fe/Zn C4 / Zn JS500	$\equiv$	$\equiv$ (0.41)	$\equiv$ (0.66)
6 weeks	Fe / Fe/Zn C4	$\succ$	$\equiv$ (0.15)	$\equiv$ (0.05)
	Fe / Zn JS500	$\equiv$	$\equiv$ (0.66)	$\equiv$ (0.74)
	Fe/Zn C4 / Zn JS500	$\prec$	$\equiv$ (0.31)	$\equiv$ (0.11)

Comment: Step 6 could lead one to speculate that perhaps the corrosion process tends to stop as time goes on. This might be a consequence of the presence of increasing quantities of corroded materials which "protect" the alloy. Such speculation should however be corroborated by chemical and physical knowledge before taken seriously.

5. Some statistical and modeling issues.

In the literature on localized corrosion, attention has been focused on Gumbel rather than general EV modeling. The two parameters Gumbel model may be better than the three parameters EV one when the number of measurements is small, provided the Gumbel model fits the data. We performed some simulations to compare the behavior of the maximum likelihood estimators in the two models for small sample sizes, see Appendix 3. One result was that the numerical maximum likelihood routine we used (the Splus routine "nlminb") 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. However, if sufficient data is available, we prefer 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.

The first two steps in the methodology presented in this paper (Figure 2) are tests of homogeneity and replication and use the Gumbel distribution even in cases where this distribution may not fit perfectly. The idea is that we believe the experiment has been carefully designed, but still want some rough safeguards to detect if things have not turned out as intended.

The confidence intervals in this paper use the delta method and, for Figure 10, a parametric bootstrap method. In some circumstances, specifically when extreme quantiles are estimated, the likelihood function can be quite asymmetric. Profile likelihood methods may then be preferable, even if they require heavier calculations.

6. Summary and conclusions.

In this paper a general methodology for comparing two treatments with extreme value distributed responses is given (for an overview see the flowchart, Figure 2). The methodology has been successfully applied to pit corrosion for magnesium alloys, with the responses being the maximum pit depth in a sector of a test plate. It is motivated by needs of the automotive industry. However, we believe that it also is a useful tool for other corrosion problems and 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 flowchart to such situations.

Several extensions, such as simultaneous comparison of more than two treatments and the introduction of regression models, e.g. describing the influence of time, are of interest. More generally, many interesting questions concerning designed experiments with EV or GP distributed responses remain. We plan to continue our work on these problems.

Acknowledgement: We want to thank Malte Isacsson 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. Our thanks finally to Nader Tajvidi for constructive discussions and comments.

Appendix 1: Details for test procedures.

1. Homogeneity tests. Recall the setup of Section 2: for each treatment n experimental units are used. Each such unit is divided into k blocks (above the blocks were called "sectors"), and the maximum pit depth in each block is measured. We further assume that these k blocks can be classified into p groups, $1 < p \leq k$, (e.g. by their location on the test units) with similar exposures. The objective is to test if the units are homogeneous, i.e. if there in fact are no systematic differences between the groups.

1.1. Likelihood ratio tests: Consider each treatment separately. A Gumbel distribution with parameters $\hat{\theta}_j = (\hat{\sigma}_j, \hat{\mu}_j)$, $(j = 1, \dots, p)$ is fitted separately to each group and the parameters estimated by maximum likelihood. Let $\ell(\hat{\theta})$ denote the log-likelihood function $\ell(\hat{\theta}) = \sum_{j=1}^{p} \ell_j(\hat{\theta}_j)$, where $\ell_j(\hat{\theta}_j)$ is the log-likelihood associated to the group j. Under the hypothesis of homogeneity, the p groups form a sample of size nk, with common parameters $\hat{\theta} = (\hat{\sigma}, \hat{\mu})$. If $\ell(\tilde{\theta})$ denotes the corresponding log-likelihood function, then under the hypothesis of homogeneity, one has approximately

$$D_H := 2[\ell(\hat{\theta}) - \ell(\tilde{\theta})] \sim \chi^2_{2(p-1)}, \tag{1}$$

and a test of homogeneity may be performed by rejecting this hypothesis if the value of D_H is larger than an appropriate quantile of the chi-squared distribution with $2(p-1)$ degrees of freedom.

1.2. Randomization tests: We use three different test statistics for testing homogeneity via a randomization procedure. The first two are non-parametric, while the third one is based on the Gumbel likelihood. Let $n_j, j = 1, \dots, p$ be the number of observations per unit that belong to group j. Denote by $\bar{X}_{ij}, i = 1, \dots, n, j = 1, \dots, p$, the mean value of these n_j observations for unit i, and let $\bar{X}_{\cdot j} = 1/n \sum_{i=1}^{n} \bar{X}_{ij}$, for

$j = 1, \dots, p$. Then our first test statistic,

$$T_1 = \sum_{i=1}^{n} \sum_{j=1}^{p} (\bar{X}_{ij} - \bar{X}_{.j})^2,$$

is a measure of both homogeneity and replication based on mean values. It is aimed at detecting deviations in location. The second test statistic is

$$T_2 = \max_{j=1,\dots,p} D_j - \min_{j=1,\dots,p} D_j,$$

where D_j is the range of the nn_j observations which belong to group j. This statistic measures the possible spread heterogeneity between the different groups of blocks. The third test statistic considered is the likelihood ratio statistic D_H defined in (1). This randomization test has the same aim as the likelihood ratio test from **1.1**, but may be used when the number of observations is small and the asymptotic chi-square distribution hence may fit badly.

The p-values for the randomization tests were obtained by drawing 500 random permutations of the nk observations. For each permutation, the values were distributed over the k blocks in the n units, to form a dataset of the same structure as the original one. The test statistic was then computed for each of the 500 datasets obtained in this way. The p-value was the proportion of these 500 randomized values of the test statistic which were larger than the actually observed value of the test statistic.

2. Replication tests: Suppose now that the analysis above has not lead us to doubt the assumptions of homogeneity, and that we hence believe that measurements from different groups have the same distribution. Then k homogeneous measurements of maximum pit depth are available for each experimental unit, and can be used to test if n units be considered as replicates.

One way to test it is to use the following likelihood ratio test. Suppose that each unit can be modelled by a Gumbel distribution with parameters estimated by maximum likelihood, say $G_i = \text{Gumbel}(\hat{\theta}_i)$, where $\hat{\theta}_i = (\hat{\sigma}_i, \hat{\mu}_i)$, $(i = 1, \dots, n)$. Denote $\ell(\hat{\theta})$ the log-likelihood $\ell(\hat{\theta}) = \sum_{i=1}^{n} \ell_i(\hat{\theta}_i)$, where $\ell_i(\hat{\theta}_i)$ is the log-likelihood associated to unit i. Under the hypothesis of homogeneity, the n units form a sample of size nk, which can be modelled by a Gumbel$(\tilde{\theta})$. If $\ell(\tilde{\theta})$ denotes the corresponding log-likelihood, then under the hypothesis of replicates, one has approximately

$$D_R := 2[\ell(\hat{\theta}) - \ell(\tilde{\theta})] \sim \chi^2_{2(n-1)}, \tag{2}$$

and the hypothesis that units are replicates is rejected if D_R is larger than the appropriate quantile in a chi-square distribution with $2(n-1)$ degrees of freedom.

A randomization test that units are replicates may be done exactly as in **1.2** but using the statistic D_R defined by (2) instead of D_H.

Appendix 2: 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 sector, X, follows a Gumbel distribution with parameters (μ, σ), and suppose that the unit of interest consists of k independent blocks. Then the maximum per unit, X_u, follows a Gumbel distribution with parameters $(\mu + \log k, \sigma)$. Analogously, if X has an EV distribution with parameters (ξ, σ, μ), then X_u 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 3: Performance of maximum likelihood estimation for extreme value parameters.

Simulations were performed to investigate the small sample behavior of the maximum likelihood estimators in the EV and Gumbel models. Five sample sizes were used, $n = 5, 8, 10, 15$, and 20. For each sample size, 500 samples were generated from an EV distribution with shape parameters $\xi = -0.25, -0.1, 0, 0.1, 0.25$, and with location and scale parameters fixed (without loss of generality) to $\mu = 0$ and $\sigma = 1$. No numerical convergence problems occured for the Gumbel distribution. For small sample sizes, maximum likelihood estimation of the parameters of the EV distribution sometimes failed (Table 5). Figure 11 shows boxplots of the L_1-errors between the empirical and theoretical distribution functions for the convergent cases. (If F_n denotes the distribution function with parameters obtained by maximum likelihood, and if F is the theoretical distribution function, the L_1-error is $\int_{-\infty}^{\infty} \mid F_n - F \mid (t)\ dt$.) From Figure 11 can be seen that the Gumbel distribution does not lead to any substantial increase of the L_1-errors. However, it may still influence extrapolation to extreme quantiles substantially.

Please insert Figure 11 about here.

Appendix 4: Confidence intervals by parametric bootstrap.

A parametric bootstrap was used to construct confidence intervals for the ratio of the return periods of 2 treatments (see Figure 10) as an alternative to the delta

Table 5: *Proportion of convergent maximum likelihood estimation in the EV model in 500 simulated samples from EV distribution, for different shape parameters* ξ *and sample sizes* n. *The parameters* σ *and* μ *are 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

method. In it the maximum likelihood estimates $(\hat{\xi}, \hat{\sigma}_1, \hat{\mu}_1, \hat{\sigma}_2, \hat{\mu}_2)$ of the parameters $(\xi, \sigma_1, \mu_1, \sigma_2, \mu_2)$ of the observed data were used to simulate 1000 samples of size 24 from an EV distribution G_1 with parameters $(\hat{\xi}, \hat{\sigma}_1, \hat{\mu}_1)$, and 1000 samples of size 24 from an EV distribution G_2 with parameters $(\hat{\xi}, \hat{\sigma}_2, \hat{\mu}_2)$. For each pair of samples (with one sample coming from G_1, and the other one from G_2) parameters $(\xi, \sigma_1, \mu_1, \sigma_2, \mu_2)$ were estimated by maximum likelihood and used to compute an estimate of $\bar{G}_2(x)/(\bar{G}_1(x) + \bar{G}_2(x))$, for $0.4 \leq x \leq 1.2$. The $(\alpha/2)$- and $(1-\alpha/2)$-quantiles of the 1000 estimated values give the parametric bootstrap confidence interval.

References

Aziz, P. M. (1956). Corrosion **13**, 495.

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

Embrechts, P. Klüppelberg, C. & Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance.* Springer, Berlin.

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.

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.

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

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.

Captions for figures

Figure 1: Magnesium plates.

Figure 2: Flowchart of the main steps for a comparison between two treatments.

Figure 3: Numbering of the sectors in terms of their location, with 1 on top of the incline and 4 at the bottom.

Figure 4: Pit depth maxima by sectors. The first column provides the 2 weeks data, the second one concerns 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 5: Gumbel plots of maximum pit depths per sector, for the three replicates and for the nine treatments. 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 6: Gumbel plots of maximum pit depths per sector, for simulated i.i.d. observations of an EV distribution with parameters $(\xi, \sigma, \mu) = (0.02, 0.1, 0.4)$.

Figure 7: Gumbel plots with Gumbel and EV fits of maximum pit depths per sector. First row shows the 2 weeks-Fe data, and second row the 4 weeks-Fe/Zn C4 data.

Figure 8: Gumbel plot with EV fit corresponding to the maximum pit depth per unit, for Zn JS500 bolts, after 2 weeks.

Figure 9: 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 10: Estimation of the ratio of the probabilities of exceeding in terms of pit depth per unit (.....) with two associated 90% confidence intervals (——), obtained by the delta method (the smoother one) or parametric bootstrap. The first row concerns the 2 weeks data, and the second row the 6 weeks data. Intervals given by the delta method have been truncated when they included impossible values.

Figure 11: Boxplots of the L_1-errors between theoretical and empirical distribution functions, estimated by maximum likelihood using EV and Gumbel distributions.

Captions for tables

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

Table 2: Maximum likelihood estimates of ξ for the magnesium data. Standard deviations are given between parentheses.

Table 3: P-values for the likelihood ratio test of the equality of the shape parameters ξ_1 and ξ_2 of the EV fits for two treatments. Abbreviation " w" states for "weeks".

Table 4: Paired comparisons based on the ratio of the return periods for the 2 and 6 weeks magnesium data. The sign ' $\equiv$' means that there is no significant difference, '$\prec$ ' (or ' $\succ$') means that the first treatment is significantly worse (or better) than the second treatment for some range of pit depths, at the confidence level 0.95 obtained by the delta method. P-values for the likelihood ratio tests for the validity of the submodels in parentheses.

Table 5: Proportion of convergent maximum likelihood estimation in the EV model in 500 simulated samples from EV distribution, for different shape parameters ξ and sample sizes n. The parameters σ and μ are equal to 1 and 0, respectively.

Figure 1: *Magnesium plates*

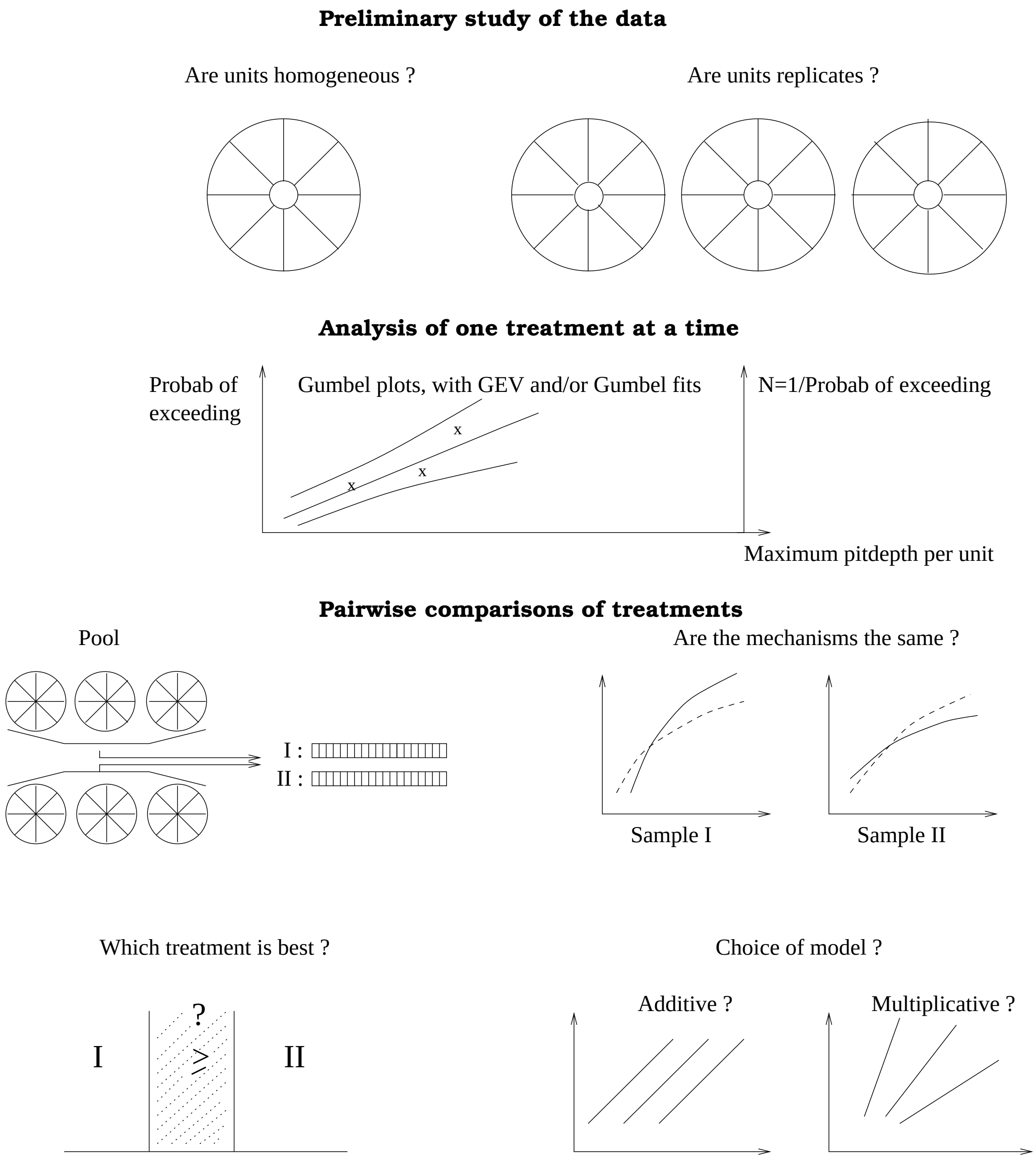

Figure 2: *Flowchart of the main steps for a comparison between two treatments.*

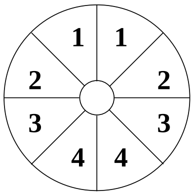

Figure 3: *Numbering of the sectors in terms of their location, with 1 on top of the incline and 4 at the bottom.*

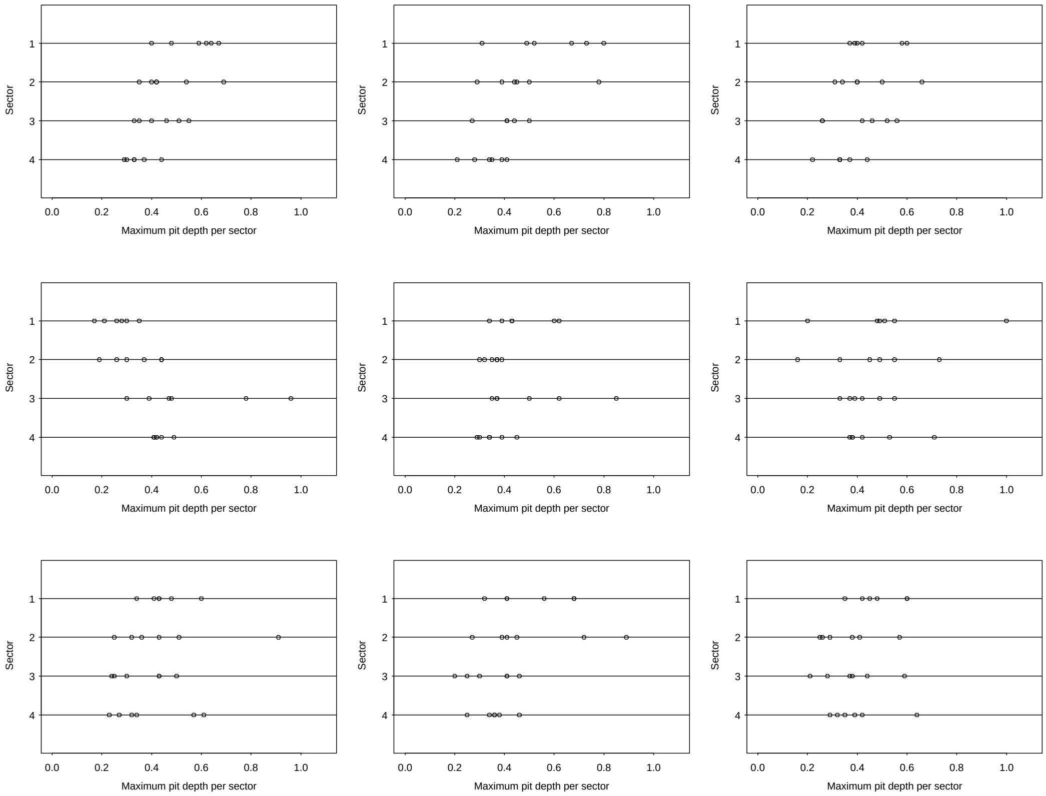

Figure 4: *Pit depth maxima by sectors. The first column provides the 2 weeks data, the second one concerns the 4 weeks data, and the third one the 6 weeks data. Rows show from top to bottom "Fe", "Fe/Zn C4" and "Zn JS500".*

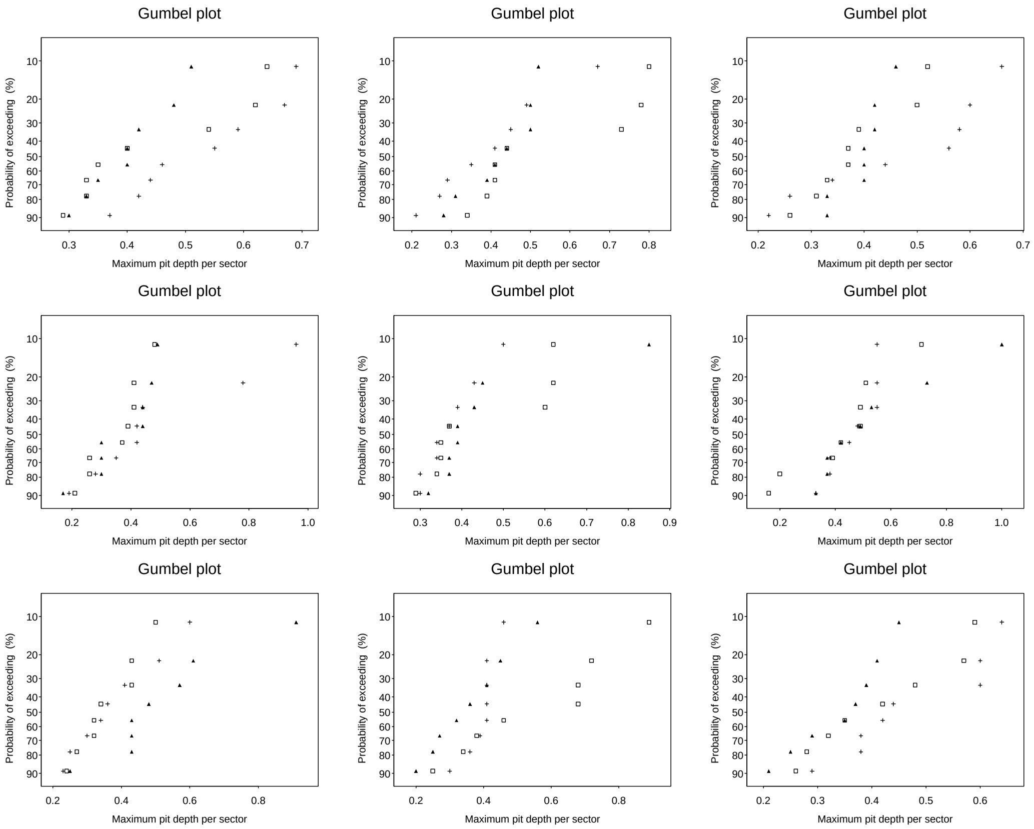

Figure 5: *Gumbel plots of maximum pit depths per sector, for the three replicates and for the nine treatments. The first column provides the 2 weeks data, the second one the 4 weeks data, and the third one the 6 weeks data. Rows show from top to bottom "Fe", "Fe/Zn C4" and "Zn JS500".*

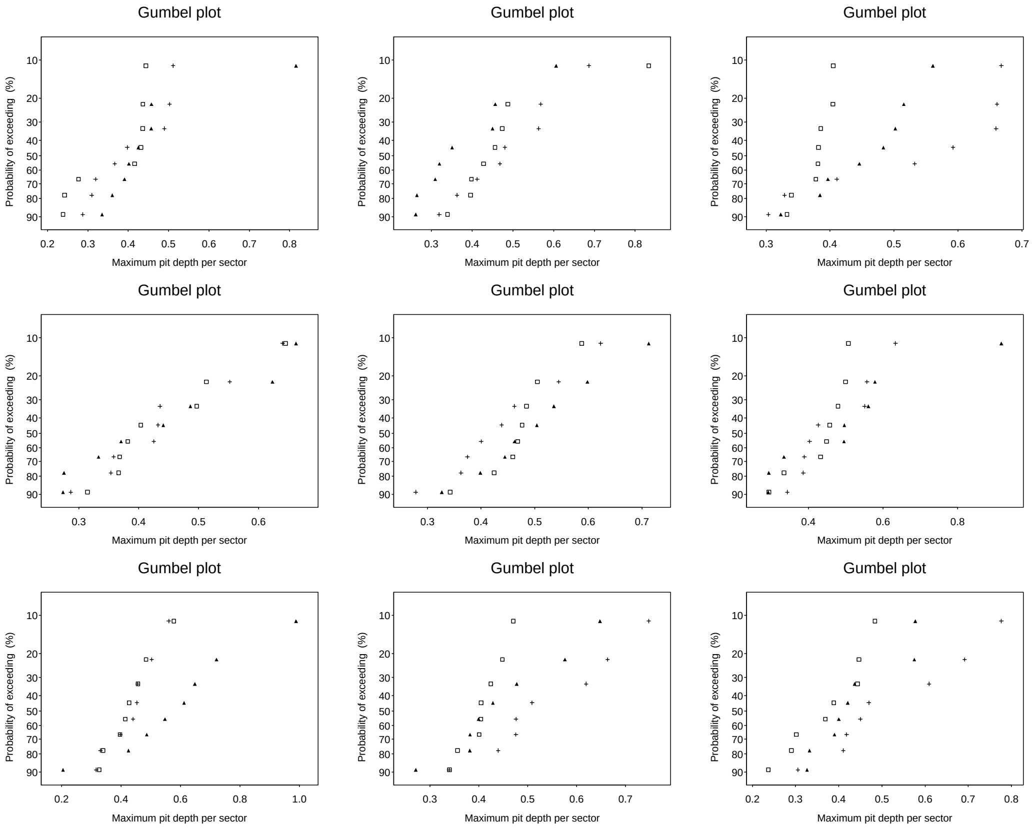

Figure 6: *Gumbel plots of maximum pit depths per sector, for simulated i.i.d. observations of an EV distribution with parameters* $(\xi, \sigma, \mu) = (0.02, 0.1, 0.4)$.

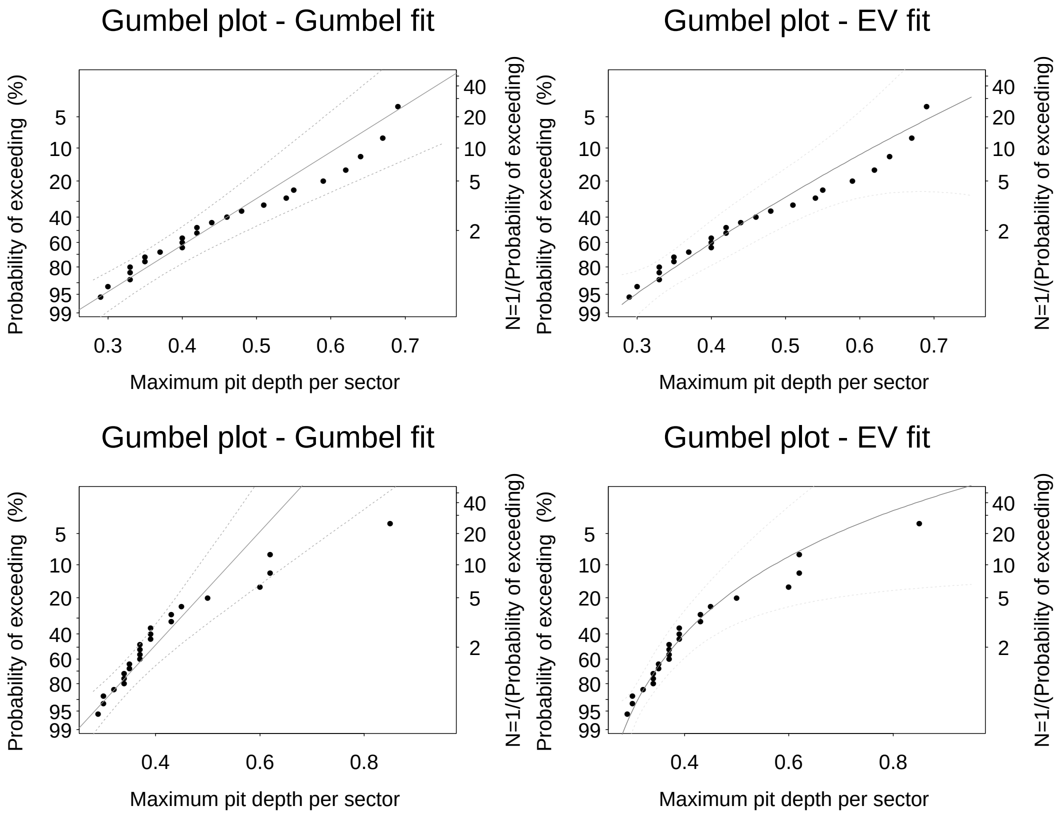

Figure 7: *Gumbel plots with Gumbel and EV fits of maximum pit depths per sector. First row shows the 2 weeks-Fe data, and second row the 4 weeks-Fe/Zn C4 data.*

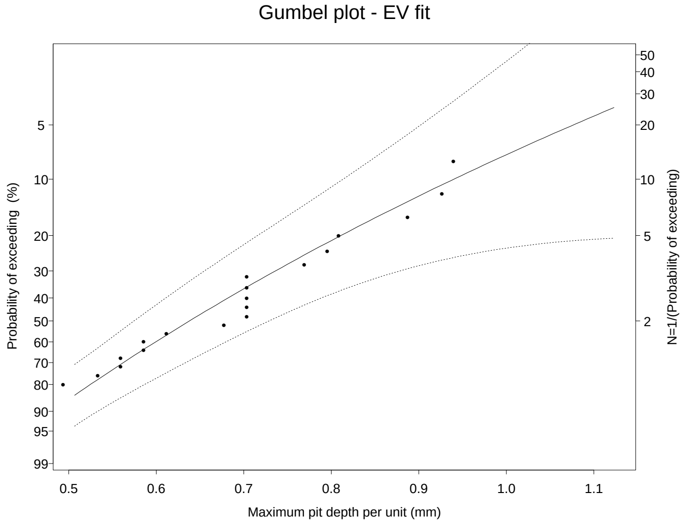

Figure 8: *Gumbel plot with EV fit corresponding to the maximum pit depth per unit, for Zn JS500 bolts, after 2 weeks.*

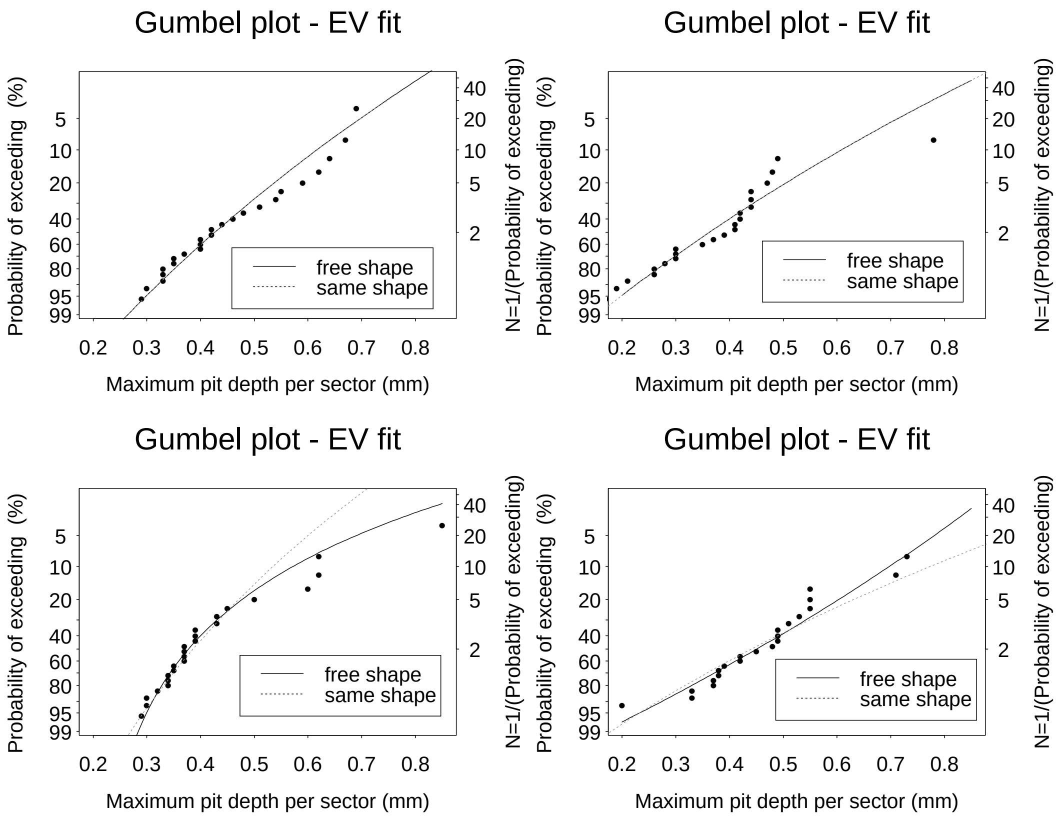

Figure 9: *Gumbel plots with EV fits of maximum pit depths per sector, for pairs of treatments (called Treatment 1, Treatment 2), with shape parameters assumed to be equal or free. Each row shows one pair. The first column shows Treatment 1, where the fitted lines are indistinguishable and the second 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.*

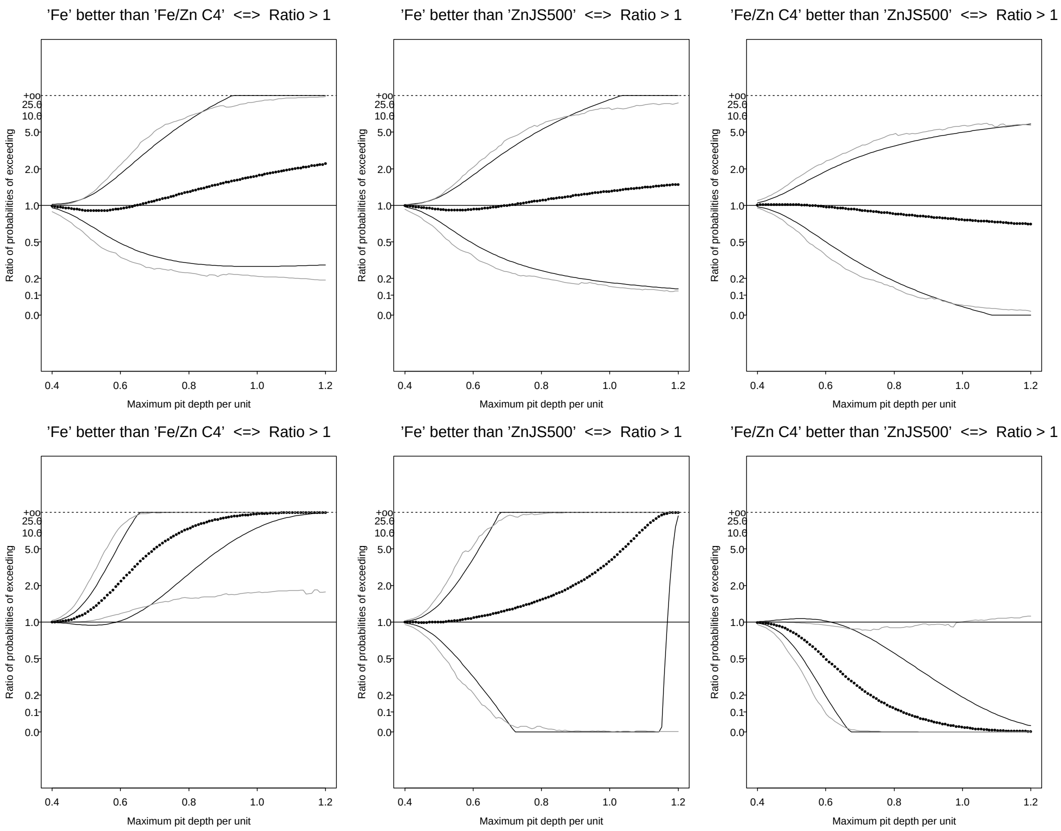

Figure 10: *Estimation of the ratio of the probabilities of exceeding in terms of pit depth per unit (•••••) with two associated* 90% *confidence intervals (——), obtained by the delta method (the smoother one) or parametric bootstrap. The first row concerns the 2 weeks data, and the second row the 6 weeks data. Intervals given by the delta method have been truncated when they included impossible values.*

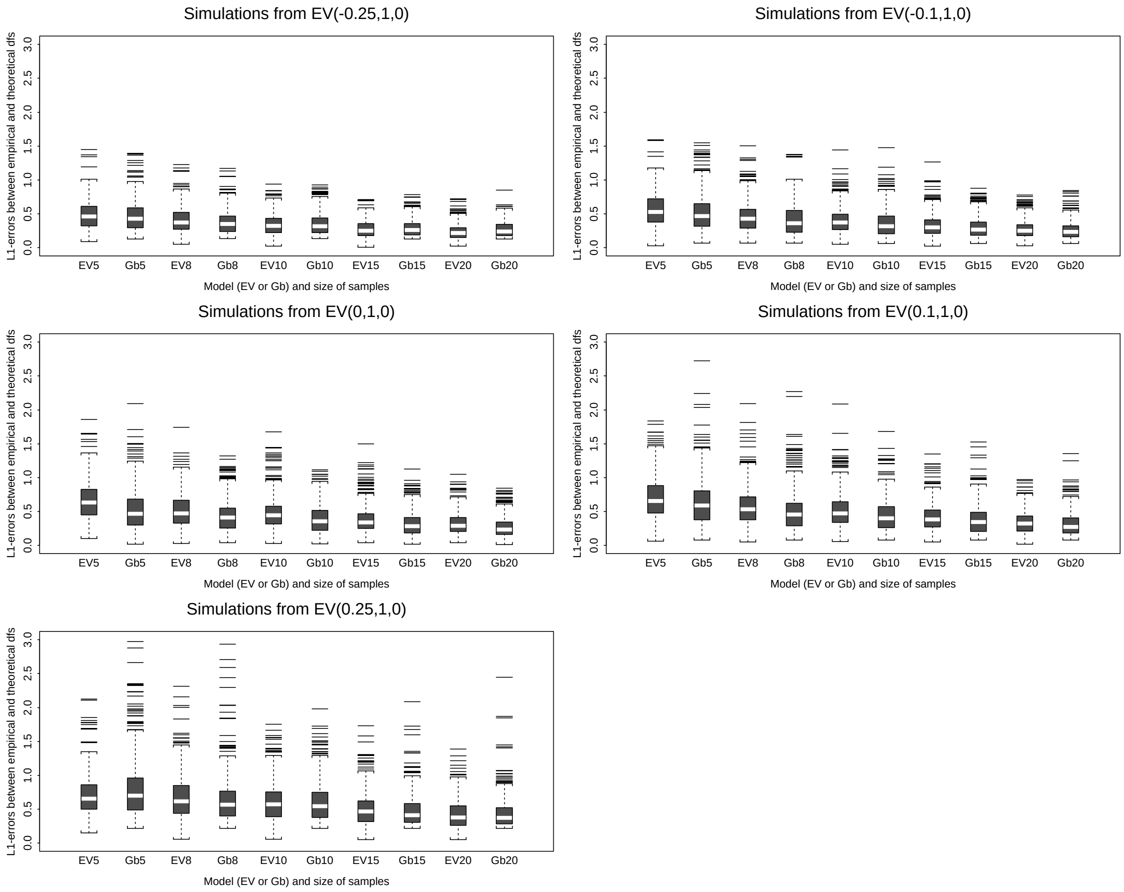

Figure 11: *Boxplots of the L_1-errors between theoretical and empirical distribution functions, estimated by maximum likelihood using EV and Gumbel distributions.*