Pitting Corrosion: Comparison of Treatments With Extreme-Value–Distributed Responses

A.-L. FOUGÈRES

CNRS, Département GMM
Institut National des Sciences Appliquées
31077 Toulouse, France
(*fougeres@insa-toulouse.fr*)

S. HOLM and H. ROOTZÉN

Department of Mathematics
Chalmers University of Technology
S-412 96 Gothenburg, Sweden
(*rootzen@math.chalmers.se*)

In this article 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 and makes judicious choice of alloys and surface treatments necessary. Standard methods for evaluating corrosion tests 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) that 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 does not give credible extrapolation into extreme tails.

KEY WORDS: Comparison of treatments; Designed experiment; Extreme value; Pitting corrosion.

1. INTRODUCTION

In this article 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 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. Thus a key issue is to reduce pitting corrosion via coating, surface treatment, choice of alloy, or use of isolating washers.

Standard methods for evaluating 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 rather the risk of perforation (i.e., the depth of the deepest pit) that is of interest (see, e.g., Isacsson, Ström, Rootzén, and Lunder 1997). The second is that the standard ANOVA assumption of homogeneity of variances typically is not satisfied by pit depth measurements, and normality does not give credible extrapolation into extreme tails. Extreme-value (EV) statistics has appeared as a theoretically well-founded alternative way to analyze data on pitting corrosion (see, e.g., Shibata 1996). This approach has been suggested by Aziz (1956) and Gumbel (1958), and many subsequent authors. Much of this work was developed and promoted by Kowaka (1994). Likelihood, generalized Pareto, and EV distribution methods for analysis and extrapolation were proposed in a series of articles by Scarf, Laycock, and Cottis (see, e.g., Scarf and Laycock 1994 and references therein). However, the literature does not seem to include advice on how to check experimental conditions or how to compare corrosion-reducing treatments.

Another application is to the fatigue limit, that is, 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 experiments with synthetic portfolios of financial instruments, where risks are evaluated from historical extreme price fluctuations over a number of time intervals.

This article reports on the first part of a continuing effort, where the distant goal is a full theory of design and analysis of experiments with EV-distributed responses. Presumably such a theory would also be likelihood-based, and incorporate results from this article, but would in particular add covariate models for the entire experiment (see, e.g., Stephenson's recent R program on covariate models for extreme values, available at *http://cran.us.r-project.org/*, package "evd"). We in fact already tried this approach in a pitting corrosion setting (Isacsson 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 parameterization, and the ability of models to respect the stochastic monotonicity implied by the nonreversibility of the corrosion process.

In this article, analysis is based on block maxima. In a complementary method, the peaks over threshold method, analysis uses not only maxima, but also 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 distributions and EV distributions with generalized Pareto distributions. However, in corrosion testing, measuring the pits is a major part of the experimental effort and the part that experimenters like the least. Hence the choice of methods is determined by measurement convenience and not by statistical

© 2006 American Statistical Association and
the American Society for Quality
TECHNOMETRICS, MAY 2006, VOL. 48, NO. 2
DOI 10.1198/004017005000000490

consideration. 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 EV 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 that arise, and Section 6 contains our conclusions. Some technical issues are relegated to appendixes.

2. STATISTICAL EXTREME VALUE THEORY

Statistical EV theory models and analyzes data obtained as the maxima of many (approximately) independent and identically distributed (iid) underlying variables. Useful recent introductions to the area have been provided by Coles (2001), who gave an up-to-date account of statistical methods, and Embrechts, Klüppelberg, and Mikosch (1997), who presented the basic theory from an econometric perspective.

The central result of extreme value theory is that the natural model for maxima is the EV distribution (sometimes also called the generalized EV distribution) with distribution function

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

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. For $\xi = 0$, the formula should be interpreted as the limiting (as $\xi \to 0$) Gumbel distribution $G(x) = \exp[-\exp\{-(x-\mu)/\sigma\}]$, and for ξ 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 iid random variables as $n \to \infty$.
- The EV distribution is the only one that is stable under change of block size, that is, such that if maxima over smaller iid blocks have this distribution, then maxima over bigger blocks have the same distribution.

Several methods for estimating the EV parameters have been proposed (see, e.g., Hosking, Wallis, and Wood 1985; Johnson, Kotz, and Balakrishnan 1994, vol. 2, chap. 22). Maximum likelihood estimation in particular gives good results when the sample size is not too small (see Sec. 5 and App. B for further discussion) and is much more general and flexible than the competitors. In this article 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," which illustrate the adequacy of the EV fit and provide easy graphical interpretation and extrapolation of results. If $X_1, \ldots, X_n$ are iid observations, then the Gumbel plot shows the graph

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

where $X_{(1)} \le \cdots \le X_{(n)}$ 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 ξ is negative, a straight line for the Gumbel case where $\xi = 0$, and a concave curve if $\xi > 0$.

3. PIT CORROSION

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

Galvanic corrosion is the consequence of an oxidation–reduction reaction. This reaction is caused by the potential difference 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 must resort to experimentation and experience to enable the manufacture of 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 datasets. 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 [Fig. 1(a)]. 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; that is, the temperature was kept constant at 35°C, and the humidity was cycled between 50% and 95% twice a day (Isacsson 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 nine 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. Thus in the terminology of design of experiments (which is somewhat incompatible with corrosion terminology),

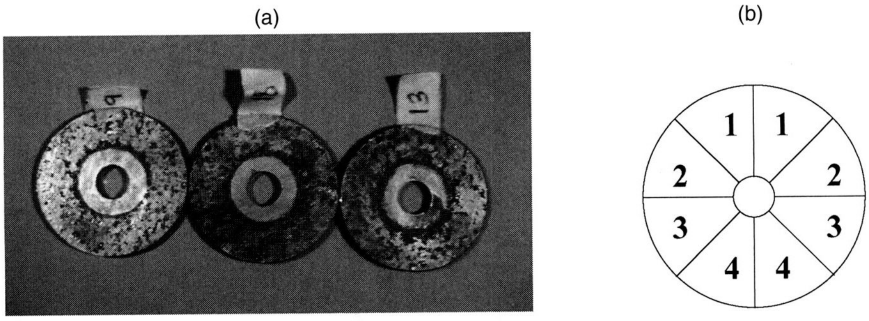

Figure 1. Specimens of Magnesium Plates (a) and Numbering of the Sectors in Terms of Their Position Relative to the Vertical Direction (b).

the treatments are the three types of bolt and the three time points. 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 (DirectRay, AGFA). The dataset obtained thus consists of an 8×27 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 as follows: 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. Hence a typical dataset consists of nk measurements of block maxima for each treatment.

Now, how should we compare the efficiency of the treatments? In the next section we propose a method for analyzing such datasets.

4. METHOD AND DATA ANALYSIS

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

The method divides the analysis into three parts: a preliminary study of the data, a separate analysis of each treatment, and pairwise comparisons of the treatments. Each part comprises several steps. For each step similar elements are provided: a graph, a parametric likelihood ratio (LR) 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 as to whether the sample sizes are large enough to make the LR tests sufficiently accurate.

In each step we first describe the method for a general situation, 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 iid. The experiments are designed to achieve this, and we expect the measurements to pass the test. However, if they do not, then this may indicate a need to improve the experimental setup. It also would mean that one cannot proceed with the following steps in the way outlined here; modifications are needed.

Step 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, then 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 an LR test. Sample sizes for this are often small; if they are below, say, 20, then it is prudent to corroborate the LR test by randomization tests. We use three such tests. The first test is based on the Gumbel LR statistic; the other two tests are completely nonparametric and are based on statistics measuring location heterogeneity and dispersion heterogeneity (see Fougères et al. 2002 for more details).

Data Analysis. 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 1(b). 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 Fig. 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 sector group 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 1–4. The p values for the more specific parametric LR tests are smaller than those 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.

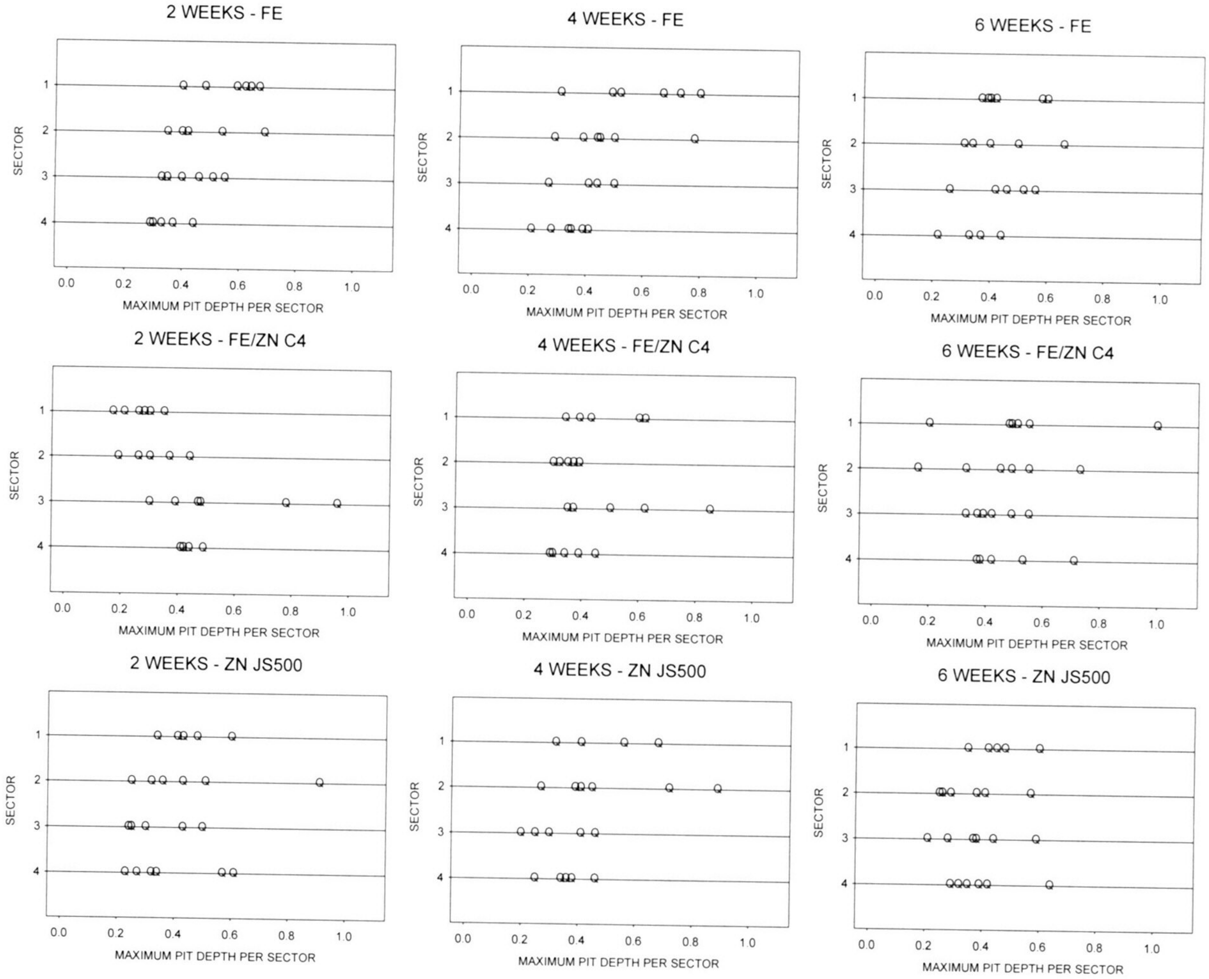

Figure 2. Pit Depth Maxima by Sectors. The first column gives the 2 weeks data; the second, the 4 weeks data; the third, the 6 weeks data. Rows show, from top to bottom, Fe, Fe/Zn C4, and Zn JS500.

Step 2: Are Units Replicates? The graphical test for systematic differences between experimental units (i.e., if units are "replicates") is a Gumbel plot (see Sec. 2) in which each unit has its own symbol. This is complemented by an LR test of the hypothesis that the separate parameters for the different experimental units are in fact the same. Because sample sizes were small (<20), we also performed randomization tests in the same way as for Step 1.

Data Analysis. Figure 3 presents the Gumbel plots for each set of three replicate units. The results of LR and randomization tests are reported in Table 1, columns 5 and 6. No consistent pattern that 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. Thus the analysis is aimed only at safeguarding against gross deviations. In particular, in the subsequent analysis we use the more flexible EV distribution and do not restrict ourselves to the Gumbel model. However, sample sizes in Steps 1 and 2 typically are too small (less than, say, 20; see App. B) for successful likelihood estimation of EV parameters, and we believe that for the present data with values of the shape para-

Table 1. p Values of Tests in Steps 1 and 2

		Homogeneous				*Replicates*	
		LR	*Randomization-disp*	*Randomization-loc*	*Randomization-LR*	*LR*	*Randomization-LR*
2 weeks	Fe	.01	.24	.60	.01	.13	.12
	Fe/Zn C4	.00	.27	.26	.00	.40	.55
	Zn JS500	.37	.66	.38	.39	.16	.17
4 weeks	Fe	.08	.52	.54	.17	.24	.38
	Fe/Zn C4	.01	.40	.71	.11	.41	.59
	Zn JS500	.13	.42	.09	.23	.01	.01
6 weeks	Fe	.31	.81	.80	.50	.02	.03
	Fe/Zn C4	.13	.24	.47	.52	.12	.44
	Zn JS500	.32	.63	.41	.40	.30	.34

NOTE: Columns 1–4 test whether sectors are homogeneous; columns 5 and 6 test whether plates with the same treatment are replicates.

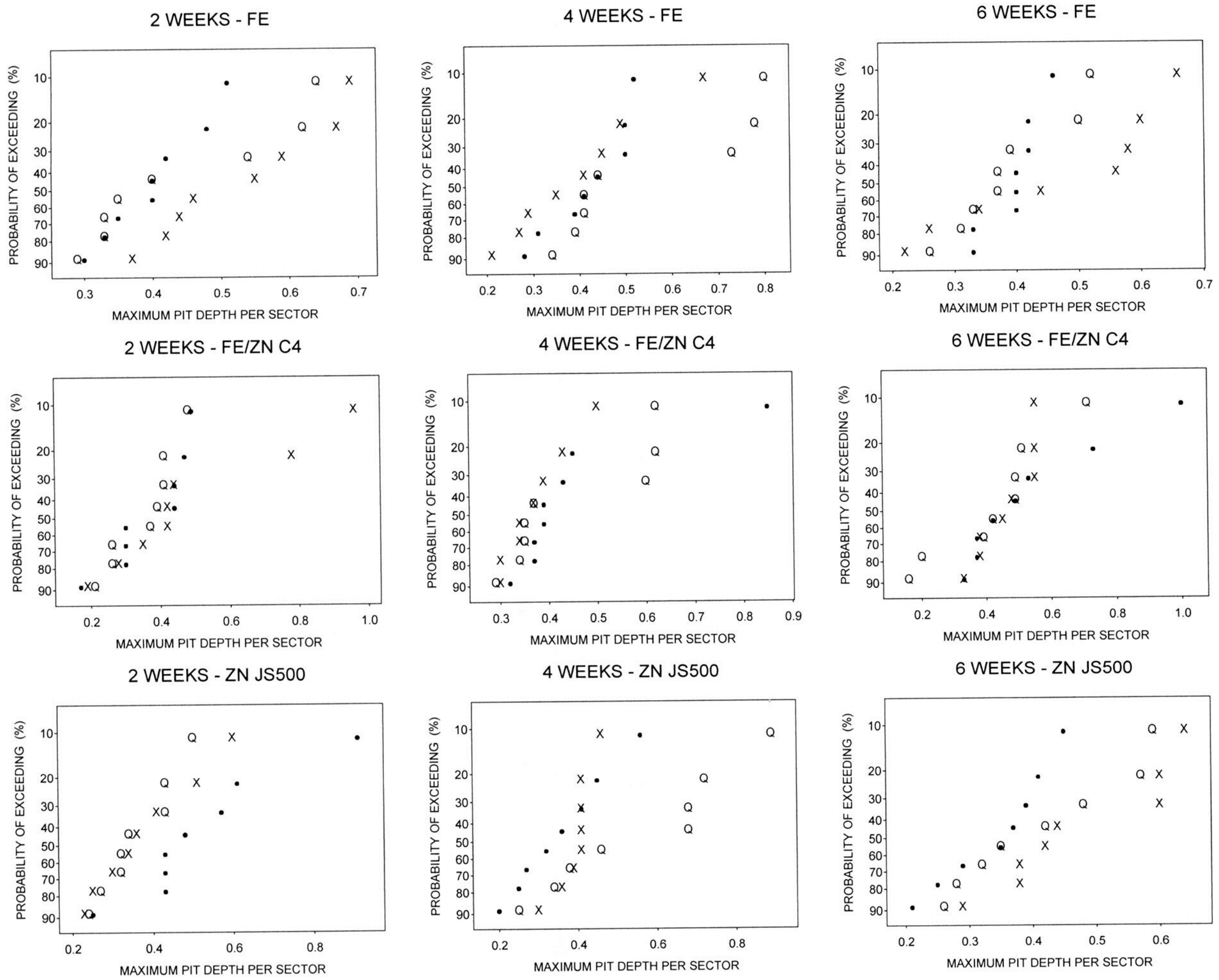

Figure 3. Gumbel Plots of Maximum Pit Depths per Sector, for the Three Replicates and the Nine Treatments. Here • is the first sample, Q is the second sample, and X is the third sample. The first column provides the 2 weeks data; the second, the 4 weeks data; and the third, the 6 weeks data. The rows show, from top to bottom, Fe, Fe/Zn C4, and Zn JS500.

meter close to 0 (Table 2), LR tests based on approximation by a Gumbel distribution will detect gross deviations.

Corroboration by graphical and randomization tests may be prudent. In particular, note that the randomized version of the LR test gave correct p values regardless of whether or not the data actually came from a Gumbel distribution. 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 which the Gumbel plot was nonlinear and the randomized LR test would be preferred.) As a further precaution, we also performed 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.

Table 2. Maximum Likelihood Estimates of ξ for the Magnesium Data

	Fe	*Fe/Zn C4*	*Zn JS500*
2 weeks	.084 (.260)	.088 (.148)	.130 (.204)
4 weeks	.027 (.161)	.384 (.205)	.091 (.163)
6 weeks	−.120 (.167)	−.079 (.105)	−.098 (.203)

NOTE: Standard deviations are in parentheses.

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 that we are interested in guarding against. Of course, even if deviations are not expected, if they would occur, then 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 use given reason to doubt homogeneity and that units are replicates. For the rest of the analysis, we then pool all of the observations that stem from the same treatment, and assume that those observations are iid.

Steps 3–5: 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 that 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 whether a Gumbel distribution is sufficient to describe the data from the separate treatments. It is relevant only sometimes. A reason to perform this step could be that experience from similar situations indicates 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 linked to the Gumbel limit distribution for maxima.

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

Step 3: Standardization to Meaningful Units. The raw data are the maximum pit depths in the sectors. However, as noted earlier, 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 App. A).

Data Analysis. The maximum pit depth per plate is the interesting quantity rather than the maximum pit depth per sector. We hence standardized to plates as units wherever possible in the following steps.

Step 4: EV Fit versus Gumbel Fit. As discussed earlier, in some situations it may be reasonable to check the fit of the Gumbel distribution. We use graphics and an LR 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, because standardization is model-dependent and would yield different scales on the x-axes for the Gumbel and EV distributions. Now consider, for example, the 2 weeks Fe data (Fig. 4, top). For these, the Gumbel model gives very good fit, in fact with p value equal to .99 in the LR test of the Gumbel distribution. But the same conclusion does not apply in all the cases, as for example for the 4 weeks Fe/Zn C4 data (Fig. 4, bottom). We hence preferred to use the EV distribution for all the main Gumbel plots in Step 5.

Step 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

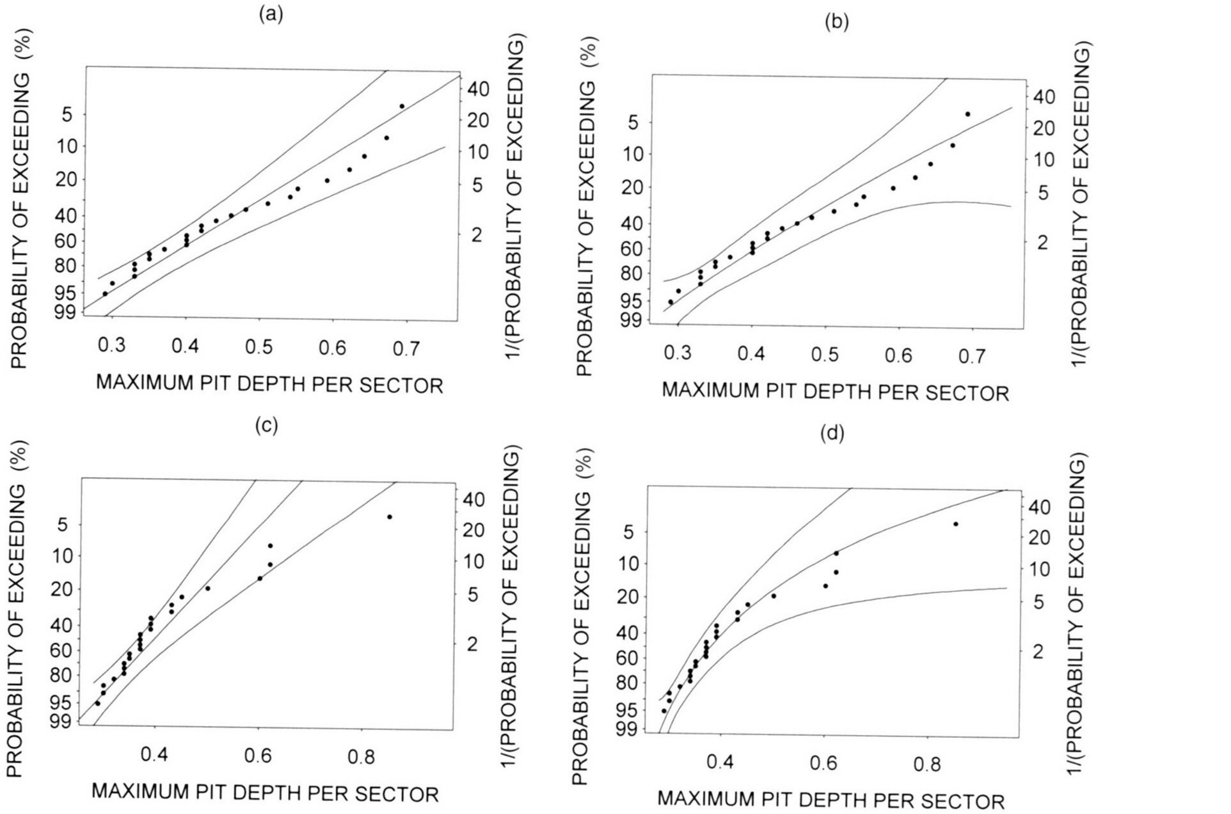

Figure 4. Gumbel Plots With Gumbel and EV Fits of Maximum Pit Depths per Sector, for the Magnesium Data, With Associated 95% Confidence Intervals. (a) and (b) The 2 weeks Fe data. (c) and (d) The 4 weeks Fe/Zn C4 data.

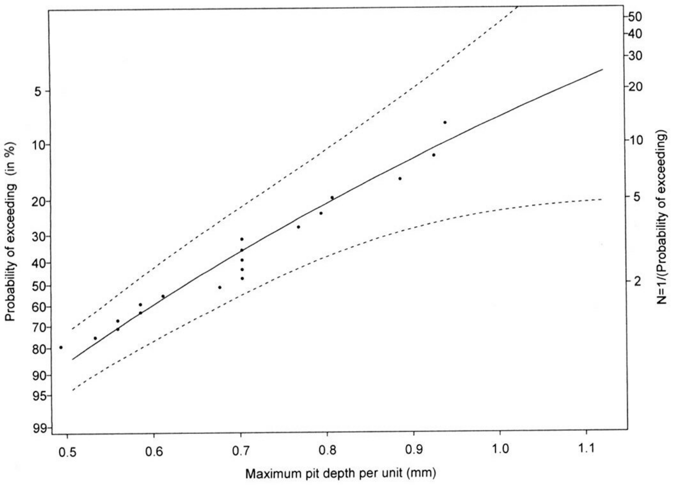

Figure 5. Gumbel Plot With EV Fit Corresponding to the Maximum Pit Depth per Unit, for the Magnesium Data With JS500 Coated Zinc Bolts, After 2 Weeks.

depth exceeding the level (in percent) [this is $100(1 - i/(n+1))$ for the ith 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 datasets. 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.

For example, the answer to what is the expected number of perforated units if one has 1,000 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 $1{,}000 \times .0716 = 71.6$. Preferably this point estimate should be complemented by a confidence interval, which in the same way can be read from the graph as $(21, 230)$. 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 (although, of course, they are more computationally demanding).

Similarly, to find out how thick the plates should 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), whereas nonconvergence was rare ($<1\%$) for sample size 20 or larger (see App. B). Thus the Gumbel distribution may be the only viable alternative for small sample sizes; however, of course it should be used only if it fits reasonably well.

Steps 6 and 7: Pairwise Comparisons of Treatments

In Step 6 we check whether 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, whereas the other one may be best in another region. Because of this, it is possible for the model to discern between situations with many shallow pits, and other, potentially more dangerous, situations with few, but deep, pits.

Step 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 (ξ_i, σ_i, μ_i), $i = 1, 2$ (cf. Step 4). 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 also in the shape of the distribution. On a more qualitative (and, from an engineering viewpoint, more important) scale, if the shape parameter ξ of the EV distribution is negative, then there is an upper bound for the possible pit depths, whereas a zero or positive ξ 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 grounds 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 equal and with free shape parameters. It is also checked by LR tests of the hypothesis $\xi_1 = \xi_2 = \xi$.

Data Analysis. The estimates of the shape parameter ξ were positive for the 2 and 4 weeks data, corresponding to an unbounded distribution, for all types of bolt, whereas 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 is, of course, quite speculative, however. 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 4 weeks of exposure and the other had 6 weeks of exposure, and the fit obtained with free parameters looks somewhat different than obtained when the shape parameters are equal. However, the LR test did not reject the hypothesis of equality of shape parameters (Table 3). Nevertheless, in this article we confine our attention to comparisons for 2 weeks and 6 weeks of exposure.

Step 7: Which Treatment Is Best? This question is answered here via pairwise comparisons of treatments. Typically, several or all pairs are compared. This may sometimes lead to considerations of "multiple inference" (see the end of Sec. 5).

Now consider a pair of treatments and assume that previous analysis has not contradicted that the corrosion mechanisms for

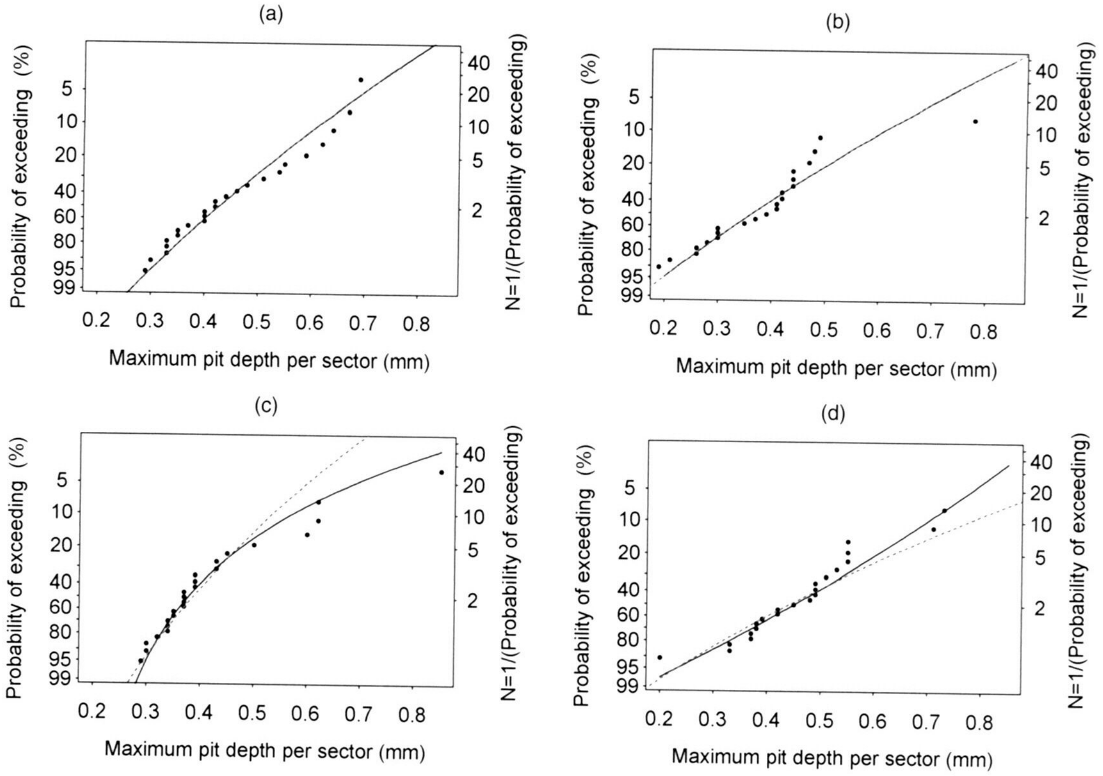

Figure 6. Gumbel Plots With EV Fits of Maximum Pit Depths per Sector, for Pairs of Treatments (called treatment 1 and treatment 2), With Shape Parameters Assumed to be Equal (······) or Free (——). Each row shows one pair. (a) and (c) Treatment 1, where the fitted lines are indistinguishable; (b) and (d) treatment 2. (a) and (b) represent 2 weeks Fe, and 2 weeks Fe/Zn C4; (c) and (d) represent 4 weeks Fe/Zn C4 and 6 weeks Fe/Zn C4.

the two treatments in the pair are same. It is hence assumed that $\xi_1 = \xi_2 = \xi$ and that 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 >1. (A stronger statement would be that the ratio is >1 for all x; however, neither the present data nor scientific knowledge of the corrosion process seemed a sufficient basis 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, one obtained by the delta method, and the other obtained from a standard parametric bootstrap are included in the plot. 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.

Table 3. p Values for LR Tests of Equality of the Shape Parameters ξ_1 and ξ_2 for Pairs of Treatments

		Fe		Fe/Zn C4			Zn JS500		
		4w	*6w*	*2w*	*4w*	*6w*	*2w*	*4w*	*6w*
Fe	2w	.85	.50	.99	.36	.55	.89	.98	.58
	4w		.52	.78	.16	.58	.69	.78	.63
	6w			.35	.05	.82	.32	.36	.92
Fe/Zn C4	2w				.24	.35	.86	.99	.46
	4w					.03	.40	.26	.09
	6w						.32	.38	.93
Zn JS500	2w							.88	.41
	4w								.47

NOTE: 2w, 2 weeks; 4w, 4 weeks; 6w, 6 weeks.

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 data 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 the Fe bolts than with the Fe/Zn C4 bolts. In the same way, the Zn JS500 bolts were found to be better than the 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 EVs. However, this led to very similar results, which are 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 in data can cause large changes in inferences. One should be cautious in the interpretation of such cases.

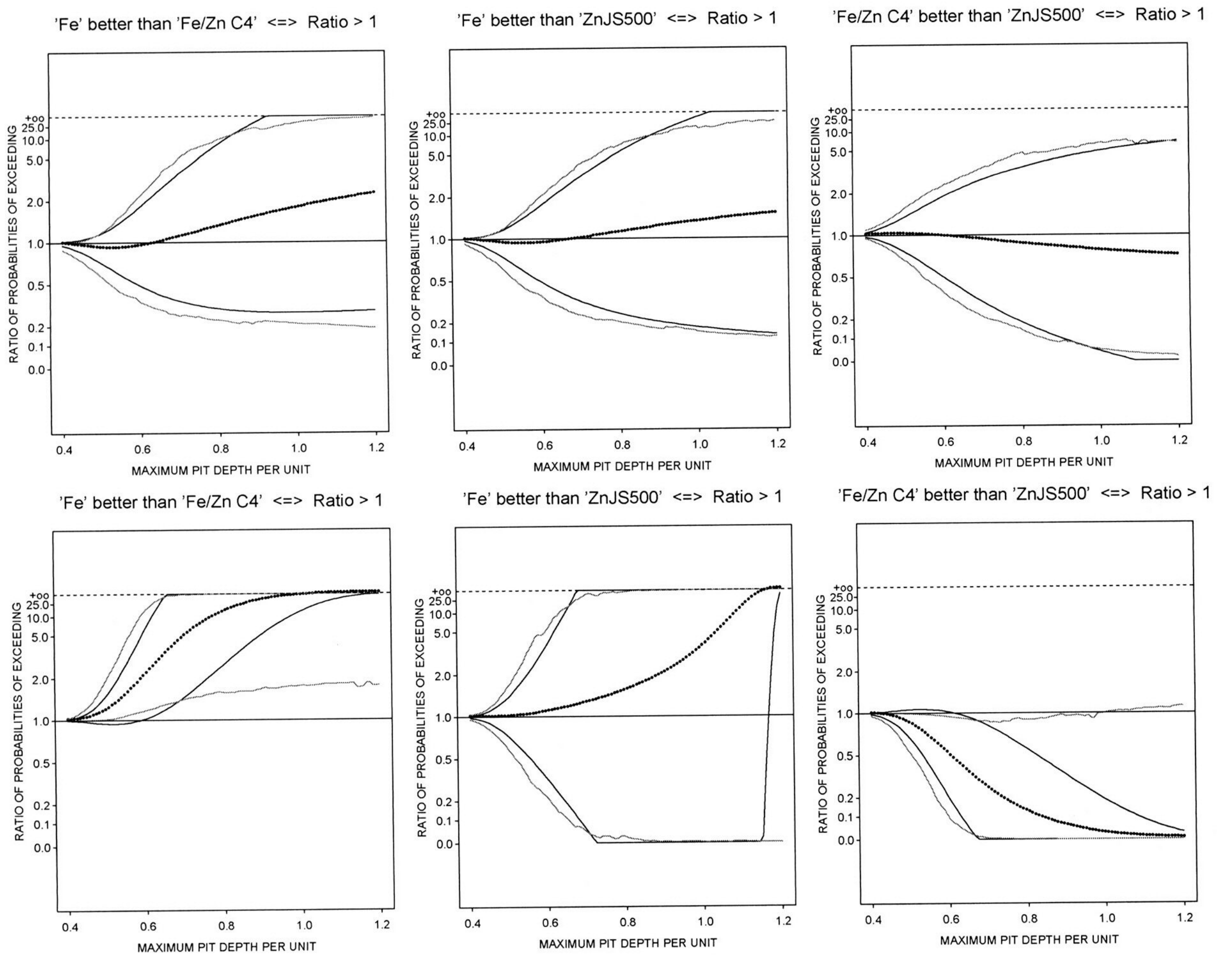

Figure 7. Estimation of the Ratio of the Return Periods in Terms of Pit Depth per Unit (------) With Associated 90% Confidence Interval (——). The irregular lines are the boostrap confidence intervals. The first row provides the 2 weeks data; the second row, the 6 weeks data.

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 that 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 hinder 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 being taken seriously.

5. SOME STATISTICAL AND MODELING ISSUES

In the literature on localized corrosion, except for work by Laycock, Scarf, and Cottis (see, e.g., Scarf and Laycock 1994; Cottis, Laycock, and Scarf 1990), attention has been focused mainly on Gumbel rather than general EV modeling. In this article we prefer the more flexible EV family. In particular, it can indicate whether pits do not continue to grow indefinitely (for $\xi < 0$). Of course, this can be synonymous with a significant cost reduction. A price for this increased flexibility is that the EV fits require slightly larger sample sizes than the 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 that we used (the S–PLUS 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 App. B). This problem was also observed by Drees, de Haan, and Li (2005, tables 2 and 4), in a slightly different context.

The first two steps in our method are tests of homogeneity and replication. Sample sizes in the Volvo experiment are very small at those steps ($n = 6$ or 8). That is 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 earlier, in our experience Gumbel-based LR tests are more sensitive than nonspecific randomization tests. The randomized version of the LR test, if available, should be preferred over the simpler asymptotic LR test, particularly in cases of imperfect Gumbel fit. The aim of the steps is to provide rough safeguards to detect whether the experimental conditions have turned out to not be as intended.

The confidence intervals in this article use the delta method and, for Figure 7, a parametric bootstrap method. Sometimes—particularly when extreme quantiles are estimated—the likelihood function can be quite asymmetric. Profile likelihood methods are then preferable, but these 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 obtained 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 goodness of fit is assessed by looking at Gumbel plots with estimated distribution lines and through an LR test. In our example, comparisons using the additive and multiplicative did not lead to more significant results than the full model.

Here we have assumed so far that the purpose of the experiment was explorative/hypothesis-generating. However, if many tests or many confidence intervals are used, then 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 that for the individual comparisons. Accordingly, if one wants to make formal inference with a controlled overall significance level, then multiple inference methods must 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 has, say, six different treatments, then one obtains an overall 5% confidence level for all 6 × 5/2 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 must be 67% wider, and for 16 treatments, they must be, 75% wider. Similarly, for testing, treatment as multiple tests with a predetermined level of significance gives the corresponding scale changes in the power function (see Hsu 1996).

6. SUMMARY AND CONCLUSIONS

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

The approach uses graphical methods throughout 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 models (see, e.g., Coles 2001). It would be straightforward to translate our method to such situations.

ACKNOWLEDGMENTS

The authors thank Malte Isacsson for initiating this work and Isacsson, Gunnar Ström, and Zheng Tan for many stimulating discussions and ideas. The magnesium corrosion experiment was performed and first analyzed by Jenny Andersson, and the authors also acknowledge her help. They also thank Nader Tajvidi for constructive discussions and comments and two referees and an associate editor for very useful and thought-provoking comments that led to substantial improvement of the article. This research was supported by VCC/Ford, by the Knut and Alice Wallerberg foundation, and by the TMR Spatial and Computational Statistics Network.

APPENDIX A: STANDARDIZATION TO UNITS

The EV distribution is preserved after taking the maximum of iid variables, as mentioned in Section 2. More precisely, assume that the maximum per block, 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_k, follows a Gumbel distribution with parameters $(\mu + \sigma \log k, \sigma)$. Analogously, if X has an EV distribution with parameters (ξ, σ, μ), 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 B: PERFORMANCE OF MAXIMUM LIKELIHOOD ESTIMATION FOR EXTREME VALUE PARAMETERS

Simulations were performed (with the S–PLUS optimization routine "nlminb") to investigate the small-sample behavior of the maximum likelihood estimators in the EV and Gumbel models (Table B.1). 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 B.1. 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, then the percentage of cases where there is convergence becomes higher. However, for small sample sizes, such as 5 or 10, many of the estimates then are so far off as to be useless.

[Received January 2004. Revised April 2005.]

Table B.1. Proportion of Convergent Maximum Likelihood Estimation in the EV Model

$n\backslash\xi$	−.25	−.1	0	.1	.25
5	.452	.520	.536	.536	.506
8	.722	.784	.838	.870	.892
10	.818	.896	.932	.952	.964
15	.942	.980	.988	.992	1
20	.996	.988	1	.996	1

NOTE: For each sample size n and shape parameter ξ, 500 samples were simulated. The parameters σ and μ were equal to 1 and 0.

REFERENCES

Aziz, P. M. (1956), "Application of the Statistical Theory of Extreme Values to the Analysis of Maximum Pit Depth Data for Aluminum," *Corrosion*, 12, 495–506.

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

Cottis, R. A., Laycock, P. J., and Scarf, P. A. (1990), "Extrapolation of Extreme Pit Depths in Space and Time," *Journal of the Electrochemical Society*, 137, 64–69.

Drees, H., de Haan, L., and Li, D. (2005), "Approximation to the Tail Empirical Distribution Function With Application to Testing Extreme Value Conditions," *Journal of Statistical Planning and Inference*, in press.

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

Fougères, A.-L., Holm, S., and Rootzén, H. (2002), "Pitting Corrosion: Comparison of Two Treatments With Extreme Value Distributed Responses, Extended Version," Technical Report, Mathematical Statistics, Chalmers, 2002:33, available at *http://www.math.chalmers.se/Stat/Research/Preprints/index.cgi*.

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

Hosking, J. R. M., Wallis, J. R., and 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*, London: Chapman & Hall.

Isacsson, M., Ström, M., Rootzén, H., and Lunder, O. (1997), "Galvanically Induced Atmospheric Corrosion on Magnesium Alloys: A Designed Experiment Evaluated by Extreme Value Statistics and Conventional Techniques," Technical Paper 970328, The Engineering Society for Advancing Mobility Land Sea Air and Space.

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

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

Murakami, Y., and Beretta, S. (1999), "Small Defects and Inhomogeneities in Fatigue Strength: Experiments, Models and Statistical Implications," *Extremes*, 2, 123–147.

Murakami, Y., and 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," *International Journal of Fatigue*, 11, 299–307.

Scarf, P. A., and Laycock, P. J. (1994), "Applications of Extreme Value Theory in Corrosion Engineering," *Journal of Research of the National Institute of Standards and Technology*, 99, 313–320.

Shibata, T. (1996), "Statistical and Stochastic Approaches to Localized Corrosion," *Corrosion*, 52, 813–830.

Takahashi, R., and Sibuya, M. (1996), "The Maximum Size of the Planar Sections of Random Spheres and Its Application to Metallurgy," *Annals of the Institute of Statistics and Mathematics*, 48, 127–144.