CORROSION SCIENCE AND TECHNOLOGY, Vol.31, No.3(2002), pp.228~233

Program for Calculating the Extreme Value Statistics on the Spreadsheets Software(EVAN-II)

Masahiro Yamamoto and Toshio Shibata*

Surface Treatment Lab. Steel Research Labs. NSC
20-1 Shintomi Futtsu Chiba, 293-8511, JAPAN
**Faculty of Engineering, Department of Mechanical Engineering,*
Fukui University of Thechnology

The extreme value statistics is widely used for predicting the lives of the various structures and the industrial plants in JAPAN. The program named EVAN, calculating the extreme value statistics, had been developed and sold by JSCE from 1989. However, a long time has passed through the development and it did not match a recent computer generation. And it was restricted only in Japanese language. So, a new program (EVAN-II) was developed. Since the program was arranged for a corrosion engineer to calculate extreme value statistics easily, generally used spreadsheet software Excel^{Tm} was chosen for the developing platform. Gumbel distribution and GEV (Generalized Extreme Value) distribution are prepared for the extreme value analysis. Both MVLUE (Minimum Variance Linear Unbiased Estimator) method and MLH (Maximum Likelihood) method are applied to calculate the parameters for Gumbel distribution. The prediction of the extreme value of the total population can be carried out using return period. GEV distribution is so-called generalized these three extreme value distribution, such as Gumbel, Weibull and Fréchet distribution. The shape parameter (k) of GEV distribution determines a type of distribution among them. If k is nearly equal to zero, GEV distribution is equivalent to Gumbel distribution. Thus, a test of this hypothesis indicates the fitness criterion for Gumbel distribution.

Keywords *: extreme value statistics, life prediction, corrosion failure, computer program*

1. Introduction

Corrosion phenomena are characterized in two different stages. One is deterministic process, and another is probabilistic process. The data of general corrosion are treated as former one. The accurate data are obtained by the attentive experiment. And the data of localized corrosion are treated as latter one. It is well known that the localized corrosion data are widely scattered and the stochastic treatment is effective.[1] Among some statistical approaches, extreme value statistics was widely used for analyzing the localized corrosion data of plant materials.[2] Gumbel had given a comprehensive treatment of extreme value statistics.[3] Aziz had introduced the application of extreme value statistics for pit depth of aluminum alloys.

The analysis of the localized corrosion by extreme value statistics was commonly used in Japan. Kowaka *et. al.* introduced the application of the life prediction techniques using extreme value statistics.[5] And also, the software named EVAN for calculating extreme value statistics was developed and sold by JSCE (Japan Society of Corrosion Engineering).[6]

However, a long time has passed through the development of this software and it did not match a recent computer generation. And it was restricted only in Japanese language. So, a new software named EVAN-II was developed. Since the software was arranged for a corrosion engineer to calculate extreme value statistics easily, generally used spreadsheet software Excel^{Tm} was chosen for the developing platform.

This software was created by the cooperation of the members of JSCE (Japan Society of Corrosion Engineering) and the Committee of Life Prediction of Industrial Plant Materials.

2. Characteristics of EVAN-II

Evan-II was coded by VBA (Visual Basic Application) language on spreadsheet software EXCEL^{TM}. So, it requires the computer environment with ready for EXCEL. And it requires the version of EXCEL 98 or upper.

Two types of the extreme value analysis are prepared. One is Gumbel distribution, and another is GEV (Generalized Extreme Value) distribution. In Gumbel distri-

bution, two types of parameter estimation methods are prepared, MVLUE (Minimum Valiance Linear Unbiased Estimator) method and MLH (Maximum Likelihood) method. And maximum value estimation can be done using return period(T).

In GEV distribution, parameters are estimated by PWR (Power Weighted Moment) method. And a fitness test of Gumbel distribution can be done.

2.1 Gumbel distribution

Gumbel proved that maximum value was dependent on the sample size and increased unlimitedly with increase of the sample size using extreme value statistics. The theory proves that the distribution of the extreme value approaches asymptotically to three types, Gumbel, Fréchet, and Weibull distribution. The first type of the extreme value distribution (Gumbel) for the largest value is often called the extreme value distribution. It is also called doubly exponential distribution from its formula.

$$F(x)=\exp\left[-\exp\left\{-\frac{(x-\lambda)}{\alpha}\right\}\right] \tag{1}$$

Where F(x) indicates cumulative probability distribution, α is the scale parameter and λ is the location parameter. If standardized valuable, y, defined as;

$$y = {}^{(x-\alpha)}\!/_{\lambda} \tag{2}$$

Equation (1) can be expressed as;

$$F(y) = \exp[-\exp(-y)] \tag{3}$$

When F(x) is determined, α and λ are calculated either MVLUE or MLH methods. In this software, average rank method is used for calculation of F(x).

2.2 MVLUE method

Generally, LUE (Linear Unbiased Estimator) method is a convenient tool for estimating the distribution parameters from comparatively small number of data. For Gumbel distribution, MVLUE method is utilized. Put the data point n and data x_i as follows,

$$x_1 \le x_2 \le \cdots \le x_n \tag{4}$$

The distribution parameters of Gumbel distribution are calculated by next equations.

$$\lambda=\sum_{i=1}^{n} a_i(N,n)\cdot x_i \quad , \quad \alpha=\sum_{i=1}^{n} b_i(N,n)\cdot x_i \tag{5}$$

Where, N is a number of total data and n is a number of significant data. Usually, N equals to n, but in the case of truncated data included, N is greater than n. For example, pit depth measurement was done and an error of these measurements regarded as 0.5 mm. N is a total number of samples measured and n is a number of samples whose pit depth is larger than 0.5 mm. (N-n) is number of truncated samples whose pit depth are smaller than 0.5mm.

The coefficients, $a_i(N,n)$ and $b_i(N,n)$ in the equation (5), are calculated from the complicated matrix equation with parameter N and n. Tsuge had calculated these coefficients up to N =45.[7] This software includes these parameters and can be calculated by MVLUE method up to N = 45.

2.3 MLH method

The likelihood function for data $x=(x_1, x_2, \cdots, x_n)$, which fit the distribution $G(x, \theta)$, are determined by,

$$L(\theta)=\prod_{i=0}^{n} g(x_i;\theta) \tag{6}$$

this equation is solved with the condition $\partial L/\partial\theta=0$. For Gumbel distribution,

$$\begin{cases} \lambda = \bar{x} - \dfrac{\sum_{i=1}^{n} x_i e^{-x_i/\lambda}}{\sum e^{-x_i/\lambda}} \\ \alpha = -\lambda \log\left(\sum_{i=1}^{n} e^{-x_i/\lambda}\Big/ n\right) \end{cases} \tag{7}$$

these two simultaneous equations are derived. The Newton-Raphson method is used for solving these equations. Number of iteration and convergence limit is a factor of this calculation.

2.4 Maximum value estimation using return period

Suppose that the given data set x from the total population fits Gumbel distribution and the distribution parameters α and λ are determined. The extreme value of the total population is calculated using return period (T). T can be expressed as:

$$T = S / s \tag{8}$$

Where S is a size of total population and s is a size of measured data. And T is related to cumulative distribution function, F,

$$F(y) = 1 - 1/T \tag{9}$$

Using this equation, the extreme value is calculated.

For example, water service pipe was analyzed. Total length of the pipe was 25 m and a 1.6 m long pipe was sampled from the line and was cut into 8 section of 0.2 m. The maximum pit was measured in each sample pipe. Then α and λ were calculated to 0.10 and 1.40, respectively.

T is expressed, 20 / 0.2 = 125 and $F(y)$ = 0.992. And the extreme value was estimated by the following equations.

$$y = -\log(-\log F) \quad , \quad x = \lambda + \alpha \cdot y \tag{10}$$

Thus, the extreme value was calculated to 1.96 mm. In this software, the extreme value can be estimated from the input values of S and s.

2.5 GEV distribution

GEV (Generalized Extreme Value) distribution is widely used for modeling extremes of natural phenomena, such as hydrology. GEV distribution is described as next formula.

$$G(x;\mu,\sigma,k) = \exp\left[-\left\{1-k\left(\frac{x-\mu}{\sigma}\right)^{1/k}\right\}\right] \tag{11}$$

Where μ, σ and k are distribution parameters and x is a random variable. These parameters are calculated by PWM (Power Weighted Moment) method.

GEV distribution has following characteristics. When $k < 0$, this equation corresponds to Fréchet distribution. When $k > 0$, this equation corresponds to Weibull distribution. And when $k = 0$, GEV distribution reduces to Gumbel distribution. Namely, Gumbel distribution is a particularly simple special case of GEV distribution, and it is often useful to test whether a given set of data is generated by Gumbel distribution than GEV distribution. This is equivalent to testing whether k equals to zero in GEV distribution.

Here, estimated parameter k is supposed to fit normal distribution. The hypothesis (H_0) whether the distribution does not fit Gumbel distribution in 10 % significance level describes,

$$\left|k/\sqrt{0.5633/n}\right| > 1.654 \Leftrightarrow \text{Rejection of } H_0. \tag{12}$$

Hosking *et al.* introduced the effectiveness of this test.[8]

In this software, fitness test can be carried out using within 10 to 50 % significance level.

3. Usage

3.1 Data input

This software was written by VBA language on Excel. The usage of macro language of Excel is referred the Excel help contents. If you access this software, then the item; "Extreme Value Analysis", is inserted into the top menu items of Excel.

The data, which will be analyzed, are input to a new spreadsheet on Excel. Fig. 1 shows an data input method. The input data are available in the forms of column vector, line vector and box style. But an empty cell is not accepted.

When you have finished the data input, and click the item; "Extreme Value Analysis", the pull-down menu is appeared. You can choose either "Gumbel distribution" or "GEV Analysis" in the menu. And then obey the direction displayed on your PC.

If the item "Gumbel Distribution" is chosen, the followed dialog box (Fig. 2) is opened. You can choose maximum value and minimum value distribution. Two kinds of parameter estimation method are available, MLH method and MVLUE method.

If MLH is chosen, two parameters, convergent condition

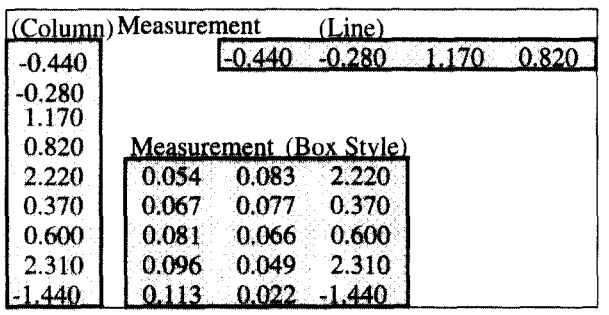

Fig. 1. Data Input Method

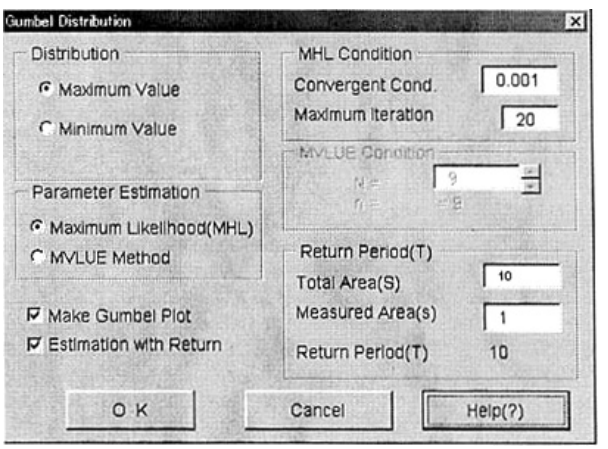

Fig. 2. Menu for Calculating Gumbel Distribution parameters.

and maximum iteration number, can be selected. In ordinary case, the default value is utilized, but in the case of divergence, the convergent condition or maximum iteration number may be changed to larger value.

If MVLUE method is chosen, the total number, N, and measured number, n, can be selected, as mentioned above. Gumbel plot is also available when the item of "Make Gumbel Plot" is checked. If you estimate the extreme value of the total population from the sampling data, the estimation method using return period (T) is prepared. If the box, "Estimation with Return Period" is checked, total area (S) and measured area (s) are required. Then the extreme value will be calculated.

3.2 GEV distribution

When "GEV Analysis" is selected, the menu shown in Fig. 3 is displayed. Maximum and minimum value distribution can be chosen and can also be selected whether GEV plot will be made or not.

Each parameter is calculated by PWM (Power Weighted Moment) method. If $k < 0$, the GEV distribution represents the Frechet distribution and $k > 0$, then GEV distribution represents Weibull distribution.

In the case of $k = 0$, GEV distribution represents Gumbel distribution. And fitness test can be carried out as mentioned above.

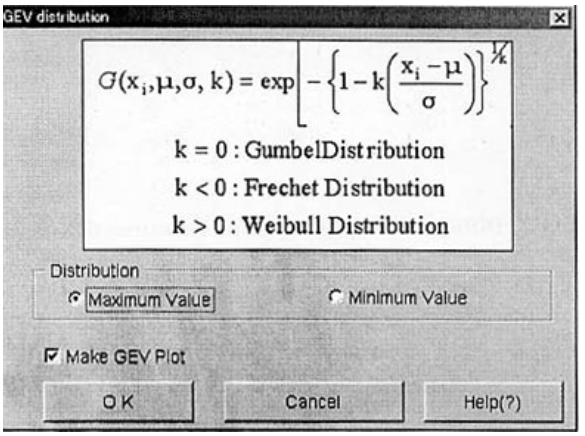

Fig. 3. Menu for GEV Calculaten

4. Analytical results

4.1 Gumbel distribution

Table 1 shows the data analyzed. And the analytical result is shown in Fig. 4. In this case, parameters were calculated by MVLUE method.

i	F(i)	x	y	Ai(N,n)	Bi(N,n)		Maximum value distributionN= 15, n= 15	
1	0.9375	64	2.740493	0.018894	0.048648	3.372523	α	7.744041
2	0.875	60	2.013419	0.024887	0.052334	2.855996	λ	37.88305
3	0.8125	50	1.571953	0.030482	0.053604	1.564681	T	60
4	0.75	48	1.245899	0.035984	0.053267	1.306418	ymax	4.085953
5	0.6875	44	0.981647	0.041329	0.051534	0.789892	Xmax	69.52483
6	0.625	43	0.755015	0.047217	0.048415	0.66076	σ	7.583402
7	0.5625	43	0.552752	0.05314	0.043798	0.66076		
8	0.5	43	0.366513	0.059401	0.037452	0.66076		
9	0.4375	43	0.190339	0.066128	0.028988	0.66076		
10	0.375	38	0.019357	0.073495	0.017779	0.015102		
11	0.3125	37	-0.15113	0.081767	0.002773	-0.11403		
12	0.25	32	-0.32663	0.091384	-0.01793	-0.75969		
13	0.1875	31	-0.5152	0.103196	-0.04829	-0.88882		
14	0.125	31	-0.7321	0.119314	-0.09877	-0.88882		
15	0.0625	31	-1.01978	0.153184	-0.27361	-0.88882		

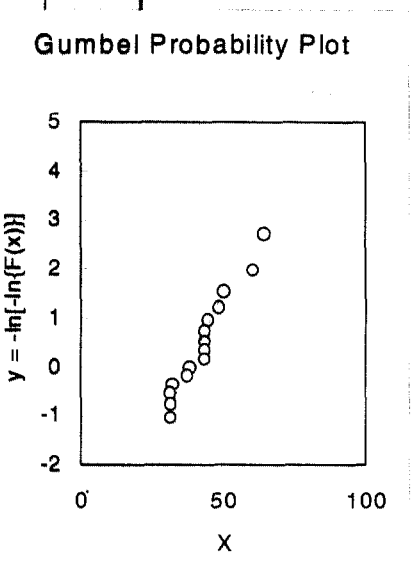

Fig. 4. Analytical Result of Gumbel Distribution

Table 1. Measurements for analysis.

31,	31,	32,	37,	43,
43,	43,	44,	48,	60,
64,	31,	43,	38,	50,

Table 2. Difference between parameter estimation method.

Estimation Method	MVLUE Method	MLH method
α (μm)	7.744	7.641
λ (μm)	37.883	38.034
Return Period(T)	60	60
Estimated Maximum Depth (μm)	69.524	69.256
Error of Estimation(μm)	7.583	7.211

The columns name F(i) and x are cumulative frequency and measured data, respectively. And the columns named Ai(N,n) and Bi(N,n) are MVLUE coefficients. The parameters of distribution, α and λ are shown in the right hand side of the table. The maximum value was estimated using return period, and the calculated results are also displayed in the right hand side columns.

Table 2 shows the difference of calculation method. The same data listed in Table 1 were used. The results calculated by both MVLUE and MLH methods are compared. The calculated α, λ estimated maximum depth and error of estimation are not so different and these are in the range of an error of measurement. This result indicates both methods can be used for analysis. But MLH method does not deal with truncated data, MVLUE is recommended for truncated data included.

4.2 GEV distribution

GEV analysis was performed by the same data and the results are shown in Fig. 5.

The columns name F(i) and x are cumulative frequency and measured data, respectively. And the columns named alpha0 to 2 are the calculation intermediates. and k, σ and μ are the calculated parameters for GEV distribution. Fitness test was also done and results are summarized. In this case, calculated k value is about -0.008. This value is relatively small and fitness test indicates that these data fit Gumbel distribution in 50 % significant level. And the result also indicates that GEV plot was not performed and suggested to use Gumbel plot.

5. Summary

Program for calculating the extreme value statistics (EVAN-II) has been developed on the spreadsheet software. This program was aimed to develop for many corrosion engineers to calculate the extreme statistics easily.

Parameter estimation for Gumbel distribution can be carried out either MVLUE or MLH method and the calculated errors of both methods are under the experimental errors. Fitness test for Gumbel distribution can be done using GEV distribution. The sample data fitted Gumbel distribution proved a benefit of this procedure.
This program is prepared for distribution by JSCE through the Internet.

References

1. T. Shibata and T. Takeyama, *Corrosion*, **33**, 243 (1977).
2. T. Shibata, *ISIJ International*, **31**, 115 (1991).

i	F(i)	x	y(i)	Alpha0	Alpha1	Alpha2		Maximum value distribution	
1	0.0625	31	-1.01544	31	0.043333	0.058211	-1.82306	C(PWM)	-0.00107
2	0.125	31	-0.72986	31	0.11	0.3751	-1.82306	k(PWM)	-0.00837
3	0.1875	31	-0.51409	31	0.176667	0.967544	-1.82306	σ	8.870661
4	0.25	32	-0.32619	32	0.243333	1.894756	-1.71033	μ	47.17174
5	0.3125	37	-0.15104	37	0.31	3.5557	-1.14667	Test for Gumbel ?	
6	0.375	38	0.019358	38	0.376667	5.391356	-1.03394	Obay the Gumbel distribution	
7	0.4375	43	0.190491	43	0.443333	8.451411	-0.47029	Significant level; 50%	
8	0.5	43	0.367076	43	0.51	11.1843	-0.47029	GEV Plot is not given because of smaller k.	
9	0.5625	43	0.554033	43	0.576667	14.29941	-0.47029	You may use Gumbel Plot.	
10	0.625	43	0.757406	43	0.643333	17.79674	-0.47029		
11	0.6875	44	0.985691	44	0.71	22.1804	-0.35755		
12	0.75	48	1.252419	48	0.776667	28.95413	0.09337		
13	0.8125	50	1.58234	50	0.843333	35.56056	0.318833		
14	0.875	60	2.030481	60	0.91	49.686	1.446144		
15	0.9375	64	2.772167	64	0.976667	61.04818	1.897069		
				42.53333	24.33644	17.42692			

Fig. 5. Analytical result of GEV distribution.

3. E. J. Gumbel, "Statistics of Extreme", Columbia Univ. Press, New York, (1958).
4. P. M. Aziz, *Corrosion*, **12**, 495t (1956).
5. M. Kowaka, H. Tsuge, M Akashi, K Masamura, and H. Ishimoto, "Introduction to Life Prediction of Industrial Plant Materials", Allerton Press, Inc. New York, (1994).
6. JSCE 60-1 Technical Committee, Working Groupe (T.Shibata et.al.) , Computer program "EVAN", Maruzen Publ., Tokyo (1989)
7. Tsuge, *J. Soc Mater. Sci. Jpn.*, **36**, 35 (1987)
8. J. R. Hosking, J. R. Wallis and E. F. Wood, *Technometrics*, **27**, 251 (1985).