Three-Parameter Weibull Generator for Replacing Missing Observations

Miloslav Nosal[a] , Eva-Marie Nosal[b]
[a]*Department of Mathematics & Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4*
[b]*SOEST, Department of Geology and Geophysics, University of Hawaii, 1680 East-West Road, POST 813, Honolulu, HI 96822, USA*
nosal@ucalgary.ca

Abstract

A three-parameter Weibull distribution model was developed and verified to simulate the underlying frequency distributions of observations with missing values through the entire range of observations or near and below detection limit of the monitoring equipment. The corresponding Weibull pseudo-random generator then can be used to estimate and replace the missing observations so that the original distribution of the observations is preserved. This method works very well with any empirical data set with missing observations, distribution of which can be well approximated by a three-parameter Weibull distribution. To demonstrate its applicability, the method was used to estimate and synthetically generate the distribution of hourly ozone (O_3) concentrations using the single, weekly mean values obtained from a passive (sodium nitrite absorbent) sampler. The estimation of the Weibull distribution parameters for the synthetic generation of (unavailable) data was based on the data derived from a co-located continuous O_3 monitor.

Keywords: Three-parameter Weibull distribution, Weibull parameter estimation, missing data replacement, observations below detection limit, Kolmogorov-Smirnov distance, ambient ozone distribution, passive sampling

1. Introduction

The problem of systematically missing observations in any field monitoring or measurements is widespread among a majority of empirical sciences. This situation particularly arises in monitoring, when a substantial proportion of observations falls below the detection limit of the corresponding monitors. In empirical practice, such observations are often replaced by zeros or by values equal to the half of the detection limit. This may be acceptable if the percentage of missing values is very small but may seriously jeopardize logical and computational integrity of the monitoring data if this proportion is significant. Another serious problem arises in regression analysis, when a systematic replacement of missing data by a constant may result in undesirable co-linearity among predictors. For this reason we have developed a more sophisticated approach which guarantees statistical consistency of the replaced data. This method consists of replacing missing observations by (pseudo) randomly generated synthetic values which are statistically indistinguishable from the original observations in the sense of homogeneity of the frequency distributions. It is demonstrated that this method works very well in any case when the original empirical frequency distribution can be well approximated by a three-parameter Weibull distribution. It is a known fact that three-parameter Weibull distribution family is extremely flexible and can fit very well an extremely wide range of empirical observations.

Another area of applications of this methodology deals with situations, when a great majority (or all) required data is unavailable (missing or inaccessible due to the cost of monitoring). Yet the general form (or theoretical distribution family) of the underlying frequency distribution of the required observations may be known from elsewhere. In such situations it may be sufficient to estimate the parameters of the known distribution family using a simplified (or cheaper) measuring or monitoring methodology. Using these estimated parameters and the known distribution form, a pseudo-random generator can be used to generate a synthetic data set with statistical properties, which are statistically indistinguishable from the original (unavailable) data. To demonstrate applicability of this approach, this method has been used to generate synthetic hourly ozone concentrations (monitoring of which is very costly) based only on weekly observations of total ozone from (inexpensive) passive ozone samplers. These synthetically generated observations were then compared with the actual co-located hourly (continuous) ozone monitoring observations. It will be demonstrated, that frequency distributions of the actual monitoring of hourly ozone concentrations are statistically indistinguishable from the synthetically generated data. In the recent years there has been a significant increase in the use of passive samplers for measuring ambient ozone (O_3) levels in ecological effects assessment (Krupa and Legge, 2000). There are a number of advantages and limitations associated

with that approach. Nevertheless, the use of passive samplers has been attractive because they are inexpensive, easy to employ and do not require electricity to operate (Manning et al. 1996). However, these data on the mean or total O_3 concentrations over the sampling period can not explain the underlying plant exposure dynamics of continuous ozone concentrations. For this reason, it is essential to have available a complete frequency distribution of the hourly ozone concentrations. Due to its highly skewed asymmetrical shape, the frequency distributions of the hourly ambient O_3 concentrations are best described by the three-parameter Weibull family. In this context, although maximum O_3 concentrations at a given location will vary from year to year, the overall profile and form of the frequency distribution of hourly concentrations will remain the same. Nosal et al. (2000) developed and successfully applied a stochastic, Weibull probability generator for replacing many sequentially missing hourly O_3 values in ambient ozone measurements. It will be shown that it is possible to use the Weibull generator and the passive (weekly) sampler data to synthetically simulate the frequency distributions of (unavailable) hourly O_3 concentrations. The corresponding Weibull distribution parameters can be obtained from a co-located (or near-located) continuous O_3 monitor and corrected with respect to the observed weekly passive ambient ozone monitoring measurements.

2. Estimating parameters of the general three-parameter Weibull distribution family

Estimating the parameters of the three-parameter Weibull distribution family is intrinsically a very difficult task. Nosal et al. (2000) used Monte Carlo methods and APL (Array Processing Language) to investigate the performance of the gradient random search minimization procedure for fitting a Weibull distribution to a given data set using minimum Kolmogorov-Smirnov distance approach. To verify the performance if this approach, the underlying three parameters of the Weibull distribution can be numerically specified and a random sample from the specified distribution can be generated using a pseudo-random number generator based on the specified distribution. Then, using the above- described method, parameters the three-parameter Weibull distribution can be estimated. Goodness of fit and performance of such estimation can be observed by comparing the numerical value of the estimated parameters to the true value of the parameters being estimated.

The above described verification of the performance of estimation of three parameters of the Weibull distribution has been carried out for the following three sets of parameters given in Table 2.1.

	Shape	**Scale**	**Location**
Set 1	1.0	2.0	0.0
Set 2	2.0	1.0	0.0
Set 3	1.5	2.0	0.3

Table 2.1 Selected parameters of the Weibull distribution

The estimation verification has been carried out in the following steps:

Step 1 For each sample size n = 10, 20 , 50, 100, 500, 1000 and each parameter Set1, Set2 and Set3 above, generate n Weibull pseudo random numbers from the distribution with the specified parameters.

Step2 For each sample from Step 1 above (each combination of parameters and sample sizes), estimate the Shape, Scale and Location parameters of the corresponding best fitting Weibull distribution. For this purpose, use the gradient random search minimization procedure for fitting a Weibull distribution to a given data set by minimizing the Kolmogorov-Smirnov distance.

Step 3 Repeat the Step 2 above 20 times and find the mean and the standard deviation of the 20 estimated parameters (Shapes, Scales and Locations) and resulting Kolmogorov – Smirnov minimum distances.

Table 2.2 summarizes results of these simulations. Values in the table represent means of 20 Kolmogorov – Smirnov minimum distances computed in Step 3 between the empirical frequency distribution of the sample data generated in Step 2 and three-parameter Weibull distribution with parameters estimated in Step 2. The last row gives the critical K-S value for ? = 0.05 and corresponding sample size n for testing the goodness of fit between the sampling Empirical Cumulative Distribution Function ECDF(x) and corresponding best fit Weibull Theoretical Cumulative Distribution Function WCDF(x).

$\mathbf{H_o}$**:** ECDF(x) = WCDF(x)
$\mathbf{H_A}$**:** ECDF(x) ? WCDF(x)

	n = 10	n = 20	n = 50	n = 100	n = 500	n = 1000
Set 1	0.13321	0.09603	0.06453	0.04452	0.02052	0.01449
Set 2	0.14012	0.09461	0.06592	0.04578	0.03141	0.01519
Set 3	0.13764	0.09734	0.06261	0.04926	0.02113	0.01445
critical KS value	**0.40925**	**0.29408**	**0.28841**	**0.13403**	**0.05473**	**0.03870**

Table 2.2a Comparison the computed minimu m K-S values with critical K-S values.

It should be noted, that none of the computed (mean) K-S values exceed their corresponding critical values and hence none of the null hypotheses can be rejected. In another words, the three-parameter Weibull dis tributions with parameters of Shape, Scale and Location, estimated in Step 2, fit well the corresponding data generated from three-parameter Weibull distributions with parameters given in Table 2.1. Thus, we can conclude, that the gradient random search minimization procedure for fitting a Weibull distribution to a given data set by minimizing the Kolmogorov-Smirnov distance performs very well indeed.

The means of the 20 estimated triplets of Weibull distribution parameters have been compared to the given values of the Shape, Scale and Location for each Set 1, Set 2 and Set3. These given parameters have been used for generation of the corresponding samples. The tables 2.2b, 2.2c and 2.2d summarize the corresponding results.

Weibull parameters	n = 10	n = 20	n = 50	n = 100	n = 500	n = 1000
Shape = 1	1.17459	1.13027	1.10875	1.06638	1.01947	1.00583
Scale = 2	2.07374	2.00929	2.02605	1.99210	2.05564	2.01180
Locat. = 0	0.13321	-0.09603	-0.06959	-0.02398	-0.01299	-0.00571

Table 2.2b Set 1: Comparison the given Weibull parameters with the means of estimated parameters

Weibull parameters	n = 10	n = 20	n = 50	n = 100	n = 500	n = 1000
Shape = 2	1.93419	1.99847	2.06888	2.12422	2.05850	2.02601
Scale = 1	0.95575	0.96375	0.99267	1.03515	1.02266	1.01111
Locat. = 0	0.07864	0.01120	0.00767	-0.02233	-0.02790	-0.00816

Table 2.2c Set 2: Comparison the given Weibull parameters with the means of estimated parameters

Weibull parameters	n = 10	n = 20	n = 50	n = 100	n = 500	n = 1000
Shape= 1.5	1.42711	1.40596	1.54607	1.53618	1.54331	1.53680
Scale = 2	2.00202	1.95542	2.06534	2.01945	2.02332	2.02065
Locat.= 0.3	0.27756	0.35573	0.24380	0.32273	0.26274	0.30481

Table 2.2d Set 3: Comparison the given Weibull parameters with the means of estimated parameters

From the tables 2.2b, 2.2c and 2.2d we can see that the actual given parameters of Shape, Scale and Location of the corresponding Weibull distribution are very close to the means of the corresponding parameters, estimated from the synthetic samples generated from these distributions. Furthermore, we should notice, that the accuracy of the estimation increases with the increasing sample sizes. This is yet another confirmation that the estimation of the Weibull distribution parameters using the minimum Kolmogorov-Smirnov distance works very well.

The main advantage of the above-described method of estimation is its ability to estimate Weibull distribution parameters when a systematic part of the sample is missing (e.g. all observations below certain value).

This situation arises when all the values around and/or below detection limit are rejected as spurious or missing observations. In this case, the resulting sampling empirical distribution function is truncated at the lower end. However, using the above-described method, the Weibull parameters still can be estimated very efficiently. In order to demonstrate this fact, a similar verification procedure has been performed with truncated distributions representing missing data below the detection limit.

Since computations of truncated distributions represent a major effort, slightly simpler and computationally more manageable approach has been used in this case. For the reasons of computational complexity, only a two-parameter Weibull distribution with Shape = 1 and Scale = 2 has been selected and n = 1000 observations from this distribution has been generated. Then, various percentages p (p = 5%, 10%, 20% 30%, 40%) of smallest data have been removed (treated as missing data). This procedure simulates data measurements or monitoring when all observations below a specified threshold (e.g. detection limit) are rejected and treated as missing data. For each data set, Weibull parameters (Shape and Scale) have been estimated from thus (lower end) truncated empirical distribution. This procedure (generation, truncation and estimation) has been repeated 20 times and means of corresponding 20 estimates of the Shape and Scale have been computed. The Table 2.3 summarizes these results.

Percentage of missing data	Estimated Shape	Estimated Scale	Minimum K-S distance
0%	1.02195	2.02134	0.01625
5%	1.02253	2.02187	0.01627
10%	1.01967	2.02295	0.01625
20%	1.02476	2.02198	0.01550
30%	1.02743	2.01934	0.02893
40%	1.01603	2.01000	0.01448

Table 2.3. Estimated Weibull distribution parameters from truncated distributions

Table 2.3 shows, that the above-described gradient random search minimization procedure for estimation of parameters and fitting Weibull distribution to a given data set by minimizing the Kolmogorov-Smirnov distance works very well. The application of this method allows replacement of missing and/or spurious observations near and/or below the monitoring detection limit (systematically truncated data) by synthetic generated values while preserving the original frequency distribution of the (non-truncated) data. This method is obviously superior from the point of integrity and consistency of data frequency distributions to standard methods of treatment of missing observations such replacement by zeros or replacement by half the detection limit.

3. Application of three-parameter Weibull distribution to ozone data

In the present study hourly (continuous monitoring) and weekly mean (co-located passive sampler) ambient O_3 concentration data were obtained at Paradise (46.47 N, 121.44 W; 1650m MSL; located in Mount Rainier National Park; 100 km SE of Seattle, WA; Brace and Peterson, 1998). These data sets have been collected during the plant growing seasons of 1996 and 1997, and have been analyzed according to the following steps:

STEP 1: Empirical Cumulative frequency Distribution Function ECDF (x) of continuous ozone monitoring data at Paradise (1996) was computed.

STEP 2: Based on the published literature (Nosal, 2000), the cumulative distribution function of the three-parameter Weibull distribution was used as the theoretical distribution to fit the data. The Weibull Cumulative Distribution Function, denoted by $WCDF(x;c,a,x_0)$, is given by the formula

$$WCDF(x;c,a,x_0) = \int_{-\infty}^{x} ca^{-1}\left[(x-x_0)/a\right]^{c-1} \exp\left\{-\left[(x-x_0)/a\right]^{c}\right\} \quad dx\ ;\ (x_0 < x) \qquad (1)$$

where, c is the Shape of the distribution, α is the Scale and ξ_0 is the Location (Johnson and Kotz , 1970).

STEP 3: In order to fit the Weibull distribution, estimation of its parameters (Shape, Scale and Location) was performed using the minimization of the Kolmogorov–Smirnov (KS) distribution distance described in Section 2 above. The KS distribution distance was computed as (Daniel, 1990):

$$KS = \sup_x \left| WCDF\left(x; c, a, x_0\right) - ECDF\left(x\right) \right| \quad (2)$$

The best estimates of the Shape, Scale and Location of the Weibull distribution are values of the parameters, which minimize the KS distribution distance. The actual minimization of the KS distance was achieved using a multidimensional random search algorithm optimized by the partial derivatives of KS with respect to the estimated parameters.

STEP 4: The main purpose here was to verify the theoretical, statistical methodology regardless of any practical considerations. For this reason the parameters of the best Weibull distributions obtained from Paradise (1996) continuous monitoring data in STEP 3 were used to generate synthetic data and their distributions were compared with Paradise (1996) continuous monitoring data distribution.

STEP 5: After obtaining best estimates of Weibull parameters for all ten time periods of Paradise (1996) data, a cubic regression model of those parameters was built using the passive monitoring data as the predictor variable. This model can be written as:

$$y_i = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + e_i \quad (3)$$

where, x is the passive monitoring data, y is the Weibull distribution parameter being predicted, β' values are the regression parameters and ε' values are random error terms. (Cook and Weisberg ,1999). The

STEP 6: Using the regression model from STEP 5, new Weibull distribution parameters for Paradise (1996) were computed using the 1996 passive monitoring data as predictors. Subsequently the Weibull random numbers generator was used to produce new synthetic data for Paradise 1996.

STEP 7: The Weibull generated synthetic data cumulative distributions for Paradise 1996 obtained in STEP 6 were compared with the empirical continuous ozone monitoring distribution functions for Paradise 1996 from STEP 1 using the Kolmogorov–Smirnov distance.

4. Results and Discussion

Table 4.1 provides a summary comparison using the Kolmogorov-Smirnov test for goodness of fit between the actual, continuous monitoring and Weibull modeled synthetic O_3 data generated from the numerical parameters describing the frequency distributions of the continuous data (STEP 3, previous section). In Table 4.1, the differences in the mean O_3 concentrations between the actual and the synthetic data are within 1.0 or 2.0 ppb and the significance values in predominant number of cases are > 0.05. These p values show that the frequency distributions of the measured and the synthetic data are statistically indistinguishable at the 95-99% level depending on the sampling period or week, thus substantiating the utility of the method. Subsequently hourly O_3 frequency distribution parameters from the actual continuous measurements were coupled with a cubic regression model (STEP 5 in the previous section) to derive the passive sampler parameters (Table 4.2). Here, the R^2 values for the shape, scale and location parameters as predictors were modest, 48.0, 64.9 and 65.3% respectively. These results appear to be the products of a non-systematic bias (based on the data from different sampling periods or weeks) in the mean values obtained from the co-located continuous monitor and the passive sampler. However, in this particular case, the mean values from the continuous monitor were always higher, ranging from 1.25 to 14.40 ppb or 3.4 to 30.3% (Table 3). The efficiency of the passive sampler is highly dependent on the face velocity, being low at low wind speeds and high at high wind speeds (Krupa and Legge, 2000). Probability analysis of the relationship between the expected and the simulated data on a weekly basis showed an extremely close fit (Figure 1). In this case, the two data points outside the specified 95% confidence intervals are within the accepted 5% occurrence of outlier values in the statistical test.

The mean values from the synthetic O_3 frequency distribution data (Syn) were only - 3.4 to + 9.07 ppb or -7.0 to + 11.66% (Con - Syn) in variance from the corresponding continuous (Con) monitoring data (Table 4.3). This is in comparison to actually observed random sampling bias of the passive sampler in the present case (-1.25 to -14.40 ppb or -3.4 to - 30.3%, Table 4.3). The overall results predicted by the Weibull model would have been even better if the variance of the measured passive sampler mean O_3 concentrations compared to the corresponding means from the co-located continuous monitor had been less than what was observed in the present case.

Acknowledgements

Some numerical computations involving fitting the Weibull distributions to synthetic data and parameter estimations were performed by Ms. Linda Fung. Applications of the Weibull generator to ambient ozone data were suggested by Dr. S.V.Krupa.

References

Bain, L.J., Antle, C.E., 1967. Estimation of parameters in the Weibull distribution. Technometrics 9, 621-627.

Birnbaum, Z.W., 1952. Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. Journal of American Statistical Association 47, 425-441.

Brace, S., Peterson, D.L., 1998. Spatial patterns of tropospheric ozone in the Mount Ranier region of the Cascade Mountains, U.S.A. Atmospheric Environment 32, 3629-3637.

Cook, D., Weisberg, S., 1999. Applied Regression Including Computing and Graphics, John Wiley, New York.

Daniel, W.W., 1990. Applied Nonparametric Statistics, Duxbury Advanced Series in Statistics, PWS-Kent Publishing Company, Boston, MA.

Dubey, S.D., 1965. Asymptotic properties of several estimators of Weibull parameters. Technometrics 7, 423-434.

Fung, L., 1988. Estimation of Parameters in Weibull Distribution, M.Sc. Thesis. Department of Mathematics and Statistics, University of Calgary.

Johnson, N.L., Kotz, S., 1970. Continuous Univariate Distributions. John Wiley & Sons, New York.

Krupa, S.V., Legge, A.H., 2000. Passive sampling of ambient, gaseous air pollutants: an assessment from an ecological perspective. Environmental Pollution 107, 31-45.

Krupa, S.V., Nosal, M., Peterson, D.L. 2001. Use of passive ambient ozone (O_3) samplers in vegetation effects assessment, . Environmental Pollution 112, 202 – 209.

Lehman, E.H., 1962. Shapes, moments and estimators of the Weibull distribution. Transactions of IEEE-Reliability 11, 32-38.

Manning, W.J., Krupa, S.V., Bergweiler, C.J., Nelson, K.I., 1996. Ambient ozone (O_3) in three Class I wilderness areas in the northeastern USA: measurements with Ogawa passive samplers. Environmental Pollution 91, 399-404.

Nosal, M., Legge, A.H., Krupa, S.V., 2000. Application of a stochastic, Weibull probability generator for replacing missing data on ambient concentrations of gaseous pollutants. Environmental Pollution 108, 439-446.

Table 4.1 Summary statistics of the Kolmogorov-Smirnov (KS) test for goodness of fit between continuously monitored and synthetic data on hourly O_3 concentrations (**ppb**, weekly). The synthetic data were generated from the frequency distribution parameters derived from the **continuously monitored** data. Paradise: **1996**

Period			Continuous O_3		Synthetic O_3	
(week)	KS-distance	Sig[a].	Mean	SD	Mean	SD
1	0.0734	0.738	36.69	8.00	36.54	8.22
2	0.0802	0.646	36.58	12.28	36.77	11.20
3	0.0875	0.536	47.68	12.57	47.23	12.04
4	0.1627	0.022	31.04	11.65	31.16	10.41
5	0.0997	0.355	54.82	10.47	53.53	8.11
6	0.0654	0.859	30.24	8.07	29.78	6.58
7	0.0726	0.710	47.54	11.69	47.37	11.71
8	0.0843	0.566	35.13	9.61	34.74	7.88
9	0.1487	0.042	48.45	13.50	49.78	10.69
10	0.1314	0.104	33.77	8.00	33.22	6.09

[a]Values >0.05 mean that the continuous and the synthetic O_3 frequency distributions are indistinguishable from each other at 95-99% significance level for a given period.

Table 4.2 Weibull, O_3 frequency distribution parameters. Paradise: **1996**

Continuous O_3			Synthetic O_3[b]		
Shape	Scale	Location	Shape	Scale	Location
3.77582	29.6912	9.6107	2.25140	17.3791	17.2160
2.58708	28.9324	9.9916	3.21751	36.0428	8.4731
3.81179	46.3462	6.0169	3.77764	44.3082	7.8221
1.63553	18.2648	13.9906	2.25333	17.3371	17.2137
3.31141	25.6723	31.0000	3.57361	27.5355	28.9631
2.39776	16.6648	14.9145	2.35236	19.1874	16.1572
3.13373	40.4560	11.8518	2.89808	30.0478	10.8570
1.83463	14.9759	20.9992	2.40380	21.2551	15.2183
4.49046	46.5007	8.3589	3.89073	43.2654	11.8336
2.30455	14.3008	19.9961	2.66430	25.4468	12.9766

[b]Synthetic frequency distribution parameters were calculated with the cubic regression model using the passive O_3 values as the predictors

Table 4.3 Summary statistics of the Kolmogorov-Smirnov (KS) test for goodness of fit between continuously monitored and synthetic data on hourly O_3 concentrations (**ppb**, weekly). The synthetic data were generated from the frequency distribution parameters derived from the **passive sampler** data. Paradise: **1996**

			Continuous O_3		Synthetic O_3		Passive O_3
Week	KS-distance	Sig.	Mean	SD	Mean	SD	Mean
1	0.2786	0.000	36.69	8.00	32.41	6.41	24.43
2	0.1932	0.004	36.58	12.28	39.73	11.57	35.33
3	0.0917	0.386	47.68	12.57	48.14	11.88	41.27
4	0.2875	0.000	31.04	11.65	32.68	7.29	25.07
5	0.0968	0.391	54.82	10.47	53.45	7.90	49.15
6	0.2774	0.000	30.24	8.07	33.20	8.15	27.63
7	0.3626	0.000	47.54	11.69	38.47	10.30	33.14
8	0.1081	0.261	35.13	9.61	34.01	8.20	21.31
9	0.1257	0.126	48.45	13.50	51.85	11.42	44.21
10	0.2571	0.000	33.77	8.00	35.13	8.88	31.13

Table 4.4 Model verification: Summary statistics of the Kolmogorov-Smirnov (KS) test for goodness of fit between continuously monitored and synthetic data on hourly O_3 concentrations (**ppb**, weekly). The synthetic data were generated from the frequency distribution parameters derived from the **passive sampler** data. Paradise: **1997**

			Continuous O_3		Synthetic O_3		PassiveO_3
Period (week)	KS-distance	Sig.	Mean	SD	Mean	SD	Mean
1	0.2718	0.000	30.06	7.48	33.46	7.62	23.00
2	0.3924	0.000	29.52	3.42	33.31	7.82	27.05
3	0.2381	0.000	30.04	5.28	33.27	8.59	26.05
4	0.2009	0.002	33.90	10.28	33.92	8.11	26.29
5	0.2289	0.000	40.65	7.20	37.62	9.81	30.51
6	0.1848	0.006	36.04	11.80	39.54	11.76	34.01
7	0.1123	0.233	32.75	9.55	34.48	9.91	29.35
8	0.2092	0.001	31.21	7.18	34.38	8.28	28.81
9	0.2133	0.001	30.02	5.89	33.31	7.75	25.26
10	0.2033	0.002	31.57	7.17	34.81	8.69	27.71

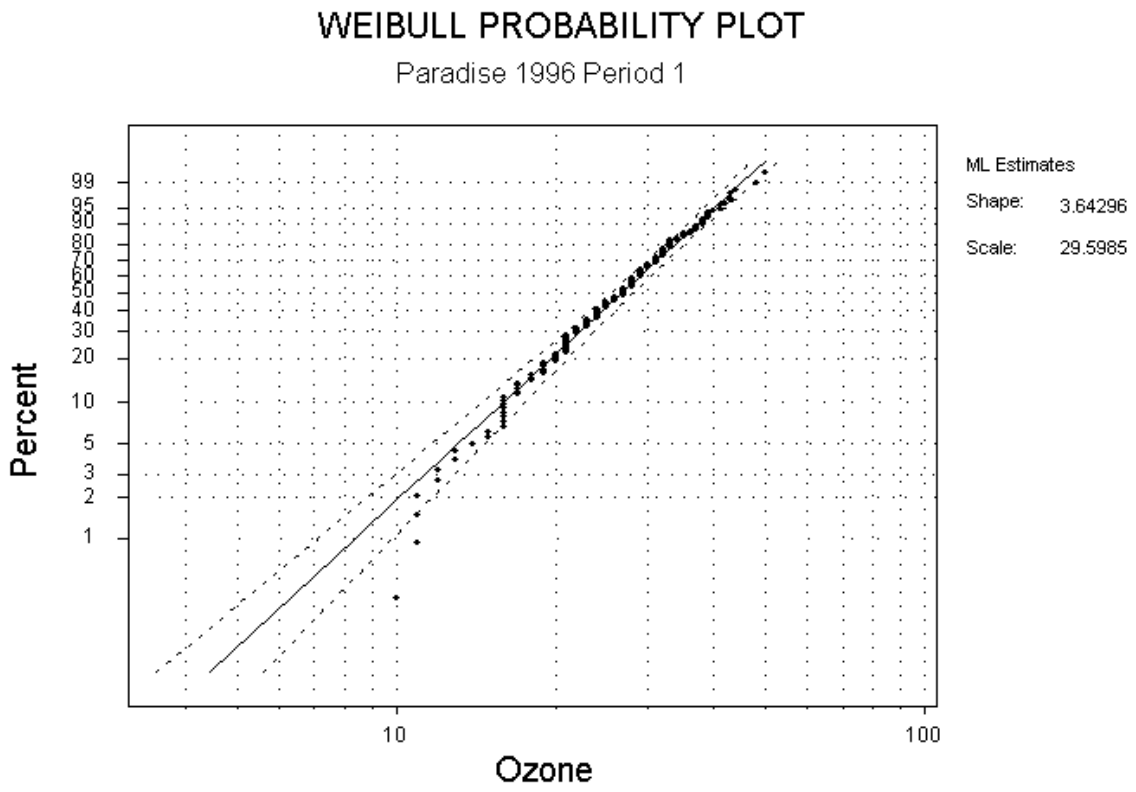


Fig.1. An example of the percent Weibull probability distribution of the synthetic or simulated hourly O_3 concentrations (**ppb**) derived from a passive sampler weekly mean value (June 26 - July 3). The horizontal axis represents the ordered statistic of the O_3 levels. The solid line in the figure shows the expected, the solid circles show the distribution of the modeled hourly concentrations and the two dotted lines represent the 95% confidence intervals. The two data points outside the 95% confidence interval are within the 5% of occurrence of outlier values allowed in the test.