PROSIDING

SEMINAR KEBANGSAAN PENGOPTIMUMAN BERANGKA DAN PENYELIDIKAN OPERASI KE - 2 (PBPO - 2)

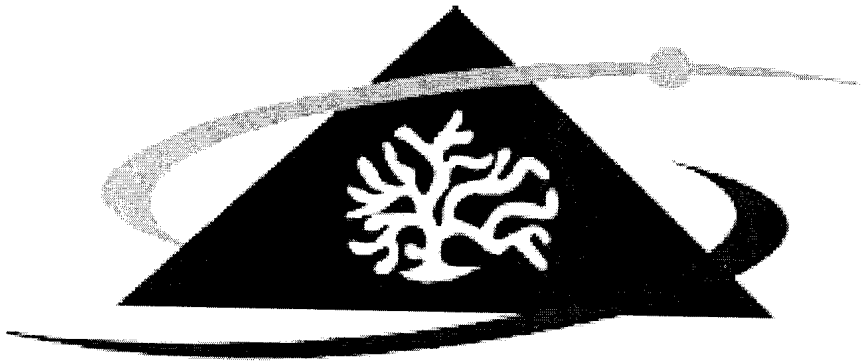
Penggabungan Idea Dalam Kepelbagaian Penyelidikan Sains Matematik Menjana Pembangunan Pengoptimuman Berangka dan Penyelidikan Operasi

Suntingan

Ismail Bin Mohd
Mustafa Bin Mamat
Ilyani Binti Abdullah
Mohd Lazim Bin Abdullah
Goh Khang Wen
Farikhin
Zulhanif
Hadi Sumadibrata
Muhammad Iqbal Al-Banna Bin Ismail

13 - 14 DISEMBER 2008

UNIVERSITI MALAYSIA TERENGGANU



UNIVERSITI MALAYSIA TERENGGANU

PROSIDING

SEMINAR KEBANGSAAN PENGOPTIMUMAN BERANGKA DAN PENYELIDIKAN OPERASI KE-2 13 – 14 DISEMBER 2008

Anjuran:
Kumpulan Penyelidikan Informatik dan Permodelan Matematik
Universiti Malaysia Terengganu

dengan kerjasama

Jabatan Matematik
Fakulti Sains dan Teknologi
Universiti Malaysia Terengganu

Prosiding Seminar Kebangsaan Pengoptimuman Berangka dan Penyelidikan Operasi Ke-2

ISBN 978-983-2888-89-5

PENERBIT UMT

KANDUNGAN/CONTENT

On the Analysis of Nondetect (Left-censored) Environmental Data

Abdul Kudus[1,2] and Noor Akma Ibrahim[2,3]
[1]Department of Statistics, Bandung Islamic University
[1,2]Institute for Mathematical Research - Universiti Putra Malaysia
[2,3]Department of Mathematics - Universiti Putra Malaysia
: [1]akudus69@yahoo.com, [2]nakma@putra.upm.edu.my

Abstract. It is common that some observations of environmental measurements such as pollutant levels are recorded as below the detection (or reporting) limits of instrumentation. A sample of data contains nondetect (left-censored) observations if some of the observations are reported only as being below some censoring level. This practice, however, creates special problems in the analysis of the data. Although nondetect results in some loss of information, we can still use data that contain nondetect for graphical and statistical analysis. In this paper we discuss various statistical methods for dealing with nondetect data, i.e. graph creation and distribution parameters estimation. Those include simple substitution of detection limits, maximum likelihood estimators, and probability plotting.

Introduction

Environmental data such as chemical concentrations in soil, air, and water frequently contain values that are below the detection limit (DL). They are recorded as below specified analytical reporting limits due to measurement capacities or economical/practical concerns. In a spreadsheet or file as received from the laboratory these data will most often be marked as "<DL", where DL is the actual value of the detection limit (e.g., 0.001 ppm).

Statistically, a data set with nondetect observations recorded as being below a certain limit is called "left censored". Type I left censoring is most often encountered; that is, each detection limit is fixed and a random (but known) number of censored observations occur below each limit. The data in a given sample may be subject to a single detection limit (singly censored) or multiple detection limits (multiply censored). Although this results in some loss of information, we can still use data that contain nondetects for graphical and statistical analysis. Inevitably, more importance is placed on extracting information from nondetect observations.

We explore some implications of these nondetects in the summary, analysis, and interpretation. Alternative approaches for handling nondetect data are examined, with practical considerations being a key element in the selection of appropriate methodology.

There are several approaches to dealing with data where some values are "<DL". Options are:

- Delete the whole variable or all samples with values "<DL" from data analysis;
- Mark all observations "<DL" as missing;
- Model a distribution in the interval [0, DL], and assign an arbitrarily chosen value from this distribution to each sample <DL;
- Try to predict a value for this variable in each sample via multiple regression (imputation) techniques using all other analytical results; or
- Set all values marked "<DL" to an arbitrarily chosen low number, e.g., half the DL.

None of these solutions is ideal. To delete samples from data analysis is not acceptable, it will shift all statistical estimates towards the "high" end, although there is information that the concentration in a considerable number of samples is low. The same happens if the values are marked as "missing".

To assess air quality data, it is useful to determine the probability distribution that best fits the data, since a distribution provides a better characterization of the data than point estimates alone. For instance, probability distributions can be used to assess the likelihood of observing data points above hazard threshold limits established by regulatory bodies (e.g. Wild *et al.* 1996). Knowledge of the underlying data distribution permits analyses of observations below the analytical limit of detection that are more precise than commonly used distribution-free substitution methods.

Common candidate probability distributions used for modeling environmental contamination data include two-parameter lognormal, gamma, four-parameter beta, and Weibull distributions (Holland and Fitz-Simmons 1982; Gilbert 1987). Given a choice between these distributions, the Weibull is preferred since it has many advantages over the other three. These advantages include the explicit computability of percentiles, maximum likelihood equations that can be solved by simple iteration methods, and a distribution function that can be linearized for least-squares estimates of parameter values. Of the remaining three probability distributions, the gamma is the most difficult to work with. Disadvantages include the necessity of partial differentiation of gamma functions for the calculation of maximum likelihood estimators (MLEs). The beta distribution is only appropriate for data that are bounded by both lower and upper limits (Gilbert 1987). Because of these limitations of the gamma and beta distributions, all statistical analyses presented here focus on Weibull and lognormal distributions.

Section 2 discussed statistical modelling of environmental data by parametric model. The methods for estimating distribution parameter which are fitted by using complete and nondetect observation are thoroughly described. Section 3 presents the actual analysis of Malaysia's air quality data, as well as a comparison of methods used to estimate parameters for simulated and actual left-censored data. Section 4 discusses results and offers some concluding remarks.

Statistical Background and Methods

Let $X_{(1)} \leq \ldots \leq X_{(c)} \leq X_{(c+1)} \leq \ldots \leq X_{(n)}$ be an ordered random sample of size n from a particular distribution, where $X_{(1)}, \ldots, X_{(c)}$ are censored on the left. Let n be the total sample size, m be the number of non-censored (fully measured) observations, and $c = n - m$ be the number of left-censored observations, where $n = m + c$.

If X_i is censored on the left, then X_i is not observed, but its left-censored limit DL is observed, and it is understood that $X_i < DL$ for $i = 1, \ldots, c$. All methods described can be used to estimate the parameter of distribution such as lognormal and Weibull.

The Lognormal model

The lognormal distribution, which is a simple transform of a normal distribution, is often considered the default in environment analysis. By considering its relationship with normal distribution, the central limit theorem is applicable upon logarithmic transformation.

Maximum likelihood estimation method

Assume that the random variable X_i can be described adequately by a lognormal distribution. In the other hand ln-transformed of X will normally distributed. Let $Y_i = \ln(X_i)$ for $i = c+1, \ldots, n$ and let $Y_{censor} = \ln(DL)$. The general maximum likelihood function for any distribution with parameter vector θ is given by

$$L(\theta \mid \underline{x}) = \binom{n}{c}\left[P(X < DL)\right]^c \prod_{i=c+1}^{n} f(x_i) \qquad (1)$$

where f and $P(X < DL) = F(DL)$ denote the pdf and cdf of the population, respectively. The likelihood is product of (1) the probability of c observations out of n being less than DL and (2) the product of values of the pdf evaluated at the uncensored observations. In the case of lognormal model above, we have

$$L(\mu_y, \sigma_y) = \binom{n}{c}\left[\Phi(\xi)\right]^c \left[\frac{1}{\sqrt{(2\pi)}\sigma_y}\right]^m \exp\left(\frac{\sum_{i=c+1}^{n}(y_i - \mu_y)^2}{-2\sigma_y^2}\right) \qquad (2)$$

where $\xi = \dfrac{y_{censor} - \mu_y}{\sigma_y}$, Φ is the cumulative distribution function of a standard normal variate, μ_y is the mean and σ_y is the standard deviation of the ln-transformed data.

The log-likelihood function for a sample of the type under consideration is given by

$$l(\mu_y, \sigma_y) = \ln\binom{n}{c} - m\ln\sqrt{2\pi} - m\ln\sigma_y + c\ln\Phi(\xi) - \frac{1}{2\sigma_y^2}\sum_{i=c+1}^{n}(y_i - \mu_y)^2 \qquad (3)$$

Equating to zero the first partial derivatives of the log-likelihood function with respect to μ_y and σ_y yields:

$$\frac{c}{\sigma_y}\frac{\phi(\xi)}{\Phi(\xi)} - \frac{\sum_{i=c+1}^{n}(y_i-\mu_y)}{\sigma_y^2} = 0 \tag{4}$$

and

$$\frac{c}{\sigma_y}\frac{\phi(\xi)}{\Phi(\xi)}\xi + \frac{m}{\sigma_y} - \frac{\sum_{i=c+1}^{n}(y_i-\mu_y)^2}{\sigma_y^3} = 0 \tag{5}$$

The maximum likelihood estimates, $\hat{\mu}_y$ and $\hat{\sigma}_y$, of μ_y and σ_y are formulated by Cohen (1959) as follows

$$\hat{\mu}_y = \bar{y} - \hat{\lambda}(\bar{y} - y_{censor}) \tag{6}$$

$$\hat{\sigma}_y = \sqrt{s^2 + \hat{\lambda}(\bar{y} - y_{censor})^2} \tag{7}$$

Thus, the maximum likelihood estimates of the mean μ_y and the standard deviation σ_y of the censored data are based on:

- the mean $\bar{y}$ and variance s^2 of the m ln-transformed observations which are numerically known,
- the detection limit, DL,
- a positive constant $\hat{\lambda}$ which provided by Cohen (1959).

To compute the estimated mean $\hat{\mu}_x$ and standard deviation $\hat{\sigma}_x$ of the original lognormal data set x_i, the following back-transformation is required:

$$\hat{\mu}_x = \exp\left(\hat{\mu}_y + \frac{1}{2}\hat{\sigma}_y^2\right) \tag{8}$$

$$\hat{\sigma}_x = \sqrt{\hat{\mu}_x^2\left[\exp\left(\hat{\sigma}_y^2\right) - 1\right]} \tag{9}$$

Probability-plot regression method

It is possible to estimate the mean and standard deviation of a censored lognormally distributed data set based on a linear relationship of the ln-transformed uncensored values versus the normal scores z_i. The regression estimates of μ_y and σ_y are found by computing the least-squares estimates in the following linear model:

$$y_{(i)} = \mu_y + \sigma_y\Phi^{-1}(p_i) + \varepsilon_i \tag{10}$$

where p_i denotes the Blom plotting position associated with the i^{th} largest values and is defined by (Kroll and Stedinger, 1996)

$$p_i = \frac{c}{n} + \frac{m}{n}\left(\frac{i-0.375-c}{n+0.25-c}\right), \; i = c+1, \ldots, n \tag{11}$$

and $\Phi^{-1}(p_i)$ denotes the inverse cumulative normal distribution evaluated at p_i.

The Weibull model (complete data)

While widely used, the lognormal distribution does not always provide the optimal representation of contamination data (cf. El-Shaarawi and Viveros 1997). A suitable alternative is the Weibull distribution (Gilbert 1987). In the following, we will use the two-parameter Weibull probability density function (PDF) in the terminology

$$f(x \mid \alpha, \beta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}} \tag{12}$$

where α and β are scale and shape parameters, respectively. The expected value and variance of x in terms of these parameters are

$$E(X)=\alpha\Gamma\left(1+\frac{1}{\beta}\right) \text{ and } Var(X)=\alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\Gamma^2\left(1+\frac{1}{\beta}\right)\right] \tag{13}$$

Maximum likelihood estimation method

For the case of all uncensored data, the log-likelihood function is:

$$l(\alpha,\beta\,|\,\underline{x})=n\ln\beta-n\beta\ln\alpha+(\beta-1)\sum_{i=1}^{n}\ln x_i-\frac{1}{\alpha^{\beta}}\sum_{i=1}^{n}x_i^{\beta} \tag{14}$$

By differentiating log-likelihood function (14) with respect to α and β in turn and equating to zero, the maximum-likelihood estimators, $\hat{\alpha}$ and $\hat{\beta}$ satisfy the equations:

$$\frac{1}{\hat{\beta}}-\ln\alpha+\frac{\sum_{i=1}^{n}\ln x_i}{n}+\frac{1}{n\hat{\alpha}^{\hat{\beta}}}\left(\ln\hat{\alpha}\sum_{i=1}^{n}x_i^{\hat{\beta}}-\sum_{i=1}^{n}x_i^{\hat{\beta}}\ln x_i\right)=0 \tag{15}$$

and

$$\hat{\alpha}=\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\hat{\beta}}\right)^{\frac{1}{\hat{\beta}}} \tag{16}$$

In principle, a standard iterative search method, such as the Newton-Raphson procedure, can be used to solve (15) and (16).

Linear regression method

An alternative to maximum likelihood estimation of the Weibull parameters is simple linear regression (LR). The regression is indeed linear, since the cumulative distribution function (CDF) of the Weibull distribution can be logarithmically transformed as

$$\ln(-\ln(1-F(x_i)))=\beta\ln x_i-\beta\ln\alpha \tag{17}$$

In actual data analyses, the true CDF is not known. It is typically approximated non-parametrically by

$$F(x_i)=(i-0.5)/n \tag{18}$$

where i is the rank of the observation, n is the total sample size, and 0.5 is a continuity correction factor (see Johnson and Wichern 1992). The continuity factor is included in the equation to prevent singularities that would result in (17) at $F(x_i) = 0$ and 1. Alternative forms for $F(x_i)$ include $\left(i-\frac{3}{8}\right)/\left(n+\frac{1}{4}\right)$ and $i/(n+1)$ (Gan and Koehler 1990). In the LR method, the confidence interval (CI) for β is estimated from the CI of the regression slope parameter, while the CI for α is derived by using normal approximation for the intercept.

The fits of candidate probability distributions to the data were assessed in our analysis with Kolmogorov-Smirnov test of fit. Because of the logarithmic relationships between normal and lognormal and between Weibull and Gumbel distributions, a log transformation of the data allows us to assess efficiently the fit of both the lognormal and Weibull distributions to the data.

Non-detects

The presence of non-detect observations complicate all aspects of statistical evaluation and must be taken into account, especially when they occur in relatively large numbers. Dealing with non-detects is essentially a problem of left data censoring with a fixed truncation point that equals the limit of detection. One common method for dealing with the problem is to set the non-detects equal to some fill-in constant such as zero, the limit of detection, or a value somewhere in between, such as

$$nondetect=\frac{DL}{2} \tag{19}$$

Substitution by (19) implicitly assumes that non-detects are uniformly distributed below the DL, where (19) is

interpreted as the expected value of the non-detects.

In general, using replacement values for non-detects produces bias in parameter estimates, and the direction and magnitude of this bias depend on the estimated parameter values (El-Shaarawi and Esterby 1992). For this reason, methods accounting for the underlying distribution of the data are recommended over simple substitution (Helsel 2005). Two such methods are a maximum likelihood procedure for left-censored data first considered by Cohen (1959), and a log-probability regression technique developed by Travis and Land (1990). Indeed, Helsel (1990) states that these two methods give unbiased results when the data sufficiently fit the assumed distribution and the sample size is large.

In most investigations of left-censored estimation methods, the data are assumed to be lognormally distributed, or they are transformed to normality (e.g. Stoline 1991). In contrast, little has been written about estimating parameters for the left-censored Weibull. Since the data may be best represented by this distribution, we explored means of dealing with non-detects in Weibull distributed data. Two strategies were pursued. The first is based on maximum likelihood estimation, and the second on the linearized representation of the Weibull in (17).

Maximum likelihood estimation method

As a method for estimating left-censored Weibull parameters, we derived MLEs for such data. By considering the general likelihood function (1) and using Weibull distribution, the likelihood function takes the form

$$l(\alpha,\beta \mid \underline{x}) = \ln\binom{n}{c} + c\ln\left\{1-\exp\left[\left(\frac{DL}{\alpha}\right)^{\beta}\right]\right\} + m\ln\beta - m\beta\ln\alpha + (\beta-1)\sum_{i=c+1}^{n}\ln x_i - \frac{\sum_{i=c+1}^{n} x_i^{\beta}}{\alpha^{\beta}} \qquad (20)$$

where c is the number of non-detects. Partial differentiation of (20) yields the estimating equations:

$$\frac{c\left(\frac{DL}{\hat{\alpha}}\right)^{\hat{\beta}}\ln\left(\frac{DL}{\hat{\alpha}}\right)\exp\left\{-\left(\frac{DL}{\hat{\alpha}}\right)^{\hat{\beta}}\right\}}{1-\exp\left\{-\left(\frac{DL}{\hat{\alpha}}\right)^{\hat{\beta}}\right\}} + \frac{m}{\hat{\beta}} - m\ln\hat{\alpha} + \sum_{i=c+1}^{n}\ln x_i + \frac{1}{\hat{\alpha}^{\hat{\beta}}}\left(\ln\hat{\alpha}\sum_{i=c+1}^{n} x_i^{\hat{\beta}} - \sum_{i=c+1}^{n} x_i^{\hat{\beta}}\ln x_i\right) = 0 \qquad (21)$$

and

$$\frac{c\left(\frac{DL}{\hat{\alpha}}\right)^{\hat{\beta}}\hat{\beta}\exp\left[-\left(\frac{DL}{\hat{\alpha}}\right)^{\hat{\beta}}\right]}{\hat{\alpha}\left\{1-\exp\left[-\left(\frac{DL}{\hat{\alpha}}\right)^{\hat{\beta}}\right]\right\}} + \frac{m\hat{\beta}}{\hat{\alpha}} - \frac{\hat{\beta}}{\hat{\alpha}^{\hat{\beta}+1}}\sum_{i=c+1}^{n} x_i^{\hat{\beta}} = 0 \qquad (22)$$

Equations (21) and (22) are not separable, but they can be solved simultaneously to yield $\hat{\alpha}$ and $\hat{\beta}$.

Linear regression method

The second method for determining the left-censored Weibull parameters is based on the linear transformation of the Weibull CDF, as presented in equation (17). This method is particularly well suited for left-censored data since the information necessary for determining parameter values is contained in the cumulative probabilities of the observed data points, which are independent of the precise location of the non-detects. Thus, the cumulative function of the group of non-detects can be estimated by the rank of the data even though their individual values are unknown. Specifically, the censored points are expected to lie somewhere below the limit of detection on the regression line of $\ln(-\ln(1-F(x_i)))$ versus $\ln x_i$, whose slope and intercept are computed from the measured values (Gilbert and Kinnison 1981). This method is computationally simpler than the maximum likelihood technique presented above; however, results may be biased due to the log transformation of the data (e.g., Helsel 1990).

Analysis of Malaysia's Air Quality Data

To illustrate the practical application of estimators discussed in the preceding sections, the following example

has been selected. The Malaysia Air Quality Data Report published by the Department of Environment (DOE) contains concentration of some air quality which monitored across country. There are 28 sites managed by DOE in Peninsula, Sabah and Sarawak. Three pollutants concentrations are selected, i.e. particulate matter (PM10), nitrogen dioxide (NO_2) and ozone (O_3).

We Performs a Kolmogorov-Smirnov goodness-of-fit test to test the hypothesis that a random sample comes from lognormal or Weibull distribution. The result shows that all three pollutants concentration considered here are well modeled by the Weibull distribution. The lognormal cannot be rejected for all of the three pollutant concentration, but its p-value is inferior compared to the Weibull as presented in Table 1.

Table 1. Goodness of fit test for pollutant concentration

	PM10		NO_2		O_3	
	statistic	p-value	statistic	p-value	statistic	p-value
Lognormal	0.111	0.500	0.168	0.057	0.121	0.500
Weibull	0.110	0.865	0.127	0.745	0.137	0.731

Weibull parameter estimates for all three pollutant concentrations are shown in Table 2. The MLEs estimates are given as a comparison to the estimates obtained by linear regression. Inspection of the table indicates that there is good agreement between the maximum likelihood and the LR estimators across all three pollutant concentrations.

Table 2. Weibull parameter estimates for pollutant concentration data.

	PM10		NO_2		O_3	
MLE						
$\hat{\alpha}$ (95% CI)	31.162	(27.753,34.571)	0.010	(0.007,0.012)	0.034	(0.029,0.040)
$\hat{\beta}$ (95% CI)	3.647	(2.651,4.644)	1.527	(1.080,1.974)	2.588	(1.757,3.420)
LR						
$\hat{\alpha}$ (95% CI)	31.078	(30.426,31.730)	0.010	(0.009,0.010)	0.034	(0.033,0.036)
$\hat{\beta}$ (95% CI)	4.021	(3.767,4.274)	1.484	(1.391,1.578)	2.651	(2.415,2.887)

Test of parameter estimation methods for data sets with non-detects

The completeness of the set of pollutant concentrations gives us the opportunity to test the effects of non-detects on parameter estimates. To this end, we artificially truncated the data to simulate the effects of higher limits of detection. Three parameter estimation techniques were considered: the ML method given by (21) and (22); the substitution method (SM), for which maximum likelihood estimates are obtained upon estimating the non-detects as shown in (19); and the LR method (17), where parameters are determined from a regression of data points above the DL. Ideally, a good parameter estimation technique for censored data should yield results in agreement with those obtained from the complete data. The standard method for measuring the deviation of parameter estimates from a reference value is given by the root mean square error:

$$RMSE = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i - \theta\right)^2} \tag{23}$$

where θ is the value of the reference parameter, $\hat{\theta}$ is the parameter estimate, and B is the total number of censored data samples. In our simulation examples, the reference parameters are set equal to estimates obtained from the complete data set, except for simulated data which use its parameter value.

As a preliminary assessment for the truncation method that was unaffected by the selection of a particular data set, parameters were estimated for random data taken from simulated Weibull distributions. We generate 200 random data sets each had a sample size of n = 1000 with Weibull parameters were α = 31 and β = 4. The ten lowest values were progressively censored until one half of the observations was removed, giving a total of 50 subsamples for each simulated data set. The

left truncation point for the ML method was set equal to the lowest observation in each censored data set.

Table 3. Minimum, maximum and RMSE of the Weibull parameter estimates for all censored subsamples.

	Simulated Weibull (α=31,β=4)	PM10	NO_2	O_3
MLE				
$\hat{\alpha}$ [min,max]	[27.364, 30.961]	[27.658,31.069]	[0.007,0.010]	[0.029,0.034]
RMSE	1.249	1.361	0.001	0.002
$\hat{\beta}$ [min,max]	[2.261, 3.978]	[2.248,3.570]	[0.883,1.491]	[1.517,2.512]
RMSE	0.816	0.697	0.326	0.564
SM				
$\hat{\alpha}$ [min,max]	[27.535, 30.946]	[27.749,30.941]	[0.009,0.010]	[0.031,0.034]
RMSE	1.495	1.694	0.000	0.002
$\hat{\beta}$ [min,max]	[2.518, 3.960]	[2.445,3.463]	[1.490,1.603]	[1.914,2.459]
RMSE	0.907	0.810	0.008	0.460
LR				
$\hat{\alpha}$ [min,max]	[31.054, 36.353]	[31.528,36.712]	[0.010,0.014]	[0.035,0.044]
RMSE	2.751	3.050	0.002	0.006
$\hat{\beta}$ [min,max]	[4.306, 9.293]	[4.450,6.679]	[1.868,2.705]	[2.865,4.728]
RMSE	2.911	1.915	0.943	1.102

The ranges and RMSEs of the parameter estimates for data from the simulated Weibull are summarized in column one of Table 3. In general, all three methods produced a small range of parameter estimates and low RMSEs. The ML estimates gave the lowest RMSEs with values of 1.249 and 0.816 for $\hat{\alpha}$ and $\hat{\beta}$, respectively. SM estimates had deviations that were slightly higher than the ML estimates, while estimates from linear regression produced largest RMSEs, 2.751 for $\hat{\alpha}$ and 2.911 for $\hat{\beta}$. Parameter estimates for all the subsamples are depicted graphically in Figure 1 for all three methods.

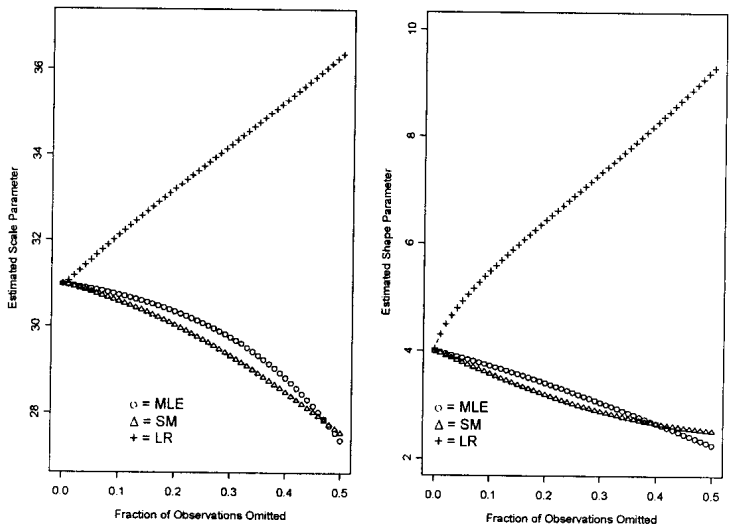


Figure 1. Estimates of the scale (α) and shape (β) parameters for the simulated Weibull distribution (α=31, β=4) data.

Based on the above results one could be tempted to conclude that the MLE method is slightly more reliable means for estimating parameters from left-censored data compared to the rest.

Ranges and RMSEs for the Weibull parameter estimated from the left-censored pollutant concentration data are also given in Table 3. These data were progressively censored with up to one half of the observations removed, giving up to 14 subsamples containing between 1 and 14 non-detects. In general, LR parameters gave the largest deviations, while the RMSEs for the ML and the SM parameters were approximately equal within each pollutant data set. Again, the substitution method by equation (36) most likely gave good estimates. However, the RMSE of the parameter estimates for these pollutants gives an incomplete picture. Figure 2 shows the parameter estimates for PM10 subsamples. In this graph, the estimates are diverging as the truncation value increase. The same result is also found for O_3 subsamples (see Figure 3). The slightly different pattern shown by parameter estimates in the NO_2 subsamples, where SM method gives fairly consistent estimates. Even though ML and LR estimates still diverging (see Figure 4).

Figure 5 illustrates the subsample means for all three pollutant concentrations calculated by (30) with the ML parameters. From this figure, it is evident that estimates of the mean are less sensitive to data censoring than the parameters themselves.

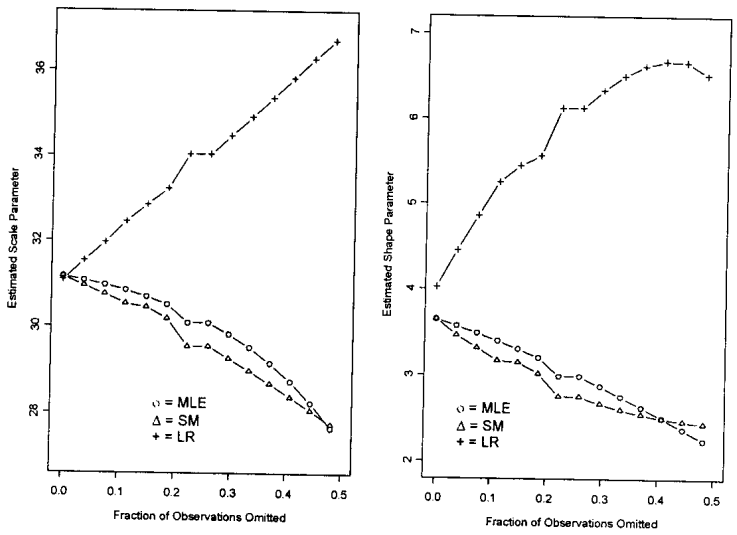


Figure 2. Estimates of the scale (α) and shape (β) parameters for the PM10 pollutant data.

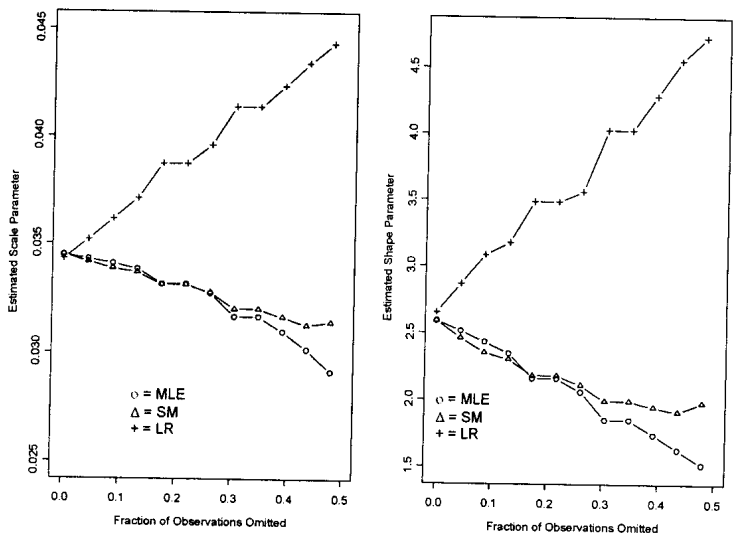


Figure 3. Estimates of the scale (α) and shape (β) parameters for the O_3 pollutant data.

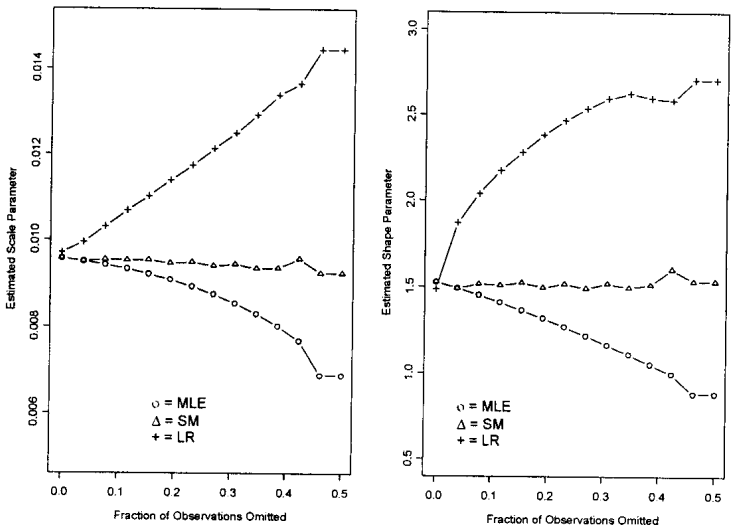


Figure 4. Estimates of the scale (α) and shape (β) parameters for the NO_2 pollutant data.

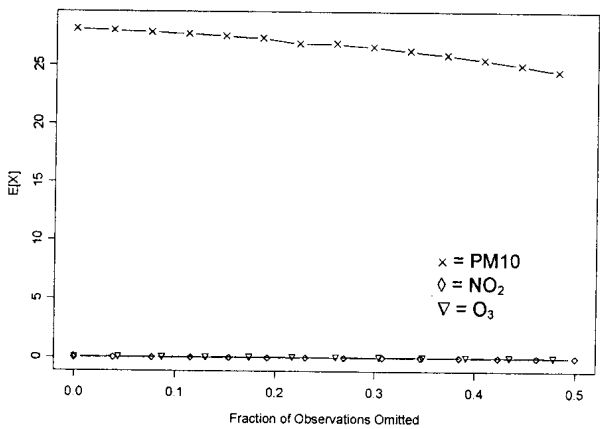


Figure 5. Estimates of the mean according to (13) for pollutant concentration data obtained from ML parameter estimates for all left-censored subsamples.

Conclusion

In this paper we presented methods for dealing with nondetect observations. By assuming data follow lognormal or Weibull distribution, MLE and parameter estimation based on linear regression method are discussed. Since lognormally distributed data can be viewed as normally distributed in the log-transformed data, then we can work with the normal distribution setting for analyzing such data. This was accomplished by first determining the probability distributions that appropriately fit the data, and second, by estimating the

parameters of the best-fitting distributions. For the first part, we demonstrated that Kolmogorov-Smirnov test provide an efficient and effective means for determining the fit of given probability distributions to the data. In the case of the lognormal, this was most easily accomplished by assessing the fit of the normal distribution to the natural log of the observations. When the normal distribution was applied to the log of pollutant concentration and the Weibull distribution was applied to the original pollutant concentration data, it was determined that the Weibull distribution provided a better fit than the lognormal for all three pollutant concentration considered.

The Weibull parameters for the pollutant concentration data were estimated by two methods; from maximum likelihood equations and regression on the linearized cumulative distribution function. Of these methods, estimates from linear regression are analytically the simplest to obtain. However, because of the logarithmic transformations involved, the lowest few concentrations may bias the regression parameters.

Because environmental data often contain non-detection points, we addressed the problem of estimating Weibull parameters for left-censored samples. As one possible solution, we proposed Weibull MLEs derived from the maximum likelihood equation for left-censored data. Another methods considered were regression on the linearized distribution function based on measured observations only, thereby ignoring the non-detects and substitution method. Tests on artificially left-censored random data taken from simulated Weibull distribution indicated that MLE and substitution methods gave smaller RMSE than linear regression method.

References

Cohen, A. C. (1959). Simplified estimators for the normal distribution when the samples are singly censored or truncated. *Technometrics* 1, 217-237.

El-Shaarawi, A. H. and Esterby, S. R. (1992). Replacement of censored observations by a constant: an evaluation. *Water Research* 26, 835-844.

El-Shaarawi, A. H. and Viveros, R. (1997). Inference about the mean in log-regression with environmental applications. *Environmetrics* 8, 569-582.

Gan, F. F., Koehler, K. J. and Thompson, J. C. (1991). Probability plots and distribution curves for assessing the fit of probability models. *The American Statistician* 45, 14-21.

Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring.* New York: Van Nostrand Reinhold.

Gilbert, R. O. and Kinnison, R. R. (1981). Statistical methods for estimating the mean and variance from radionuclide data sets containing negative, unreported or less-than values. *Health Physics* 40, 377-390.

Helsel, D. R. (1990). Less than obvious: statistical treatment of data below the detection limit. *Environmental Science and Technology* 24, 1766-1774.

Helsel, D. R. (2005). *Nondetects and Data Analysis: Statistics for Censored Environmental Data.* John Wiley and Sons, New York.

Holland, D. M. and Fitz-Simmons, T. (1982). Fitting statistical distributions to air quality data by the maximum likelihood method. *Atmospheric Environment* 16, 1071-1076.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis.* Englewood Cliffs, NJ: Prentice Hall.

Kroll, C. N. and Stedinger, J. R. (1996). Estimation of moments and quantiles using censored data. *Water Resour. Res.* 32(4), 1005-1012.

Stoline, M. R. (1991). An examination of the lognormal and Box and Cox family of Transformations in fitting environmental data. *Environmetrics* 2, 85-106.

Travis, C. C. and Land, M. L. (1990). Estimating the mean of data sets with nondetectable values. *Environmental Science and Technology* 24, 961-962.

Wild, P., Hordan, R., LePlay, A. and Vincent, R. (1996). Confidence intervals for probabilities of exceeding threshold limits with censored log-normal data. *Environmetrics* 7, 247-259.