JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION
VOL. 33, NO. 3 AMERICAN WATER RESOURCES ASSOCIATION JUNE 1997

ANALYZING CENSORED WATER QUALITY DATA USING A NON-PARAMETRIC APPROACH[1]

Nian She[2]

ABSTRACT: In the analysis of water quality data, samples with concentrations reported below the limit of detection (LOD) are referred to as Type I censored on the left. A variety of procedures have been proposed for estimating descriptive statistics from left-censored data. Usually, the estimation is carried out by either replacing the LOD with a constant between 0 and the LOD, or assuming the data follow a normal or lognormal distribution. In this paper, a simple transformation is proposed to convert multiple left-censored water quality data to right-censored data. The transformed cumulative distribution is similar to a survival function, and enables use of survival analysis techniques for left-censored data. In particular, the product limit method (Kaplan-Meier estimator) is applied to estimate descriptive statistics from the transformed data. The performance of the Kaplan-Meier estimator is compared with maximum likelihood, probability plotting, and substitution methods by Monte Carlo simulations. The Kaplan-Meier estimator performs as well as or better than these more familiar methods. Finally, the Kaplan-Meier estimator is used to analyze some priority pollutant data collected in sediment from the central basin of Puget Sound.
(KEY TERMS: limit of detection; censored data; Kaplan-Meier estimator; non-detects; product-limit estimator.)

INTRODUCTION

Water quality monitoring data are frequently reported as below the limit of detection (LOD). When measurements of pollutant concentration are less than the "method detection level" (MDL), it is reported as "not detected" (ND). The MDL is defined as the minimum concentration of a substance that can be measured and reported with 99 percent confidence that the analyte concentration is greater than zero and is determined from the analysis of a sample in a given matrix containing the analyte. The MDL depends on the analytical method used to analyze the specific pollutant. For each analytical method, the MDL is determined based on the confidence of analytical signal discerned from noise within the sample. Usually LOD is obtained by multiplying a factor by MDL. Sometimes the LOD is also called "report detection limit" (RDL). In this paper, LOD and RDL will be used interchangeably.

In statistical terminology, water quality data with below LOD results are referred to as "censored at LOD on the left," because the "left" or lower portion of the cumulative distribution cannot be observed. LOD is the largest concentration quantified in a sample. When only the largest values of an experiment can be observed, the data are said to be "censored from the left." By contrast, when only values smaller than a given value can be observed, the data are "censored from the right." Both left and right-censored data are classified into two types: Type I and Type II. Type I censored data has a random number of censored observations, and Type II censored data, a fixed number. Water quality data are Type I censored, because the proportion of samples that are below LOD cannot be controlled (Miller, 1981).

Many non-parametric methods have been used to analyze Type I censored data for biomedical applications, but few have been applied to environmental sciences. Applications of these non-parametric methods to censored water quality data are rare because most of the statistical theorems and software packages were developed for right-censored failure time or survival data only. Environmental researchers have focused on parametric methods for censored data, which assume the underlying distribution is known. Gilliom and Helsel (1986) compared the performance

[1]Paper No. 95173 of the *Journal of the American Water Resources Association* (formerly *Water Resources Bulletin*). **Discussions are open until December 1, 1997.**

[2]Environmental Analyst, Drainage and Wastewater Utilities, Engineering Department, City of Seattle, 660 Dexter Horton Bldg., 710 Second Avenue, Seattle, Washington 98104.

of the maximum likelihood, probability plotting and substitution methods for a wide range of "parent distributions" with single LOD and tested the robustness of the methods to departure from the normality. Helsel and Cohn (1988) extended the work of Gilliom and Helsel (1986) to multiple LODs. El-Shaarawi and Naderi (1991) and El-Shaarawi and Esterby (1992) discussed the maximum likelihood method for estimating mean and standard deviation by replacing LOD with a constant between 0 and LOD, assuming that the underlying distribution of the data is either normal or lognormal. Newman (1992) and Loaiciga *et al.* (1992) developed a maximum likelihood method to estimate parameters for various distributions, including uniform, exponential and Weibull distributions. Shumway *et al.* (1989) considered using the Box and Cox (1964) transformation to normalize the data.

The only non-parametric method commonly used for water quality research is the substitution method, which replaces ND with a random number between 0 and LOD, usually one half of LOD. Gleit (1985) evaluated the performance of replacing ND with a random number between 0 and LOD for normally distributed water quality data by using Monte Carlo simulations. Several authors (Gilliom and Helsel, 1986; Helsel and Cohn, 1988; Helsel, 1990) showed that the substitution method did not perform as well as the maximum likelihood and log-probability plotting methods for several specified distributions lumped together. These methods were evaluated based on the ability to replicate true population statistics. Departures from true values are measured by root-mean-square error (rmse), which combines bias and lack of precision. Methods with lower rmse are considered better. Their results are difficult to interpret because only a lumped rmse over all distributions is given. No information about the performance of each method corresponding to each individual distribution is provided. Moreover, most left-censored water quality data have very complex distributions, which are usually unknown. Descriptive statistics, such as mean and standard deviation, estimated from the sample will be highly inaccurate if the underlying distribution departs far from the assumed distribution. Thus there is a need for non-parametric methods to analyze left-censored water quality data.

This paper demonstrates a more general and robust non-parametric method developed from survival analysis which may be used to analyze multiple left-censored water quality data. A "non-parametric" method does not depend on the underlying distribution of the data, and is fairly robust when the distribution departs from normality. To make survival analysis methods directly applicable to left-censored water quality data, a simple transformation is used to covert the left-censored data to right-censored data, given that the transformation from time to space-based uncertainty is appropriate. In particular, the product limit method (Kaplan-Meier estimator) from survival analysis is applied to the transformed data to estimate the mean and standard deviation. Once these parameters are estimated, they are transformed back to the original scale without loss of any information. The following sections describe how to transform left-censored water quality data to right-censored data and how to use the Kaplan-Meier estimator. The performance of the Kaplan-Meier estimator is then compared with other more familiar methods, such as the maximum likelihood, probability plotting, and substitution methods using Monte Carlo simulations. Finally, the Kaplan-Meier estimator is used to analyze organic priority pollutant data from 20 monitoring stations in the Central Basin of Puget Sound.

DATA TRANSFORMATION

Suppose that n water quality samples are collected and analyzed. When the concentration of a pollutant can not be measured precisely or is below the MDL, the observed concentration is said to be censored on the left at LOD. If d_c denotes the value of LOD, then dc may be a fixed number, or may vary from sample to sample depending upon the magnitude of noise within the samples. Let X_0 denote true concentration and assume that the n samples are independent identically distributed (iid) and are randomly taken from an unknown distribution. Then what is observed is either a precise concentration or an LOD. The observation can be written as $X = \max(X_0, d_c)$. Therefore, the n observations can be written as n pairs of random variables $\{(X_1, (\delta_1), (X_2, (\delta_2), (, (X_n, (\delta_n)\}$, where

$$\delta_i = \begin{cases} 1 \text{ if concentration is observed} \\ 0 \text{ if LOD is observed} \end{cases}, 1 = 1, 2, \ldots, N \quad (1)$$

For example, a pair (23, 1) indicates that a concentration of 23 units was observed, while (134, 0) means that the detection limit of 134 units was observed. Note that in this example the limit of detection is greater than the concentration precisely measured. This is a typical phenomenon in many water quality monitoring samples as well as in sediment samples.

Using definition (Equation 1), an observation (X_i, δ_i) can be thought of as randomly selected from an unknown distribution F(x):

$$F(x) = P(X \le x) = P(-X \ge -x)$$

$$= P(A - X \ge A - x) = P(Y \ge y) \quad (2)$$

where Y = A - X, y = A - x, and A is a large positive number.

Let S(y) = P(Y ≥ y), where S(y) is called the survival function (Miller, 1981). Then, it can be seen from Equation (2) that F(x) = S(y), and the distribution of the left-censored random variable X is equal to the survival function of the right-censored random variable Y. So, the estimation of the descriptive statistics from the left-censored data is equivalent to estimate the descriptive statistics from the right-censored data by a simple transformation Y = A - X. Visually, the survival function S(y) is obtained by flipping the distribution function F(x) 180 degrees to the left and then shifting it A units to the right. The transformed data set is denoted by $\{(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)\}$.

KAPLAN-MEIER ESTIMATOR

Let $(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)$ be the ordered pairs such that $Y_1 \le Y_2 \le, \ldots \le Y_n$ and δ_n is the value of δ associated with Y_i. Assuming no ties for now, let $p_i = P\{Y > y_i \mid Y > y_{i-1}\}$, which is the conditional probability that a measurement after the transformation will be greater than the i^{th} ordered observation given that it is greater than the $(i-1)^{th}$ ordered observation.

Based on Kalbfleisch and Prentice (1980), the maximum likelihood estimate for $\hat{p}_i$ is

$$\hat{p}_i = \begin{cases} 1 - \dfrac{1}{m_i} & \text{if } \delta_i = 1 \text{ (concentration is observed)} \\ 1 & \text{if } \delta_i = 0 \text{ (LOD is observed)} \end{cases}$$

where m_i is the "risk set" or number of transformed observations greater than Y_i. The product limit estimate of the survival function (Kaplan-Meier estimator) when no ties are present is (Miller, 1981),

$$\hat{S}(y) = \prod_{y_i \le y} \hat{p}_i = \prod_{y_i \le y} \left(1 - \frac{1}{m_i}\right)^{\delta_i} . \tag{3}$$

Now suppose there are tied measurements, and only r distinct measurements $Y_1' < Y_2 < \ldots < Y_r'$ are recorded. Let n_i = number of transformed observations $\ge Y_i'$, d_i = number of uncensored measurements occurring at Y_i', and

$$\delta_i' = \begin{cases} 1 \text{ if at least one transformed observation at } y_i' \text{ is uncensored} \\ 0 \text{ if all observations at } y_i' \text{ are censored} \end{cases}$$

The Kaplan-Meier estimator with ties is then

$$\hat{S}(y) = \prod_{y_i' \le y} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i'} \tag{4}$$

Note that if the last transformed observation y_n is censored, then for $\hat{S}(t)$ as defined above

$$\lim_{y \to \infty} \hat{S}(y) > 0.$$

Sometimes, it is preferable to redefine $\hat{S}(y) = 0$ for $y \ge y_n$ or to treat it as undefined for $y \ge y_n$ if $\delta_n = 0$ (Miller, 1981). Also note that the Kaplan-Meier estimator changes only at non-censored value of y. Notice that when there are no censored data, the Kaplan-Meier estimator is reduced to 1 minus the experical cumulative distribution.

The asymptotic variance of $\hat{S}(t)$ when no ties are present is given by Greenwood's formula (Miller, 1981):

$$\hat{\mathrm{Var}}(\hat{S}(y)) = \hat{S}^2(y) \sum_{y_i \le y} \frac{\delta_i}{(n-i)(n-i+1)} \tag{5}$$

with ties the variance becomes

$$\hat{\mathrm{Var}}(\hat{S}(y)) = \hat{S}^2(y) \sum_{y_i' \le y} \frac{\delta_i' d_i}{n_i(n_i - d_i)} \tag{6}$$

Using this estimated variance, confidence intervals for $\hat{S}(y)$ can be obtained via the usual normal approximation methods since it can be shown that $\hat{S}(y)$ is asymptotically normally distributed (Miller, 1981).

The Kaplan-Meier estimator can be used to estimate descriptive statistics. The estimated mean, median and asymptotic variance of the transformed data are given respectively by

$$\hat{\mu} = \int_0^\infty \hat{S}(y)dy \tag{7}$$

$$\hat{\xi}_{\frac{1}{2}} = \hat{S}^{-1}\left(\frac{1}{2}\right) \tag{8}$$

$$\text{A var}(\hat{\mu}) = \sum_{i=1}^{r} \left(\int_{y_i}^\infty \hat{S}(y)dy\right)^2 \frac{d_i}{n_i(n_i - d_i)} \tag{9}$$

where $\hat{S}^{-1}$ is the inverse of $\hat{S}$. Since $\hat{S}$ is a step function, $\hat{S}^{-1}(p)$ may not have a unique solution for $0 \le p \le 1$. Petersen (1983) defines $\hat{S}^{-1}(p) = \inf\{t: \hat{S}(y) \ge p\}$, where inf indicates the infimum. The median, for example, is the infimum of the set of values for y for which $S(y) \ge 0.5$. Since S(y) is left continuous, the median will be the value of y at which S(y) jumps from being greater than (or equal to) 0.5 to be less than 0.5. In other literature (Miller, 1981), $\hat{S}^{-1}\left(\frac{1}{2}\right)$ is defined as the midpoint of the interval of the solutions.

In Equations (7) and (9) the integrals are Lebesgue integrals. For the step functions, the integrals can be replaced by summations. For mathematical convenience, the Lebesgue integral will be used in the rest of the paper.

COMPARISONS OF KAPLAN-MEIER ESTIMATOR WITH OTHER FAMILIAR METHODS

The Kaplan-Meier estimator is one of the most commonly used non-parametric methods in survival analysis. Detailed descriptions and applications can be found in standard textbooks (Miller, 1981; Kalbfleisch and Prentice, 1980). The question now remains as to how well the Kaplan-Meier estimator performs compared with other familiar methods of analyzing left-censored water quality data? To answer this, the Monte Carlo simulations are used to compare the performance of the Kaplan-Meier estimator (KM) with the maximum likelihood (ML), probability-plotting (PL) and substitution (HA) methods for the left censored data generated from lognormal and gamma distributions, which are distributions frequently used to represent water quality data. The relative performance of ML, PL and HA were discussed by Gilliom and Helsel (1986), Helsel and Cohn (1988) and Helsel (1990).

Let $\theta = \theta(F)$ be a true mean, or a true standard deviation of the simulated data. Note that θ is dependent on the underlying distribution F. For the Kaplan-Meier estimator, we first estimate $\gamma = \gamma(F)$ by $\hat{\gamma} = \gamma(\hat{F})$, where γ is the mean or standard deviation of the transformed data and $\hat{F}$ is the estimated cumulative distribution function (cdf). Having calculated $\hat{\gamma}$, the untransformed estimate of the mean is then given by $\hat{\theta} = A - \hat{\gamma}$. The variance of the transformed and untransformed data are the same because var(A-X) = Var(X).

The performance of each method is evaluated using its root-mean-square error (rmse), defined as

$$\eta(F) = \left[\sum_{j=1}^{N}\left(\frac{\hat{\theta}-\theta}{\theta}\right)^2 \Big/ N\right]^{\frac{1}{2}} \tag{10}$$

and its relative bias given by

$$\beta(F) = \left[\sum_{j=1}^{N}\left(\frac{\hat{\theta}-\theta}{\theta}\right) \Big/ N\right] \tag{11}$$

where N is the number of simulations of each method. η and β are both dependent on the underlying distribution F.

Each method was tested on samples generated from lognormal and gamma distributions with mean set to 1.0 and coefficients of variation CV = 0.25, 0.5, 1.0 and 2.0 (Figure 1). One thousand repetitions of sample size n = 21 were generated from each distribution. Three percentiles of each distribution were randomly chosen as LOD, and were denoted as q_1, q_2, and q_3. For computational convenience, q_i, i = 1, 2, 3 were chosen only from the 10th, 20th, . . ., 80th percentile. One third of each sample was randomly assigned to q_1, q_2, and q_3 in the following way: any value falling below q_1 was assigned to q_1; any value falling between q_1 and q_2 was assigned to q_2; and any value falling above q_2 was assigned to q_3.

Tables 1 and 2 compare the performance of KM with ML, PL and HA methods. For the lognormal distribution that HA has the highest rmse and relative bias when CV = 0.25 and 0.5, but it has the smallest rmse and relative bias when CV = 1.0 and 2.0. The rmse and relative bias for ML and KM are quite close. For the gamma distribution, KM has the smallest rmse and relative bias when CV = 0.25 and 0.5, and both KM and HA have smaller rmse and relative bias than ML and PL when CV = 1.0 and 2.0. This indicates that the performance of KM is almost as good as ML for the lognormal distributions, and is as good as HA for the gamma distributions, but is better than ML and PL for the gamma distributions.

The comparisons show that the performance of the Kaplan-Meier estimator is as good as or better than these familiar methods under both lognormal and non-normal assumptions. It is found that ML and PL are sensitive to underlying distributions; HA is sensitive to the shape of the distributions; but KM is relatively robust. Though the HA sometimes works a little bit better than KM for the extremely right skewed distributions, replacing each LOD by a constant alters the underlying variation attributable to the sample, and has no statistical theoretical basis. By contrast, KM is well justified in theory and in

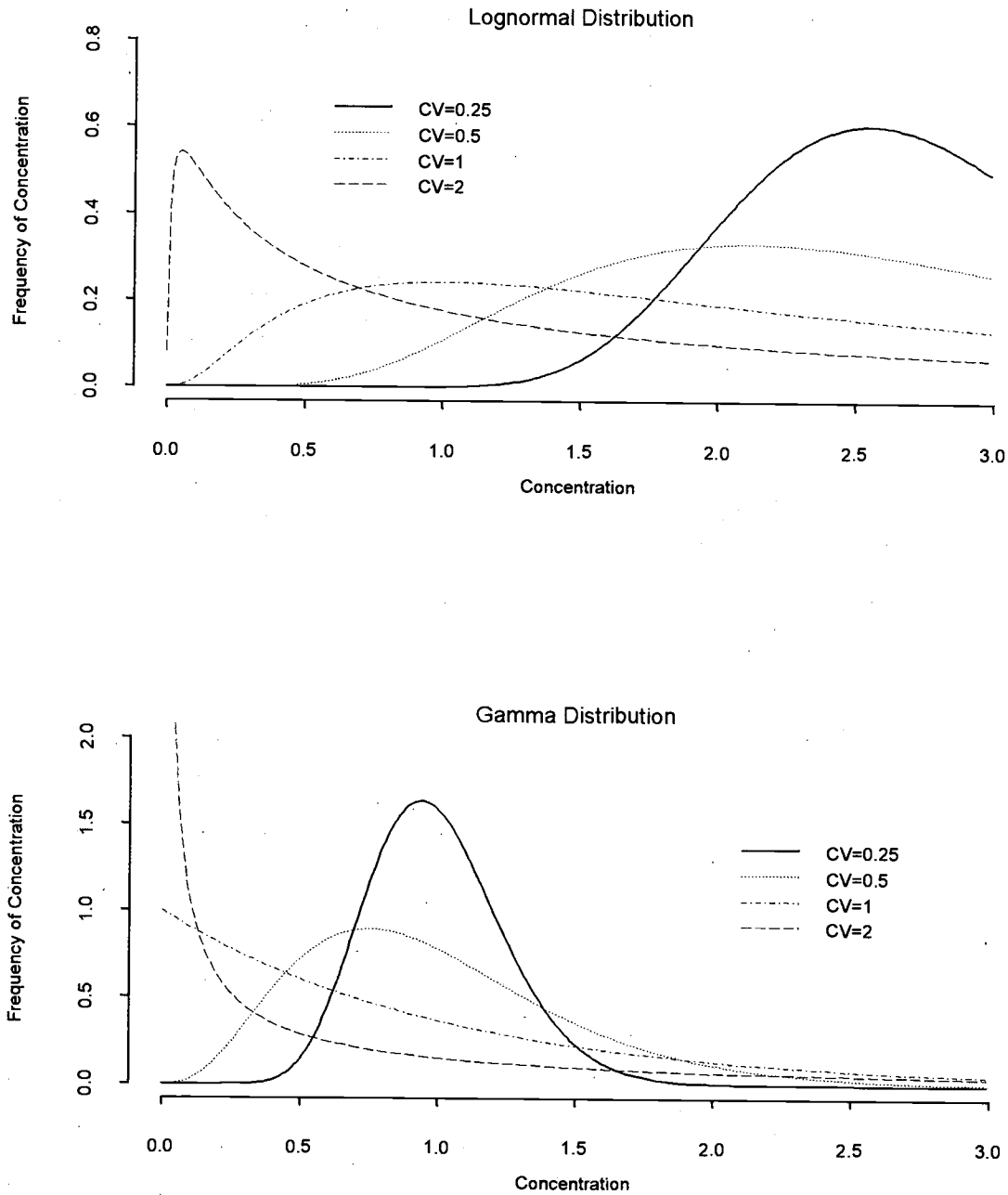

Figure 1. Probability Density Functions Used in Simulation.

biomedical applications. Thus, in practice it should be an attractive alternative to HA.

The Kaplan-Meier estimator has at least two advantages over other familiar methods. First, KM is robust when the distributions depart from normality, because actual observed data are used rather than a fitted distribution above the LOD. Second, the estimation of descriptive statistics can be directly performed, thereby avoiding bias induced from the log or other non-linear transformations. One disadvantage of KM is that it is difficult to obtain statistical inference for some parameters. For instance, the standard error of the median is equal to the asymptotic variance of $\hat{S}(y)$ divided by the square of an unknown density function f, i.e., $\text{se(median)} = \frac{A\,\text{var}(\hat{S})}{f^2}$. Usually, it is calculated by using complicated normal approximation, and requires a large sample size.

TABLE 1. The Standard Error of Estimation Methods.

Method	CV = 0.25 Mean	CV = 0.25 Standard Deviation	CV = 0.50 Mean	CV = 0.50 Standard Deviation	CV = 1.00 Mean	CV = 1.00 Standard Deviation	CV = 2.00 Mean	CV = 2.00 Standard Deviation
Lognormal Distribution								
ML	0.080	0.219	0.160	0.212	0.313	0.204	0.614	0.216
HA	0.206	0.695	0.201	0.264	0.244	0.150	0.441	0.164
KM	0.081	0.201	0.163	0.215	0.319	0.256	0.635	0.276
PL	0.097	0.254	0.194	0.247	0.278	0.241	0.744	0.252
Gamma Distribution								
ML	0.079	0.119	0.131	0.213	0.229	0.160	0.568	0.809
HA	0.145	0.198	0.147	0.245	0.043	0.028	0.444	0.661
KM	0.076	0.113	0.135	0.223	0.101	0.002	0.414	0.651
PL	0.093	0.132	0.146	0.243	0.241	0.157	0.478	0.730

TABLE 2. Bias of Estimation Methods.

Method	CV = 0.25 Mean	CV = 0.25 Standard Deviation	CV = 0.50 Mean	CV = 0.50 Standard Deviation	CV = 1.00 Mean	CV = 1.00 Standard Deviation	CV = 2.00 Mean	CV = 2.00 Standard Deviation
Lognormal Distribution								
ML	-0.055	0.104	-0.111	0.095	-0.216	0.089	-0.415	0.101
HA	-0.198	0.679	-0.166	0.182	-0.100	0.012	0.056	-0.027
KM	-0.054	0.141	-0.110	0.140	-0.214	0.126	-0.416	0.144
PL	-0.074	0.148	-0.151	0.141	-0.293	0.132	-0.570	0.144
Gamma Distribution								
ML	-0.056	0.120	-0.115	0.142	-0.229	0.160	-0.487	0.097
HA	-0.137	0.369	-0.126	0.254	-0.043	-0.028	0.014	-0.113
KM	-0.048	0.101	-0.081	0.065	-0.101	0.002	-0.075	-0.115
PL	-0.073	0.156	-0.139	0.170	-0.241	0.157	-0.367	0.040

A WORKING EXAMPLE

In this section, the methods discussed in the previous sections are used to analyze water quality monitoring data collected by the King County Department of Metropolitan Services (Metro) from 1988 to 1990 in the central basin of the Puget Sound, Washington. This monitoring program was designed to achieve two goals: (1) to demonstrate that the wastewater discharge from Metro's Renton secondary treatment plant (which has the deepest effluent outfall in the world) complies with the National Pollutant Discharge Elimination System (NPDES) permits; and, (2) to detect any environmental impact due to the secondary effluent discharge.

Twenty benthic sampling stations were chosen as compliance monitoring sites. The selection of these stations was based on effluent plume behavior models which predict that effluent, discharged at the outfall, will be trapped between 100 and 150 meters below the surface. The portion of the benthic community potentially impacted by the trapped plume was defined by the area of the seafloor between the 100 and 150 meter contours in the vicinity of Elliott Bay. These stations were located in the area of highest probable contact with the discharged effluent and the seabed.

Sediment samples were collected annually from these twenty stations in the summer. Since the stations were so deep, the seasonal effects may be ignored. The spatial correlation among stations was not a primary concern in this study because the data were only compared with the baseline data in the same locations. However, if the post-discharge conditions were found to be significantly different from baseline conditions, a "red flag event" would be

declared. This procedure activates an archive retrieval process, which can yield information on possible spatial gradients of pollutants, and thereby indicate whether or not the outfall discharge is the causative factor for the "red flag event". The temporal rate of environmental change may be assessed by evaluating relevant archived samples collected from these same sites during previous monitoring periods.

Pyrene is one of the 76 priority pollutants listed by Metro in its compliance monitoring program. Table 3 lists Pyrene concentrations collected by Metro from these 20 compliance monitoring stations from 1988 to 1990. The negative sign on the right hand side of the number indicates that the observation is below the limit of detection (LOD), or left-censored. Concentrations of pyrene vary from 28 (μg/l to 2982 μg/l, and 11 data points (about 20 percent of the data) are below the LOD, which varies depending on sample size, percentage of solids and dilution factors (Table 3). Clearly, the data are multiply censored.

TABLE 3. Pyrene Concentration (μg/l) Collected from 20 Monitoring Stations in the Central Basin of Puget Sound from 1988-1990.

Reported Data	Transformed Data	Reported Data	Transformed Data
28-	2972+	105	2895
31	2969	107	2893
32	2968	110	2890
34	2966	111	2889
35-	2965+	117-	2883+
35-	2965+	119	2881
40	2960	119	2881
47	2953	122-	2878+
48	2952	122	2878
58-	2942+	132	2868
59	2941	133	2867
63	2937	133	2867
64	2936	138	2862
64	2936	163-	2837+
67	2933	163-	2837+
67	2933	163-	2837+
67	2933	163	2837
72	2928	174-	2826+
73	2927	187	2813+
84	2916	190	2810
86	2914	222	2778
86-	2914+	238	2762
87	2913	273	2727
94	2906	289	2711
98	2902	306	2694
100	2900	333	2667
103	2897	459	2541
103	2897	2982	18

Notes: - sign indicates that the data is censored at left.
+ sign indicates that the data is censored at right.

To use the Kaplan-Meier estimator the data listed in Table 3 are transformed to right-censored data by multiplying each value by -1 and then adding a value of 3000 determined by Equation (2). The estimated survival function, $\hat{S}$, is plotted in Figure 2, and the estimated cumulative distribution function, $\hat{F}$, is shown in Figure 3. Note that $\hat{F}$ is constructed by flipping $\hat{S}$ 180 degrees to the left first, then shifting it 3000 units to the right. The mean and standard deviation are estimated by using Equations (7) and (9). The estimated mean is then transformed back to original scale by subtracting 3000. Recall that the standard deviations of transformed and untransformed data are the same.

For comparisons, the data were also analyzed using ML, HA and PL. Table 4 lists the estimated mean, standard deviation and standard errors. Observe that the four methods agree closely when pyrene concentrations were transformed into the log-scale. Without making the log transformation, a large difference would be expected between parametric and non-parametric methods because the distribution of the data is heavily skewed to the right with a long tail, while the mean and standard deviation are very sensitive to outliers and are dependent on the assumed distribution of the data. KM has the smallest standard error, followed by HA and PL. It is not surprising that ML has the largest standard error because of the large outlier.

Differences among the four methods are largely attributable to the large outlier. If the extreme concentration of 2982 μg/l is deleted, then the estimated means from KM, ML, HA and PL are 111.94, 105.97, 111.07 and 109.16, respectively. By plotting the uncensored data against the corresponding normal scores (QQ-plot), it can be seen that the data appear to fit a lognormal distribution (Figure 4). But deleting outliers to force the data to fit an expected distribution is not usually a good idea. For example, unusually low ozone concentrations on the Antarctic ozone hole had been collected for approximately 10 years prior to its actual discovery; however, these data were automatically discarded as outliers in routine data checking process (Helsel and Hirsch, 1992). Outliers may be the most important points in the data set, and should be investigated further.

SUMMARY

The Kaplan-Meier estimator, though widely used in survival analysis, formally applies only if the data is randomly censored to the right. It was shown that with a simple transformation that converts left-censored data to right-censored data, the Kaplan-

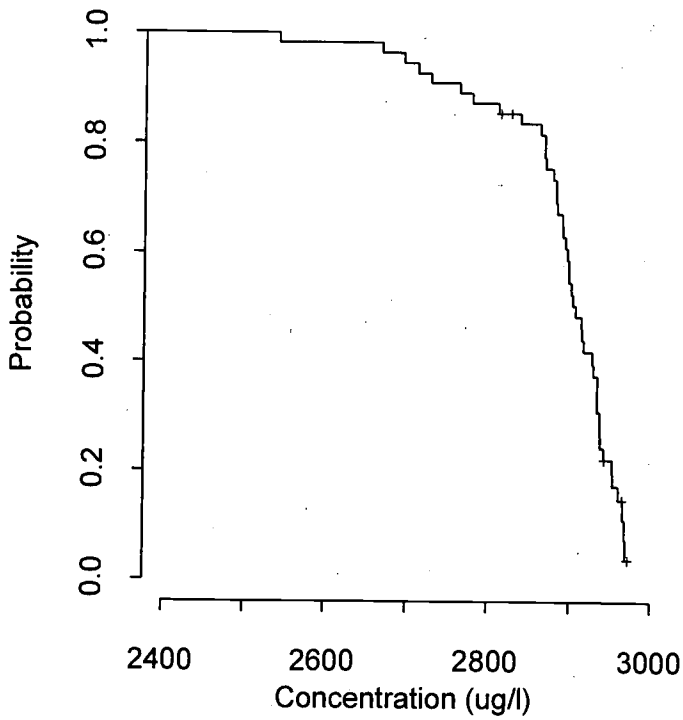

Figure 2. Kaplan-Meier Estimator of Transformed Data.

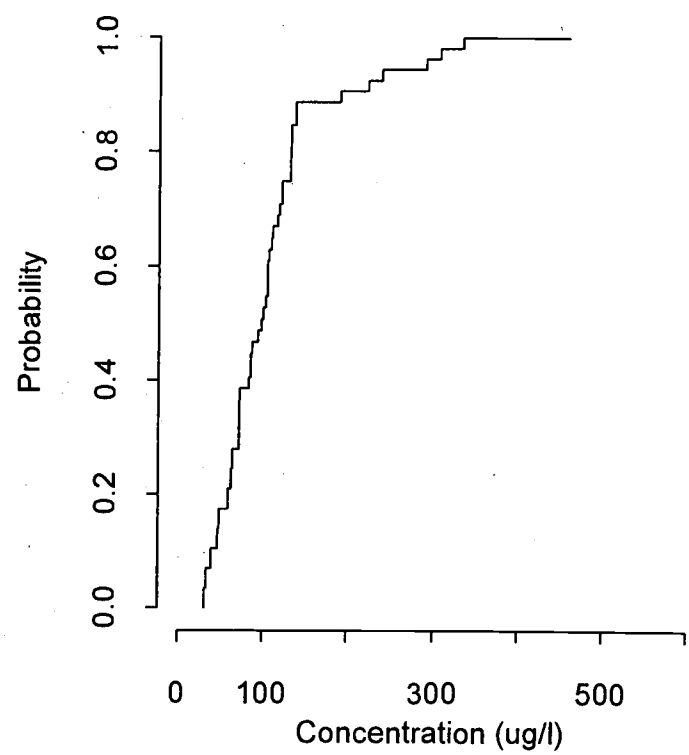

Figure 3. Estimated Cumulative Distribution of Pyrene.

TABLE 4. Mean and Standard Error of Pyrene Concentration.

Method	Log Scale		Original Scale	
	Mean	Standard Error	Mean	Standard Error
ML	4.514448	0.86431	103.7995	438.5382
HA	4.549426	0.861949	162.3393	393.0764
KM	4.523628	0.830923	163.1926	389.5784
PL	4.552082	0.806234	151.8903	399.7561

ML = Maximum likelihood estimate.
HA = Substitution method.
KM = Kaplan-Meier estimator.
PL = Probability plotting method.

Meier estimator can be applied to analyze left-censored water quality data. It was also shown that the Kaplan-Meier estimator performs as well as or better than the maximum likelihood, probability plotting and substitution methods for the left-censored data generated from the specified distributions. This makes the Kaplan-Meier estimator an attractive alternative to these familiar methods because it is non-parametric and quite robust when the distribution departs from normality. In many cases, assuming the water quality data follows a specified distribution is inadequate and may yield misleading results.

The key for the Kaplan-Meier estimator to work under any distribution is that the censoring must be "random." That is, the probability that the measurement of an object is censored can not depend on the value of the censored variable. Thus, the censoring mechanisms of each water quality data set must be understood individually to judge whether the censoring is, or is not, likely to be random. If we consider the field of survival analysis from a broader perspective, we note a number of deficiencies with respect to censored statistical problems in water quality studies. Most importantly, survival analysis assumes the upper limits in a given experiment are precisely measured, while in water quality studies, LOD are frequently obtained by multiplying a factor to MDL. The factor is more or less subjective depending on the noise level of the sample.

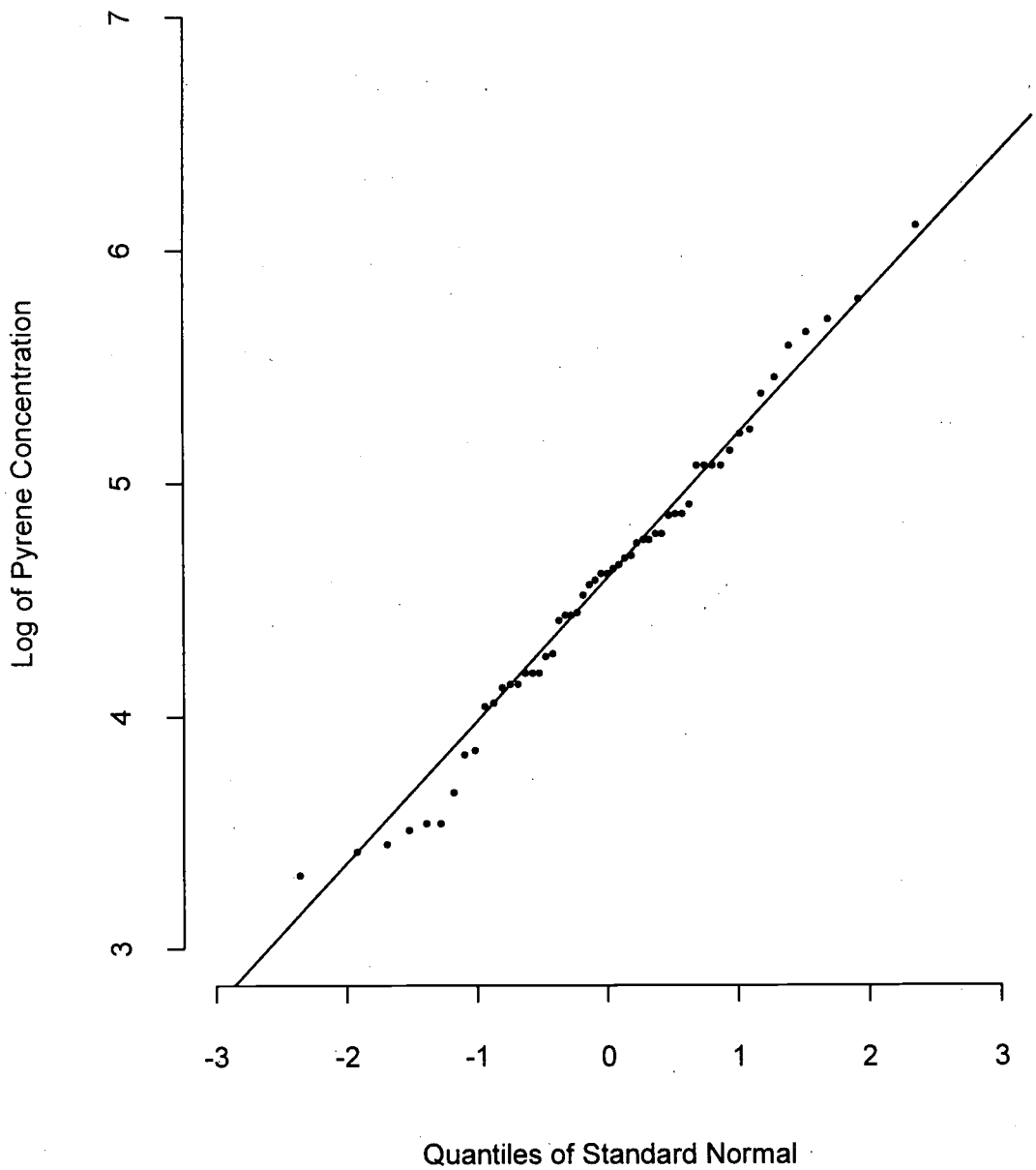

Figure 4. QQ-Plot of Pyrene Concentration in Log-Scale.

Finally, it should be noted that making a simple transformation to convert left-censored data to right-censored data is just for computational convenience. Most theorems and algorithms applied to right-censored data can be similarly derived for left-censored data. However, to estimate location parameters, it is much easier to use existing software packages. Once the estimated parameters are obtained from transformed data, they can be transformed back to original scales without adding any bias from the log or other non-linear transformations.

This paper demonstrated that one of the most popular non-parametric methods in survival analysis is applicable to water quality data with ND results. Numerous methods in survival analysis, such as K sample comparisons, regression and goodness of fit, can be easily applied to left-censored water quality data. It is hoped that the non-parametric approach to water quality data will aid environmental statisticians in applying contemporary statistical methods to water quality studies.

ACKNOWLEDGMENTS

The author expresses sincere appreciation to Dr. Jonathan Frodge, Water Quality Planner, Water Pollution Control Division, Department of Natural Resources, Metropolitan King County; and Dr. Richard Henrichson, School of Fisheries, University of Washington for their valuable comments and suggestions.

LITERATURE CITED

Box, G. E. P. and D. R. Cox, 1964. An Analysis of Transformations (with discussion). Journal of the Royal Statistical Society, Ser. B, 39:211-252.

El-Shaarawi A. H. and S. R. Esterby, 1992. Replacement of Censored Observations by a Constant: An Evaluation. Water Resources. 26(6):835-844.

El-Shaarawi A. H. and A. Naderi, 1991. Statistical Inference from Multiply Censored Environmental Data. Environmental Monitoring and Assessment 17:339-347.

Gilliom, R. J. and D. R. Helsel, 1986. Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 1. Estimation Techniques. Water Resources Research 22:135-146.

Gleit, A., 1985. Estimation for Small Normal Data Sets with Detection Limits. Environmental Science and Technology 19:1206-1213.

Helsel, D. R., 1990. Less Than Obvious, Statistical Treatment of Data Below the Detection Limit. Environ. Sci. Technol. 249(12): 1767-1774.

Helsel, D. R. and T. A. Cohn, 1988. Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. Water Resources Research 24:1997-2004.

Helsel, D. R. and R. M. Hirsch, 1992. Statistical Methods in Water Resources. Elsevier. Amsterdam, The Netherlands.

Kalbfleisch, J. D and R. L. Prentice, 1980. The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York, New York.

Loaiciga, H. A., J. Michaelsen, and P. Hudak, 1992. Truncated Distributions in Hydrologic Analysis. Water Resources Bulletin 28(5):853-863.

Miller, R. G. J., 1981. Survival Analysis. John Wiley and Sons, Inc. New York, New York.

Newman, M. C., 1992. Enhancing Toxicity Data Interpretation and Prediction of Ecological Risk with Survival Prediction of Survival Time Modeling. Proceedings of Society of Environmental Toxicology and Chemistry, p. 121.

Petersen, A. V., 1983. Kaplan-Meier Estimator. *In:* Encyclopedia of Statistical Sciences, S. Koltz and N. L. Johnson (Editors). John Wiley and Sons, New York, New York, Vol. 4, pp. 346-352.

Shumway, R. H., A. S. Azari, and P. Johnson, 1989. Estimating Mean Concentrations Under Transformation for Environmental Data with Detection Limits. Technolometrics 31(3):347-356.