# **Appendix B:** Selection Criteria for Screening Data Used in Source Characterization Report

# **Appendix B:** Selection Criteria for Screening Data Used in Source Characterization Report

The first purpose of this section is to identify data that can be used to characterize contaminant concentrations. To use these data with confidence requires that the reported sample concentrations are representative of the environmental media and location from which samples were taken. Concentration data can be used to track down sources of contaminants in the environment. Concentration data also can be used for the evaluation of contaminants in relation to standards or criteria for the protection of aquatic life and human health. Some degree of confidence in the data is needed for either of these purposes. Concentration data along with data such as that found in the TRI, NPRI, PCS, and SRDS can be used to corroborate environmental release information. This type of analysis relies on the weight of evidence from multiple data sources.

The other purpose of this section is to identify data that can be used for the computation of loads. Identification of the contaminant loading sources to Lake Erie can provide information on when, where, and how contaminants enter Lake Erie and the comparative magnitude of loading from streams and point sources relative to atmospheric deposition.

### Minimum Criteria for Estimating Concentrations and Loads

The objectives of this report are to characterize concentrations and loads. Available data with which to achieve these objectives were collected by many agencies for various purposes. To determine which data are suitable to characterize concentrations and loads requires a screening procedure and selection criteria. The basis for the data-selection criteria is a review of published literature on techniques for analyzing environmental data and techniques for estimating fluvial loads. The objective of screening criteria is to extract as much suitable data as possible from the available data sets so that concentrations can be characterized and loads can be computed with confidence.



Minimizing errors and maximizing confidence in the information presented in this report is a primary goal because management actions might arise from conclusions drawn from the data. There can be many sources of error in any reported data. Variability in an estimate of a concentration or load is dependent on the sampling errors and nonsampling errors in the data. Nonsampling errors can be random or nonrandom. Random nonsampling errors tend to cancel each other out in large data sets (Iman and Conover, 1983) and so will not be considered a serious problem for the data sets discussed in this report.

Biases (nonrandom errors) in the data may not cancel each other out and elimination of bias is important to data quality. Biases can be minimized by the use of selection criteria that provide the analyst with only data applicable to the purposes of the data analysis. Sampling error is the other type of error that is an important consideration. Sampling theory dictates that the magnitude of the error in any data is inversely proportional to the square root of the number of samples (Richards, 1999, in press). To reduce this error by one-half requires that four times as many samples be collected. This knowledge translates into certain minimum sample sizes for data sets intended to be used to characterize contaminant concentrations and compute loads. The questions are 1) how many samples are enough, and 2) how much confidence in the sample estimates is desired or needed?

### Selection Criteria

One consideration is how well the samples represent the environment from which they were obtained such as point sources, the lake, connecting channels, tributaries, sediments, fish, or airshed. Even for simple descriptions of concentration and loading data, it is important that the samples collected represent the range of environmental conditions. For example, the concentrations of contaminants that are primarily delivered during runoff and high streamflows will be underestimated when samples are collected only during low or moderate streamflows. In much the same way, contaminant concentration in rain is dependent on

rainfall volume. Samples collected from streams, lakes, and the atmosphere at daily, weekly, monthly, or seasonal frequencies were deemed suitable and were included in data analysis for this report. Only where data are reported for a representative number of locations across the range of environmental conditions, were they deemed suitable for lakewide assessment purposes.

For aquatic sediments, single or multiple surficial sediment samples were deemed to be most representative of recently deposited sediments and associated contaminants. Another consideration is the period of record of data collection. Only water quality collected from October 1, 1985, to September 30, 1996, were inventoried for selected contaminants. The distribution of data sites is also a consideration.

### Minimum Criteria for Concentrations

The distribution of contaminant concentrations in hydrologic environments can be highly variable. Estimates of the mean, median, and range of concentrations cannot be described adequately with very small sample sizes (Helsel and Hirsch, 1995). Summary statistics used for this report are measures of center of the data (median or mean), the variability of the data (variance and standard deviation), the symmetry of the data distribution (kurtosis), and estimates of data quantiles and extremes (minimum, maximum or some large or small percentiles) (Helsel and Hirsch, 1995). For purposes of this report, concentration data sets with a sample size of at least 10 in which no sample results are reported below the limits of detection are deemed suitable for the description of contaminant concentrations. A sample size of 10 provides sufficient information for computation of median, mean, estimates of variability, and percentiles of the distribution.

A further complication is that the concentrations of certain contaminants are often reported as being censored or "below the detection or reporting limit." Censored data may present an interpretation problem. For example, censored data may be of limited use for evaluating the presence or absence of a contaminant if the reporting limit is higher than an environmentally relevant concentration. Concentration and (or) loading data can be used for evaluating a discharge or permit limit. To a lesser degree, concentration data can be used for evaluating compliance with a standard or criteria for the protection of aquatic life or human health. Data censoring is considered severe at the level of 50 percent or more (Helsel and Hirsch, 1995). At censoring levels greater than 50 percent, the median concentration, for example, may have to be estimated because it is not a detected value.



Statistical techniques that substitute values for censored data can be used to overcome the detection limit problem. The MLE (Maximum Likelihood Estimate; Cohen, 1959) is one technique that substitutes values for censored data based on what is known about the distribution of the data reported above the detection limit and the percentage of data below the detection limit. The MLE is a favored method to compute the median and other percentiles of a data set because it is less biased compared to simpler techniques that substitute zero, one-half, or the detection limit value for censored data (Helsel and Hirsch, 1995). The MLE works best with sample sizes greater than 25 (Helsel and Hirsch, 1995). If the MLE is used with lognormally distributed data, estimates of the mean and standard deviation may require some adjustment (Gilliom and Helsel, 1986). For purposes of this report, the MLE is the desired method for addressing censored data.

Data sets with censored data were judged to be suitable for the computation of statistical summaries if the detection frequency is at least 50 percent for sample sizes of 25 to 49, and at least 25 percent for sample sizes of 50 or more (Gleit, 1985). Even at sample sizes of 50, the estimated mean and standard deviation can be biased by 50 to 100 percent when there is a low percentage of detected values (Helsel and Hirsch, 1995).

The minimum criteria to characterize contaminant concentrations from point sources can differ somewhat from nonpoint sources such as tributaries. Tributaries are predominately influenced by event-based phenomena and hence the need for more stringent screening criteria to avoid bias. Point sources, on the other hand, are process-based and hence are characterized by relatively constant flows and concentrations. The minimum number of observations needed to characterize concentrations of contaminants discharged from point sources remains 10. However, if data are censored (reported below detection), 25 to 50 percent or more of the observations should be reported above the detection limit. However,

with small data sets, the computation of the MLE can produce unexpected results. Therefore, an MLE based on 50% or more observations above the detection limit would still be preferable.

### Minimum Criteria for Loads

Contaminant loads are computed using two types of data: 1) concentration data and 2) ancillary data such as streamflow, effluent discharge, rainfall, or dryfall. A load is a measure of the rate of transport of a known mass of a contaminant expressed in kilograms or tons per unit of time, either per day or per year. The most desirable situation for computing annual loads is if samples and measurements are taken concurrently each day. Daily loads are computed and summed for the year. The availability of daily values for computation of annual loads is uncommon because of funding constraints on monitoring programs. In the absence of daily values for loads, statistical techniques are available that can make up for limited data. Load estimators based on statistical regressions are techniques commonly used to estimate daily loads from samples collected at less than daily frequencies (Cohn, 1988; 1994; Cohn *et al.*, 1992; Richards, in press).

Load estimators for tributaries and connecting channels require sample data be collected at sufficient frequencies with regard to important variables such as streamflow and season (Cohn, 1994; Richards, in press). Most stream load estimators in use today can accommodate some degree of censored values. The MVUE (Minimum Variance Unbiased Estimator) (Cohn *et al.*, 1992; Cohn, 1994) can be used to compute tributary load estimates when at least 25 percent of the sample data are above the detection limit and when there are 50 or more samples with at least 25 samples collected per year. The AMLE (adjusted maximum likelihood estimator) (Cohn *et al.*, 1992; Cohn, 1994) requires the same sampling frequency with at least 20 samples above the detection limit. Realistically, an estimator technique may perform quite well with anywhere from as few as 30 to as many as 75 or more samples. Many samples are more desirable than few samples. For smaller sample sizes, the percentage of censored data must be kept to a minimum of 50 percent.



For purposes of this report, data sets suitable for the computation of contaminant loads from non-point sources such as tributaries were judged to be best represented by a sample size of at least 50. In addition, concentrations of contaminants should be detected in at least 25 percent of the samples. Samples applicable to the computation of loads should have been collected at or near a daily streamflow gage. The samples also must be collected over a range of low to high streamflows representative of the stream at the sample-collection site.

Referring to the prior discussion on point sources, the minimum number of reported observations needed to compute loads from point sources remains 10. The minimum criteria established to compute loads discharged from point sources were at least 25 percent of the observations above the detection limit.