A CAUTIONARY NOTE ON THE USE OF SPECIES PRESENCE AND ABSENCE DATA IN DERIVING SEDIMENT CRITERIA

KATHERINE VON STACKELBERG* and CHARLES A. MENZIE
Menzie-Cura & Associates, 1 Courthouse Lane, Suite 2, Chelmsford, Massachusetts 01824, USA

(Received 22 March 2001; Accepted 13 August 2001)

Abstract—In recent years, a variety of approaches to deriving sediment quality guidelines have been developed. One approach relies on establishing an empirical relationship between the concentration of a contaminant in sediment and the condition of some biological indicator, for example, combining measured sediment concentrations of contaminants combined with data on colocated benthic species to measure in situ community effects of contamination. Biological threshold concentrations derived in this manner are being considered or have already been adopted by some regulatory agencies as a means for deriving sediment guidelines (e.g., Canada’s Provincial Sediment Quality Guidelines). In order to test the validity of this method, we constructed several Monte Carlo simulations to illustrate that the methodology used to develop these guidelines is flawed by the effects of sampling and statistical artifacts that emerge from undersampling a lognormal density function. As a case study, this paper will present the screening level concentration method used by the Ontario Ministry of the Environment (Toronto, ON, Canada) and provide the results of several probabilistic exercises highlighting these issues. We present a word of caution on the applicability of methods that rely exclusively on statistical and mathematical relationships between invertebrate data and sediment concentrations to derive sediment quality guidelines.

Keywords—Sediment quality guidelines Probabilistic Benthos Screening-level concentration method Monte Carlo

INTRODUCTION

This paper identifies several important limitations to using colocated data on the presence of benthic invertebrate species and sediment chemistry to assess sediment quality and deriving sediment screening levels or criteria. The use of benthic invertebrate data to evaluate water and sediment quality has a long history, and many useful assessment tools have been developed [1–3]. Differences in the tolerances of benthic invertebrates to environmental stressors such as toxic chemicals can be revealed in shifts in the community structure and by the presence or absence of species as demonstrated by carefully designed studies. When combined with physical and chemical data on sediments, information on benthic community structure can identify possible causal relationships. However, statistics based on species presence/absence data can be misleading if based on limited observations. In such cases, species that are found at elevated concentrations of toxic chemicals in sediments may appear to be tolerant, while those that occur only at lower concentrations may be judged as intolerant. As we will demonstrate, these apparent differences among species can occur as statistical artifacts when species have been undersampled, as represented by the less common taxa. If colocated chemistry data are matched with species presence/absence data, the sample size (number of samples) for chemistry data will be greater for common species than for rare ones. These differences in sample size can directly affect the statistics on chemical concentrations derived for each species. This, in turn, can mislead the analyst who is exploring relationships between chemical and biological data.

The potential for undersampling or differential sampling of species arises from the fact that species vary in their distributions and density throughout the environment. This well-recognized phenomenon is due to a number of biological, chemical, physical, geographic, and stochastic factors [4,5]. The importance of these factors is well recognized for aquatic insects [6] and is evident in the mosaic patterns of abundance reflected in species distributions [7]. Typically, in a collection of benthic samples obtained on a localized or regional basis, some species are found at a majority of sites and other species at only a few sites. Because no inherent relationship exists between species presence and absence and potential threshold effects, the screening-level concentration (SLC) method becomes a statistical exercise in which more common species are well represented and less common species are undersampled or subject to artifacts associated with small sample sizes.

To evaluate the potential statistical artifacts—and potential misinterpretations—associated with undersampling and/or differential sampling of species, we examined and compared actual and simulated data sets with colocated species presence/absence data and chemical concentrations. For initial calculations, we chose to work with the data set used to derive Canada’s Provincial Sediment Quality Guidelines [8,9]. These numerical guidelines are based on the SLC method [10,11], which compares measured concentrations of sediment contaminants with data on the presence/absence of various benthic species. The SLC methodology has been adopted or is being considered by several U.S. and Canadian regulatory agencies in developing sediment-screening criteria.

We also constructed simulated data sets from a national database on chemicals in sediments developed by the National Oceanic and Atmospheric Administration (NOAA) [12]. Using this chemistry database, we simulated the presence and absence of hypothetical benthic species by drawing random sam-

* To whom correspondence may be addressed (kvon@menziecura.com).
Fig. 1. Conceptual schematic of the screening level concentration (SLC) method [10-11].

METHODS

Data set of Persaud et al. and the SLC method

In developing SLCs, Persaud et al. [10] selected 100 representative (as determined by Persaud et al.) benthic species and documented their presence or absence at 250 sediment sampling locations. The sediment samples were analyzed for various contaminants, including most of the heavy metals and a number of organochlorines, such as polychlorinated biphenyls and DDT. For each location at which a benthic species was found, the authors plotted a corresponding sediment concentration for each chemical. This resulted in a distribution of sediment concentrations for each of the 100 species. For less common species, only a few associated measurements of chemicals were taken, while for common species many measurements were taken (up to a maximum of 250). Figure 1 presents a conceptual schematic for how this was done. The top four graphs in Figure 1 show the general shape of these distributions for each species. The hatched line in the top four graphs shows the 90th-percentile concentration for each species. This concentration will be different for each species for which a concentration is available.

The authors combined the individual species’ distributions in the following way. From the chemical plots for each species, Persaud et al. [10] chose the 90th-percentile concentration as a conservative estimate of the upper bound on the tolerance range for that species. They then combined the individual species’ 90th-percentile concentrations to construct a distribution of 90th percentiles for each chemical across all species. This is shown conceptually in the bottom graph of Figure 1. They define the fifth percentile from the final combined 90th-percentile distribution as the lowest effect level (LEL) and the 95th percentile from this distribution as the severe effect level (SEL). This is shown by the hatched lines in the bottom graph of Figure 1.

Simulations based on the NOAA database

We used the National Status and Trends Program database [12] to derive distributions of cadmium and DDT in sediments. Environmental data tend to follow an approximate lognormal distribution, characterized by a skewed right tail, as was the case for both cadmium and DDT, as shown in Figure 2. These distributions are likely to be similar in shape to those of contaminants found in sediments in the Persaud et al. [10] data set, although presumably specific values differ.

To construct species-specific sediment concentration distributions, we simulated a benthic data set of 105 hypothetical
species by sampling at various levels of intensity from the sediment concentration distributions shown in Figure 2. We designated seven categories of species, ranging from rare (occurring in only 10 out of 250 samples) to common (occurring in all 250 samples). Other groups of 15 species occurred in 20, 50, 100, 150, 200, and 250 samples, yielding 105 species in total. Sampling of hypothetical species from sediments reflecting the NOAA cadmium and DDT data sets was carried out using a Monte Carlo analysis method set up in the Excel add-in Crystal Ball® (Microsoft, Redmond, WA, USA). This permitted random sampling at varying levels of intensity from the lognormal data set. That is, for the 15 rare species, found at only 10 locations, the data set was sampled 10 times (15 Monte Carlo simulations of 10 iterations each to represent 15 rare species), while for the 15 most common species, the data set was sampled 250 times (250 iterations) to obtain species-specific distributions. We then treated these data in the manner described previously for the SLC method to obtain 90th percentiles for each of the species. Finally, we calculated the overall fifth and 95th percentiles from the chemical-specific 90th-percentile distributions. This exercise (case 1) assumed no a priori biological threshold of sensitivity to a given contaminant.

Simulated data set with imposed threshold concentrations

To examine the characteristics of a data set for which chemical effects on benthic organisms were known to be actually occurring, we conducted a second set of simulations. We assumed a priori knowledge of a threshold concentration above which the species would not typically be found by imposing a range of hypothetical threshold concentrations for each simulated species and contaminant (cadmium and DDT). Sampling of the species from the cadmium and DDT data sets was accomplished as described previously using a Monte Carlo analysis. If a sampled cadmium or DDT value exceeded assigned species-specific thresholds, the species was designated as absent in that sample and that data point excluded from all further analyses. The fifth and 95th percentiles of the distribution for all species were calculated as described previously for the SLC method.

To evaluate the effects of sampling under the assumption of known effects, three sets of simulations were conducted for cadmium and DDT. In the first set of simulations (case 2), a range of known tolerances was assumed across each of the 15 species in each category. That is, each category of species showed both low and high threshold concentrations, so that each of the 15 species had two different assigned thresholds. In the second set of simulations (case 3), rare species were assigned low thresholds, while common species were assigned high thresholds, but within each category of species, the assigned threshold was the same. This was done to evaluate the stability in the resulting 90th percentiles. In the third set of simulations (case 4), rare species were assigned high thresholds and common species low thresholds.

RESULTS

Data set of Persaud et al.

Figure 3 shows the species-specific 90th-percentile values of cadmium, copper, lead, and zinc sediment concentrations versus the number of sites at which a species was observed (sample size) for the Persaud et al. [10] data set. These graphs reveal a common pattern. Variance decreases as a function of sample size. Consequently, the highest and lowest species-specific 90th-percentile concentrations typically occur for the species with the fewest observations. Greater scatter exists in the derived percentiles for less common species (those for which fewer observations were made). This scatter diminishes for species for which 100 or more observations were made. The highest and lowest species-specific 90th-percentile values plotted in Figure 3 typically occur at the left of each plot and are associated with the more rare species. The SLC methodology estimates the LEL (fifth percentile) and the SEL (95th percentile) from the tails of these distributions. As can be seen from Figure 3, these effect-level values would be substantially influenced by statistics for species for which fewer observations were made and greater scatter exists. As we will show using the simulated data sets, this scatter is likely a statistical artifact related to sample size rather than actual biological effects being experienced by the organisms.

Data sets simulated from the NOAA database

The 90th-percentile concentration values for the 105 simulated species sampled using the NOAA cadmium and DDT data sets are presented in Figure 4. This simulation does not presume any toxic effect of either contaminant on the presence or absence of invertebrates but simply indicates the effect of different sampling intensities. The pattern in Figure 4 resembles those presented for the actual data sets in Figure 3. The 90th percentiles for rare species (data set sampled 10–20 times) show more scatter than the 90th percentiles for more common species (data set sampled 200–250 times). Because data for each species represent random sampling from the same cadmium or DDT distributions, the observed scatter for species
Flaws of presence/absence method for sediment guidelines

Environ. Toxicol. Chem. 21, 2002 469

Fig. 3. Observed data from Ontario, Canada dataset [10].

with fewer observations reflects uncertainty in statistical estimates and not variability among species. The higher uncertainty, reflected in greater scatter, is associated with under-sampling the distribution or small sample sizes. If large numbers of observations were made for each of the simulated species, their statistics would converge to the same distribution.

Other statistics are also subject to sample size effects. For example, the maximum cadmium concentration observed in this set of simulations increases with increased number of observations. Again, this is a statistical artifact that reflects sampling the same distribution of sediment cadmium at increasing levels of intensity. The more samples taken, the greater the likelihood that the tails of the distribution are represented. In the case of the simulated data set, it would be incorrect to assume that the different maximum values obtained for species represent different tolerance levels. Maximum values and the plotted 90th-percentile values simply reflect the effects of sampling a distribution at varying degrees of intensity. Alone, these statistics provide no insight into the toxic effects of cadmium or DDT.

Cadmium and imposed threshold ranges

The results of this analysis are shown in Table 1. For this set of simulations, we imposed presumed known threshold concentrations for each category of species and eliminated all values exceeding the imposed thresholds. Following elimination of the exceedances, we calculated the 95th and fifth percentiles (SEls and LELs) across all species for the range of threshold concentrations using the method described previously. Table 1 provides the results for the following cases: (1) assuming no threshold, (2) assuming a range of thresholds from low to high within one species, (3) assuming a low threshold for rare species and a high threshold for common species, and (4) assuming a high threshold for rare species and a low threshold for common species. The table shows the estimated fifth and 95th percentiles from the distribution of 90th percentiles. The overall column provides the results across all species as compared to the results for the individual (rare to more common) category of species.

The overall calculated SELs and LELs for the cadmium data set did not differ substantially despite different assumed thresholds. However, the LEL decreased from 0.5 in the no-threshold case to 0.2 (the lowest assumed threshold), indicating that the LEL is sensitive to assumptions made at the low end of the distribution.

The simulations in which we assigned thresholds indicate that the resulting statistics may not have the same meaning as intended by the investigator. For example, the use of 95th and fifth percentiles to derive SELs and LELs implies that these values are protective of the specified percentiles of species. As we have shown, these percentiles do not have that meaning. In fact, in our simulations, the SEL appears to provide a conservative estimate, and the LEL is even more conservative, providing values lower than the lowest assigned tolerance level.

DDT and imposed threshold ranges

Because of the strongly skewed nature of the underlying frequency distribution for DDT, a different pattern emerges. Table 2 presents the results for the following cases: (1) assuming no threshold, (2) assuming a range of thresholds from low to high within one species, (3) assuming a low threshold
for rare species and a high threshold for common species, and (4) assuming a high threshold for rare species and a low threshold for common species. For example, the imposed threshold for the species sampled 100 times under case 3 is 2.0 µg/g. The table shows the estimated 5th and 95th percentiles from the distribution of 90th percentiles. The overall column provides the results across all species as compared to the results for the individual (rare to more common) category of species.

For DDT, the overall estimated LEL was not substantially different across imposed thresholds, but the SEL did change. When a 2,500-µg/kg threshold was imposed on the rare species (10 observations or Monte Carlo iterations), the resulting LEL was 38.4 µg/kg and the SEL 627.4 µg/kg in contrast to the most common species (250 observations of Monte Carlo iterations), for which the estimated LEL was 130.7 µg/kg and the estimated SEL 188.3 µg/kg. Overall, the estimated LELs were essentially the same between the two different methods for imposing thresholds, but the SELs differed from 181.1 to 322.3 µg/kg.

Overall, the LELs ranged from 3.4 to 4.8 µg/kg for the lowest imposed thresholds for the least common species and assuming no threshold at all, respectively. The SEL ranged from 181.1 to 470.3 µg/kg. The lowest SEL was observed for the case in which the least common species exhibited the lowest a priori threshold.

DISCUSSION

Species presence/absence data should be used with caution when interpreting the effects of chemicals in the environment. As we have shown in this paper, a potential exists for misidentifying some species as tolerant and others as intolerant when undersampling from a lognormal distribution. Results should not be interpreted in the absence of an ecological and toxicological framework. Some statistics derived from presence/absence data can lead to misleading interpretations regarding potential effects. In our simulations in which we use a method currently employed to derive sediment guidelines, we are able to interpret results in light of the known tolerance levels that we assigned to the data set. In one case, the contaminant was presumed to have no toxicity, and in the other
cases the toxicity was species dependent over some range. Collectively, these simulations showed that a seemingly meaningful pattern can emerge when, in fact, the chemical is not having an effect. They also showed that when effects were occurring, the interpretation of resulting data (as toxicity thresholds) could be different than the manner in which the effect was actually manifest. In part, the potential for misinterpretation is due to undersampling of species (small sample sizes) and the statistical attributes of lognormal distributions (or other right-skewed distributions) of chemical data.

The distributions of contaminant concentrations in Persaud et al. [10] are obtained from sampling at numerous different locations, including both relatively contaminated and uncontaminated sites. This makes the statistics obtained from the exercise particularly difficult to interpret. If all the sites were relatively uncontaminated, then the resulting statistics might be indicative of background conditions; however, this is not the case. In addition, the presence and/or absence of particular species is considered representative for a larger area on the basis of the samples taken. It could turn out that collecting additional samples would reveal the presence of a previously absent species in the general area being characterized.

Other factors not considered in this paper could also be important. Relationships between habitat size, sample size, and the presence and number of species has long been recognized in ecology [4,5]. This relationship can be seen by plotting the absolute number of individuals per species against the number of species in each abundance class. The resulting plot is strongly skewed to the right, indicating that many more moderately rare species than moderately common ones exist. Gray [13] documents an increase in abundance of eight indicator species classified as neither rare nor common under conditions of slight pollution, changing the expected right skew of the plot. Distributions of contaminants from locations displaying this altered pattern may be misleading.

Certain species are more common than others for a variety

| Table 1. Results of Monte Carlo simulations for cadmium using the National Oceanic and Atmospheric Administration [12] data set |
|--------------------------------------------------|--------------------------------------------------|
| Cadmium<sup>a</sup> | No. of observations/Monte Carlo iterations<sup>b</sup> |
| | 10 | 20 | 50 | 100 | 150 | 200 | 250 | Overall |
| 1. No threshold assumed | | | | | | | | |
| 5th | 0.9 | 0.4 | 1.3 | 1.4 | 1.5 | 1.4 | 1.5 | 0.5 |
| 95th | 2.5 | 2.1 | 2.1 | 2.2 | 2.5 | 2.0 | 1.8 | 2.2 |
| 2. Assumed threshold<sup>c</sup> = range from 0.2 to 7.0 µg/g | | | | | | | | |
| 5th | 0.4 | 0.4 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.2 |
| 95th | 2.1 | 2.0 | 2.0 | 2.0 | 1.7 | 2.0 | 1.6 | 2.0 |
| 3. Assumed threshold<sup>c</sup> = | | | | | | | | |
| 5th | 0.2 | 0.5 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 |
| 95th | 0.2 | 0.5 | 0.9 | 1.3 | 1.6 | 1.9 | 1.8 | 1.7 |
| 4. Assumed threshold<sup>c</sup> = | | | | | | | | |
| 5th | 0.9 | 0.9 | 1.1 | 1.0 | 0.8 | 0.4 | 0.2 | 0.2 |
| 95th | 2.2 | 2.2 | 1.8 | 1.4 | 0.9 | 0.5 | 0.2 | 2.0 |

<sup>a</sup>The number of observations represents the number of iterations from the probability density function in the Monte Carlo simulation with particular threshold ranges imposed. Fifteen simulations for each of 10, 20, 50, 100, 150, 200, and 250 iterations were run.

<sup>b</sup>Number of Monte Carlo iterations to represent rare to more common species. For example, 10 = 10 iterations or rare species, 20 = 20 iterations or less rare species, and so on, up to 250 species, the most common species.

<sup>c</sup>Imposed threshold concentration in µg/g or ppm.

| Table 2. Results of Monte Carlo simulations for DDT using the National Oceanic and Atmospheric Administration [12] data set |
|--------------------------------------------------|--------------------------------------------------|
| DDT<sup>a</sup> | No. of Observations<sup>b</sup> |
| | 10 | 20 | 50 | 100 | 150 | 200 | 250 | Overall |
| 1. No threshold assumed | | | | | | | | |
| 5th | 38.4 | 45.0 | 51.0 | 106.4 | 33.5 | 14.2 | 4.3 | 4.8 |
| 95th | 627.4 | 449.0 | 316.3 | 273.5 | 226.8 | 237.9 | 190.7 | 470.3 |
| 2. Assumed threshold<sup>c</sup> = 5 to 2,500 µg/kg | | | | | | | | |
| 5th | 6.1 | 9.6 | 12.2 | 10.3 | 11.0 | 13.2 | 12.7 | 4.4 |
| 95th | 167.3 | 179.8 | 253.1 | 161.4 | 169.2 | 207.6 | 184.1 | 196.0 |
| 3. Assumed threshold<sup>c</sup> = | | | | | | | | |
| 5th | 5.1 | 20 | 60 | 120 | 200 | 1,000 | 2,500 |
| 95th | 1.1 | 8.6 | 28.1 | 47.2 | 61.2 | 118.3 | 130.7 | 3.4 |
| 4. Assumed threshold<sup>c</sup> = | | | | | | | | |
| 5th | 5.1 | 18.0 | 43.9 | 83.4 | 102.5 | 201.9 | 188.3 | 181.1 |
| 95th | 2,500 | 1,000 | 200 | 120 | 60 | 20 | 5 |

<sup>a</sup>The number of observations represents the number of iterations from the probability density function in the Monte Carlo simulation with particular threshold ranges imposed. Fifteen simulations for each of 10, 20, 50, 100, 150, 200, and 250 iterations were run.

<sup>b</sup>Number of Monte Carlo iterations to represent rare to more common species. For example, 10 = 10 iterations or rare species, 20 = 20 iterations or less rare species, and so on, up to 250 species, the most common species.

<sup>c</sup>Imposed threshold concentration in µg/g or ppm.
of reasons. For example, one species that was infrequently found in the Ontario study, *Branchiura sowerbyi*, is known to be rare [14]. The presence or absence of this particular species is difficult to attribute to a measured contaminant concentration, as the species is infrequently found for other reasons. Every organism exists within a fundamental niche, defined as the set of all environmental conditions that permit it to exist [4]. Further, the availability of a number of substrates within a microhabitat increases the likelihood that more species will coexist, which in turn affects the absence or presence of any given species. Ecological bases for rareness or commonness exist that should be considered when evaluating presence/absence data with respect to specific environmental stressors.

Cairns and Dickson [15] and others have demonstrated the dangers of using presence/absence data to the exclusion of other measures in environmental studies or parameters in environmental modeling. Interpretation of presence/absence data is particularly difficult in studies with numerous and disparate sampling locations. In any two given samples in which a species is present in one and absent in the other, a number of environmental factors may be responsible, including historical zoogeographic rather than proximate environmental reasons [4]. A more robust method of evaluating impact involves the use of a temporal control to establish the status of the species before impact. Several U.S. Environmental Protection Agency reports [1,2] describe a number of methods for evaluating benthic communities. Several authors have demonstrated that reduced species richness, reduced community biomass, reduced deep-dwelling species biomass, and increased opportunistic species biomass are often indicators of contaminant impact [16–18]. In particular, species richness (calculated as the mean number of species per sample) provides an important indicator of benthic health [1,2,18].

The emergence of weight-of-evidence approaches that take into account multiple lines of evidence have proven to be useful for evaluation of risks posed by chemicals in sediments [18]. These are most useful when they incorporate information on the toxicity of the chemical(s), an understanding of toxic mechanisms, chemistry, physical conditions of the sediments, and biology. Integrative assessments combine many different attributes of the localized system to determine environmental quality [18–20].

Many state, provincial, and federal or national agencies would like to implement screening-level values for chemicals in sediments. We recommend that the screening of sediments incorporate key factors that affect exposure and toxicity rather than relying on a single metric such as bulk chemical values. These factors vary among classes of contaminants and among habitats. In selecting them, consideration should be given to receptor relevance, geochemical factors that may modify exposure on a site-specific basis (e.g., acid-volatile sulfides), and other factors that may confound interpretation of results (e.g., history of physical disturbance). A wide range of tools can be used to this end. The challenge is selecting a suite of tools that are simple enough and conservative enough to be used for screening-level purposes. The simulations we conducted show that sediment criteria or values derived through the use of absence/presence data should not be considered as valid screening levels.

**REFERENCES**


