Weight of Evidence (WOE): Quantitative Estimation of Probability of Impact

Eric P. Smith
Ilya Lipkovich
Keying Ye

Dept. of Statistics
Virginia Tech

Corresponding Author
Eric P. Smith

Dept. of Statistics
Virginia Tech
Blacksburg, VA 24061-0439
email:epsmith@vt.edu
Phone: 542-231-7929
Fax: 540-231-3863

Updated 12/03/01, 1/14/02, edited 2/5/02
Corrected 2/10/02

Running head: Quantitative weight of evidence for environmental assessment

Abstract

Environmental decision-making is complex and often based on multiple lines of evidence. Integrating the information from these multiple lines of evidence is rarely a simple process. We present a quantitative approach to the combination of multiple lines of evidence through calculation of weight of evidence, with reference conditions used to define a not impaired state. The approach is risk-based with measurement of risk computed as the probability of impairment. When data on reference conditions are available, there are a variety of methods for calculating this probability. Statistical theory and the use of odds ratios provide a method for combining the measures of risk from the different lines of evidence. The approach is illustrated using data from the Great Lakes to predict the risk at potentially contaminated sites.

Keywords: Bayesian statistics, odds ratio, hazard ranking, combining information, risk assessment, reference conditions

Introduction

Environmental decision-making is often based on multiple sets of information or lines of evidence. By line of evidence we mean a set of information that pertains to an important aspect of the environment. For example, in the sediment quality triad (Chapman 1996), there are three lines of evidence, the toxicity line, the biological field line and the chemistry line. It is difficult to combine the information from these multiple sources into a single measure for decision-making. Weight of evidence (WOE) is sometimes used as an approach for combining the information, however it is rarely used in a quantitative manner. This paper discusses a quantitative approach to WOE. A statistical approach is taken in which the likelihood of the data is calculated under two different scenarios and a decision made based on the ratio of the likelihood. Our view is that there are two states, and we must decide which of the states is true. Examples of pairs of states common to environmental decision making are (impaired, not impaired), (remediate, don't remediate), (list, don't list), etc. The view we take is that interest is in a single site, we collect a sample at that site and based on the sample, the site is impacted or the site is not impacted. For practical reasons, we assume the simple case that there is ample information on reference conditions and interest is in evaluating a single new location. This gives us the ability to obtain a precise estimate of the probability.

Estimating weight of evidence (WOE)

A quantitative approach to WOE is based on the concept of statistical weight of evidence. This idea dates back to work by Alan Turing in World War II (for a more general discussion of history and concepts of statistical weight of evidence, see Good 1988). In this approach, there are two states and we must decide which state is more likely given the data. We can view the outcomes as the site is or is not impacted. Without observing the information, we may have opinions or insights into the condition of the locations. This insight might be based on previous data (condition in previous years) or be from sites that are close in space. This information may be used to form a prior opinion or probability of impairment. After the data are collected, we process the data to evaluate the site. This leads to a Bayesian approach in which the data are used to update the prior information. The lack of prior information suggests a frequentist approach where the data alone are used to make a decision. The approach may be based on a single line of evidence or multiple lines. The individual lines of evidence are usually evaluated separately and by combining them we hope to make a stronger inference.

Statistical WOE is based on a quantitative evaluation of the data and requires a model that describes the data. In the simplest approach, there are two states and we must decide which state is more likely given the data. Because it is not always easy to describe impact, an alternative approach is to evaluate the risk of impairment of a site considering the baseline risk of a not impaired site. If we view the possible outcomes as the site is or is not impacted, the risk is then the probability that the site is impacted or *P(impact)* where we evaluate this probability after information is collected.

The odds are a way to evaluate how big the probability is relative to the baseline risk. Odds of impact are defined as

$$\text{odds(impact)} = \text{P(impact)/P(no impact)} \qquad (1)$$

Although this problem is analogous to tossing a coin, estimating the probability of impact is not easy since “impact” is not an observable attribute of a sample. The state (impact or no impact) must be inferred based on information that is collected on both impacted and unimpacted sites. A reasonable approach is based on Bayes rule. With no data the probabilities would be estimated based on prior information that may come from previous studies. We more generally would collect data to improve these estimates. Given data we have to calculate probabilities of impact or no impact. However, even with data we do not have probabilities of impact, only probabilities associated with observations given a model for impacted sites and not impacted sites. For example, if there is ample information on sites that are not impacted we may compare our data to that and estimate the probability the data come from that distribution. We do not have the probability of no impact only the probability that the data come from that distribution. A representation of what we want to calculate is

$$\frac{P(impact \mid data)}{P(no\ impact \mid data)}$$

Given data and information about the different groups we can calculate how likely the data are given they are from one of the groups. This probability is written as P(data|impact) or P(data|no impact) and must be computed based on a statistical model for the data (possibly different models for each group). For example, we might specify that the model for the impact sites for dissolved oxygen is normal with mean 4 and for no impact sites is normal with mean 7. If we obtain a sample with dissolved oxygen equal to 6 and we know the variance of the dissolved oxygen measurements, we can calculate how likely the observation is to have come from each group by calculating the value of the density under each model. Bayes theorem may be used to calculate the ratio in terms of the probability of the data since

$$\frac{P(impact \mid data)}{P(no\ impact \mid data)} = \frac{P(data \mid impact)P(impact)}{P(data \mid no\ impact)P(no\ impact)}$$

where *P(impact)* is the probability the site is an impact site before collecting the measurements and is referred to as the prior probability. If the prior probabilities are taken to be equal the result is

$$\frac{P(impact \mid data)}{P(no\ impact \mid data)} = \frac{P(data \mid impact)}{P(data \mid no\ impact)}$$

The quantity on the right hand side of the equation is referred to as the likelihood ratio or a Bayes Factor in this case and measures the likelihood of the data given the site is in the impact class versus the no impact class. (More generally the Bayes Factor also involves parameters that are treated as random and integrated out of calculations, see Kass and Raftery, 1995). For environmental problems there may not be simple approaches for estimating these quantities. Building a model for the impacted or unimpacted sites requires information on how data are distributed for these types of sites and other factors that might influence the observations. Sites classified as unimpacted are often viewed as reference sites. It may not be possible to obtain these sites or there may be covariates that must be considered. Calculation of the likelihood of the data under impact requires a definition of the impact or model of the data that we might expect if the observation came from the impact group. One would have to have different models for different types of impacts (for example chemical toxicity versus sedimentation). The models should depend on the strength of the impact, and may vary in space and time. These models may involve a good deal of work to describe. One approach to calculating the ratio given a lack of information on impact is to calculate the odds as *p/(1-p)* where *p* is the probability of the data given there is an impact.

WOE is a measure of how much an observed feature in the data adds to or subtracts from the evidence of impact. Numerically it has been defined as (Good 1988)

$$\textit{Weight of evidence} = 10 \log(\textit{likelihood ratio}).$$

In our applications, this would correspond to the weight of evidence for one line of evidence. A natural consequence of using logs of ratios is that the weight of evidence from different lines may be added together to get an overall weight of evidence. For an individual line of evidence values of the ratio are interpreted as follows (Good, 1988):

Bayes Factor	Weight of Evidence	Strength of Evidence
<5	<6.9	Weak
5-10	6.9-10	Moderate
10-100	10-20	Moderate to strong
>100	>20	Strong

These rules are guidelines in much the same way that p-values are guidelines and other authors have suggested alternative views (see Kass and Raftery, 1995). From a hypothesis testing perspective the weight of evidence measures the strength of the evidence against the null hypothesis.

Numerical calculation of weight of evidence is not common to statistics. The reason is that testing of hypotheses and interpretation more common approach of likelihood ratio testing and calculation of Bayes Factors. Using the natural log scale and twice the log of the Bayes Factor leads to the same scale as likelihood ratio testing in general statistical theory (where –2log(likelihood ratio) is used to test hypotheses) and deviance measures in generalized linear models (McCullagh and Nelder, 1989). Thus, for applications of WOE in environmental problems a user may choose to summarize results in terms of a WOE measure that is based on the probability of impairment. Alternatively the user may simply calculate and report the actual probability. The value of the use of Bayes Factors and odds is that these may be combined easily over the different lines of evidence.

The calculation of WOE for multi-attribute environmental studies involves three general stages of analysis. We assume that the researcher has available the information required for making the decision. Thus decisions have been made about what information needs to be collected or this information has already been collected. The three stages of the analysis are the preprocessing stage, the processing stage and the combination of the information over the lines of evidence.

Data preprocessing

The initial step in the analysis is the preprocessing of the data. Preprocessing involves selection of the variables to be used in the analysis and scaling or transforming these variables. Variables are selected to provide relevant statistical and scientific information on differences between control and impact. Scaling and transformation are often used to meet assumptions required for analysis. The assumptions needed depend on the model used to calculate the probability. Two methods for this calculation are to assume a model (parametric approach) or to calculate the probability using a nonparametric approach. In the parametric approach we select a probability model for the data. Here we use a normal or Gaussian model to estimate the risk. Then there are a several assumptions that need to be evaluated for the statistical model to provide a good estimate:

1. Normality of the reference data.
2. Independence of samples in the reference set.
3. Homogeneity of variance in the reference set.

Since a normal distribution is used to calculate probability of impairment the validity of this assumption is required for the estimate of risk to be accurate. In the normal model, probability is directly related to the standardized distance to the mean. If the normal model is not reasonable then the estimate may be poor and misleading. Problems such as skewness and outliers may lead to inaccuracies in the probability estimate. As environmental data often are not normal, we try to achieve normality via choice of a suitable transformation of the variables or use a method based on the distribution of the data (i.e. logistic regression assumes a binomial distribution). The logarithm is typically used as a transformation with contaminant concentrations. Independence of the reference data is required to provide a good estimate of the variances and covariances in the contaminants. This assumption is best met through choice of the reference locations and sampling occasion. Sites that are spatially close and repeated samples at the same site that are temporally close should be avoided. Homogeneity of the variance in the reference set is required to produce a good estimate of the variances and covariances. An alternative approach would be to allow heterogeneity but include this in the model in some manner. A potential concern here is with multiple sets of reference sites. If the multiple sets are treated as a single set then the estimates of the variances are likely to be smaller than data collected from different sites. Hence detection of impairment is potentially overly sensitive. A useful strategy with multiple samples from a collection of sites is to try to match the test site with similar reference sites rather than to use all of the sites. This might involve forming clusters of reference sites and matching the test site with a cluster or using an auxiliary set of measurements (such as sediment type). Other potential problems include measuring a single site multiple times and the time of sampling. Selecting reference sites is a difficult and important problem.

The data transformation need not be a transformation of individual contaminants but may also be on the set of measurements. For example, it is common to analyze composites of variables rather than individual variables. Two common approaches are to use principal components (PC) or correspondence analysis (CA) to form new variables. These two methods are useful when the dimensionality of variable space is high and variables are highly correlated. In the case of the sediment quality triad, the PC transformation would typically be applied to the sediment toxicity and metal chemical variables, while correspondence analysis axes can be used for species composition data. Another possible data transformation could be computing some univariate index. For example, composition data can be represented by diversity measures or an index of biological integrity.

When the normal assumption is not valid a possible approach is to use a distance measure and build a nonparametric estimate of the probability of impairment (details are given below). When a nonparametric method is used, there are also assumptions that must be considered. The nonparametric approach that we use involves a measure of the distance from the site to the reference sites. A probability model is then developed based on how close reference sites are to each other. For this approach to work we have similar assumptions:

1. The reference site data are from a common distribution
2. The reference sites are independent
3. The distance measure is appropriate for detecting change.

If the data for the reference sites come from a common distribution, then a single distance measure will produce reasonable estimates of how similar the test site is to the reference sites. When there are different sets of reference conditions the distance measure would have to be computed with respect to the different distributions or with respect to the set of reference conditions most similar to the test site. Independence of samples implies that equal weight may be given to each of the samples from the reference sites. The choice of distance measure is a critical step as the distance measure defines the measure of impact. One important consideration in the selection of the distance measure is the weight given to the variables used in computing the distance (Smith 1998).

The choice of how to preprocess the information is critical to the analysis as it defines the deviations that are of interest. One should be aware of the limitations associated with these choices. For example, if information on a large number of variables is collected then one has a better chance of detecting a broad scale impact. If the impact is only observed through one variable, the other information becomes of low utility for the detection of impact. Thus there is a need to have a clear idea of what types of impact are to be detected. For example, if a chemical is only toxic to fish then measuring abundance of benthic macroinvertebrates will not be risk informative. It is critical in selecting an approach to be aware of what types of changes will or will not be well detected in the analysis and how likely the analysis is to detect changes of important magnitude. Methods such as power analysis are useful for evaluating variables and their importance in the decision process. Also transformation of the variables is often needed. For example, chemical data are often collected in environmental studies. The user needs to decide if the original or standardized data are used. If standardized the method of standardization needs to be chosen. Options might include an overall standardization,

standardization relative to a reference group or standardization in terms of toxic units. Choice of standardization will change the magnitude of distances between observations.

Data processing: Estimating probability of impact

In the data processing step, the test site is compared with reference conditions in order to obtain a measure of the degree of impairment. This may be an indirect or direct evaluation. For example, with biotic indices (Smoger and Angermeier 1999), there is often a calibration step in which the metrics that make up the index are scaled based on reference conditions. A direct evaluation is based on comparing the reference and test measurements. A statistical approach is to compute the probability of impairment through the use of the distance between reference and test measurements. The use of distance in ecological and environmental impact assessment has a long history that will not be explored here. For example, distance from control forms the basis of outlier detection methods and control chart approaches that view water quality analysis as a quality control problem. Distance also is central to many multivariate methods used to assess ecological change such as correspondence analysis (using chi-square distance, Legendre and Legendre, 1998) or multidimensional scaling (Smith et al, 1990, Gray et al, 1990).

The estimation of the probability of impact is based on comparison of the site in question with reference sites and for certain statistical models, the probability is directly related to distance from the reference. For example, with a single measurement and mode of evaluation, the calculation of WOE requires estimation of the distribution and calculating the probability that the observation is from that distribution. In the simplest case, the probability distribution of the measurement is treated as a normal distribution. Then the two-tailed probability may be calculated by first computing

$$d(x_0,\mu) = (x_0 - \mu)/\sigma$$

Next, the quantity is used to calculate the probability. Since the parameters are unknown, they are estimated so one would use

$$\hat{d}(x_0,\mu) = (x_0 - \overline{x})/s$$

This idea then suggests a more general approach: first transform the measurement to try to obtain a normal distribution for the reference site data, then use the estimated parameters from the normal distribution to compute the distance and then the probability. This approach is useful in that the direction of deviation may be taken into account. The disadvantage is that when there are several variables, the individual estimates may be correlated and this correlation is not taken into account in the distance measure.

The possibility of correlated measurements suggests a multivariate distance measure be used. Suppose the measurements at a particular site are given by $\mathbf{x_0}$ where the convention of bold indicates a vector or set of measurements. This vector may be chemical measurements, toxicity information or biological information. The measurements are typically manipulated or transformed in such a way that the distribution of similar vectors for the reference sites follows a normal distribution. Then the squared distance is calculated as the Mahalanobis (Rencher 1995) distance

$$d^2(\mathbf{x_0},\boldsymbol{\mu}_\mathbf{R}) = (\mathbf{x_0} - \boldsymbol{\mu}_\mathbf{R})'\Sigma_R^{-1}(\mathbf{x_0} - \boldsymbol{\mu}_\mathbf{R})$$

where $\boldsymbol{\mu}_\mathbf{R}$ is the mean for the reference sites and Σ_R is the covariance matrix for the reference sites. The squared distance is estimate by (provided the variance covariance matrix is invertible)

$$\hat{d}^2(\mathbf{x_0},\boldsymbol{\mu}_\mathbf{R}) = (\mathbf{x_0} - \overline{\mathbf{x}}_\mathbf{R})'S_R^{-1}(\mathbf{x_0} - \overline{\mathbf{x}}_\mathbf{R})$$

where the sample mean from the reference sites $\overline{\mathbf{x}}_\mathbf{R}$ and the sample covariance matrix S_R are substituted into the distance measure. This distance measure gives less weight to variables that have high variance. Similarly, highly correlated variables do not contribute as much to the variance as uncorrelated variables (Rencher 1995).

From the distance measure a probability is calculated. This is easy for multivariate normal data as the probability is determined from the distance to the mean. Other distance measures are also feasible. The advantage to using the Mahalanobis distance is the relationship with the multivariate normal model. From the distance measure, the probability of obtaining a more extreme measurement is easily obtained (as described below). The disadvantage to the distance is that extreme values may occur from extremely good as well as extremely bad sites since the comparison is with the mean. Another possible measure would be

$$d^2(\mathbf{x_0}, m_\mathbf{R}) = (\mathbf{x_0} - m_\mathbf{R})'\Sigma_R^{-1}(\mathbf{x_0} - m_\mathbf{R})$$

where m_R is the smallest possible value in the reference set (such as the detection limit or zero). Then the distance would not be large for sites with small concentrations of contaminants. This approach would make sense when the mean does not represent optimal or good conditions. A model based estimate of probability associated with this distances is difficult to obtain although a model free estimate (nonparametric) is possible given an adequate sample size for reference sites. When the lower value is zero analytical approaches are possible. The distribution of $d^2 = \mathbf{x}_0{}'\mathbf{S}_R^{-1}\mathbf{x}_0$ is related to the non-central T^2 distribution and the non-central F distribution so probabilities may be evaluated when the observations are normal.

The impairment probabilities for each line of evidence can be computed based on the multivariate normal model or, if sample sizes are sufficient, through a data based estimate. If multivariate normality is assumed, then the probability associated with the distance measure may be calculated using a chi-squared approximation or through Hotelling's T^2 distribution (Rencher 1995). The squared distance is estimated by

$$\hat{d}^2(\mathbf{x_0}, \boldsymbol{\mu}_\mathbf{R}) = (\mathbf{x_0} - \overline{\mathbf{x}}_\mathbf{R})'S_R^{-1}(\mathbf{x_0} - \overline{\mathbf{x}}_\mathbf{R})$$

where $\overline{\mathbf{x}}_\mathbf{R}$, is the mean for the reference sites and S_R is the within cluster variance-covariance matrix for the reference sites (when no clusters are specified, we of course just use the common mean and variance matrix). The probability of impairment is then defined as one minus the right tail probability of the chi-squared distribution with number of degrees of freedom equal to the number of variables. The probability of impairment is then approximated based on the chi-squared distribution of $X^2 = \frac{n}{n+1}(\mathbf{x}_o - \overline{\mathbf{x}}_R)'\Sigma_R^{-1}(\mathbf{x}_o - \overline{\mathbf{x}}_R) \sim \chi_r^2$ where Σ_R is the true covariance matrix for the reference set, and r is number of degrees of freedom equal to the rank of Σ_R (typically equal to the number of variables, p)

A more accurate estimate of the probability is obtained using Hotelling T^2 distribution. The statistic used is $T^2_{obs} = \frac{n}{n+1}(\mathbf{x}_o - \overline{\mathbf{x}}_R)'\mathbf{S}_R^{-1}(\mathbf{x}_o - \overline{\mathbf{x}}_R)$, where *k* is number of groups and *n* is the number of reference sites. To compute the p-values we can use the relationship between T^2 and F distribution

$$T^2_{p,n,\alpha} = \frac{p(n-k)}{n-p-k+1} F_{p,n-p-k+1,\alpha}$$

(Mardia et al. 1979, p. 74). To estimate the probability of impact we first compute

$$F_{obs} = \frac{n-p-k+1}{p(n-k)} T^2_{obs}$$

then calculate the probability of a more extreme *F* value using the $F_{p,n-p-k+1}$ distribution (i.e compute the probability of impairment as Pr(impairment)=Pr($F_{p,n-p-k+1} > F_{obs}$).

An alternative approach is to use a nonparametric estimate of the probability based on the empirical distribution of distances in the reference set. If there is a large enough set of reference sites then an individual reference site may be compared with the other reference sites to build an empirical or data based probability model. The idea is to treat one reference site as a test site and compare it with the other reference sites. By repeating this for all reference sites a distribution of distances is obtained and then the actual test site may be evaluated with that distribution. The distances are defined as above except that the cluster means and within cluster variance matrix are computed with removal of the observation for which the distance is computed. Thus, for the jth observation, we compute

$$\hat{d}^2(\mathbf{x}_j, \boldsymbol{\mu}_\mathbf{R}) = (\mathbf{x}_j - \overline{\mathbf{x}}_{\mathbf{R},-j})' S^{-1}_{R,-j}(\mathbf{x}_j - \overline{\mathbf{x}}_{\mathbf{R},-j})$$

where the term *–j* indicates that the observation was removed from the reference set prior to the computation of the parameter estimate. Although the statistically savvy reader will recognize this analysis as computationally intensive, it is possible by employing the fast updating formulas for the variance-covariance matrix and its inverse (for details see Thisted 1988, p. 52). The probability of impairment is computed as the proportion of reference sites whose “leave-one-out” distance is smaller than that of the test site in question.

Combining estimates

Given estimates of probability of impairment for each line of evidence the problem now becomes combining these together to produce a single weight of evidence estimate. We again assume that the estimate of impact is based on the reference conditions. There are several options available for making the combined estimate. Two approaches involve combining the probabilities for the lines and combining the odds for each of the lines. There are several possibilities for combining information across the different lines by combining the probabilities. The general equation for this over the *k* lines of measurement is calculated as the probability of the union of impairment events (in sediment toxicity *k*=3 corresponding to lines of evidence associated with toxicity, chemistry and biology). A site is declared impaired if it is impaired on any of the modes. Thus,

$$P(\text{site impaired}) = 1 - \prod_{\text{line}=1}^{k} (1 - P(\text{site impaired, line i}))$$

For our weight of evidence problem,

$$\begin{aligned} \text{P(site impaired)} &= 1 - \text{P(site not impaired on any of the 3 lines of evidence)} \\ &= 1 - \{\ [1\text{-P}(\textit{species composition" impaired"})] \\ &\quad \times [1\text{-P}(\textit{toxicity" impairment"})] \\ &\quad \times [1\text{-P}(\textit{chemical " impairment"})] \end{aligned}$$

The formula indicates that the overall probability of site impairment is based on first calculating the probability that the site is not impaired over the different lines, multiplying these together to get an overall probability the site is not impaired, then subtracting from one to get the probability of impairment. Note that with two lines this is the more familiar equation that gives the probability as the union of events (under independence)

$$\text{P(site impaired)} = \text{P(impaired line 1)} + \text{P(impaired line 2)} - \text{P(impaired both lines)}$$

and is based on computing the individual probabilities and subtracting off the joint probability. There are also some variations on this approach that might be used, similar to that used to estimate joint toxicity using individual estimates of toxicity. The above product approach works if the measurements are independent, which is a difficult assumption. One adjustment is to make bounding assumptions. The most extreme estimate of impairment is to assume no intersection and results in simply a sum i.e.,

$$P_{sum}\text{(site impaired)} = \sum_{\text{lines}} \text{P(site impaired, line i)}$$

A drawback to the use of this equation is that the probability may exceed one. An average represents an alternative. The least extreme case occurs when there is complete dependence. In this case, one or more of the events is contained inside another event. The estimate of the impairment probability is then greater or equal to the maximum probability across the lines

$$P_{\max}\text{(site impaired)} = \max(P_i)$$

where P_i is the probability of impairment on line i.

Another approach is based on combining odds ratios. We can convert the individual probability of impairment from each line into an odds ratio, $O_i = \frac{P_i}{1-P_i}$ and compute the final odds ratio by using the Bayes rule of updating evidence as the product of the odds ratios corresponding to each line, i.e., $O = \prod_{i=1}^{k} \frac{P_i}{1-P_i}$. The advantage of this approach is that it allows one to easily incorporate the prior odds of impairment (O_{prior}) of the site by simply multiplying $O = O_{prior} \prod_{i=1}^{k} O_i$. We can account for the effect that the evidence from different sources (lines) may overlap by introducing weights as follow. Let $O = O_{prior} \prod_{i=1}^{k} O_i^{W_i}$, where the weight W_i for a given line should be the number between 0 and 1, with smaller values reducing the influence of the corresponding odds ratio. The weight factors can be computed as $W_i = 1 - R^2(D_i \mid D_1,..,D_{i-1})$, where $R^2(D_i \mid D_1,..,D_{i-1})$ is the coefficient of multiple correlation for the regression of the reference distance matrix (extended in a single vector with only upper diagonal elements) associated with *i*-th line regressed on similarly defined distance matrices associated with the lines already accounted for in the combined odds ratio product. Therefore, the weights will account for the relationships among various sources of evidence and effectively down-weight the evidence that does not add much to the information about the disparity between the test and reference sites that was already taken into account. The computations obviously will depend on the order of lines of evidence in the product of odds ratios; therefore we have to use some expert information about relative importance of different lines. For example, with our three lines of comparison, we can use first the species composition data, then toxicity and finally chemistry data.

In cases where univariate scores are calculated for each measurement within a line of evidence, these would first have to be combined. Estimating impairment probability based on multivariate distances makes deviations from the center of reference sites (or the reference clusters) equally unfavorable in any direction. In many cases however, the underlying univariate measures (for example in the case of metals data) have well-defined directions that would cause impairment of the sites. For such a data set we can compute the impairment probability as *one minus upper tail probabilities* from the underlying univariate tests based on normal approximation for the quantities $\hat{d}(x_0, \mu) = (x_0 - \overline{x}) / s$. These univariate probabilities are then combined using the intersection principle

$$P(\text{site impaired}) = 1 - \prod_{i=1}^{\text{m}} (1 - P(\text{site impaired, variable } i))$$

This probability however should be adjusted so as to take into account the effect of multiple testing as follows

$$P_{adj}(\text{site impaired}) = 1 - \chi^2_{2m}(-2Log(1\text{-}P(\text{site impaired}))) ,$$

where χ^2_{2m} is the cumulative distribution function of a chi-squared distribution with *2m* degrees of freedom. This is a standard meta-analytic procedure based on the fact that the null distributions of individual *p*-values from *m* independent tests are independent *Uniform*(0,1) random variables (for details see Hedges and Olkin 1985, p.37).

Example

As an example we consider data from the Great Lakes that was obtained via the BEAST software (Reynoldson *et al*. 1998; 2000). The reference data consisted of 146 reference samples collected in 1992. Although more data are available on different lines of evidence, these data have information from all three lines of evidence. In addition, there were twenty-five samples taken from Collingwood Harbour that we use as the predictive or test sample. These twenty-five samples were taken at nine locations in 1992, 1995 and 1997 (Table 1). Collingwood Harbour is located in the south shore of Nottawasaga Bay, in the southern extension of Lake Huron's Georgian Bay. The site is of interest as it was identified as an Area of Concern (AOC) by the International Joint Commission but was then de-listed in November 1994, following remediation (for details, see http://www.on.ec.gc.ca/glimr/raps/huron/collingwood/intro.html). Contaminants in the sediment resulted from use of the harbor as a location for ship repair with greatest contamination near the shipyard and in the east and west slips. The nine sampling locations are located as follows: 6703, 6704 and 6705 are located in the harbor with 6703 being farthest from the shipyard and 6705 closest. Sites 6706-6708 are located in the east slip and sites 6709-6711 are located in the west slip. For a map of the locations and additional details on remediation history see http://www.ijc.org/boards/wqb/cases/collingwood/collingwood.html.

Several analyses are provided below. The different analyses are compared to illustrate how different assumptions will lead to different estimates of impairment. This is to show the importance of the assumptions and models used for the analysis. First, we consider comparisons of test sites with the entire group of reference sites. No transformations are made. The second model uses transformed chemistry variables. The third model is similar to the second but uses scores from the first two (nontrivial) correspondence analysis axes for the biological data.

Additional models are based on principal components reductions, transformations of the toxicity data and empirical measures of probability. Descriptions of each model are in Table 1. We first comment on the choices that were made.

Chemical Data

Graphical displays of the chemical data for the reference sites suggested the data were not normal. Distributional plots suggested skewness of the distributions and odd observations. The chemical data were preprocessed using a log transformation for all variables. Figure 1 displays the scatterplot matrix for the transformed data. We note that there is evidence of correlations between the chemicals. Also, there are some locations that have very low or non-detectable concentrations. This suggests groupings may be useful for analysis. Two approaches for further processing the data were considered. The first uses the Mahalanobis distance measure and the multivariate normal distribution. An alternative is to first combine the chemical data into a single number. This may be done using toxic units or a principal components analysis. In the principal components analysis, a linear combination is formed and the combination used as the principal variable for analysis. The data from the Great Lakes **reference** sites were standardized treating the data as one group, to give each chemical equal weight. A principal components analysis of the log transformed standardized chemistry data resulted in two components. The two components explained 75% of the variation in the data. A plot of the loadings is given in Figure 2 and suggests the first component consists of all the chemicals except arsenic while the second was comprised primarily of arsenic. Note the loadings are not as high for lead and cadmium. The model might not accurately describe impairment probability at sites with high levels of these contaminants. The test sites were scored on the components using the standardization from the reference sites. The site scores are graphed in Figure 3. A symbol is used to separate the test sites from the references sites. In addition, the reference sites were divided into six different groups and the group is also indicated in the graph. Note that most of the test sites fall to the upper end of the plot suggesting extreme chemical composition. Also there is a split in the display of points. This is associated with the arsenic variable that is at the limits of detection for many of the sites.

Biological Data

There were several possibilities considered for the analysis of the biological data. First one could use the original or transformed data and then use the distance measure approach. Second, the data could be converted to an index then the index used with the distance approach. A third approach is to compute multivariate axes using correspondence analysis and then compute distances using scores from these axes. In the analysis below, we used the first approach only in model 1 and third approach in the remainder of the models. The reason for discarding the first approach was that the effect of the contaminants is to decrease abundance of the majority of the taxa. Due to large variances, the method is only capable of detecting large increases rather than large decreases. A more detailed description is described below.

Toxicity Data

Plots of the toxicity data suggested skewness of the measurements. As with the chemical data, the log transformation greatly reduced the skewness. We did not consider combining the toxicity data into a single component, as the correlations between the different tests were not high. Analysis was based on T^2 tests or distances with probabilities computed based on empirical distances. We also considered analyses using log transformed variables.

Combined Estimates

Table 1 compares overall impairment estimates from six different models. What is apparent is that considerable differences in the estimates can occur as a result of different analysis options. For example, for the first sample, the probability is estimated as zero using the first model while the second model estimates probability of impairment as close to 1.0. The difference in estimates is a result of the way the species data are treated. In the first model, the biological data are treated as normal while in the second model, the biological data are first transformed using correspondence analysis and the first two components used in the analysis. The new variables from the correspondence analysis are treated as having a normal distribution and the T^2 approach applied. It is instructive to compare the individual estimates of impairment probability from the different lines of evidence for the two models. For model 1, the individual estimates are 0.99 for toxicity, 0.49 for chemistry and 0.00 for community structure. Note that even though the estimate for toxicity is practically 1.0, the community structure estimate is close to 0 and dominates the weight of evidence analysis. For model 2, the estimates are 0.99 for toxicity, 0.43 for chemistry and 0.25 for community structure. In this case, the toxicity measure dominates the analysis and results in an overall estimate of 0.99. The difference in the overall estimate is mostly a result of the contribution from the biological data. This example also illustrates the importance of considering the individual components as well as the overall estimate. The overall estimate for model 2 reflects a different view of the biological data. The distance measure based on untransformed data (model 1) will detect change if the abundance at the test site is different from the reference mean relative to the variance. Variance in abundance tends to be relatively large for most taxa hence detection of change is made difficult. The coefficient of variation for individual taxa abundance ranges from 1.5 to 15 with a mean of 7.5. Since the standard deviation is greater than the mean, it will be difficult to detect a difference unless there is a large change in abundance for a single taxon. Since it might be expected that the effect of metals contamination is to reduce abundance, the problem is made more difficult as zero is less than one standard deviation away from all the mean abundances. For example, the high probability for 6704 in 1997 results from high abundance of *Tubificidae*. The high value for 6711 in 1992 results from a high value of *Caenidae*. Interestingly the mean abundance at this site is only 0.6. While this is not high, the mean for that taxon in the reference sites is only 0.11. Sample 6708 in 1992 (which is similar to 6711 in 1992) is not considered impaired even though almost all taxa have zero abundance (only two taxa are present)! Thus, this modeling approach may be useful for detecting effects that cause an increase in abundance (perhaps due to sewage or deposition) but is poor for detecting change due to toxicity.

Correspondence analysis tends to produce better results, as it will produce “average” taxa. Because the primary effect of the metals is reduction in abundance, the effect is more readily detected using the correspondence analysis scores, as the sites with low abundance tend to have smaller scores than the correspondence analysis reference site scores.

Another factor that seems to be important is whether or not the variables are transformed. For example, compare samples 6704 in 1997 and 6706 in 1995 for models 3 and 3a. The difference in models is only a transformation of the toxicity data. Thus the increase in probability results from higher probability from the toxicity results. In the first case, the probability associated with the toxicity changes from 0.76 to 1.0 and in the second from 0.62 to 0.99. The increased probability of impairment from the toxicity line of evidence results in a large change in the overall probability of impairment for these two samples. Similar results hold when chemistry variables are transformed. Other modeling changes do not seem to have as great an effect on the estimates. For example, model 4 uses a subset of the reference sites for comparisons. Estimates for this model are not that different from those for model 3. The use of empirical estimates of probability rather than model based (i.e. based on T^2) results in a moderate change for some of the samples. The use of principal components in the analysis of the metals data can lead to differences. For example, sample 6706 in 1995 suggests much lower impairment using the component approach than the full variables approach (compare estimates for model 2 and 3).

A difference in the estimates from different models suggests the importance of considering how different lines of evidence contribute to the estimate. Probabilities of impairment for the three lines of evidence and the combined estimate are displayed in Figure 4 for the twenty-five samples in the test set using model 6. In model 6, the probabilities for metals were calculated by combining probabilities calculated from principal component (two components) distances. The distances were calculated by first selecting a cluster of reference observations based on covariates and the estimates of probability were based on distances within the reference cluster. For the toxicity data we calculated the multivariate distance and computed the probability based on the overall empirical distance distribution using all reference distances. Finally, for the biological (community structure) data, correspondence analysis was used and distance and probability evaluated on the first two axes using the empirical distance approach. Figure 4 displays the overall impairment probabilities as well as estimates from individual lines. We note that for most of the sites, the probability of impairment is high. Exceptions are for 6703 in 1995, 6704 in 1992, 6704 in 1995, 6705 in 1995 and 6706 in 1995. Sites 6707 through 6711 appear impaired in all years. The individual impairment probabilities are highest for metals and sediment toxicity tests. The 1992 estimates are variable across the different lines of evidence. The biological impairment probabilities are not generally high and perhaps are a result of not comparing abundance with a more closely matched subset of reference sites. Most of the metal data indicate high levels for 6703 and 6704 but are generally high for the other sites.

The results given here are summaries of the different components of each line of evidence. A further, valuable component of an analysis would be to study which components were important to the individual line of evidence. For example, the ten toxicity tests are all not equally important to the impairment estimate for toxicity. One approach would be to remove individual components then to evaluate the effect on the impairment probability. In this way the components may be evaluated in terms of importance to the probability estimate (see for an example, Smith *et al*. 1990).

Discussion

We have presented an approach for estimation of the probability or risk of impairment for a site based on multiple lines of evidence. Many variations on the approach are possible based on different distances and different summarization of the information. Decisions about distance measures and summarization are best made in the initial stages of the study. Although the results of the weight of evidence analysis summarize impairment in terms of a single number, this approach is generally restrictive. A single measure attempts to summarize the multivariate degree of impairment. It is certainly possible that several different scenarios will lead to the same or similar measure of impairment. Hence it is important to consider the individual lines of evidence as well as the data themselves. Graphical display of the data is necessary to check for problems and assess assumptions. Biological and environmental evaluation of the collected data is also necessary to verify that the evaluation is scientifically correct as well as statistically valid.

We envision the above approach to be most useful for comparing and ranking sites. As Figure 4 illustrates, it is possible to display the estimates of impairment for different lines and a combined estimate for several sites/times in a single display. The information plotted may be ordered in time or space to look for change in impairment probabilities. This might be useful for restoration/recovery studies. The information for different sites may also be compared to identify sites with greatest risk or trends in impairment. For example, if the sites were located along a toxic gradient, then one would expect to see increases in the toxicity estimate of impairment and this could be displayed on the diagram.

Although we have focused on a statistical approach based on estimation of the probability of impairment, other approaches are available for obtaining a combined estimate of impairment.

One common approach is an index-oriented method. In this approach, the numerical values are combined, possibly after a transformation. One example of this approach is Wildhaber and Schmitt (1996). They combine data over biological, toxicological and chemical lines. To do this the chemistry data are standardized by calculating the ratio of the bioavailable component of the contaminant to the chronic toxicity water quality criterion. The toxicological data are standardized by adjusting the test endpoint for the control endpoint. Values may then be averaged over each of the lines of evidence. The biological data are not evaluated in that manner, rather tolerance values are used and a biological index is computed as a tolerance weighted average. The values are then combined over the three lines of evidence. A variation on this approach is given in Soucek *et al.* (2000) and Cherry *et al.* (2001). In these papers the data are not combined within each line of evidence. Rather, important metrics from each line are selected then the metrics are combined into an overall index.

An important aspect of the analysis is the choice of distance measure. Our approach is to use a distance measure that is directly related to a probability distribution. It is therefore important that the assumptions of the distribution be checked so the distance measure is directly related to probability. It is also important that the distance measure reflects impairment in that the farther away from the center, the more impaired the site. One approach to achieve this is to base distance on ordinated data rather than on actual observations. For example, when dealing with chemical data an approach is to calculate principal components for the data from the reference sites. Then decide if the location of the test site is indicative of impairment. A distance measure that reflects a directional distance would be appropriate. Distance might be calculated as zero if the site is on the safe side of the distribution and the ordinary distance measure used if the site is located on the not-safe side. Principal components would be computed based on standardized data to give all the chemicals equal weight in the derivation of the components.

Acknowledgements

This research was supported in part by funds from the International Lead Zinc Research Organization, Inc. The authors are thankful for encouragement from Allen Burton and for comments from Peter Chapman, Valerie Forbes and Trefor Reynoldson on previous drafts. Their struggles to get statistical minds to think environmentally were greatly appreciated.

References

Chapman PM. 1996. Presentation and interpretation of Sediment Quality Triad data. Ecotoxicology 5: 327-339

Cherry DS, Currie RJ, Soucek DJ, *et al.* 2001. An integrative assessment of a watershed impacted by abandoned mined land discharges. Envir Poll 111:377-388

Gilbert RO. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York, NY, USA

Good IJ. 1988. Statistical evidence. In: Kotz S and Johnson NL (eds), Encyclopedia of Statistics vol. 8, pp 651-655. John Wiley and Sons, New York, NY, USA

Gray JS, Clarke KR, Warwick RM, Hobbs G. 1990. Detection of initial effects of pollution on marine benthos: An example from the Ekofisk and Eldfisk oilfields, North Sea. Mar Ecol Prog Ser 66:285-299.

Hedges LV, Olkin I. 1985. Statistical Methods for Meta-Analysis. Academic Press, London, UK

Kass RE, Raftery AE. 1995. Bayes factors. Jour Amer Stat Assoc 90:773-795

Legendre P, Legendre L. 1990. Numerical Ecology, 2nd ed, Elsevier, Amsterdam, Holland

Mardia KV, Kent JT, Bibby JM. 1979. Multivariate Analysis. Academic Press, London, UK

McCullagh P, Nelder JA. 1989. Generalized Linear Models, 2nd ed, Chapman and Hall, London, UK

Rencher A. 1995. Methods of Multivariate Analysis. John Wiley & Sons, New York, USA

Reynoldson, TF, Day, KE. 1998. Biological guidelines for the assessment of sediment quality in the Laurentian Great Lakes. NWRI Report 98-232, Dept of Fisheries and Oceans Canada, Burlington, ON, Canada

Reynoldson TB, Day KE, Pascoe T. 2000 The development of BEAST: A predictive approach for assessing sediment quality in the North American Great Lakes. In: Wright JF, Sutcliffe DW, Furse MT (eds), Assessing the Quality of Freshwaters: RIVPACS and Other Techniques, Chapter 11, pp 165-180. Freshwater Biological Association, Ambleside, UK

Smith EP. 1998. Randomization methods and the analysis of multivariate ecological data. Environmetrics 9:37-51

Smith EP, Pontasch KW, Cairns J Jr. 1990. Community similarity and the analysis of multispecies environmental data: A unified statistical approach. Water Res 24:507-514

Smogor RA, Angermeier, PL. 1999. Relations between fish metrics and measures of anthropogenic distrurbance in three IBI regions in Virginia. In: Simon TP (ed), Assessing the Sustainability and Biological Integrity of Water Resources Using Fish Communities, pp.585-610. CRC Press, Boca Raton, FL, USA

Soucek DJ, Cherry DS, Currie RJ, *et al*. 2000. Laboratory to field validation in an integrative assessment of an acid mine drainage-impacted watershed. Environ Toxicol Chem 19:1036-1043

Thisted RA. 1988. Elements of Statistical Computing. Chapman and Hall, London, UK

Wildhaber MW, Schmitt CJ. 1996. Hazard ranking of contaminated sediments based on chemical analysis, laboratory toxicity tests, and benthic community composition: Prioritizing sites for remedial action. J Great Lakes Res 22:639-652

Figure 1. Scatterplot matrix of log transformed chemical reference measurements. The following abbreviations are used: LAS=log(arsenic), LCD=log(cadmium), LCR=log(chromium), LCU=log(copper), LNI=log(nickel), LPB=log(lead) and LZN=log(zinc).

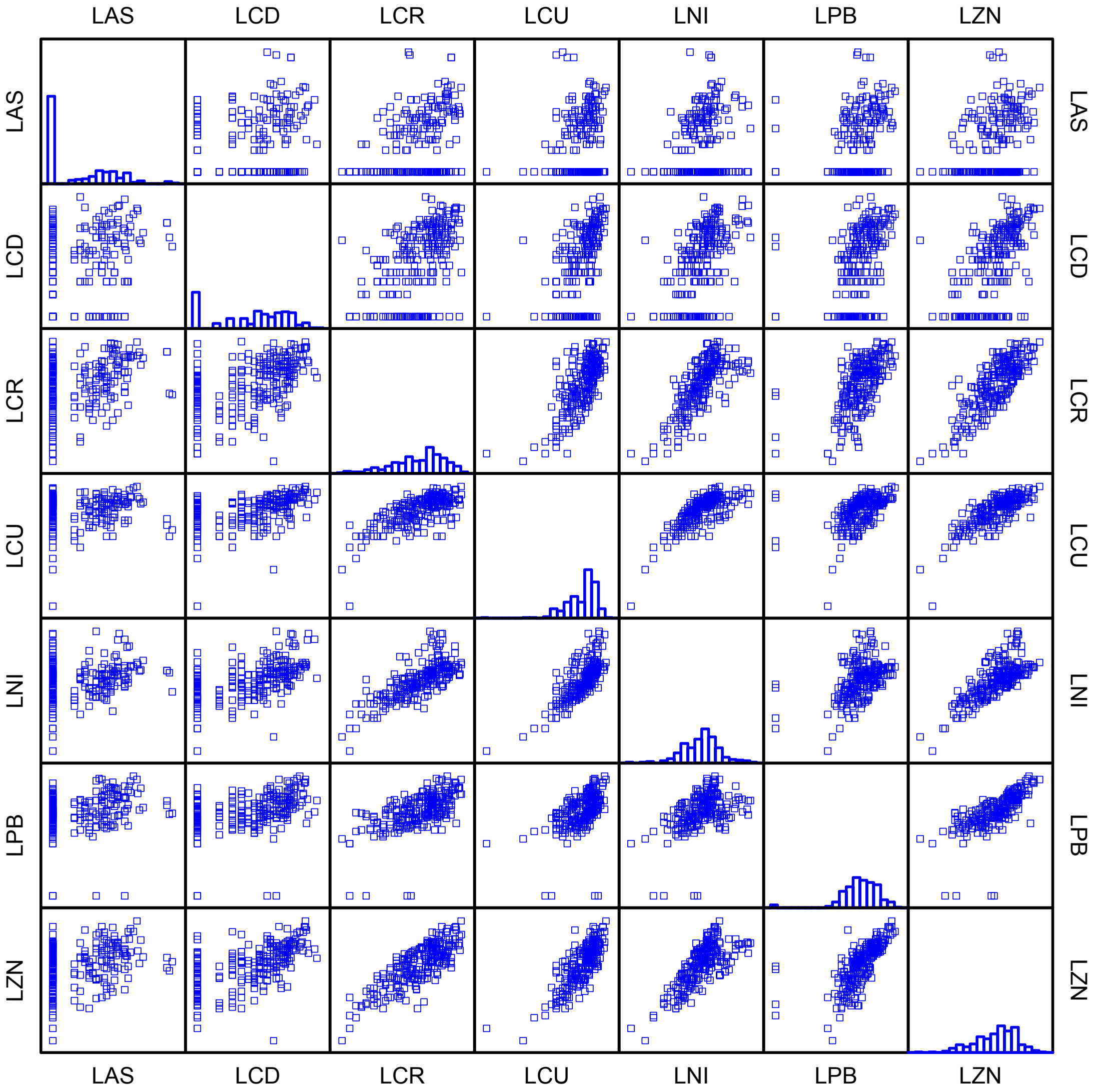

Figure 2. Plot of loading for the principal components for the metal data. The coordinates (loadings) are: LAS (0.36, 0.92), LCD (0.70, -0.06), LCR (0.89, 0.04), LCU (0.86, -0.22), LNI (0.87, -0.11), LPB (0.73, 0.03) and LZN (0.93, -0.05).

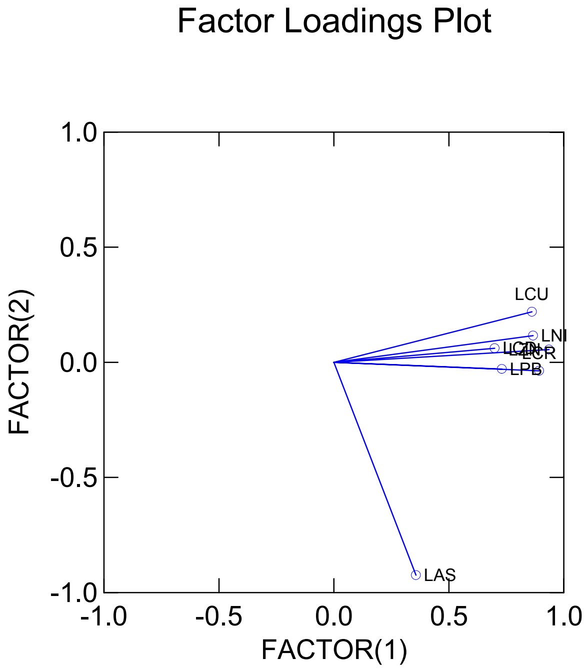

Figure 3. Plot of site scores for reference and test sites. Symbols are coded based on whether they are reference or test (r=reference, test=test) and the number after the “r” indicates the reference cluster.

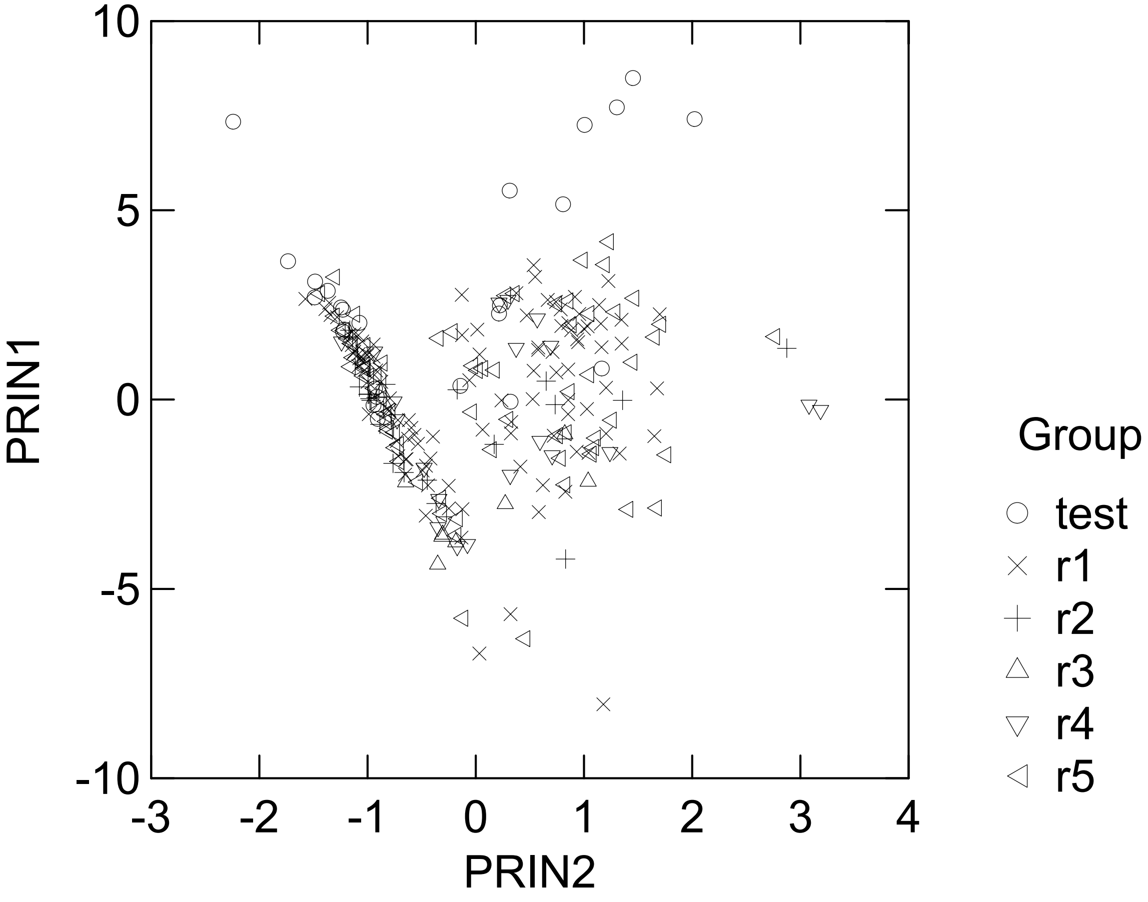

Figure 4. Plot of probabilities for three lines of evidence and overall estimate for sites in test data set. The center corresponds to zero and tick marks represent tenths. The first four numbers in the site/sample label indicate the site number while the last 2 provide the year of sampling.

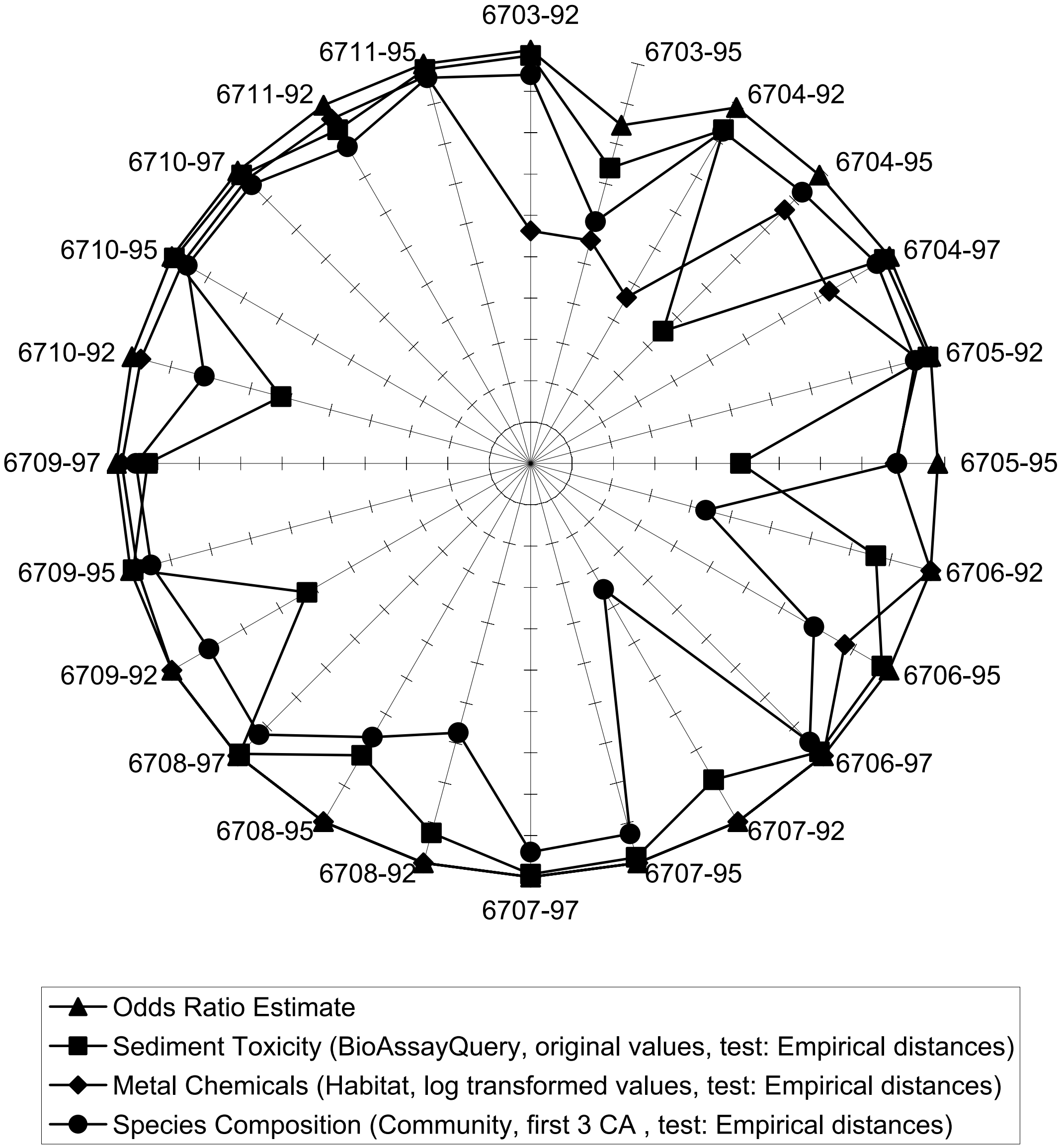

Table 1. Comparison of impairment probabilities using different analysis options. Hotelling T^2 used in procedure 1-4, empirical distance approach used for models 5 and 6.

Procedure 1: No transformations

Procedure 2: Log transformation for metals, correspondence analysis (2 axes) for biological data,

Procedure 3: Principal components on log transformed metals (2 axes), correspondence analysis for biological data, no transformation for toxicity data

Procedure 4: Same as 3 except that log transformed toxicity data used

Procedure 5: Same as 3 except metals data are compared with a subset of the controls

Procedure 6: Same as 4 using empirical distances to determine probabilities

Procedure 7: Same as 5 but using log transformed toxicity data and 3 correspondence analysis axes.

Site-year	Procedure 1	Procedure 2	Procedure 3	Procedure 4	Procedure 5	Procedure 6	Procedure 7
6703-92	0.0000	0.9928	0.9910	0.9999	0.9909	0.8602	0.9993
6703-95	0.0000	0.6670	0.6290	0.4040	0.6412	0.6558	0.8461
6704-92	0.1535	0.8469	0.8609	0.9712	0.8631	0.6707	0.9933
6704-97	0.9997	0.7485	0.3586	0.9998	0.3849	0.3527	0.9856
6704-95	0.0000	0.5664	0.0126	0.0064	0.0146	0.0210	0.9999
6705-92	0.0000	1.0000	0.9997	1.0000	0.9996	0.9887	1.0000
6705-95	0.0000	0.5570	0.1745	0.1745	0.1624	0.3282	0.9829
6706-92	0.0000	1.0000	0.9998	1.0000	0.9998	0.9989	1.0000
6706-97	0.0005	1.0000	0.9974	1.0000	0.9972	0.9882	0.9992
6706-95	0.0000	0.8196	0.0184	0.9449	0.0095	0.0068	1.0000
6707-92	0.0000	1.0000	1.0000	1.0000	1.0000	0.9995	1.0000
6707-97	0.0000	1.0000	0.9953	1.0000	0.9948	0.9902	1.0000
6707-95	0.0000	1.0000	0.9983	1.0000	0.9981	0.9951	1.0000
6708-92	0.0000	1.0000	0.9998	1.0000	0.9998	0.9992	1.0000
6708-97	0.9976	1.0000	1.0000	1.0000	1.0000	0.9998	1.0000
6708-95	0.0001	1.0000	0.9999	0.9999	0.9999	0.9984	1.0000
6709-92	0.0000	1.0000	0.9009	0.9148	0.8937	0.9468	1.0000
6709-97	0.0000	1.0000	0.9762	0.9957	0.9735	0.9653	1.0000
6709-95	0.0000	1.0000	0.9732	1.0000	0.9699	0.9407	0.9999
6710-92	0.0000	0.9972	0.6238	0.7503	0.5986	0.9019	0.9964
6710-97	0.9500	0.9999	0.9844	0.9999	0.9827	0.9377	1.0000
6710-95	0.0002	0.9998	0.9363	1.0000	0.9231	0.7486	1.0000
6711-92	1.0000	0.9875	0.6084	0.9936	0.5837	0.7876	0.9996
6711-97	0.9977	0.9999	0.7001	0.5816	0.6761	0.8147	1.0000
6711-95	0.0000	0.9997	0.9298	0.9996	0.9165	0.7990	0.9996

Table 2. Prediction of impairment from different lines of evidence using empirical distances with log transformed metals, log transformed toxicity data and two correspondence analysis axes (Procedure 7 from Table 1).

Site-year	**Odds Ratio Estimate**	**Sediment Toxicity (Toxicity, log transformed values, test: Empirical distances)**	**Metal Chemicals (log transformed values, test: Empirical distances)**	**Species Composition (Community, first 3 CA , test: Empirical distances)**
6703-92	0.9993	0.9863	0.5622	0.9399
6703-95	0.8461	0.7397	0.5579	0.6052
6704-92	0.9933	0.9315	0.4635	0.9270
6704-97	0.9856	0.4521	0.8670	0.9270
6704-95	0.9999	0.9863	0.8326	0.9657
6705-92	1.0000	0.9932	0.9700	0.9614
6705-95	0.9829	0.5068	0.8798	0.8841
6706-92	1.0000	0.8630	1.0000	0.4378
6706-97	0.9992	0.9795	0.8755	0.7897
6706-95	1.0000	0.9863	1.0000	0.9528
6707-92	1.0000	0.8836	1.0000	0.3519
6707-97	1.0000	0.9863	1.0000	0.9270
6707-95	1.0000	0.9932	1.0000	0.9399
6708-92	1.0000	0.9247	1.0000	0.6738
6708-97	1.0000	0.8151	1.0000	0.7639
6708-95	1.0000	0.9932	1.0000	0.9270
6709-92	1.0000	0.6233	1.0000	0.8970
6709-97	1.0000	0.9932	0.9828	0.9485
6709-95	0.9999	0.9247	0.9871	0.9528
671-92	0.9964	0.6233	0.9742	0.8155
671-97	1.0000	0.9932	0.9700	0.9571
671-95	1.0000	0.9863	0.9700	0.9528
6711-92	0.9996	0.9315	0.9614	0.8841
6711-97	1.0000	0.9863	0.9700	0.9657
6711-95	0.9996	0.6849	0.9785	0.9657