Appendix E: Statistical Analysis Methods for Biological Assessment

Introduction

A central premise of biological assessment is comparison of the biological resources of a waterbody to an expected reference condition (Figure E-1). Impairment of the waterbody is judged by its departure from the expected condition. This approach presumes that the purpose of management is to prevent, identify, and subsequently repair anthropogenic damage to natural resources. Biological assessment of waterbodies is predicated on our ability to define, measure, and compare an assessment endpoint between similar systems. This guidance outlines analytic methodologies to perform two tasks shown in Figure E-1:

- Characterization of the biological expectation.
- Determining whether a site deviates from that expectation.

Figure E-1. Graphical representation of bioassessment. Assessment sites a and b are compared to an ideal biological expectation. Site a is near to its expectation; Site b deviates from it and is considered to be impaired.
All of the methods considered here use the same general approach: sites are assessed by comparing the assemblage of organisms found at a site to an expectation derived from observations of many relatively undisturbed reference sites (Figure E-1). The expectations are modified by classifying the reference sites to account for natural variability, and each assessment site is classified using non-biological (physical, chemical, geographic) information. Biological variables are tested for response to stressors by comparison of reference unimpaired sites and known impaired sites. A set of "rules" are developed from this information, which are then used to determine if the biota of a site deviate from the expectation, indicating that the site is impaired.

Several analytic methods have been developed to assess the condition of water resources from biological data, beginning with the saprobien system in the early 20th century to present-day development of biological markers. This appendix outlines three methods for analyzing and assessing waterbody condition from assemblage and community-level biological information:

1. Multimetric assessment using an index that is the sum of several metrics. This is the basis of the Index of Biotic Integrity (IBI) (Karr et al. 1986), the Invertebrate Community Index (ICI) (Ohio EPA 1990); the Rapid Bioassessment Protocol (USEPA 1989b); and state indices developed from these (e.g., Southerland and Stirling 1995, Barbour et al. 1996a, Barbour et al. 1996b).

2. Multimetric assessment using an index that is developed from a multivariate discriminant model to discriminate reference from impaired sites. This is the basis of stream bioassessment in Maine (Davies et al. 1993), and of the estuarine invertebrate indices developed by the EMAP-Estuaries program (USEPA 1993e, USEPA 1994h, USEPA 1994i, Engle et al. 1994).

3. Assessment using multivariate ordination of species abundances. This methodology has been used widely in assessment of streams in Britain (e.g., Wright et al. 1984); assessment of marine macroinvertebrates in the North Sea (e.g., Warwick and Clarke 1991); and in assessment of benthic macroinvertebrates in the Great Lakes (e.g., Reynoldson et al. 1995).

Many other methods are possible, as well as permutations of the three methods above, all of which are beyond the scope of this document. The three approaches were selected because:

- They use community and assemblage data.
- The methods are not restricted to any one assemblage. The examples all use benthic macroinvertebrates and fish (freshwater and estuarine), but any other assemblage could also be used, such as phytoplankton, zooplankton or macrophytes.
- The methods are general, and have been used by many agencies in many areas. The examples used to illustrate the methods have also been carried out over wide geographic areas with many sites, demonstrating the generality of the methods.
- The examples used to illustrate the methods are concise. Methods were fully documented, and have been carried to completion, that is, assessment of biological impairment and non-impairment.
• The optimal analysis methodology should also be cost-effective and easy to communicate to managers and the public. Both the multimetric index and the discriminant model index (approaches 1 and 2) are easy to apply in a continuing operational monitoring program because data from an individual site are entered into a formula, and the site's deviation from reference conditions can be known immediately (Gerritsen 1995). The ordination approach (3) requires reanalysis of the entire reference data set for each new batch of monitoring sites. The multimetric index (approach 1) is the easiest to explain to managers and the public because it does not rely on specialized concepts such as multivariate statistics. The ordination approach (3) may be most cost-effective if the biological survey is a single event—a large number of sites are surveyed once, and there is no plan to continue monitoring or to survey new sites.

Characterization of Reference Conditions

Reference conditions establish the basis for comparison and for detecting impairment of waterbodies. They should be applicable to an individual waterbody, such as a stream or lake and also to similar waterbodies on a regional scale (USEPA 1996a).

Classification Tools

The objective of classification is to group similar waterbodies together, so that reference conditions will reflect reasonable expectations for assessing waterbodies. There are two fundamental approaches to classifications: a priori or rule based, where known rules are applied to classifying objects; and a posteriori, or data-based, where rules for classifying objects are derived from data obtained from the objects (waterbodies) themselves (Conquest et al. 1994).

For example, a rule-based classification may divide mountain and lowland streams by elevation or stream gradient. An a posteriori classification would examine data from all streams, and determine if there is a basis for separating them into two or more classes (not necessarily including elevation or gradient). The a posteriori approach requires a relatively large sample of reference sites to derive the classes and rules, with both biological and physical-chemical data from each site.

The basic assumption of classification is that physical habitat and water quality largely determine the composition of biological communities in waterbodies. Therefore, if waterbodies are classified adequately, reference biological community types should correspond to the classification. Classification is often an iterative process of refining the classification scheme as new data are obtained, until a satisfactory classification emerges that accounts for variation in the reference site biological data.

Several statistical tools can assist in site classification, but there is no set procedure. If a priori classification is based on well-developed prior knowledge, then graphical analysis of biological data, followed by any necessary modifications and tests of the resultant classification, may be sufficient.
If a rule-based classification is not self-evident, then it may be necessary to develop an alternative classification from the data using one or more analytical classification approaches. These methods include several cluster analysis methods, and several approaches to ordination analysis, including principal components analysis (PCA), correspondence analysis (CA) and its variants, and non-metric multidimensional scaling (NMDS).

In statistical terminology, each site is a sample unit (SU) (Ludwig and Reynolds 1988). Ideally, sample units should be independent, which is generally achievable in small streams and in many lakes, where each waterbody can be a separate sample unit. Large lakes and reservoirs, large rivers, and estuaries may include several sample units within the same waterbody. For large and complex lakes and estuaries, it may be necessary to define a site as a contiguous basin or embayment. Any portion of the waterbody that is partially isolated from the rest by bottom topography or water motion should be considered a separate site and sampled accordingly. This also applies to the three zones of large reservoirs (riverine, transition, and forebay) and to salinity zones of estuaries (e.g., fresh, mesohaline polyhaline), which have different biological communities and dynamics even though they are not hydrologically isolated (Thornton 1990b). Thus, large waterbodies (including large reservoirs) may comprise several sites or SUs. Sites (SUs) are considered independent and are kept separate in analysis; no "average" is estimated for a multiple-site waterbody. Multiple sites are not strictly independent and will need to be considered carefully in reference condition characterization and in metric response evaluation.

Large rivers may be more problematic in that sites on a river are serially linked by water flow. Sites are defined as river reaches of some minimum length that exhibit some (but not complete) independence. Sample units (reaches) may be defined by length (e.g., a set length or a multiple of stream widths), as the reach between major tributaries, or as segments downstream of major impacts and discharges (e.g., urban areas).

**Graphical Analysis**

A key graphical display is box-and-whisker plots (Figure E-2). These show population attributes of the data: central tendency, spread, and outliers. In the display used here, the central point is the median value of the variable; the box shows the 25th and 75th percentiles (interquartile range); and the whiskers show values within the inner fences (Figure E-2). Points beyond the fences may be considered outliers or extreme values. Box-and-whisker plots are simple, straightforward, powerful, and the interquartile ranges are used to evaluate whether there is a real difference between two areas and whether a metric is a good candidate for use in assessment. Graphing the data should always be a first step in data analysis.
Figure E-2. Box and whisker diagram (after Tukey 1977). The box is the interquartile range (25th-75th percentile). Inner fences are the quartiles ± 1.5 x interquartile range; outer fence is 3 x interquartile range. Ends of whiskers are the most extreme observations within the inner fences.

Statistical methods used by biologists are frequently tests of whether two or more populations have different means using t-tests, analysis of variance, or various nonparametric methods. However, the fundamental problem of biological assessment is not to determine whether two populations (or samples) have a different mean, but to determine whether an individual site is a member of the least-impaired reference population (Figure E-1). If it is not, then a second question is how far it has deviated from that reference. Therefore, biological assessment requires the entire distribution of a metric, which is effectively displayed with a box-and-whisker plot.

**Ordination Methods**

The purpose of ordination analysis is to reduce the complexity of many variables (for example, abundance of 200 species from 50 sites) into fewer variables, such that the sites and the species are ordered on the new variables (Figure E-3). The new variables
are called the principal axes of the analysis; the first axis accounts for the most variation in the original data, the second accounts for somewhat less variation, and so on. Typically, only the first two to four axes of the analysis are presented because higher axes contribute little to the variance explained and because one cannot present or conceptualize more than three axes simultaneously.

**Figure E-3. Ordination.** The relationship of species 1 and species 2 can be described by translating and rotating the axes, so that most of the variance is on the first axis. In this 2-dimensional example, the observations have been reduced to a single dimension, the first axis, which is a linear combination of species 1 and species 2.

**Principal Components Analysis** - One of the most commonly used ordinations is principal components analysis (PCA). In PCA, the new variables (principal axes) are linear combinations of the original data; that is, the relationship between each principal axis and the abundance of each species can be expressed as a straight line, as in simple linear regression (Jongman et al. 1987). Thus, PCA is a multivariate extension of linear regression (Figure E-3), making the assumption that a variable will have a maximum value at one end of a principal axis and minimum value at the other. Because the principal axes can be seen as environmental gradients to which the species respond, ordination is also called gradient analysis (Jongman et al. 1987).

The procedure of PCA is an eigenanalysis of the correlation matrix among variables in the original data matrix. The variables may be species abundance, calculated assemblage metrics, or environmental (chemical and habitat) variables. Eigenanalysis results in as many eigenvalues as there are rows (or columns) in the correlation matrix, and each eigenvalue and corresponding eigenvector describes an axis of the ordination. The eigenvalue of an axis is the variance accounted for by that axis. Often, only the first two or three axes explain significantly more variance in the original data than a random axis. Rules for determining the number of significant axes are explained in Jackson (1993b). Details of formulas and calculations for PCA, as well as variations of PCA, are in Ludwig and Reynolds (1988).

Because PCA is linear, and assumes multivariate normal distributions, data transformations are often necessary. Species abundance data usually have many zeros
in the data matrix, and no transformation will normalize them. PCA is not useful for species abundance data, although it can be made to work well for data that are normal or can be transformed to a normal distribution (e.g., environmental variables, assemblage attributes such as number of taxa, etc.).

**Correspondence Analysis Family** - A problem with linear ordinations such as PCA is that species do not always respond linearly to gradients; in fact, a unimodal response to environmental gradients is much more common (Jongman et al. 1987). A unimodal response is one in which a species has peak abundances at certain optimal values of an environmental variable (for example, pH or nutrient concentration) and abundances are lower at both higher and lower values of the environmental variable. There are many examples of environmental optima for aquatic organisms; optima are supported by uptake kinetics, and they form the basis for resource-based competition and seasonal succession (e.g., Tilman 1982).

Multivariate ordination based on unimodal responses to environmental gradients is called correspondence analysis. As in PCA, correspondence analysis also seeks new variables to explain the species abundances on fewer axes and is frequently "detrended" to eliminate a mathematical artifact from its calculation (Jongman et al. 1987).

Ordination can also be done to develop associations between the species abundances and measured environmental variables. In this case, both species abundance and the environmental variables are related to the principal axes and the whole procedure can be regarded as a multivariate, multiple regression. The linear form is called canonical correlation (CC); the unimodal form is called canonical correspondence analysis (CCA). Because it assumes unimodal responses, CCA is thought to be a realistic and robust multivariate ordination (ter Braak 1986, Palmer 1993).

In CCA, each species, site, and environmental variable has a score on each of the principal axes. Results of CCA are presented graphically by plotting the scores on two of the axes (usually the first two) (Figure E-4). Plotting site scores with environmental variable scores shows the relationship between the sites and the environmental variables and can also show clustering of sites.
Nonmetric Multidimensional Scaling - Nonmetric multidimensional scaling (NMDS) is increasing in use in ecological application because it offers several advantages over other ordination methods. Because the ordination works on a matrix of distance ranks, it is distribution-free and hence unaffected by non-normality and nonlinearity in the data (Ludwig and Reynolds 1988). It is robust and produces interpretable ordinations from different ecological data sets. The disadvantages of NMDS are that it is iterative and subject to local minima (SYSTAT 1992) and that no canonical form has yet been developed. It is possible, however, to estimate correlations of environmental (explanatory) variables with the axes of NMDS.

Like cluster analysis, NMDS uses a distance metric among sample units (sites), and results can be sensitive to the choice of the distance metric (Jackson 1993a). Bray-Curtis distance and the relative distance metrics (relative Euclidean distance and chord distance) tend to work best (Kenkel and Orloci 1986, Ludwig and Reynolds 1988).

The objective of NMDS is to obtain a "best fit" between the dissimilarity measures and the distances calculated in ordination space. The dissimilarities have as many dimensions as there are sites, but the ordination reduces these to a smaller number, usually 2 or 3. The procedure is to rank the distances in the similarity matrix from smallest to largest, then to calculate an initial starting ordination (termed the initial
configuration) directly from the dissimilarity matrix. Intersite distances are calculated from the initial configuration, ranked, and compared to the ranked dissimilarities. A best solution is sought iteratively, changing the configuration so that the two rankings (dissimilarities and configuration) become more similar. Goodness of fit of the configuration to the dissimilarities is measured by Kruskal's stress coefficient (Ludwig and Reynolds 1988) or Guttman's coefficient of alienation (SYSTAT 1992). Iterations stop when stress or alienation reaches a minimum value.

NMDS is available on many commercial statistical software packages. Distance measures used by ecologists, especially Bray-Curtis distance and chord distance, are not usually available in these packages and must be calculated separately. Relative Euclidean distance is also only rarely available; however, if an input matrix of percent abundances of species is used, then Euclidean distance will yield relative Euclidean distance.

Results from NMDS are a final configuration, consisting of coordinates for each site in the 2 or 3 dimensional ordination. As in other ordinations, points close to each other in the ordination space (Figures E-3 and E-4) represent sites with similar species composition.

Classification Analysis - Classification, or the placement of objects into categories, is an innate human activity. A wide variety of formal classification procedures have been developed (see Gauch 1982 for a review). Only two will be discussed here, cluster analysis and two-way indicator species analysis (TWINSPAN).

Cluster Analysis - Cluster analysis is known as an agglomerative classification, that is, it successively builds clusters until all objects have been joined in a single cluster. Cluster analysis begins with a matrix of intersite dissimilarities. The smallest dissimilarity in the matrix is selected and those two sites are joined in a cluster. The algorithm then calculates the dissimilarities between the new cluster and all other sites or clusters. Again, the smallest dissimilarity is selected, the two objects are joined, and the process repeats itself until all objects are joined. Results can be shown in a dendrogram (Figure E-5), where the bars connecting clusters represent the dissimilarity between them. Final clusters are identified by choosing a cutoff dissimilarity value. The cutoff dissimilarity value clearly affects the number of clusters (Figure E-5): it may range from one to the number of sites. The number of clusters should be small, and should explain as much variance of the biological data as possible.

Classification with cluster analysis is not as straightforward and objective as is implied by a dendrogram produced by a mathematical algorithm. First, several algorithms may be used for recalculating dissimilarities among agglomerated clusters of sites, and each algorithm may produce different results. A favored algorithm for ecological data is the unweighted pair-group method (UPGMA) (Ludwig and Reynolds 1988, Reynolds et al. 1995). Second, the dissimilarity measure affects results. As in NMDS analysis, relative dissimilarity measures (relative Euclidean, chord distance) and Bray-Curtis distance work best for species-abundance data (Ludwig and Reynolds 1988). Finally, as noted above, selection of a distance cut point for defining clusters is subjective (Figure E-5).
Figure E-5. Dendrogram from cluster analysis. Cutpoints a, b, c are at distances 0.75, 1.0, 1.25, respectively, and result in 5, 4 and 2 clusters, respectively.

Two-way Indicator Species Analysis (TWINSPAN) - TWINSPAN was developed by Hill (1979), and is a divisive technique. Instead of building up clusters from individual sites, divisive methods start with the entire data set and divide it into two. The division process is repeated until a specified number of clusters are obtained (Gauch 1982). TWINSPAN first ordinates the data, then divides the sample into two clusters near the middle of the first ordination axis. Ordination is by reciprocal averaging, which is a variation of correspondence analysis. New ordinations are repeated on each daughter cluster, and the daughters are in turn divided on their first ordination axis. TWINSPAN is only available in specialized software packages.

Discriminant Model - The objective of a discriminant model is to predict community type, or community composition, from non-biological data. Development of such a model requires a data set with both biological and non-biological data, and testing of the model requires a second, similar data set. Discriminant analysis is best illustrated with a simple example (e.g., Ludwig and Reynolds 1988, Johnson and Wichern 1992). Suppose that abundances of two species are examined in riffle and pool sites of streams (Figure E-6) and we wish to develop a model that will discriminate between riffle and pool sites, using only the biological data. As shown in the figure, pool sites tend to have greater abundances of both species. Using either species alone to form the rule would lead to frequent errors. Discriminant analysis finds a best fit straight line to separate the groups; the heavy line of Figure 6 is the border and the hatched line perpendicular to it is the discriminant function. Sites with positive scores are more likely to be pools, and sites with negative scores are more likely to be riffles.
Figure E-6. Illustration of discriminant function analysis. Neither species A nor species B can be used alone to distinguish riffle from pool sites. Discriminant analysis estimates a linear border between the two site classes (heavy line), and a discriminant function (graduated line). The discriminant function is a linear combination of the input variables (species A and species B), and yields a probability that a site belongs to the riffle or pool class.

Discriminant function analysis involves computation of a pooled variance-covariance matrix of the groups, and solving for the coefficients of the discriminant function. Formulae and computations are shown in Ludwig and Reynolds (1988), Johnson and Wichern (1992), Pielou (1977), and other multivariate statistics textbooks. Discriminant analysis also allows calculation of multivariate distance (Mahalanobis D2) between groups, and an F-test for group differences. A limitation of discriminant function analysis is that it is linear; strong nonlinearity of the data will reduce its power to separate groups.

Rule-Based Classification: Characterization of Reference Conditions

The objective of reference characterization is to describe (characterize) each of the reference classes in terms of biological indicators and other descriptive variables. The first step is to support or reject the a priori classification, followed by modifying it to arrive at a parsimonious and robust classification; that is, one with the fewest classes that explains the most variance in the reference data set.

There is no single "best" classification nor are resources generally available to determine all possible differences between all waterbodies in a region. The key to classification is practicality within the region or state in which it will be applied; local conditions determine the classes. Classification will depend on regional experts familiar with the range of conditions in a region as well as biological similarities and differences among
waterbodies. Ultimately, classification can be used to develop a predictive model of those chemical and physical characteristics that affect the values of the biological metrics and indices in reference sites.

A useful classification scheme is hierarchical, beginning at the highest (regional) level and stratifying as far down as necessary (Conquest et al. 1994). The procedure is to classify waterbodies according to region and then to increase the stratification in the classification hierarchy to a reasonable point for the given region. Although several classification levels are possible, in practice, only one, or at most two, relevant levels would typically be used. Classification should avoid a proliferation of classes that do not contribute to assessment. One or two relevant levels of the hierarchy will yield the best classification scheme. Potential hierarchical classifications for streams, lakes, and estuaries, respectively, are given in Gerritsen (1995), USEPA (1996a), and USEPA (1997a).

**Confirmation of a priori Classification**

*Univariate Tests* - Univariate tests of classifications include all the standard statistical tests for comparing two or more groups: t-test, analysis of variance, sign test, Wilcoxon rank test, and Mann-Whitney U-test (USEPA 1996b, Ludwig and Reynolds 1988). These methods are used to test for significant differences between groups (classes) to confirm or reject the classes. They are univariate, with a single dependent (response) variable. Biological variables (metrics) may require transformation to meet assumptions of t-tests and ANOVA, or non-parametric tests (e.g., rank tests, Mann-Whitney) may be used. See USEPA (1996b) for discussions on the use of these and other univariate tests for biocriteria. Failure to confirm the classification for any single response variable does not mean that it will fail for other response variables. Because assessment is based on multiple variables (metrics or species composition), multivariate tests might be more convenient than a succession of individual tests.

*Discriminant Analysis* - Discriminant analysis can be used as a form of multivariate, one-way analysis of variance that tests differences between a set of groups based on several response variables. It is used as a test of classifications (Conquest et al. 1994), provided that the assumptions of linearity and normality are met. Many statistical software packages provide discriminant analysis.

*Gradients* - On occasion, environmental gradients might not allow formation of discrete site classes. For example, the number of zooplankton taxa in lakes is usually related to lake size (e.g., Dodson 1992). Similarly, fish and invertebrate number of taxa in streams is typically related to stream size (order, discharge or watershed area) (e.g., Ohio EPA 1987, DeShon 1995)

*Ordination* - The *a priori* classification may also be confirmed with one of the ordination methods. Sites are plotted in ordination space using different symbols for the *a priori* classes. If classes overlap completely in ordination space, then there is no apparent difference in their species composition (or other variables used in the ordination), and it may be appropriate to aggregate the coinciding classes. Species or variable scores can be plotted in ordination space to determine which contribute most to separation among classes. Correlation coefficients of environmental variables with the site scores will show...
if there are environmental gradients that are associated with the ordination and with the site classes. Examples and detailed methods for ordinations are given in Jongman et al. (1987) and Ludwig and Reynolds (1988).

**A Posteriori Classification**

This method of classification determines classes from the structure of the data, rather than from pre-existing knowledge or hypotheses. Because the principal goal of classification in biocriteria programs is to account for biological variation, the biological data (typically species composition data) are used for classifying. As with *a priori* methods, only data from reference sites are used to develop the classification (e.g., Moss et al. 1987, Wright et al. 1984, Reynoldson et al. 1995).

Test sites must also be assigned to appropriate classes, so that they can be compared to reference sites. Because anthropogenic degradation affects the biota of the waterbodies, assigning test sites to classes using their biological data may lead to incorrect classification (Figure E-7). Therefore, the classification also requires a method to assign test sites to classes, using non-biological measures that are not affected by anthropogenic degradation. Following an *a posteriori* classification, this is typically a discriminant function model that is constructed from the reference data set (Norris 1995).

![Figure E-7. Misclassification of test sites. A test site (x) that was originally in assemblage Type A has been degraded (arrow). If biological data are used to classify the test site, then it would be classified as Type B because it is now more similar to Type B. If, on the other hand, non-biological measures that are not affected by degradation are used to classify the test site, then it would be correctly identified as Type A and the degree of biological degradation could assessed.](image)
Identifying Classes

Classification is a subjective activity even when it is done with seemingly objective quantitative methods. The subjectivity is due in part to the information that will be used to decide if objects are similar or not, and in part to the methods and their variations that will be used to classify the objects. For example, we may say that Miami is similar to Havana. We may also say that, during the Cold War era, Havana was similar to Moscow. Does it then follow that Miami is similar to Moscow (SYSTAT 1992)? This example illustrates that the variables used to determine similarity (climate, economic system) profoundly affect the resultant classification.

There are several different quantitative methods to classify objects, each of which may result in different classification. Furthermore, each classification method requires subjective decisions on the similarity measure to use in the classification and on the number of classes to identify. Thus, classification remains subjective, even when done with seemingly quantitative algorithms. Classifications developed from biological data should make sense in the physical and chemical context of the habitats. A posteriori classification is developed from the biological data set. Species abundance data are examined, and groups of sites are identified that are similar to each other. Usually, this is done with a similarity (or dissimilarity) measure and a form of cluster analysis. Subjective decisions are required to select the classification methodology, the similarity measure, and the number of groups to identify.

As was stated above, the general objective of classification is parsimony of classes (few classes) to obtain a large partitioning of variance among the classes. Too few classes results in large variability within each class, and too many classes results in trivial differences among classes.

Assigning Test Sites to Classes

After reference site classes have been determined, using cluster analysis or some other a posteriori classification a model is developed to enable test sites to be assigned to one of the reference classes. This is typically a discriminant model developed from non-biological data of the reference sites. Data for the discriminant model should be measurements that are not affected by anthropogenic degradation, such as stream gradient, sinuosity, natural water chemistry, lake depth, watershed soil type, etc. (Norris 1995). The output of a discriminant model is a discriminant function that assigns sites to one of the classes. It is developed from reference site data, and should be tested with an independent reference site data set.

Multimetric Index Method

The indices currently used are variations of the Index of Biotic Integrity (IBI) for fish assemblages in streams, developed by Karr and his co-workers (e.g., Karr 1981, Karr et

Each index is the sum of several (up to 12) standardized component metric scores. Metric scores are usually on an ordinal scale of 1 to 5 (Karr et al. 1986), or 0 to 6 (USEPA 1989b) or as a percentage of the reference metric value (Maxted et al. 1994). Component metrics consist of measures such as total number of taxa, percent abundance of the dominant taxon, number of species and percent abundance of intolerant groups, and percent abundance of functional feeding groups such as planktivorous fish or invertebrate shredders.

**Metric Variability**

Metrics that are too highly variable within the reference sites are unlikely to be effective for assessment. Relative variability is often measured with the coefficient of variability, defined as the standard deviation divided by the mean (expressed as percent):

$$CV = \frac{s}{\bar{x}} \times 100$$

The CV is a measure of how large the variability is compared to the mean. Ideally the CV should be small, which can be achieved with a small variance or with a large mean value. However, some metrics might have low values under reference conditions (e.g., number of exotic species), and CV will always be large for such metrics. For example, if a sample of 10 reference sites, each with 10 taxa, includes a single site with a single exotic species, then the CV of the number of exotic species is over 300 percent. Furthermore, the multimetric approach calls for comparison of metric values to a percentile of the reference population values and is thus a distribution-free approach. Because the CV is the ratio of the sample standard deviation to the mean, it might not adequately express variability for non-normal distributions.

An alternative measure to the CV is the "interquartile coefficient," which is based on quartiles of the reference distribution and the expected change of the metric rather than its parameters (Gerritsen and Bowman 1994). In operational bioassessment, metric values below the lower quartile of reference conditions are typically judged as not meeting reference expectations (e.g., Ohio EPA 1990). The range from 0 to the lower quartile can be termed a "scope for detection." For those metrics with low values under reference conditions and high values under impaired conditions, the scope for detection is the range from the 75th percentile to the maximum possible value (e.g., 100 percent) (Figure E-8).
Figure E-8. Assessing candidate metrics that have (a) high values under reference conditions, and (b) low values under reference conditions.

The larger the scope for detection, compared to the interquartile range, the easier it will be to detect deviation from the reference condition. The "interquartile coefficient" is thus defined here as the ratio of the interquartile range to the scope for detection:

\[ C_{IR} = \frac{IQ}{Ds} \]

where \( IQ \) = interquartile range

\[ Ds = \{ \]

25th percentile (for metrics that decrease with...
The interquartile coefficient is analogous to the CV and is used similarly, but it is bidirectional and is calculated from percentiles in the same way that assessment uses percentiles. In general, an interquartile coefficient greater than 1 indicates excessive variability of a metric.

**Metric Response**

Response of metrics to stresses is evaluated by comparison of reference sites to test sites. The simplest comparison is using box-and-whisker plots of the metric distribution in reference and test sites (Figure E-8) or by univariate tests of metrics in reference and test sites. Alternatively, it may be possible to develop an empirical model of metric response to stressors.

Several approaches are available including multiple regression, canonical correlation, canonical correspondence analysis, and log-linear models (Ludwig and Reynolds 1988, Jongman et al. 1987).

Metrics are judged responsive if there are significant differences in central tendency or in variance between reference and test sites (Figure E-8). If the test sites are known to be affected by anthropogenic pollution or disturbance, then mean or median values of responsive metrics should be substantially different between reference and test sites (Figure E-8). If the test sites simply do not meet reference criteria (i.e., they might be a mix of impaired and unimpaired sites, or sites with different stressors), then the variance in the test sites should be larger than that in the reference sites (Figure E-8). If possible, it is advisable to separate test sites according to the stressors or types of impairment (e.g., habitat degradation, toxic substances, organic enrichment) so that response to each stressor can be determined.

When selecting metrics, it is important to visually examine the distribution of metrics in reference sites and in impacted sites. Metrics are selected for inclusion based on their responsiveness, typically by visual examination of box and whisker plots (e.g., Fig. E-8) or scatterplots (Barbour et al. 1996a, Fore et al. 1996). If there is no overlap of the data points, or if the overlap is restricted only to the whiskers of the box plots, then the metric responds strongly to the impairment. A strong response here implies that at least 75% of affected sites have no overlap with at least 50% of the reference sites. A minimum response strength might be defined as no overlap of the median of one site type with the quartile of the other; implying that at least 50% of affected sites are below the 25th percentile of reference sites.

Many biologists may be tempted to use statistical significance tests to select metrics, but slavish reliance on significance tests does not contribute to biological understanding (Yoccoz 1991) and may weaken a multimetric index. If sample size is small (say, n = 6 in both reference and impact sites), then significance tests (at a = 0.05) will have low power and responsive metrics may be rejected. On the other hand, if sample size is large (say, n = 30 in both site categories), then it would be possible to detect a statistically
significant difference that is biologically meaningless. In this case, metrics that do not contribute to meaningful assessment could be selected, simply because statistical significance was detected. A better measure is the expected frequency with which a metric will fall below a threshold to register impairment. Frequency can be estimated with a box and whisker plot, but not with a significance test. For example, if the median of impaired sites is below the quartile of reference sites (Figure E-4), then we estimate that impaired test sites will be below the reference quartile in at least 50% of all observations.

Metrics that are responsive to known or unknown stresses are retained for index development. Finally, responsive metrics are evaluated for redundancy, where redundancy means a tight correlation ($r>0.9$) and a linear relationship. A metric that is linearly correlated with another might not contribute new information to the assessment. Pairs of metrics with correlation coefficients greater than 0.9 should be examined carefully to determine whether they are linear and if both metrics are necessary. Often, strongly correlated metrics are calculated from the same raw data, or their method of calculation ensures correlation. For example, Shannon-Wiener diversity and percent abundance of the dominant taxon are linearly correlated in any data set. A scatterplot of the strongly ($>0.9$) correlated metrics should be examined; if there is an apparent nonlinear or curved relationship, then both should be retained. If all the points fall very close to a straight line, then one of the metrics can be safely eliminated.

**Multimetric Index Development**

Multimetric indices are typically developed by summing the metrics that proved responsive to disturbance. The first step is to standardize the different numerical scales of metrics (e.g., number of taxa; % of individuals that are predators) into unitless scores (e.g., Karr et al. 1986, Gerritsen 1995). The scores may be ordinal, or they may be a percentage of a reference value. Ordinal scores are more commonly used, and correspond to categories such as "impaired" and "unimpaired." The index is the sum (or mean) of the metric scores, and is likewise compared to index values at reference sites. Index values at reference sites are then used to establish biocriteria. Socio-political decisions must then determine the numerical values of biocriteria corresponding to aquatic life use categories.

**Metric Scoring**

Several methods may be used for scoring metrics, all of which are based on the metric distribution in reference sites. Metrics may be given ordinal scores (most often 1, 3, or 5); corresponding to impaired, intermediate, or unimpaired biota, respectively, or may be given a score which is the metric's percentage of the reference value (Figure E-9).
Figure E-9. Illustration of alternative scoring methods, using an upper percentile, a lower percentile, or a central tendency. Most common score breakdowns (5-3-1 ordinal, or percentage) are shown for each, but other ordinal scores have also been used (e.g., 6-4-2-0).

All of these require comparison to some measure of the reference value distribution: an upper percentile, a lower percentile, or a central tendency (Figure E-9). Although a central tendency of the reference sites (e.g., the mean value) may be intuitively attractive as a basis of comparison, there are two important reasons for using percentiles instead:

- An assessment methodology must be able to take into account natural variability of ecological systems. We know that aquatic biota may differ from riffle to riffle in the same unimpacted stream. Central tendency does not take into account the natural variability, and scoring criteria based on central tendency may result in lowered scores for many sites that are within the expected variability of natural, undisturbed sites.

- A second disadvantage of central tendency measures occurs when reference sites, upon which the reference condition is empirically based, are known to be affected to some degree by human activities. Reference sites are often selected to be the least anthropogenically affected in a region, but may still be subject to regional and wide-ranging impacts (e.g., USEPA 1996a). Examples include estuaries and large rivers which receive inputs from their entire watersheds, and small streams and lakes in extensively altered agricultural ecoregions (e.g., the Cornbelt Plains). Use of central tendency then reflects the general (and unquantifiable) degradation of the region and will not result in reference conditions that represent the biological potential.

Two approaches are used to develop metric expectations and scoring criteria (Simon and Lyons 1995). The first approach uses defined reference sites that meet criteria for
representative reference sites. Data from the reference sites are used to define expectations and develop metric scoring criteria (Simon and Lyons 1995). The principal scoring criterion (between meeting and not meeting reference expectations) is typically based on a lower percentile of the reference distribution; for example, the 25th percentile (Ohio EPA 1990, Barbour et al. 1996a, Barbour et al. 1996b). In this method, values above the 25th percentile are considered unimpaired (similar to reference conditions) and values below the 25th percentile are considered impaired to some degree. The range from 0 to the 25th percentile is bisected, with values in the top half receiving a score of 3 and those in the bottom half receiving a score of 1 (Figure E-9).

This approach also lends itself to scores using percent of reference value (Figure E-9). The second approach does not include definition of reference criteria, but uses information from the entire range of sites, from the most to the least affected by anthropogenic pollution and disturbance. A large and representative survey data set is required to develop the reference criteria. Reference expectations and scoring criteria are based on the best values observed for each metric, even if the best values do not occur in the least affected sites (Simon and Lyons 1995). The most common scoring method is trisection (Karr et al. 1986) using the 95th percentile of the metric distribution. Metric values from 0 (or the lowest possible value) to the 95th percentile are trisected; values in the top one-third receive a 5, values in the middle third receive a 3, and values in the bottom third receive a 1 (most impaired).

Choice of scoring method should be based on the approach used for defining reference sites, rather than on the method that will produce the most conservative or most liberal scoring. If reference sites are representative of relatively unimpaired conditions, then the lower percentile cutoff and bisection is preferred. If reference sites are not definable, then scoring criteria based on the "best" values are the only alternative.

To account for covariables such as size, the data are plotted, a locally weighted estimate is made of the appropriate percentile (95th or 25th), and the range below it is trisected or bisected (Figure E-10).
Figure E-10. Total crustacean zooplankton taxa in North American lakes (redrawn after Dodson 1992).

Additive Index

The index is the sum of the scores of the selected metrics that prove responsive to disturbance. Criteria for index values are also generated from the reference sites, just as with individual metrics. A perfect index score is unlikely in the reference sites, therefore, a reference expectation is developed for the total score. Because the index is the sum of several metrics, the Central Limit Theorem predicts that it will have a lower coefficient of variation than individual metrics, and can be approximated better by a normal distribution than can individual metrics (Fore et al. 1994). Because of these properties, multimetric indices can usually distinguish 3 to 5 statistically significant gradations of impairment, based on comparison of a single sample to the reference distribution (Fore et al. 1994, Gerritsen 1995).

Discriminant Model Index

Discriminant analysis may be used to develop a model that will divide, or discriminate, observations among two or more predetermined classes. Output of discriminant analysis is a function that is a linear combination of the input variables, and that obtains the maximum separation (discrimination) among the defined classes. The model may then be used to determine class membership of new observations. Thus, given a set of unaffected reference sites, and a set of degraded sites (due to toxicity, low DO, or habitat degradation), a discriminant function model can identify variables that will discriminate reference from degraded sites.

Developing biocriteria with a discriminant model requires a training data set to develop the discriminant model, and a confirmation data set to test the model. The training and confirmation data may be from the same biosurvey, randomly divided into two, or they
may be two consecutive years of survey data, etc. All sites in each data set are identified by degradation class (e.g., reference vs impaired) or by designated aquatic life use class. To avoid circularity, identification of reference and impaired, or of designated use classes, should be made from non-biological information such as riparian zone modification; known discharges, known contamination, toxicity, nonpoint sources, impervious surface in the watershed, land use practices, etc.

One or more discriminant function models are developed from the training set, to predict class membership from biological data. After development, the model is applied to the confirmation data set to determine its performance: The test determines how well the model can assign sites to classes, using independent data that were not used to develop the model. More information on discriminant analysis is in any textbook on multivariate statistics (e.g., Ludwig and Reynolds 1988, Jongman et al. 1987, Johnson and Wichern 1992).

**Classification of Estuaries with a Discriminant Model Index**

A straightforward *a priori* classification for estuaries was that used by EMAP-Estuaries: first, regionalization into the biogeographic provinces used by NOAA, the U.S. Fish and Wildlife Service and USEPA; second, stratification of estuaries by physical characteristics of shape and size; and third, measurement of physical covariates that affect assemblage composition, principally salinity, depth, and sediment attributes (for benthos) (USEPA 1993e).

Estuaries were classified as large estuaries, large tidal rivers, and small estuaries. Data collected in 1990 showed that large estuaries had the greatest number of taxa, and tidal rivers the fewest taxa. On that basis, the three estuary classes were retained for further analysis.

It has long been known that estuarine faunal diversity is highest at the seaward end of estuaries, in full-salinity seawater. The lowest number of taxa are found in brackish waters that are too saline for freshwater organisms, and too fresh for marine-adapted organisms. To characterize reference conditions in EMAP estuaries, it was therefore necessary to predict the number of taxa that could be found at any given salinity. Figure E-11 shows effect of the covariate, salinity, on the number of taxa captured in benthic grabs. In the example shown, the reference expectation for EMAP estuaries in the Virginian Province was a third order polynomial regression of a running average 90th percentile of the data shown, given by the line in the figure.
Figure E-11. Plot showing the regression line used to estimate salinity-adjusted species richness measures for mean number of species per site. The distribution of reference and degraded sites relative to the estimated line is shown. The line is the expected number of species based on the polynomial regression. From USEPA 1993a.

Of five possible covariates considered in EMAP-estuaries, only salinity was deemed to have a strong enough effect on benthic macroinvertebrates to justify adjusting reference expectations for it (USEPA 1993e, USEPA 1994h). The observed number of taxa per site was corrected by dividing by the expected number of taxa for the site, obtained from the regression model, to yield percent of expected number of taxa (USEPA 1993e, Engle et al. 1994, USEPA 1994h).

Index

From the set of sampled sites in the EMAP data set, reference and degraded sites were identified based on predetermined criteria. Criteria for reference sites were:

- Summertime bottom DO never less than 1 ppm.
- No contaminants observed in the sediment (1980).
- No toxicity observed in the sediment.

Reference sites had to meet all three criteria. Reference site salinities ranged from < 5 ppt to > 18 ppt. Sites were rated as degraded if they met either of the two following criteria:

- One or more hypoxic events with bottom DO < 0.3 ppm.
• The concentration of at least one sediment contaminant exceeding the ER-M value, and Ampelisca bioassay indicated toxicity (< 75% survival and significantly different from control) (USEPA 1993e).

Stepwise discriminant analysis was used to determine which metrics could best discriminate between reference and degraded sites. Because number of taxa was deemed an important indicator by itself, it was “forced” into the discriminant model. The eventual discriminant model had five variables:

• % expected number of taxa.
• No. of amphipods.
• % of abundance as bivalves.
• Mean weight per polychaete.
• No. of capitellid polychaetes.

The model correctly classified 89% of degraded sites, and 86% of reference sites, using the learning data set to test its performance. Discriminant scores were normalized to a range of 0 to 10, for ease in communication of index scores (USEPA 1993e).

The original discriminant model was developed from the 1990 EMAP-Virginian Province sampling effort. The model was subsequently tested with the EMAP-Virginian Province data set collected in 1991, an independent test (USEPA 1994h). The 1990 model failed to discriminate the 1991 data correctly, so a new discriminant model was developed using both 1990 and 1991 data sets. The revised model used the following 3 variables:

• Mean abundance of opportunistic species.
• Biomass/abundance ratio for all species.
• Mean number of infaunal species per grab.

The revised model was subsequently tested with another independent data set, the 1992 Virginian Province data. The revised model correctly discriminated the 1992 reference and degraded sites, and the model was not further modified (USEPA 1994i). The model correctly identified 83% of reference sites and 100% of degraded sites.

**Designated Aquatic Life Use Classes**

An alternative to the above methodology is to develop biocriteria directly for administrative aquatic life use classes (Davies et al. 1993). In this approach, data from a set of sites (the training set) are assigned to predetermined aquatic life use classes. The classes are determined by regulation and might be (for example): (a) pristine; (b) altered habitats, but native species maintained; (c) discharges and vegetation permitted, native communities altered, but fishable-swimmable goals met; or (d) nonattainment. Experts assign sites to one of the four classes based on the narrative descriptions of the aquatic life use classes (above) and biological data from the training set sites (Davies et al. 1993).

One or more discriminant models to predict class membership are developed from the training set. The purpose of the discriminant analysis here is not to test the classification
(the classification is administrative rather than scientific), but to assign test sites to one of the classes. An example of this approach is the biocriteria adopted by Maine for streams (Davies et al. 1993). Stream biologists assigned a training set of streams to four life use classes. A two-stage discriminant modeling process was used to develop discriminant models for assigning test streams to use classes. The first stage was a model to predict membership in each of the four classes, expressed as a probability for each. The second stage was a set of three discriminant models that predict two-way class membership (i.e., nonattainment (NA) versus A or B or C; NA or C versus A or B; and NA or C or B versus A). A selection procedure was used to select predictive variables for the models, and the second-stage models were constrained to exclude predictive variables used in the first-stage model. This approach is detailed by Davies et al. (1993).

**Multivariate Ordination Model**

The third approach that has been successfully used for development of biocriteria uses multivariate ordination to determine if test sites are different from reference sites. The comparisons are usually made graphically, in ordination space (c.f., Figures E-1, E-3, and E-9); such that if a site is outside of the area on an ordination diagram defined by reference sites, it is judged to be degraded.

Classification of reference sites is often *a posteriori*, using one of several clustering methods on the biological data. Following definition of biological clusters (reference classes), a discriminant model is developed using physical-chemical data to allow classification of test sites (Moss et al. 1987, Wright et al. 1984, Reynoldson et al. 1995, Norris 1995).

Cluster analysis must be done with great caution because there are many similarity measures and many clustering algorithms, many of which may produce different, and often unintelligible, results (Ludwig and Reynolds 1988, Jackson 1993a). In general, the best results for bioassessment purposes have been achieved with UPGMA, and with TWINSPLAN, a divisive technique (Reynoldson et al. 1995, Moss et al. 1987, Gauch 1982). The most successful similarity measures have been Bray-Curtis similarity, chord distance, and relative Euclidean distance (Kenkel and Orloci 1986, Ludwig and Reynolds 1988).

Ordination analysis is often done after determination of clusters, to see whether the identified clusters also separate in ordination space. The clusters are now treated the same as an *a priori* classification. Ordination methods most often used at this stage include correspondence analysis and non-metric multidimensional scaling (e.g., Moss et al. 1987, Reynoldson et al. 1995).

Test sites are then assigned to one of the reference classes with the discriminant model (using physical-chemical data), and test sites are compared to their respective reference population in ordination space.

The ordination approach is illustrated with the classification of benthic macroinvertebrate assemblages from Great Lakes reference sites (Reynoldson et al. 1995). Cluster analysis of benthic macroinvertebrates from 96 reference sites revealed five groups of
Cluster analysis used the Bray-Curtis distance measure, and the clustering algorithm was unweighted pair group mean averages (UPGMA) (Reynoldson et al. 1995). The sites were subsequently visually depicted with ordination by NMDS, which showed each cluster occupying a unique area in ordination space.

Following classification, a discriminant model was developed to identify class membership of sites using physical and chemical (i.e., non-biological) data. The model was developed with the reference sites as the calibration data set, for subsequent use with test sites to identify the class of benthic assemblage the test site should belong to. Sites were selected for uniform characteristics (< 2 km from shore, < 30 m depth, fine-grained sediment), and the criterion for reference sites was large distance (> 10 km) from known discharges (Reynoldson et al. 1995). First, explanatory variables for input into the discriminant function analysis were identified with correlation analysis of physical-chemical variables. Variables that were significantly correlated with any of the three ordination axes were used for the discriminant analysis. Of 25 variables examined, 18 were strongly correlated with the ordination and were input into the stepwise discriminant analysis. Of these, nine produced the best model to predict class membership. Biological assemblage group membership was correctly predicted by the discriminant model for 87% of the sites, ranging from 64%-100% for each of the five assemblage groups (Reynoldson et al. 1995).

Biological integrity of test sites was assessed by first assigning a test site to one of the five assemblage groups, based on the discriminant model applied to the site's physical-chemical data. The biological assemblage structure of the test site was then compared to assemblage structure of the reference sites of that group, by plotting the positions of reference sites and the test site in ordination space, and determining if the test site was within the region defined by the reference sites (e.g., Figure E-12). The approach was used for an assessment of benthic sites in Collingwood Harbour, Ontario, which had been contaminated with metals. Benthic macroinvertebrate assemblages were different from reference sites within two boat slips, and the authors concluded that sediment remediation was justified in the boat slips. Outer reaches of the harbor exceeded Ontario sediment metals criteria but benthic assemblages in the outer harbor were similar to reference sites of their respective classes. Because there were no discernible biological differences, the authors concluded that sediment remediation could not be justified in the outer harbor (Reynoldson et al. 1995).
Figure E-12. Assessment by ordination. Solid circles are reference sites, known impacted sites (triangles) deviate from the reference group, primarily on the first axis. Impairment may be judged by whether a site is outside the region bounding reference sites (ellipse), or by the distance between a site and the reference centroid (arrow).