

# Chapter 8.0 Data Analysis

---

## This chapter...

- describes how to create IBI and RIVPACS models
- describes how to analyze regional, site-specific, watershed and gradient study design data
- describes different reporting approaches

## Data Analysis...

- is critical in translating monitoring data into information for management action

## 8.1 Introduction

Assessment data are collected not only to help define the status of large rivers, but also to guide management decisions. Data must be translated into a format from which management decisions regarding water resources can be made. The strategies outlined in Chapters 2 and 3 on the elements of assessment and study design, provide users with approaches to clearly define questions and objectives to create an appropriate study

design. These chapters also provide approaches for developing a thorough quality assurance plan (QAP) that will allow the study to meet and quantify measurement and data quality objectives (MQOs and DQOs). The tools introduced in those chapters lead to this section which outlines some common approaches for analyzing assessment data and presenting it in a way that is most useful for decision-making.

There are a variety of materials available detailing analyses of assessment data (Reckhow and Warren-Hicks 1997, Barbour et al. 1999). This section discusses the main approaches, but interested readers should also consult the existing literature (see Barbour et al. 1999). First, this chapter discusses two of the major biological analysis strategies used in assessment: the multimetric approach and the predictive modeling approach. Then, the chapter presents analysis approaches to be used under the major study designs introduced in Chapter 3 (watershed, site-specific assessments and gradient studies) and approaches that can be used for relating assessment data to stressors and stressor sources. Last is a brief discussion of the different approaches that can be used for reporting results.

## 8.2 Biological Analysis Strategies

Water quality data can be used as stand-alone data and analyzed as individual variables. However, it is common to combine physical habitat or biological data into habitat or biological indexes that synthesize multivariate data into one variable or score (e.g., index of biological integrity (IBI) - Karr et al. 1986, Hughes et al. 1998, Barbour et al. 1999, Karr and Chu 1999; predictive models - Moss et al. 1987, Novak and Bode 1992, Hawkins et al. 2000, Wright 2000; qualitative habitat evaluation index-Rankin 1989, non-wadeable stream habitat index - Wilhelm et al. 2005). It is important to note that indices are developed for specific methods. Data derived from different methods would have to be evaluated for comparability before being applied to an existing index or a different index developed.

### **8.2.1 *Multimetric Indexes***

Multimetric indexes of biotic condition for fish and benthic macroinvertebrates have been developed for many regions of North America and Europe and are generally accepted for biological assessment of aquatic resource quality. Some examples include IBI (Index of Biotic Integrity for fish; Karr et al. 1986), RBP (Rapid Bioassessment Protocol, Plafkin et al. 1989), ICI (Invertebrate Condition Index, Ohio EPA 1987), B-IBI (benthic IBI; Kerans and Karr 1994), SCI (Stream Condition Index; Barbour et al. 1996), and others (see Chapter 7 for large river fish index examples). A multimetric index is a simple sum or average of several standardized metrics. For index development, metrics are attributes of the biota that respond to anthropogenic stressors in consistent ways and are thus, useful indicators of stress (Barbour et al. 1999). Developing a multimetric index consists of three overall steps:

1. Classifying natural biological assemblages into relatively homogeneous groups, so that the species composition can be reliably predicted by geographic location or site characteristics;
2. Identifying metrics that respond to anthropogenic stressors; and
3. Aggregating standardized, non-redundant metrics that represent aspects of diversity, composition, sensitivity, and function into an index.

Data analysis for index development consists of characterizing reference conditions that will form the basis for assessment of degradation and calibration of the index to a gradient of human influence. This is a well-documented procedure (Davis and Simon 1995, Gibson et al. 1996, Barbour et al. 1999) and is described below. Reference site selection was described in Chapter 2.

#### **8.2.1.1 *Classification of Biological Resources***

Index development requires a waterbody classification framework to partition natural variability. Classification frameworks can be geographic (e.g., ecoregions [Omernik 1995]), they may be based on continuous variables (e.g., catchment area, elevation), or they may be a combination. The framework should rely on characteristics that are intrinsic and independent of human influence (e.g., climate, topography, vegetation, soils, geology, elevation, waterbody type and size) and that account for differences in the composition of relatively undisturbed reference sites (Barbour et al. 1999, Hawkins et al. 2000b). Classification is best accomplished with reference sites that represent the range of natural conditions of the region (Chapter 2). Candidate reference sites that are based on least degraded physical habitat and water chemistry can also be used as the basis for river classification. Using quantitative criteria for reference site selection helps provide a consistent classification framework.

A result of the classification step is a set of rules that directs the partition of sites into biologically-meaningful natural classes or groups. These rules may be simple. For example, if elevation is above 2000 m, then a site belongs to the “mountain” class. Conversely, the rules may be complex, requiring multivariate discriminant equations to determine site class. Classes are initially determined by the biota of reference sites. However, because biological information is reserved for assessment, the rules generally do not use biological information.

The two basic approaches to developing rules are:

- Examining prior rules (e.g., ecoregions) with biological information. If a prior rule is found or modified that adequately explains biological variability, it is used for further index development.
- Developing posterior rules (e.g., using ordination) from a biological classification.

There is no clear distinction between prior and posterior rule development. Prior hypotheses could be applied (e.g., elevation and catchment area) to a biological classification to determine the rules for class boundaries. Rules may be fixed (the elevation example above) or probabilistic (a discriminant function). The key to classification is practicality within the region or state in which it will be applied; local conditions determine the classes.

The most common prior rules examined are geographic region, elevation or gradient, and measure of waterbody size (catchment area, stream order, surface area). As a guide for developing rules, landscape types or ecoregions are a very good start and account for much variability (e.g., Yoder and Rankin 1995, Barbour et al. 1996, Feminella 2000, Gerritsen et al. 2000a, Jessup and Gerritsen 2000). But in some landscapes, more continuous variables have done a better job accounting for variability (e.g., montane regions, Hawkins and Vinson 2000, McCormick et al. 2000, Pan et al. 2000, Van Sickle and Hughes 2000, Waite et al. 2000).

The general approach for confirming or testing prior classification rules is to examine alternative sets of prior classification rules to determine which rule yields the simplest classification of the reference sites in the data set and accounts for a substantial fraction of variability in biological composition among sites. The techniques for examining and testing the alternatives include: 1) ordination and examining prior classifications in ordination space (e.g., using unique labels for prior classes), 2) comparing prior classification to cluster analysis results using similarity analysis (Van Sickle 1997), and 3) multivariate analysis of variance (MANOVA) on prior groups.

Posterior rule classification involves using the biological data collected from reference sites to classify sites into groups based on similarity in taxonomic composition. This can be done using ordination or cluster analysis, and is the classification approach used principally in RIVPACS analysis (Hawkins et al. 2000a, Wright 2000). This approach does not use prior rules or adherence to any existing framework.

### ***8.2.1.2 Selection and Evaluation of Metrics and Formation of a Multimetric Index***

Metrics allow the investigator to use meaningful indicator attributes in assessing the status of assemblages and communities in response to perturbation. The definition of a metric is a characteristic of the biota that changes in some predictable way with increased human influence (Barbour et al. 1999). For a metric to be useful, it must have the following attributes: 1) ecological relevance to the biological assemblage under study and to the specified program objectives, and 2) sensitivity to stressors and a response that can be discriminated from natural variation. The purpose of using multiple metrics to assess biological condition is to aggregate

the information available from multiple structural and functional elements of aquatic communities into one score.

All metrics that have ecological relevance to the assemblage under study and that respond to the targeted stressors are potential metrics for testing. From this "universe" of metrics, some will be eliminated because of insufficient data or because the range of values does not sufficiently discriminate between natural variability and anthropogenic effects. In this step, investigators identify the candidate metrics that are most informative and, therefore, warrant further analysis.

Investigators should select the measures that are relevant to the ecology of rivers within a region to ensure that various aspects of the structure and function of the aquatic assemblage are addressed. Representative metrics should be selected from each of four primary categories: 1) richness measures for diversity or variety of the assemblage; 2) composition measures for identity and dominance; 3) tolerance measures that represent sensitivity to perturbation; and 4) trophic or habit measures for information on feeding strategies and guilds. Other metric categories (especially useful in fish multimetrics) include life history and reproductive strategies. Common metrics are shown in Table 8-1. Karr and Chu (1999) suggest that measures of individual health be used to supplement other metrics.

**TABLE 8-1. Some potential metrics for periphyton, benthic macroinvertebrates, and fish that could be considered for rivers. Redundancy can be evaluated during the calibration phase to eliminate overlapping metrics.**

	Richness Measures	Composition Measures	Tolerance Measures	Trophic/Habit Measures
<b>Periphyton</b>	<ul style="list-style-type: none"> <li>Total no. of taxa</li> <li>No. of common nondiatom taxa</li> <li>No. of diatom taxa</li> </ul>	<ul style="list-style-type: none"> <li>% community similarity</li> <li>% live diatoms</li> <li>Diatom (Shannon) diversity index</li> </ul>	<ul style="list-style-type: none"> <li>% tolerant diatoms</li> <li>% sensitive taxa</li> <li>% aberrant diatoms</li> <li>% acidobiontic</li> <li>% alkalibiontic</li> <li>% halobiontic</li> </ul>	<ul style="list-style-type: none"> <li>% motile taxa</li> <li>Chlorophyll <i>a</i></li> <li>% saprobiontic</li> <li>% eutrophic</li> </ul>
<b>Benthic macroinvertebrates</b>	<ul style="list-style-type: none"> <li>No. Total taxa</li> <li>No. EPT taxa</li> <li>No. Ephemeroptera taxa</li> <li>No. Plecoptera taxa</li> <li>No. Trichoptera taxa</li> </ul>	<ul style="list-style-type: none"> <li>% EPT</li> <li>% Ephemeroptera</li> <li>% Chironomidae</li> </ul>	<ul style="list-style-type: none"> <li>No. Intolerant Taxa</li> <li>% Tolerant Organisms</li> <li>Hilsenhoff Biotic Index (HBI)</li> <li>% Dominant Taxon</li> </ul>	<ul style="list-style-type: none"> <li>No. Clinger taxa</li> <li>% Clingers</li> <li>% Filterers</li> <li>% Scrapers</li> </ul>
<b>Fish</b>	<ul style="list-style-type: none"> <li>Total no. of native fish species</li> <li>No. of darter species</li> <li>No. of sunfish species</li> <li>No. of sucker species</li> </ul>	<ul style="list-style-type: none"> <li>% pioneering species</li> <li>Number of fish per unit of sampling effort corrected for drainage area</li> </ul>	<ul style="list-style-type: none"> <li>No. of intolerant species</li> <li>% of individuals as tolerant species</li> <li>% of individuals as hybrids</li> <li>% of individuals with disease, tumors, fin damage, and skeletal anomalies</li> </ul>	<ul style="list-style-type: none"> <li>% omnivores</li> <li>% insectivores</li> <li>% top carnivores</li> </ul>

It is generally not advisable to use metrics that are inherently unstable or variable due to their quantitative definition (e.g., ratio of scrapers to filterers, or ratio of EPT to Chironomidae in RBP 2; Plafkin et al. 1989). For example, ratios of two independent variables ( $x/y$ ) should never be used as metrics because they range from zero (if  $x = 0$ ) to undefined (if  $y = 0$ ). Instead, use proportions of a total ( $x / (x + y)$ ), which range from 0 to 1. Components of metric review include:

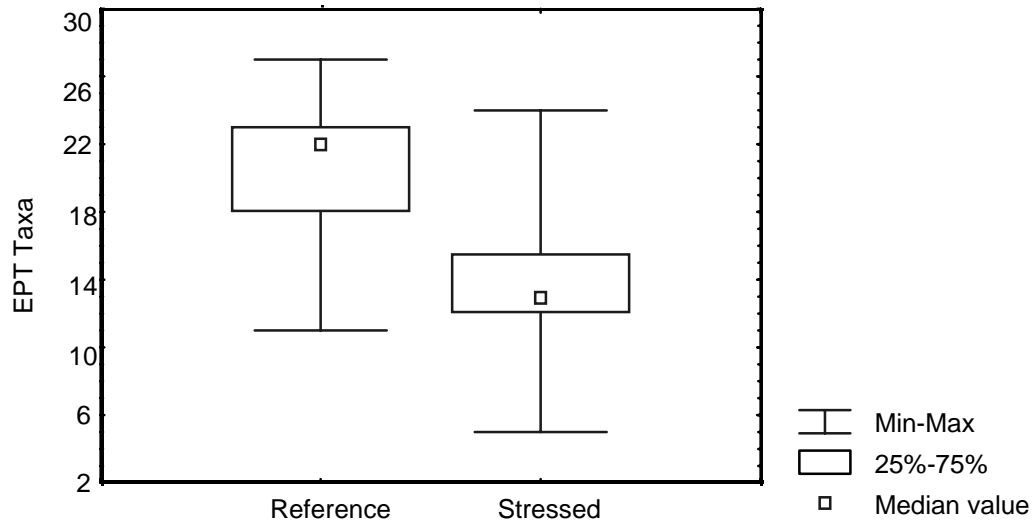
- Eliminating metrics that have too many zero values in the population of sites to calculate the metric at a large enough proportion of sites. Many zero values or a predominance of either very low values or very high values (close to 100%) indicate that the metrics may not have sufficient range to discriminate impairment. For example, the number of Plecoptera taxa (stoneflies) in even unstressed reference sites is often low, in the range of three to five genera. Although stoneflies are highly sensitive, there is not enough range (i.e., three taxa) to detect intermediate levels of impairment. This is why stoneflies are commonly grouped with mayflies and caddisflies to form the metric EPT taxa, which usually has sufficient scope (10-15 genera or more) to be a sensitive indicator.
- Using descriptive statistics (central tendency, range, distribution, outliers) to characterize metrics within the population of reference sites of each site class.
- Eliminating metrics where variability in the reference site population of a class is so large, they cannot discriminate among sites of different condition. The potential for each measure is based on containing enough information within a specific range of variability to discriminate among site classes and biological condition (reference vs degraded). Highly variable metrics (in unstressed sites) are poor indicators because their precision is low. This can also be characterized using signal/noise ratios across all sites (Kauffman et al. 1999).

It is important to understand the effects of various stressors on the behavior of specific metrics. If metric response is counter-intuitive or poorly understood on conceptual grounds, it is better to avoid using them.

The ability of a biological metric to discriminate between “known” non-stressed conditions and “known” stressed conditions (defined by physical and chemical characteristics) is crucial in the selection of core metrics for future assessments. Two general approaches to identifying responsive metrics to stressors are: 1) looking for categorical responses and 2) looking for response to gradients of stressors. The categorical approach is more common and analytically simpler, but does not provide potential diagnostic information. Examining response to gradients can only be done if measurements of the stressors exist in the data set.

*Categorical Response:* Examining categorical metric responses is based on comparing metric distributions in reference and degraded sites. The simplest comparison, and in many ways the most effective, is to examine box and whisker plots of the metric values in two groups of sites: reference sites and known “stressed” sites (defined by physical and chemical criteria, much like reference sites) (Figure 8-1). Box plots show several attributes of the distribution graphically: median, upper and lower quartiles, tails, outliers and/or minimum and maximum. Box plots of

two distributions (reference and stressed sites) show exactly how much the distributions differ or overlap with each other. Formal hypothesis tests are therefore not necessary; in fact, they are generally not meaningful because the question is not whether the reference and stressed sites differ (the subject of a hypothesis test), but whether a given metric can distinguish between them (e.g., Salsburg 1985, 1986, Yoccz 1991, Suter 1996). Metrics having the strongest discriminatory power provide the most confidence in assessing biological condition of unknown sites.



**FIGURE 8-1.** A box and whisker plot comparing the distribution of the number of EPT taxa, a common macroinvertebrate metric, in reference and stressed sites.

Discrimination efficiency (DE) is also used to evaluate metrics. The DE is the proportion of stressed sites that would be deemed different from reference if below a given threshold. For example, in Figure 8-1, if we choose the 25th percentile of EPT taxa (18 taxa) as a threshold, then all sites with fewer than 18 EPT taxa would be “different from reference”. The discrimination efficiency for EPT is then the fraction of stressed sites with EPT taxa <18; 80% in this case.

*Gradient Response:* If quantitative measures exist for stressors or sources of stressors in the data set, then it is possible to examine the response of candidate metrics to those gradients using scatterplots. Measured stressors could include habitat, water chemical measures, water column contaminants, and sediment contaminants. Measured sources include land use or known discharges. Since many stressor measurements are correlated (e.g., pH, conductivity, sulfate), it is often advantageous to define stressor axes with principal components analysis (PCA) of chemical and habitat measures (e.g., Norton et al. 2000, Gerritsen et al. 2002) or some combined disturbance gradient (Fore 2004). In using the gradient approach, those metrics exhibiting the strongest response to stressors are usually selected as candidates. Other multivariate approaches can be used to identify responsive metrics, including canonical correspondence and canonical

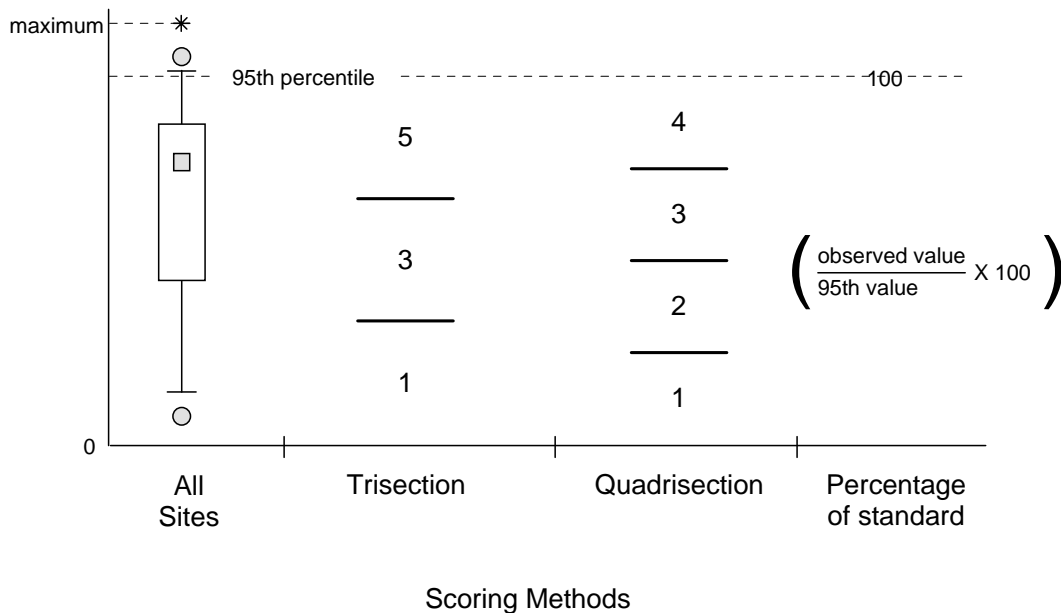
correlation analysis of metrics and environmental variables to explore the relationship of metrics to certain stressors (Griffiths et al. 2001, 2002, 2003).

The final step in the process is combining candidate metrics into a multimetric index. The index should include metrics representing richness, composition, tolerance, and trophic aspects of the assemblage, contain minimally redundant metrics, and be able to discriminate reference from impaired sites with low variability. Metrics are standardized to a common scale so that all are weighted equally, and alternative combinations of metrics are examined for discrimination.

Responsive metrics are evaluated for redundancy. A metric that is highly correlated with another metric may not contribute new information. Redundancy among candidate metrics is determined from correlation analysis. A correlation matrix (Pearson) is calculated for all remaining candidate metrics. High correlation coefficients ( $r > 0.7$ ) indicate strong linear relationships. A high correlation coefficient alone is not sufficient to eliminate one of a pair of correlated metrics (Karr 1991). Although there is no absolute threshold,  $r > 0.7$  is generally used to indicate “forbidden combinations”, and all pairs with  $r > 0.7$  are examined with scatterplots to determine if there are nonlinearities in the relationship. If the scatterplot shows a curvilinear relationship, then both metrics may be retained because each one contributes information in a different part of the range.

The purpose of an index is to provide a means of integrating information from the various measures of biological attributes (or metrics). Metrics vary in their scale—they are integers, percentages, or dimensionless numbers. Prior to developing an integrated index for assessing biological condition, it is necessary to standardize core metrics via transformation to unitless scores. The standardization assumes that each metric has the same value and importance (i.e., they are weighted the same), and that a 50% change in one metric is of equal value to assessment as a 50% change in another.

Where possible, scoring criteria for each metric are based on the distribution of values from the population of sites, which include reference rivers. For example, the 95<sup>th</sup> percentile of the data distribution is commonly used (Figure 8-2) to eliminate extreme outliers (e.g., Hughes et al. 1998, Gerritsen et al. 2000b). From this upper percentile, the range of the metric values can be standardized as a percentage of the 95<sup>th</sup> percentile value, or other percentile (e.g., trisected or quadrisectioned), to provide a range of scores. Those values that are closest to the 95<sup>th</sup> percentile would receive higher scores, and those having a greater deviation from this percentile would have lower scores. For those metrics whose values increase in response to perturbation, the 5<sup>th</sup> percentile is used to remove outliers and to form a basis for scoring.



**FIGURE 8-2.** A comparison of different methods used for standardizing metric scores. The trisection method split the score distribution into 3 categories and the quadrisection, into 4. The last approach creates a continuous range of scores from 0 to 100, and the standardization formula depends on the response of the metric to disturbance.

Alternative methods for scoring metrics, as illustrated in Figure 8-2, are currently in use in various parts of the USA for multimetric indexes. A “trisection” of the scoring range has been well-documented (Karr et al. 1986, Ohio EPA 1987, Barbour et al. 1996, Fore et al. 1996). A “quadrisection” of the range has also been found to be useful for benthic assemblages (DeShon 1995). More recent studies are finding that a standardization of all metrics as percentages (0-100) of the 95th percentile value yields the most sensitive index, because information of the component metrics is retained (Minns et al. 1994, Ganasan and Hughes 1998, Hughes et al. 1998). Index development from statewide databases for Idaho (Jessup and Gerritsen 2000), Wyoming (Jessup et al. 2002), and West Virginia (Gerritsen et al. 2000b) are supportive of this third alternative for scoring metrics. The 95<sup>th</sup> percentile scoring method is as follows:

*Scoring metrics that decrease with stress.* The 95<sup>th</sup> percentile of metric values in all samples is assigned a unitless “best” or “standard” score of 100. Values between the minimum (“worst,” usually 0) and the 95<sup>th</sup> percentile values are scored proportionally from 0 to 100 according to Equation 1:

$$score = \left( \frac{x - x_{\min}}{x_{95} - x_{\min}} \right) \times 100 \quad \text{Equation 1}$$

where,

$x$  = the calculated metric value

$x_{95}$  = the 95th percentile of this metric’s values in all samples

$x_{\min}$  = the minimum possible value, usually 0.

*Scoring metrics that increase with stress.* The 5<sup>th</sup> percentile of metric values in all samples is assigned a unitless best, or standard, score of 100. Values between the maximum (worst) value in the range and the 5<sup>th</sup> percentile value (standard, or best value) are scored proportionally from 0 to 100 according to Equation 2:

$$score = \left( \frac{x_{max} - X}{x_{max} - x_5} \right) \times 100 \quad \text{Equation 2}$$

where,

$x$  = the calculated metric value

$x_5$  = the 5<sup>th</sup> percentile of this metric's values in all samples

$x_{max}$  = the maximum observed or possible value; e.g., 10 for HBI or 100% for percentage metrics.

In some States, trisected, quadrisected, or continuous scoring is based on percentiles of the reference distribution and not the entire range (Ohio EPA 1987, Stribling et al. 1998). After identifying redundant metric pairs, possible alternative indexes that exclude one of each redundant pair are built by averaging individual metric scores across different combinations or summing metric scores. Alternative configurations are examined for discrimination efficiency. The optimal index has no redundant pairs of metrics, has a high discrimination, and a mix of metrics from the richness, composition, tolerance, and trophic categories.

### 8.2.2 Predictive Models

Multimetric indicators are very broadly used across the US, and thus a more in-depth understanding exists on the use of these analytical approaches. However, there is growing interest in simultaneous use of predictive models in conjunction with multimetric approaches. To support this trend, this section provides detailed information related to development of predictive models, much more so than for multimetric indicators in previous sections.

The River InVertebrate Prediction And Classification System (RIVPACS), developed as one bioassessment model for Britain, and AUSTRALIAN RIVER Assessment System (AUSRIVAS) are methods of bioassessment that predict an expected invertebrate assemblage in a river based on physical and chemical features of the river reach and surrounding landscape (Wright et al. 1984, Furse et al. 1984, Moss et al. 1987, Marchant et al. 1995, Wright 1995, Davies 2000, Simpson and Norris 2000, Wright 2000). These models compare the observed assemblage of macroinvertebrates at a test site to that expected in the absence of human disturbance (Observed:Expected; O/E) and assess biological condition based on a significant departure from 1.0 (where Observed = Expected). The observed assemblage is that found using standard sampling methods, whereas the expected assemblage is built using a model based on reference sites from across the sampling region. The approach is based on the concept that any site, in the absence of stressors, would likely have those taxa commonly found from physically similar reference sites. So, in essence, a site-specific reference condition is constructed for each test site based on the most probable assemblage of invertebrates expected at that test site in the absence of human disturbance. Conceptually, the expected taxa list is a weighted average of taxa frequencies found in reference sites. The weights represent the probability that a site falls in a particular group of reference sites based on physical similarity. Taxa from reference sites that

are physically very similar to a test site are weighted most. The approach has been applied successfully in the UK, Australia, and in several states in the USA (Wright et al. 1993, Hawkins et al. 2000, Paul et al. 2002).

This type of analysis proceeds in three main steps (Figure 8-3) described in detail below: 1) a cluster analysis of reference sites based on taxonomic composition to classify reference assemblage groups; 2) a discriminant analysis to develop linear models using physical variables to estimate the probability with which a test site belongs to each of the reference assemblage groups created in step 1; and 3) the prediction of the taxonomic composition of test sites based on group membership probabilities (step 2) and the frequency of taxa occurrence in each reference group.

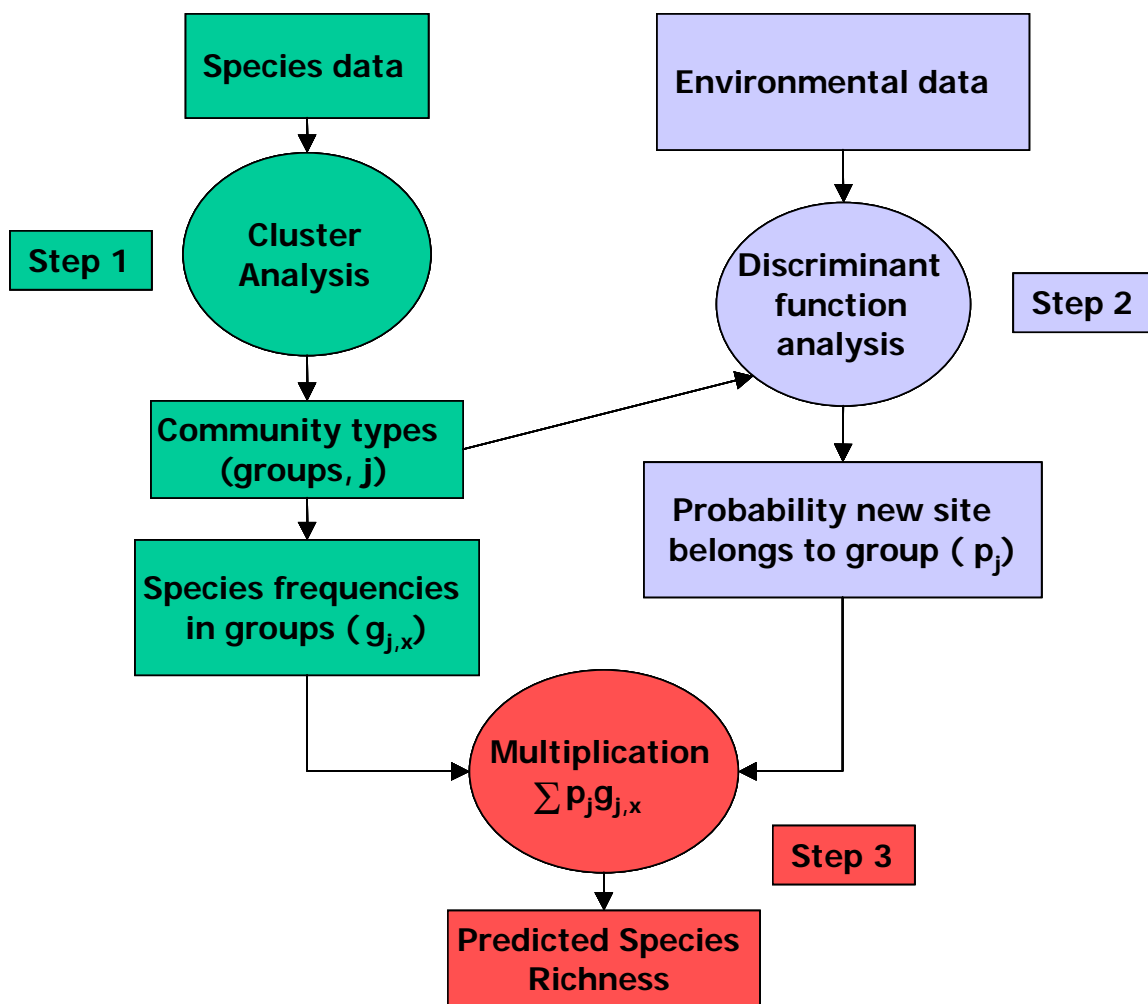


Figure 8-3. Schematic showing the three main steps involved in building RIVPACS-type bioassessment models.

### 8.2.2.1 Data Preparation for Predictive Models

RIVPACS models are built from variables considered relatively invariant to human disturbance (Wright et al. 1984, Hawkins et al. 2000, Wright 2000). Using established biogeographic factors that are minimally affected by human activity, it is possible to predict the expected assemblage for altered rivers. If alterable variables were used (e.g., nutrient concentrations, conductivity, forest cover), it would be difficult to discriminate the natural gradient from that caused by human activity; and confident prediction of an expected assemblage in the absence of human disturbance for a test site using this approach would be impossible. Commonly used variables for building RIVPACS models are shown in Table 8-2.

**TABLE 8-2. Predictor variables commonly used for building multivariate predictive models.**

Predictor Variables Used		Reference
<u>RIVPACS in United Kingdom</u>		Wright 2000
Mean depth	Slope	
Mean width	Discharge category	
Mean substratum	Mean air temperature	
Alkalinity	Annual air temperature range	
Altitude	Latitude	
Distance from source	Longitude	
<u>AUSRIVAS in Australia</u>		Simpson and Norris 2000
Longitude	Macrophyte taxa	
Latitude	Flow pattern	
Alkalinity	Macrophyte cover	
Altitude	Shading	
Distance from source	Bedrock	
Catchment area	Stream width	
Conductivity	Riffle depth	
Stream slope	Percent pebble	
Riparian width	Edge/bank vegetation	
Percent cobble	Vegetation category	
Percent boulder	Annual air temperature range	
Stream order	Percent gravel	
Discharge	Percent silt	
Percent sand	Percent clay	
<u>Models from California</u>		Hawkins et al.2000
Conductivity		
Longitude	Stream length	
Catchment area	Mean width	
Altitude	Sampling date	
Mean depth	Slope	
Latitude	Azimuth	

After a comprehensive dataset has been established, including ample reference sites across the range of natural environmental gradients sampled, the data must be prepared for analysis. As part of a preliminary analysis, all of the physical and chemical variables should be investigated graphically (e.g., frequency plot, normal quantile-quantile plot) to look for obvious lack of normality. Variables should be transformed as necessary. Common transformations include  $\log_{10}$  for chemical concentrations and arcsine square-root transformations for percentile data (which are calculated using the ratio form of the percentile – 0 to 1.0) (Zar 1999). Transformed variables should also be inspected graphically. If necessary, tests for normality (e.g., Shapiro and Wilk's test) and equal variance (Bartlett's test) can be used to check these assumptions. Again, some departure from normality and equal variance is generally acceptable, especially since no hypotheses are being tested; but predictor models are being built using these techniques.

Many multivariate predictive models use an external validation (Hawkins et al. 2000). This consists of testing the final models with an independent set of data. Remember that models are built with reference sites only, so one approach is to set aside randomly selected reference sites (approximately 20%) before constructing the models. These are labeled as validation reference sites and are used to validate the models.

Other considerations in data preparation include some assessment of sampling the temporal and spatial variability in final scores. Estimating scores through time at a set of reference sites allows investigation of temporal stability of scores. Where this has been assessed, RIVPACS scores have exhibited marked stability (C.P. Hawkins, personal communication). Similarly, estimating scores in a number of replicate reaches within a set of reference sites examines the spatial stability (essentially the sampling error) associated with the models (see error estimation below). If different teams independently conduct each replicate, it is also possible to assess inter-team sources of sampling error. These all help identify the true error associated with model estimates (Clarke 2000).

#### **8.2.2.2 Cluster Analysis**

Once the environmental data are prepared for discriminant function analysis, the biological data should be prepared for cluster analysis. The cluster analysis is essentially the classification step in RIVPACS type modeling and is run only using reference sites and only using taxa that exist in reference sites. Software programs differ as to how data are prepared for analysis. Generally a site (rows) by taxon (columns) matrix is constructed with binomial data (0 or 1) entered into each cell to indicate the presence or absence of each taxon at each site. Cluster analysis can also be run using abundance data (commonly using Bray-Curtis similarity), which are commonly transformed using log (abundance), relative abundance, or fourth-root abundance. A cumulative taxa list is used, representing the entire list of taxa collected across the study and a record entered for each taxon at each site. At this point, two important factors need to be considered: taxon resolution and the exclusion of rare and/or common taxa.

Taxonomic resolution must be consistent among samples. This does not mean that all organisms must be identified to the same taxonomic level, but that a group (e.g., Diptera) is identified the same way among all samples. Thus, Diptera may be identified to family and Ephemeroptera to genus. In many real-world samples, fragments, juveniles, early instars, and pupae are not

identifiable to the target taxonomic level. These individuals are either not included in the data analysis, or they may be identified at the next higher taxonomic level. During data analysis, it is impossible to tell if records are different species or unidentifiable (e.g., damaged, too immature, etc.) individuals of the same species. There are two ways to use these records: 1) keep the species records or 2) collapse all of the species records to a higher level (Figure 8-4). Whatever choice is made, resolution decisions have to be applied consistently. In general, rules that keep the most data are preferable, but too much lumping can mask the unique elements that distinguish sites. Imagine models built from insect records at the order level only – there are only 13 unique aquatic or semi-aquatic orders of insects to use and the sites would look very similar. On the other hand, if species resolution is used, individuals that could not be identified to species (due to cost, specimen quality, or taxonomic expertise) would be lost. There is a trade-off between comparability of taxonomy among sites and maintaining as much information as possible. Taxonomic resolution rules (species, family, operational taxonomic unit, etc.) need be applied consistently across all sites – reference and test sites. So even though the cluster analysis step of RIVPACS uses reference data – the same taxonomic rules have to be applied to all sites.

The treatment of rare and common taxa in this step of the predictive model process is important as well. In general, rare taxa (occurring at less than 5% of reference sites) are often excluded because they contribute too much unique information for only a few sites and lead to under-clustering (over-splitting) (Hawkins et al. 2000). Likewise, common taxa (occurring at more than 95% of reference sites) are often excluded at this point because they can obscure unique differences among sites and lead to over-clustering (Hawkins et al. 2000). These taxa are not eliminated from the whole process, only from the cluster analysis. They are used later in the construction of expected communities for each site. Once the data have been prepared, with rare and common taxa removed and the validation set of reference sites set aside, a cluster analysis can be performed.

In this approach, the goal of cluster analysis is to produce as many groups as possible to simulate the continuous and dynamic assemblage structure that exists across any region and to minimize the number of unique small groups that would be too hard to predict accurately without overfitting the discriminant function models. Organisms exist along continuous environmental gradients with optima under certain conditions. Of course, there are a multitude of different environmental gradients and many different taxa. Therefore, modeling the distribution of all of those taxa and all of those continuous gradients would not be a trivial exercise. The cluster analysis step is used to dissect the distributions of taxa into as many small groups of co-occurring taxa as possible, much like how one learns to approximate curves by breaking them into small pieces using integral calculus. The ultimate result is a series of unique site clusters with similar taxonomic composition.

Original List

<u>Taxon</u>	<u>Records</u>	<u>Taxon</u>	<u>Records</u>
Family:		Family:	
Baetidae	19	Scirtidae	7
Genera:		Genera:	
<i>Baetis</i>	113	<i>Elodes</i>	1
<i>Callibaetis</i>	49	<i>Prionocyphon</i>	1
<i>Centroptilum</i>	18	<i>Scirtes</i>	1
<i>Cloeon</i>	10		
<i>Heterocloeon</i>	1		



Revised List

<u>Taxon</u>	<u>Records</u>	<u>Taxon</u>	<u>Records</u>
<del>Family:</del>		Family:	
<del>    Baetidae</del>	<del>19</del>	Scirtidae	10
Genera:		<del>Genera:</del>	
<i>Baetis</i>	113	<del>    <i>Elodes</i></del>	0
<i>Callibaetis</i>	49	<del>    <i>Prionocyphon</i></del>	0
<i>Centroptilum</i>	18	<del>    <i>Scirtes</i></del>	0
<i>Cloeon</i>	10		
<i>Heterocloeon</i>	1		

**FIGURE 8-4. A table demonstrating decisions made for lumping taxa upwards or discarding higher taxa records. In the case of the Baetidae, lumping all of the genera to the family level would obscure all of the unique information stored in those five genera, represented by the 191 reference site observations. Clearly removing the 19 records keeps the most information intact. In contrast, while three individual Scirtidae genera were identified, the vast majority of individuals could only be identified to family. Throwing out the seven records in favor of keeping the three genera records would lose the seven reference sites that had Scirtidae present. Clearly, the three genera records should be lumped to family unless there is 100% certainty the seven identified to family represent different genera.**

Cluster analysis actually refers to a suite of different methods that group sites together based on their similarity with regards to many elements. Different cluster analysis approaches have been used in building bioassessment models. Approaches are split into agglomerative (lumping) or divisive (splitting) approaches. Agglomerative cluster analyses start with all of the sites separated and the sites with the greatest similarity are joined to form new groups. This is the most common type used in predictive modeling. Those groups and the remaining individual sites are then compared, and the next most similar elements are joined together – either two other sites or a third site is joined to the first group. The cluster analysis proceeds until all of the sites are grouped together into one large group. As agglomerative cluster analysis proceeds, however, there is less similarity among the elements being joined. By the end, the final group containing

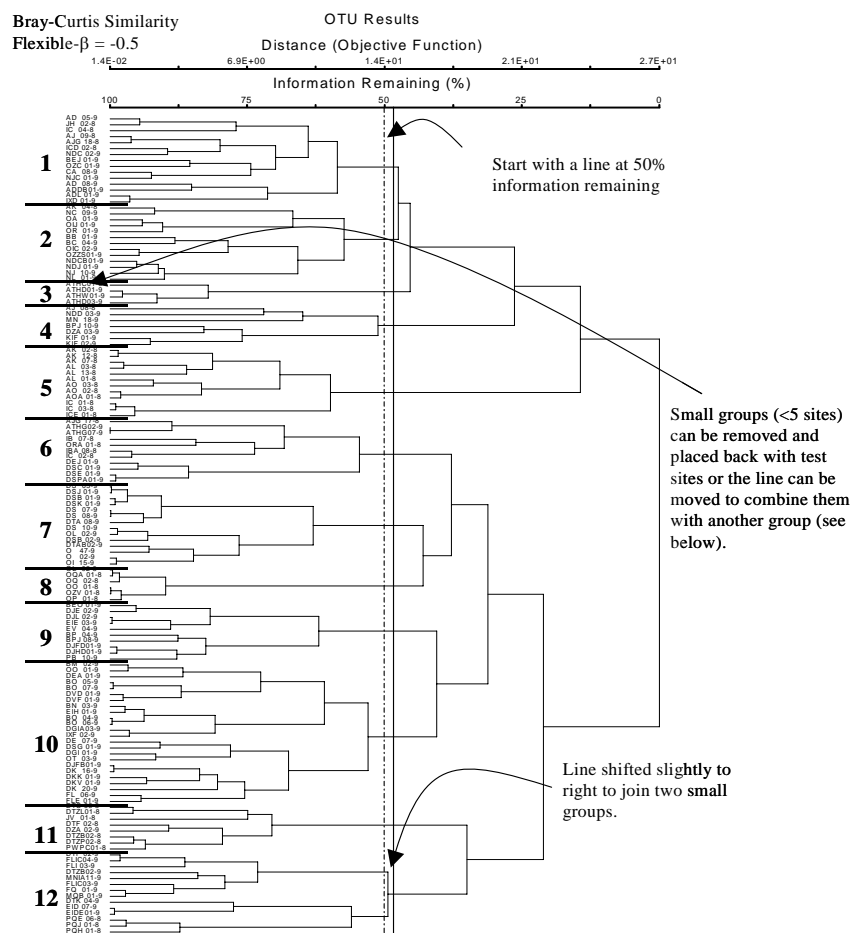
all of the sites has the lowest overall average similarity. The common representation of the process of clustering is the dendrogram – which is a graphical way of viewing the clustering of sites. The axes usually contain some indication of the amount of unique information contained at each level of clustering.

There are a variety of agglomerative approaches, differing in the similarity or dissimilarity indices used and in the rules that are used to link similar elements together during clustering. The two most commonly used similarity indices are the Bray-Curtis and Jaccard indices. The most commonly used linkage methods are the flexible-beta method (with beta commonly set at 0.25), the unweighted pair-group method with arithmetic mean (UPGMA), and Ward's method. In practice, it may be best to explore a variety of approaches (or varying beta) and select the one that gives the best overall clustering.

Divisive cluster analysis is the second approach and works in the opposite direction from agglomerative clustering. Divisive analysis starts with all of the sites grouped together and they are split into the two most dissimilar groups. These two groups are then each split into two dissimilar groups (to yield 4 groups), and so on until either some pre-selected final number of groups is reached or all of the sites are split apart (Gauch 1982).

The most common divisive technique, the two-way indicator species analysis (TWINSPAN), is based on a correspondence analysis of site similarities (Hill 1979). Correspondence analysis is an ordination method that defines an axis along which the sites are ordered in terms of their taxonomic similarity. The mid-point of that axis is located and the sites are split along it into equal halves. Then, two new canonical analyses are run on the two new groups and those groups are split in half, and the process repeats accordingly.

Once a good cluster analysis is achieved, the selection of the optimum number of clusters is made. Obviously, the final cluster (one group) will not work. Likewise, using every individual site will not work. There is a point between these two extremes that represents the optimum number of clusters (Figure 8-5). This step also relies on professional judgment. A good rule-of-thumb is to draw a line in the middle of the cluster axis (e.g., 50% information remaining or other axis value indicating 50% of variance explained) and investigate how many clusters this resolves (Figure 8-5). A cluster consists of all the sites below the stem that is intersected by the line drawn. In many cases, there will be cluster nodes very near this line. So the line can be moved up and down until an optimum set of clusters is selected. The goal is to have as many clusters as possible to resolve the continuous distribution well, while at the same time avoiding very small clusters (<5 sites). Small clusters should either be joined to the next most similar cluster if possible or simply removed and placed with the test dataset. Once the final decision on the number of groups is made, the groups are numbered and each site within a group is given a group code. Again, the ideal approach may be to select three or four final clustering strategies and test each one through the rest of the analyses to see which produces the most precise and responsive models.



**FIGURE 8-5.** A final dendrogram used with a genera only dataset. This example starts with a line drawn at 50% information remaining to delimit groups and then the line is moved slightly to join two smaller groups, resulting in the final 12 groupings. Different software will produce different axes, but generally you start where half the variance is explained. The 14 group models would have worked as well. It would be best to test both groupings.

It is not uncommon to use an independent ordination of the sites using the same presence/absence matrix as a check on the final cluster groupings (Wright et al. 1984). To do this, you use an appropriate ordination technique (e.g., non-metric multi-dimensional scaling or detrended correspondence analysis [DCA]) and give sites within each cluster group unique symbols. Visual assessment of the ordination can then be used to assess whether the groups are also unique in the new ordination space.

### 8.2.2.3 Discriminant Function Analysis

The goal of discriminant function analysis in predictive modeling is to generate a probability that a site belongs to each of the reference cluster groups generated by the cluster analysis. This probability is generated using environmental predictor variables available for each site. Discriminant function analysis (DFA) itself is a technique used when investigators have an

existing grouping structure and want to develop a model to predict the group membership of a new observation (Legendre and Legendre 1998). In some applications, we only want to know into which one group to assign a site. But in the RIVPACS approach, the desire is to generate the probability with which a new site belongs to each of the cluster groups. When a non-reference site has physical characteristics that resemble a mixture of a few different reference groups (e.g., along an ecotone), the expectation is a mixture of the most common taxa found in each of those different groups. The degree of mixture is generated using probabilities derived from discriminant function analysis. An important distinction should be made here. In this context, DFA is being used to build predictive models not to test hypotheses, so many of the statistical constraints are not applicable.

Discriminant function analysis is a mixture of MANOVA and multiple linear regression (MLR) (Statsoft 1994). Like classic ANOVA, MANOVA is a group means comparisons test that can determine if two or more groups are different with respect to many dependent variables simultaneously (Zar 1999). Its importance in discriminant function analysis is to decide if the groups identified with cluster analysis are indeed different with respect to a set of physical predictor variables. If they are not significantly different with respect to the variables, then those variables will not be much use in discriminating among the different groups.

Much like MLR, discriminant function analysis creates a set of equations that are used to predict to which group a site belongs. Unlike MLR, discriminant function analysis uses a canonical ordination approach, most like canonical correlation analysis, to construct linear equations (called discriminant functions) that are the combination of predictor variables that best discriminate among the groups. The number of discriminant functions (also called roots) is always one less than the number of groups or equal to the number of predictor variables, whichever is least. The first discriminant function explains the greatest variation among the different group means (it discriminates the best), the second function explains the second most, and so on. The coefficients in front of each predictor variable, when standardized, indicate which variable is most strongly contributing to the discrimination.

From these functions, a distance is calculated between each site and each group average. The Mahalanobis distance is often used in multivariate space. A site is assigned to the group centroid to which it is closest. But, more importantly, the probability that a site belongs to each group (which is what is needed for predictive models) is derived from the Mahalanobis distance to each group centroid. The closer a site is to a certain group centroid, the more it resembles the environmental characteristics of that group, and the higher the probability it belongs there. Sometimes, however, a site is so anomalous that the Mahalanobis distances to the centroids are very large. Most RIVPACS methods calculate a minimum distance that a site must be to any one centroid to be considered “within the experience of the model,” and these are based on a chi-square value (the 99<sup>th</sup> percentile chi-square value, degrees of freedom = number of groups - 1) (Moss et al. 1987). If none of the groups are within the critical chi-square distance to the site, the site is not assessed since a confident prediction of the probabilities cannot be made.

Most software programs will do all of the discriminant function analysis and most have stepwise options (forward, backward, or mixed) which allow users to choose criteria for selecting or removing variables until some final criterion is met. Entry and removal usually are determined

by F-to-enter and F-to-remove criteria, as in multiple regression. Similar to MLR, these F-values indicate the statistical significance of each variable to the overall discrimination; in essence, the significance with which an additional variable makes a unique contribution to the prediction of groups. Variables will be added in the order of their significance and will be added as long as they meet the criterion. As in MLR, the final model produced in a stepwise discriminant analysis may not be the global optimum. If possible, it is best to test different combinations of starting variables and see which model works best.

A novel approach for selecting an appropriate discriminant model has been developed using all-subsets modeling (Van Sickle et al. 2006). In this approach, all possible combinations of predictor variables are used and run through the calculation of O/E scores for calibration and validation reference data. The best predictor combinations are those that produce models that are the most precise (lowest standard deviation or root-mean square error of O/E scores in reference sites) while avoiding over-fitting (similar values for validation data). These models are available in the R open-source statistical programming language and offer an alternative to the traditional stepwise approach described here. One advantage of this approach is that it considers the universe of possible models, minimizing the risk of selecting locally optimal models. It also places a large value on avoiding overfit models, which is one of the more important risks when constructing these (or any) models.

Among the many statistics often generated from DFA, Wilk's  $\lambda$  is a common statistic used to indicate how well a model discriminates among groups (Pillai's trace, and Lawley Hotelling's trace are other similar statistics) (Zar 1999). Values range from 1.0 (no discrimination) to 0.0 (full discrimination). Wilks  $\lambda$  can be used to help select among the most discriminating models. The all-subsets modeling routines also use Wilks  $\lambda$  to evaluate and select the most discriminating models.

The ultimate test of model performance, in most cases, has been how well they predict the assemblage structure (i.e., how close the number of expected taxa matches the observed) of the reference sites for both the model building and validation datasets, while minimizing the risk of over-fit models. Highly discriminatory models are the goal, but over-fitting problems are also a threat. The all-subsets modeling routines include methods for evaluating the risk of over-fit models (Van Sickle et al. 2006). The value of independent set of reference validation data, however, cannot be overstated. Running the final model through the validation data will also provide an indication of model fit.

The classification of elements into distinct groups is a traditional focus of discriminant function analysis. Predictive modeling, is more interested in the group membership probabilities rather than exact group classification. However, the classification efficiency can also be investigated to look at general model fitting. In DFA, a classification matrix is a matrix of actual group membership vs predicted group membership, and is an *a posteriori* analysis, since it is looking at how well it predicts group membership of sites actually used to build the models. Therefore, it is not truly independent. In classic discriminant function analysis, the group classification functions derived from the discriminant functions are run for each site. A site is then assigned to whatever group classification score is highest. These are compared to the actual group to which each site belongs. In RIVPACS modeling, group classification efficiencies around 50% or less

are not uncommon, especially for small groups. This applies to the validation set as well, which is a more appropriate independent test of the classification efficiency. The all-subsets models actually compare DF classification efficiencies after leave-one-out cross-validation and resubstitution routines to evaluate appropriate model size and model fit.

The final step from the discriminant function analysis is the calculation of group membership probabilities, which is the final product of interest from this step. These membership probabilities were discussed above but need to be explained in detail. As described, the actual goal of the discriminant function analysis is to generate the probability with which each site belongs to each reference group. The cluster analysis was used to break the continuous distribution of communities into discrete pieces and the discriminant function analysis uses the physical characteristics of those groups, in a sense, to place a site back along that continuous gradient. Ideally, each test site would look physically just like one reference group. But what about those sites that fall somewhere among the physical characteristics of a number of groups? As mentioned earlier, those sites would have an equal probability of being in any one of the groups. Those probabilities are generated from the Mahalanobis distances. The Mahalanobis distance is a multivariate distance measure. It is the distance from any one site to the centroids of each of the different groups in multivariate space. The probability a site belongs in each group is derived from those distances – the closer a site is to one centroid, the higher the probability it belongs to that group. Many programs will calculate these probabilities using a variety of methods. In the original RIVPACS formulation, the probabilities were calculated using the formula:

$$p_j = q_j / \sum_{j=1}^k q_j, \quad \text{Equation 3}$$

where  $p_j$  is the probability a site belongs to group  $j$  (of  $k$  different groups). The value  $q_j$  is a weighted distance measure and is defined as:

$$q_j = n_j \times e^{\left( \frac{-d_j^2}{2} \right)}, \quad \text{Equation 4}$$

where  $n_j$  is the number of sites in group  $j$ , and  $d_j^2$  is the square distance (e.g., Euclidean, Mahalanobis) between the site score and each group mean discriminant function score (Moss et al. 1987). These probabilities are the important outcome of the discriminant function analysis. They are combined with taxa frequencies in each group to predict the final taxonomic composition of a site. This will be explained in the next section.

Some sites are so far from any group centroid that an accurate determination of the probabilities cannot be made. The critical distance for a site to be accurately determined is calculated using the 99<sup>th</sup> percentile chi-square distribution value based on degrees of freedom equal to the number of groups. Since the Mahalanobis distances follow a chi-square distribution, any site that does not contain a distance less than the critical distance cannot be adequately assigned a probability and is considered “outside the experience of the model”. These sites are often set aside and must wait until more reference sites with similar physical characteristics are assessed and the model is updated. If the sites are taken through the prediction analysis, any conclusions using O/E scores

generated from these sites need to be tempered by the fact that they are physically distinct from the reference groups used to construct the models.

#### 8.2.2.4 Prediction of Taxa Composition

The final step in model building is to predict the number of expected taxa for a site. Before this step takes place, rare and common taxa removed before cluster analysis are added back into the database. These taxa, while rare over all sites, may be frequently found in one group and would be an important prediction for that group. Once these are reincorporated, the prediction analysis proceeds.

As mentioned before, the predicted taxa list for a site is not based solely on the taxa composition of the one reference group to which a site is most similar. If that were the case, one could simply find the group to which the site had the highest probability of belonging and compare the observed assemblage to the average assemblage composition of that one group. If each test site looked exactly like only one reference group, this would be fine. But sites are often physically similar to several groups, because the groupings frequently reflect subtle differences among reference sites (e.g., low gradient vs high gradient reaches within one basin). The sensible thing is to predict a mixture of taxa based on: 1) which group a site is most similar to and 2) which taxa are most frequently found in those groups. Therefore, essentially, a weighted average expected assemblage composition is calculated. This is done by using the probability a site belongs to each reference group as the weight and then multiplying this by the frequency of taxa in each reference group (Moss et al. 1987).

In order to do this, the frequency of each taxon in each reference group has to be estimated. This is done by calculating the frequency with which each taxon is found in each group (Table 8-3);  $g_{j,x}$  = proportion of reference sites in group  $j$  containing taxon  $x$ . This value is calculated for each taxon in the master taxa list (over all sites). In the end, each taxon has a frequency with which it occurs in each reference group. Many taxa from the master list are not found in every group; therefore, they will have a probability of zero where they are absent; others are ubiquitous and have a value near 1.0 for every reference group.

Now that the probability of membership of any site in each reference group ( $p_j$ ) has been generated from the discriminant function analysis and the frequency of every taxon  $x$  in each reference group ( $g_{j,x}$ ), the probability of capturing ( $P_c$ ) each taxon  $x$  at any site can be calculated (Table 8-4):

$$P_{c,x} = \sum_{j=1}^k p_j \times g_{j,x}, \text{ for } k \text{ reference groups.} \quad \text{Equation 5}$$

**TABLE 8-3.** The first component of the prediction phase is to estimate average assemblage composition of reference groups. For each taxon, the fraction of reference sites containing each taxon is calculated. This is an estimate of the frequency ( $g_{j,x}$ ) with which each taxon ( $x$ ) is found in each group ( $j$ ). A sample for a few taxa is shown here. Not all the reference sites could fit in the table. But for the first taxon (*Ablabesmyia*) in group 1, 11 of the 15 sites had that taxon; therefore, its frequency at that site is  $11/15 = 0.73$ . Only 1 of the 15 sites in group 1 contained *Acroneuria*, therefore its frequency in group 1 is  $1/15 = 0.07$ . This proceeds for all taxa (even the rare ones added back in) and for all 12 groups. Note that some taxa are fairly common across the groups (*Baetis*) whereas others are frequent in only a few groups (*Acroneuria*).

		Frequencies				
Group	1	0.73	0.07	0.33	0.73	0.27
	2	0.77	0.00	0.38	0.15	0.23
	3	1.00	0.00	0.00	0.50	0.25
	4	0.29	0.14	0.14	0.57	0.00
	5	0.75	0.58	0.00	0.75	0.17
	6	0.09	0.00	0.00	1.00	0.09
	7	0.93	0.00	0.36	1.00	0.00
	8	1.00	0.00	0.50	1.00	0.00
	9	0.00	0.00	0.00	1.00	0.00
	10	0.17	0.00	0.21	1.00	0.00
	11	0.00	0.00	0.00	0.88	0.00
	12	0.21	0.00	0.29	0.93	0.00
Site	Group	<i>Ablabesmyia</i>	<i>Acroneuria</i>	<i>Anopheles</i>	<i>Baetis</i>	<i>Basiaeschna</i>
AD 05-92	1	0	0	0	0	0
AD 08-92	1	1	0	0	1	0
ADDB01-92	1	0	0	0	1	0
ADL 01-92	1	0	0	0	0	0
AJ 09-87	1	1	0	1	1	1
AJG 18-87	1	1	1	0	1	1
BEJ 01-96	1	1	0	1	1	0
CA 08-98	1	1	0	0	1	0
IC 04-87	1	1	0	0	0	0
ICD 02-87	1	1	0	1	1	1
IXD 01-92	1	1	0	1	1	0
JH 02-84	1	0	0	0	1	0
NDC 02-95	1	1	0	0	1	0
NJC 01-95	1	1	0	0	0	1
OZC 01-96	1	1	0	1	1	0
AK 04-86	2	0	0	0	0	0
BB 01-96	2	0	0	0	0	0
BC 04-96	2	1	0	0	0	0
NC 09-95	2	0	0	0	0	0
NDCB01-95	2	1	0	0	0	1

**TABLE 8-4. Having calculated the taxon frequencies ( $g_{i,x}$ , above) and the group probabilities ( $p_j$ , from the discriminant function analysis), the product of these values is used to calculate the probability of capturing each taxon at a site ( $P_c$ ). For example, for *Ablabesmyia* at site AD 05-92, the probability that a site is in each group ( $p_j$ ) is multiplied times the frequency of finding *Ablabesmyia* in each reference group ( $g_j$ ). The sum of those products = 0.713, which is the probability of capturing *Ablabesmyia* at this site. The same calculation is made for all taxa.**

Site		AD 05-92						
Group	Frequencies ( $g_{j,x}$ )			Probability of Group Membership ( $p_j$ )			( $g_{j,x})(p_j$ )	
	<i>Ablabesmyia</i>	<i>Acroneuria</i>	<i>Baetis</i>		<i>Ablabesmyia</i>	<i>Acroneuria</i>	<i>Baetis</i>	
1	0.73	0.07	0.73	0.657	0.479	0.046	0.480	
2	0.77	0.00	0.15	0.012	0.009	0.000	0.002	
3	1.00	0.00	0.50	0.136	0.136	0.000	0.068	
4	0.29	0.14	0.57	0.015	0.004	0.002	0.009	
5	0.75	0.58	0.75	0.096	0.072	0.056	0.072	
6	0.09	0.00	1.00	0.081	0.007	0.000	0.081	
7	0.93	0.00	1.00	0.000	0.000	0.000	0.000	
8	1.00	0.00	1.00	0.002	0.002	0.000	0.002	
9	0.00	0.00	1.00	0.001	0.000	0.000	0.001	
10	0.17	0.00	1.00	0.001	0.000	0.000	0.001	
11	0.00	0.00	0.88	0.000	0.000	0.000	0.000	
12	0.21	0.00	0.93	0.000	<u>0.000</u>	<u>0.000</u>	<u>0.000</u>	
Probability of Capture ( $P_c$ ) =					0.713	0.102	0.717	
$P_c = \Sigma(g_{j,x})(p_j)$								

Note that each probability of capturing a taxon is a continuous probability and not a discrete number. It is derived from the probability of group membership and the distribution of taxa frequencies. The expected number of taxa (E), then, is the sum of the capture probabilities of all the taxa at a site:

$$E = \sum_{x=1}^i P_{c_x} \quad \text{Equation 6}$$

This value is compared to the sum of the expected taxa (from the same master taxa list) actually observed (O) at the site. It is important to note that the number of observed taxa is the sum of only those expected taxa that are actually observed. The final ratio of these values (O/E), is the proportion of expected taxa actually observed at the site and is the indicator of biological condition. At relatively undegraded sites, one would expect to capture all the taxa frequently found in reference sites of comparable physical characteristics from the same region and O/E = 1.0. The lower the O/E ratio is, the fewer expected taxa actually captured.

Because the expected number of taxa is generated from a continuous frequency distribution over many reference sites within a group, capture probabilities can range from 0 to >1.0. It is possible

to have a test site with  $O/E > 1.0$ , where there are more taxa captured than expected. This reflects a site where one observes many taxa with even partial probabilities of capture (e.g. 0.4), so that the sum of observed taxa (integers) is greater than the sum of expected (fractions  $< 1.0$ ). The average reference site  $O/E$  score, however, ought to be equal to 1.0 and this is used as a check on the adequacy of the model. If the mean reference  $O/E$  is significantly different from 1.0, then there is a problem with the model and it would need to be checked.

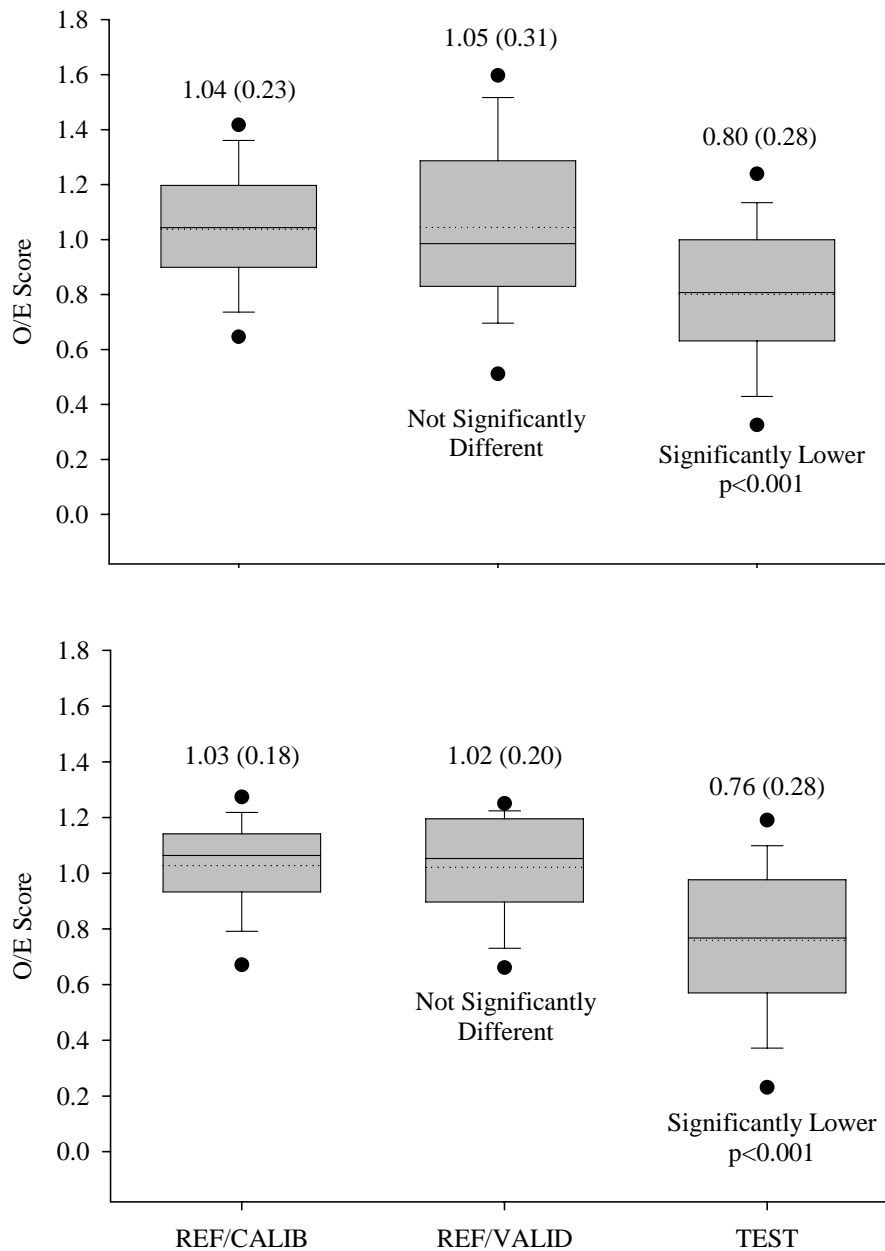
To this point, all taxa have been considered, regardless of their probability of capture at each site, which introduced some variability in comparing fractional expected data to integer observed data. Several RIVPACS-type model applications constrain the expected taxa list to only those taxa with a capture probability  $> 0.5$  (e.g., AUSRIVAS, Simpson and Norris 2000). This limits the list of taxa considered (both observed and expected) to only the most commonly expected. It is important to remember that the number of observed taxa is the sum of only those expected taxa that are actually observed. So if one only uses taxa with  $P_c > 0.5$  to estimate the expected number of taxa, one would only count actual observations of that same restricted taxa list, not all of the observed taxa.

The primary test of final model adequacy is running an independent set of validation reference sites through the model and calculating  $O/E$  scores (Hawkins et al. 2000, Simpson and Norris 2000). Therefore, a test of model robustness is that the  $O/E$  of the validation dataset is not significantly different from the  $O/E$  of the dataset used to construct the models, and neither of these means should be significantly different from 1.0 (Figure 8-6).

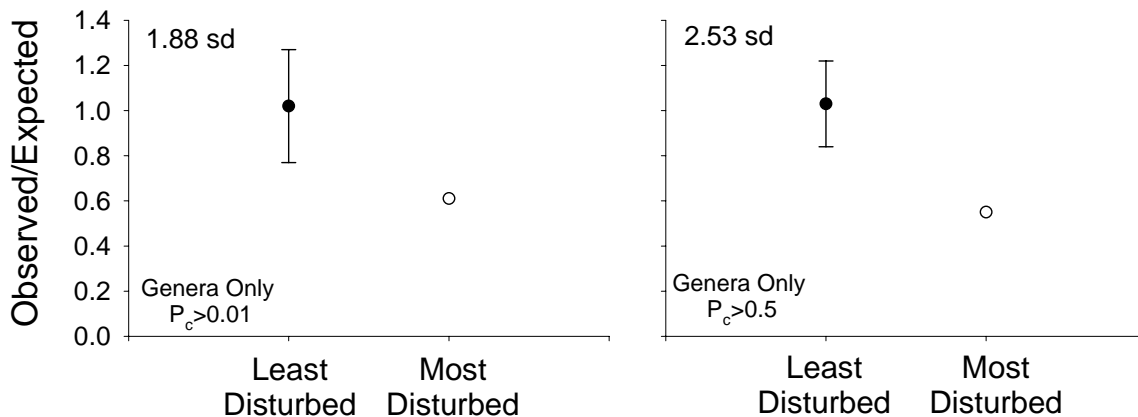
A second test of model adequacy is model precision. The objective is to create models with as low a variation about the mean reference score (1.0) as possible (i.e., to precisely predict the observed taxa). One rule of thumb is a standard deviation (or root mean square error) of mean reference  $O/E$  score of 0.15 to 0.20 or less (Figure 8-6). The lower the standard deviation is, the more precise the model and the greater the potential discrimination of degraded rivers. This also means that more degraded classes can be resolved.

A third way of assessing the model adequacy is to compare the model with a null model (Van Sickle et al. 2005). The null model for these predictive models is to simply compare the number of observed taxa at each reference site to the sum of the average frequency of taxa across all the reference sites without any clustering. In the null model, therefore, the number of expected taxa ( $E$ ) is the same for every site. In essence, the null model ignores the cluster analysis and discriminant analysis and assesses how much extra precision one adds by going through those calculations. This is fairly straightforward to do, can be done for any capture probability threshold, and is a good check on the modeling effort.

One final important test of the models is whether or not the  $O/E$  scores respond to disturbance. There are a number of ways to evaluate disturbance response (Figure 8-7). Gradient approaches evaluate the response of  $O/E$  scores to a pre-determined disturbance gradient (e.g. water chemistry, land use, habitat, etc.). Another approach would be to rank degradation classes and test whether there are significant differences in mean  $O/E$  score between the reference class and the degradation classes, using either ANOVA or some other means comparison test.



**FIGURE 8-6.** This figure shows the O/E score distributions for reference calibration, validation, and non-reference test site data. The dataset used genera only data and results for both the  $P_c > 0.01$  (top) and  $P_c > 0.5$  (bottom) taxa are shown. Mean scores are shown along with the standard deviations in parentheses. Reference validation (REF/VALID) scores were not significantly different from reference calibration (REF/CALIB) data scores, but non-reference scores (TEST) were significantly lower than both reference datasets (ANOVA, Tukey's HSD comparison). Whiskers indicate the 10<sup>th</sup> and 90<sup>th</sup> percentiles, and the boxes indicate the 25<sup>th</sup> and 75<sup>th</sup> percentiles. The solid line is the median and the dotted line the mean O/E scores. Filled circles indicate the 5<sup>th</sup> and 95<sup>th</sup> percentile values.



**FIGURE 8-7. Comparisons of the discrimination between least and most disturbed sites using  $P_c > 0.01$  and  $P_c > 0.5$  taxa. One standard deviation around the mean least disturbed O/E scores is shown and compared to the mean O/E scores for the most disturbed sites. The number of least disturbed standard deviations between the least and most disturbed mean scores is also shown. The higher this value, the greater the discrimination.**

### 8.2.3 Estimating Measurement Error

Variability has many possible sources. In dealing with bioassessment variability, the goal is to: 1) minimize variability due to uncontrolled measurement error and 2) characterize and partition the natural variability. When sampling rivers measurements (e.g., taxa richness) are often made at single points in space and time (riffle, 10-cm depth, 10 AM on 2 July). If the same measurement is made at a different place (pool, 1-m depth) or time (30 January), the measured value will be different. These two natural components of variability (space and time in this example) are called sample variability or sampling error (Fore et al. 1994). A third component of variability, called measurement error, refers to the ability to accurately measure the quantity of interest. Measurement error can be affected by sampling gear, instrumentation, errors in proper adherence to field and laboratory protocols, the choice of methods used in making determinations, and small-scale spatial variability at the sample site. The three basic rules of efficient sampling and measurement are:

1. Sample so as to minimize measurement error.
2. Characterize the components of variability that influence the central questions and reporting units.
3. Control other sources of variability that are not of interest, in order to minimize their effects on the observations.

In the example of taxa richness, it may be useful to sample each of several rivers, with a 1-m<sup>2</sup> kick net in early spring before the bulk of emergence. Many reaches are sampled in this example to examine and characterize the variability due to different reaches (the sampling unit). Each reach is sampled in the same way, in the same place, and in the same index period (time frame)

in an attempt to minimize variability due to habitat and season, which are not of interest in this particular study.

In the above example, taxa richness may vary with habitat, among rivers, and time of sampling (season, year), a fact that may be particularly important for sets of reference sites. If the spatial and temporal components of variability within rivers are large, then it is best to use either an index period sample or to estimate a composite from several determinations. For example, taxa richness may vary more between spring and fall samples within a river than among similar rivers within an ecoregion.

Measurement error is the result of methodological bias and errors: gear bias; improper use of gear or improper training; variability in use of gear; laboratory errors (chemical analysis errors); and natural variability that is not of interest and is not being sampled. Measurement error is minimized with methodological standardization; selection of cost-effective, low variability sampling methods; proper training of personnel; and quality assurance procedures designed to minimize methodological errors. Method performance standards (see Chapter 3) are designed to help ensure that these kinds of errors are consistently held to a minimum.

ANOVA is used to estimate measurement error. All multiple observations of a variable are used (from all streams with multiple observations), and streams are the primary effect factor. The root mean square error (RMSE) of the ANOVA is the estimated standard deviation of repeated observations within sites. A hypothesis test (F-test) is not of interest in this application because it tests the trivial hypothesis that sites are different from one another.

The estimated RMSE is used in the same way as standard deviation; in this case, it is an estimate of the standard deviation around a single point, and it is used to estimate a confidence interval around the point. The advantage is that a confidence interval can be estimated without replication at the site, because we are using the population estimate of standard deviations around single measurements.

Having a standard deviation, one can estimate confidence intervals around an index score or O/E score of a site (Fore 2004). For a single (non-repeated) sample, the confidence interval is:

$$CI = \pm Z_{\left(1-\frac{\alpha}{2}\right)} \frac{S_{rep}}{\sqrt{n_{rep}}}, \quad \text{Equation 7}$$

where  $S_{rep}$  is the standard deviation calculated as RMSE with ANOVA,  $n_{rep}$  is the number of replicates at the site, and  $Z_{(1-\alpha/2)}$  is the cumulative standard normal deviate ( $Z$  – score) for the chosen  $\alpha$ . This approach makes three assumptions:

- measurement error is normally distributed,
- measurement error is not affected by site class or impairment, and

- the sample standard deviation of repeated measures is an unbiased estimate of population measurement error.

The same procedure can be used to estimate variability due to season and year, if sites are resampled in multiple seasons or years.

Natural variability that is not of interest for the questions being asked, but may affect ability to address these questions, should be estimated with the RMSE method. If the variance estimated from RMSE is unacceptably large (i.e., as large or larger than variance expected among sample units), then it is often necessary to alter the sampling protocol, usually by increasing sampling effort in some way, to further reduce the measurement error. Measurement error can be reduced by multiple observations at each sample unit (e.g., multiple Ponar casts at each sampling event, multiple observations in time during a growing season or index period, depth-integrated samples, or spatially integrated samples).

Spatial integration of sample material and compositing the material into a single sample is almost always more cost-effective than retaining separate, multiple observations. This is especially true for relatively costly laboratory analyses such as organic contaminants and benthic macroinvertebrates.

For quality assurance, some effort will always be required for repeated samples so that measurement error can be estimated from a subset of sites. Repeated measurement at 10% or more of sites is common among many monitoring programs, and is recommended (see Chapter 3).

### **8.3 Site-Specific Assessments**

The next two sections deal primarily with the analytical approaches that can be used for site-specific and watershed assessments. The design considerations for these approaches were outlined in Chapter 3. Here we describe the analysis methods.

For site-specific assessments using before-after control-impact (BACI) type designs, the analytical approaches depend on which of the designs was used. In any of the analytical approaches, however, some attention to data preparation is necessary. Most of the BACI analysis approaches use a form of analysis of variance (ANOVA) or simpler parametric means comparison tests (e.g., t-test). As mentioned in the multimetric data preparation discussion above, variables that will be compared using parametric analyses must adhere to some basic assumptions. All the variables to be used, including individual metric, multimetric or O/E scores, should be explored visually for normality and equal variance. There are tests that can be used to examine these assumptions as well (e.g., Kolmogorov-Smirnov). The most important assumption is independence of observations. As mentioned above, as long as there is substantial spatial and temporal separation of sampling, independence generally should not be a problem.

The simplest test in site-assessment design is the t-test. The t-test can be used to compare two means or to compare a mean to some specific value (i.e., is the mean O/E score in samples below a discharge different from 1.0?). In the BACI designs, simple t-tests can be used to compare

pair-differences between the before and after periods. A significant t-test would suggest that mean differences changed after the treatment (e.g., impact, discharge location, or restoration activity) (Rathbun 1999, Smith 2002). The t-test can be performed using any standard packaged software and conceptual information is available in any introductory statistical text (Sokal and Rohlf 1995). Non-parametric versions of the t-test can also be used, the Mann-Whitney test being the most common. Again, these tests are explained in most texts.

In addition to the t-test, a simple ANOVA can be used to test difference in means between before and after (period) data or control and impact data (location). In this case, only two means are being compared, but sampling times are used to parse some of the variance of the model (Table 8-5). The significance is tested on one contrast alone. An extension of this simple comparison is when multiple sample sites exist either upstream and downstream, or before and after an impact. The simple ANOVA is extended by including a factor for sites, which are treated as replicates (Table 8-6). The presence of the site replicates affects the principal factor comparison (period or location) by attributing variance to the sampling location. In classic BACI designs, however, control sites are added and both before-after and control-impact contrasts are available, and the interaction term between BA and CI is the statistic of interest. ANOVA is also commonly used in this approach (Table 8-7).

The logical extension of the BACI model is to include multiple paired sampling times. The analysis is similar to repeated measures, and the ANOVA table for this BACI paired (BACIP) design adds a factor for the sampling times within each period, but the interaction is still the statistic of interest (Table 8-8). As noted above, paired samples between the control and impact site can be represented as differences between the two paired observations and the before and after period mean differences compared with a two-sample test (Stewart-Oaten et al. 1986). If a two sample t-test is used to compare differences, the analysis is similar to the interaction test (Underwood 1991, Smith et al. 1992, Smith 2002). The final version includes the incorporation of multiple control streams as well as multiple sample times (Table 8-9, Underwood 1991, 1994). Once more, the interaction test is the statistic of interest; but some have argued that more individual contrasts can also be used, for example breaking the BA x CI sum of squares into variance associated with before (B x CI) and after (A x CI). Other extensions also exist and are discussed in Underwood (1992, 1994).

#### **8.4 Watershed Assessments**

The general focus of watershed assessments is to characterize resource condition across a watershed or a broad region. For example, such assessments are routinely performed for meeting 305(b) reporting requirements under the CWA and the design options were discussed in Chapter 3. The probabilistic designs favored for this approach lend themselves to a variety of analyses related to a number of assessment elements. This section discusses a few of these analytical options.

**TABLE 8-5. An ANOVA table for the simple before-after model. A test for Location (control vs impact or upstream vs downstream) would be similar, but the location would be the principal treatment instead of Period. MS = mean squares;  $t_B$  and  $t_A$  are the number of observations before and after (Smith 2002).**

Source	SS	df	F
Period: Before-After	$SS_{BA}$	1	$MS_{BA}/MS_{times}$
Sampling times	$SS_{times}$	$t_B + t_A - 2$	
Total	$SS_{Total}$	$t_B + t_A - 1$	

**TABLE 8-6. ANOVA table for a similar design to Table 8-5, but with multiple sampling sites for each treatment. M indicates the number of sites (Smith 2002).**

Source	SS	df	F
Period: Before-After	$SS_{BA}$	1	$MS_{BA}/MS_{times}$
Sampling times	$SS_{times}$	$t_B + t_A - 2$	
Replicate sites	SSE	$(M-1)(t_B + t_A)$	
Total	$SS_{Total}$	$M(t_B + t_A) - 1$	

**TABLE 8-7. ANOVA table for the two-factor BACI design. N is the total number of observations, with multiple observations over time or space (Smith 2002).**

Source	SS	df	F
Period: Before-After	$SS_{BA}$	1	
Location: Control-Impact	$SS_{CI}$	1	
Interaction: BA x CI	$SS_{BACI}$	1	$MS_{BACI}/MS_E$
Error	SSE	$N-4$	
Total	$SS_{Total}$	$N-1$	

**TABLE 8-8. ANOVA table for the BACIP design (Smith 2002).**

Source	SS	df	F
Period: Before-After	$SS_{BA}$	1	
Times within period	$SS_{t(BA)}$	$t_B + t_A - 2$	
Location: Control-Impact	$SS_{CI}$	1	
Interaction: BA x CI	$SS_{BACI}$	1	$MS_{BACI}/MS_E$
Error	SSE	$t_B + t_A - 2$	
Total	$SS_{Total}$	$2(t_B + t_A) - 1$	

**TABLE 8-9. ANOVA table for the asymmetrical BACI design with L-1 control sites and N observations (Smith 2002).**

Source	SS	df	F
Period: Before-After	SS <sub>BA</sub>	1	
Times within period	SS <sub>t(BA)</sub>	t <sub>B</sub> + t <sub>A</sub> - 2	
Location: Control-Impact	SS <sub>CI</sub>	L-1	
Interaction: BA x CI	SS <sub>BACI</sub>	L-1	MS <sub>BACI</sub> /MS <sub>E</sub>
Error	SSE	(L-1) x (t <sub>B</sub> + t <sub>A</sub> - 2)	
Total	SS <sub>Total</sub>	N-1	

In a truly random design, the estimate of average condition is fairly straightforward and is simply calculated as the overall mean ( $\bar{y}$ ) of all the values. The variance of the mean is estimated as:

$$\text{var}(\bar{y}) = \frac{s^2}{n}, \quad \text{Equation 8}$$

where  $s^2$  is the sample variance (Rathbun 1999). Similarly, the proportion of river miles in a certain condition can be estimated from such designs as  $\hat{p}$ , the proportion of sample sites showing that condition with corresponding variance:

$$\text{var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1} \quad (\text{Rathbun 1999}). \quad \text{Equation 9}$$

For stratified random designs, individual strata means can be calculated; and an average overall mean condition across the entire study region can be calculated as a weighted average, where the percent of the resource within each strata is used as the weight. The region is split into  $K$  strata and the average condition for environmental variable  $\bar{y}$  can be estimated as:

$$\bar{y} = \frac{1}{L} \sum_{h=1}^K L_h \cdot \bar{y}_h, \quad \text{Equation 10}$$

where  $\bar{y}_h$  is the sample mean of  $n_h$  observations in stratum  $h$  calculated as:

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}, \quad \text{Equation 11}$$

$L_h$  is the length of rivers in stratum  $h$ , and  $L$  is the total river length in the population of rivers (Rathbun 1999). The variance estimate for the regional average is:

$$\text{var}(\bar{y}) = \frac{1}{L^2} \sum_{h=1}^K L_h^2 \frac{s_h^2}{n_h}. \quad \text{Equation 12}$$

The proportion of river miles in a given condition can be estimated as:

$$\hat{p} = \frac{1}{L} \sum_{h=1}^K L_h \cdot \hat{p}_h \quad (\text{Rathbun 1999}), \quad \text{Equation 13}$$

where  $\hat{p}_h$  is the proportion of sample stations from stratum  $h$  showing that condition. The variance associated with the measure is:

$$\text{var}(\hat{p}) = \frac{1}{L^2} \sum_{h=1}^K L_h^2 \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}. \quad \text{Equation 14}$$

## 8.5 Gradient Designs

The design chapter discussed the use of gradient designs to identify trends in condition variables with source or stressor data (e.g., to assess biological condition under varying levels of urbanization). These designs primarily use regression and correlation analysis approaches. Only a brief discussion is given here, and interested users are directed to texts on regression and correlation analysis. Ordinary least squares regression and correlation are the simplest designs, where one is interested in exploring or predicting a particular dependent variable response given a level of some independent variable. Ideally, these data should all meet the requirements of standard parametric statistical analyses, and transformations should be used if these assumptions are violated. Data preparation is, therefore, also an important step in these analyses. It is strongly recommended that bivariate scatter plots be used to examine bivariate relationships before running correlation or simple linear regressions. These plots are valuable in exploring the strength and nature (linear or non-linear) of the relationships among variable pairs (Reckhow and Warren-Hicks 1997), and may recommend transformations worth exploring.

In reality, it is difficult to randomize all the sites, as one is often interested in reflecting the entire potential range in source or stressor levels. One potential solution for this is to use a validation set of data. Randomly selecting 10-20% of the available data and setting it aside, building the regression models and testing their accuracy with validation data is one option. Resampling approaches (e.g., bootstrapping or jackknifing) could also be used, especially if setting aside data is not an option (e.g., sample size issues). Again, it is important to guard against the over extension of regression models used in this way. There is a temptation to link correlation or regression as used here to causation. Technically, because of limited control over the independent variables, causation is a problematic concept. Correlation and regression certainly increase insight and can contribute to strength-of-evidence arguments, but when used as data mining tools, they can often lead to spurious relationships where causation is theoretically troubling (e.g., the ratio of percent row crop to percent evergreen land cover in the riparian zone does not cause a decrease in species diversity per se). Causal pathways still often need to be identified. There is great danger in packaged software that allows large batch correlation and regression modeling. Technically, running 100 correlations will lead to five significant results (at  $p = 0.05$ ) from chance alone. Care must be taken to adjust the acceptable significance level for multiple unplanned tests (e.g., Bonferroni or Dunn-Šidák) and perhaps use this to better guide which relationships merit attention. In addition, a number of model diagnostics exist, though too

many to adequately cover in this document, but they are likely covered in most introductory statistical texts or regression analysis texts. Critical diagnostics include regression coefficients, residuals analysis, outlier analysis, and goodness of fit.

Multiple linear regression approaches can also be used. In these cases a dependent variable of interest is regressed against a number of independent variables. Single, combinatorial, or transformed independent variables can all be used. This powerful tool allows the user to compare effects of variables together and also to dissect the partial contribution of individual variables to the total response. As an example, one could explore the contribution of riparian canopy cover to the response of an index to land use alteration. In many cases, intact riparian zones contribute to higher index scores than those predicted for a given level of watershed disturbance (e.g., urban land cover) (Yoder et al. 1999). This can be assessed using partial residuals analysis in multiple linear regression.

Multiple regression models can either be forced (where a set of independent variables are used in the model) or variable selection procedures can be used where variables are added in the order with which they reduce the overall variance (forward selection, backward elimination, and stepwise approaches can be used). All possible model approaches can also be used, but run the same risk for any multiple tests approach (see above).

As with any approach, there are a number of pitfalls with multiple regression. It is very easy to generate significant multiple regression models as every added variable will reduce the error of the model. One risk is generating over-fit models (models that are unique to the modeled data, but not generally applicable). This can be avoided either using validation data or any number of tools that penalize a model for adding additional variables (e.g., Akaike's Information Criterion, AIC). Another big risk in multiple regression models is multicollinearity, or the inclusion of independent variables that are redundant. Multicollinear predictors dramatically impact the estimation of regression coefficients and may increase the risk of overfitting, which should be assessed. Removing highly correlated variables is encouraged, and diagnostics for identifying multicollinearity also exist in many software programs (e.g., variance inflation factor).

A variety of model diagnostics (in addition to those just described) exist for multiple regression; most of which are similar to those used in linear regression. They are also related to residuals analysis, outlier or leverage point analysis, and model fit. One unique diagnostic for multiple regression is partial residual analysis. Partial residual analysis examines the relationship between the response variable and a predictor when the effect of all other predictors is removed (i.e., already modeled). This approach allows the user to look at the unique contribution of individual predictors and can be done numerically and visually.

Exploratory pattern analysis across large gradients can also take advantage of the large number of multivariate statistical approaches. These methods (e.g., principal components analysis [PCA], detrended canonical correspondence analysis [DCCA], and non-metric multidimensional scaling) can be used to identify patterns in environmental stressors related to sources and to identify patterns in assemblage change across environmental gradients. These approaches are especially useful with large datasets containing many variables, like the ones being generated by many agencies. The approaches are designed to reduce the dimensionality of data to identify

prominent gradients. Users interested in multivariate statistical analyses should consult the array of resources available to guide these analyses (e.g., Manly 1994, Legendre and Legendre 1998, McCune and Grace 2002)

## **8.6 Reporting Results**

This section briefly describes strategies for report writing that have been successful in assessment programs. The topic was dealt with in Barbour et al. (1999), and here we review important elements from that discussion. Reports should be tailored to the intended audience. Technical reports intended for fairly knowledgeable audiences can include greater detail on design and methodological description, greater flexibility in use of technical graphics that may require some sophistication to interpret, and more detailed discussion of technical issues. These reporting formats are likely familiar to most technical experts in any field and, for professional manuscripts, are dictated by the intended journal.

More frequently, however, assessment information is reaching a broader, less technical or non-technical audience including water resource managers and environmentally conscientious citizens. Communicating the condition of water resources and the potential impact of human activities on those resources is an important goal of resource monitoring (Karr and Chu 1999). Effective communication is vital for conveying technical information to non-technical audiences involved in important environmental decision-making. Reporting style and formats are important for assuring this is done accurately and efficiently, and a variety of resources are available to guide reporting (e.g., USEPA Office of Environmental Information).

### **8.6.1 Graphical Displays**

The adage that “a picture is worth a thousand words” is no less true for conveying science than it is for conveying any other information. Well-designed, straightforward graphs can more effectively reveal patterns in biological response than strictly statistical tools, especially for non-technical audiences. Patterns, including outliers, may convey important information for both site assessment and diagnosis (Karr and Chu 1999). Some examples of useful graphical techniques are presented for specific program objectives:

- **Classification** – Graphs should easily convey strong differences among groups of sites within classes. Two common displays are bivariate scatter plots (Figure 8-8) from ordinations for clarity and cluster dendrograms, used to compare degree of separation of site groups based on sets of characteristics (Figure 8-9). Both are used to support classification decisions for building models.
- **Problem identification and water resource status** – Conveying information about the status of water resources of regional condition assessments is a critical task of assessment programs. This information needs to be conveyed easily and clearly. It also requires consolidating information from many samples. Pie charts (Figure 8-10), box and whisker plots (Figure 8-11), and bar charts are straightforward reporting tools for this job.

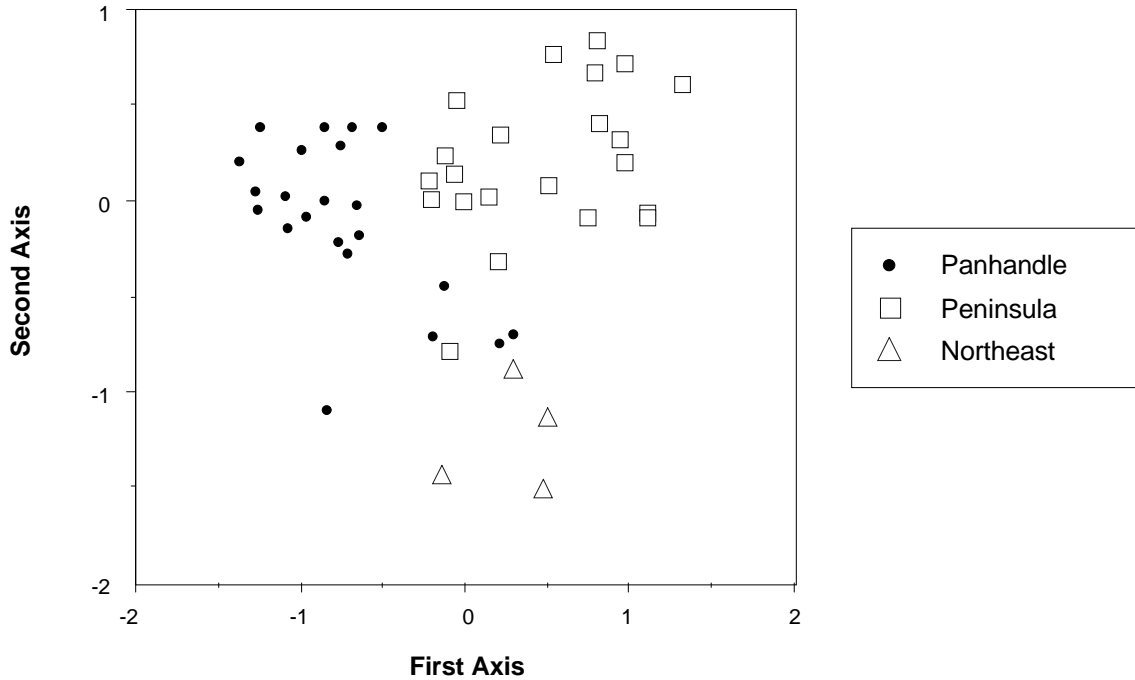
- Trend monitoring and assessment – Illustrating resource changes over space or time requires graphical displays that convey trends. Line graphs are ideal tools (Figure 8-12). Cumulative frequency curves are a bit more technical (Figure 8-13), but are also efficient ways to illustrate the percent of observations below some critical value.
- Causal diagnosis – Illustrating sources of impairment is not necessarily straightforward, as it often requires the evaluation of several variables simultaneously or in series. Indeed, the process of stressor diagnosis with environmental monitoring data is multi-faceted and challenging. However, certain graphical approaches do lend themselves to comparing diagnostic information. Bar charts, sun ray plots, and box-and-whisker plots are good options (Figures 8-14–8-16).

### **8.6.2 Report Format**

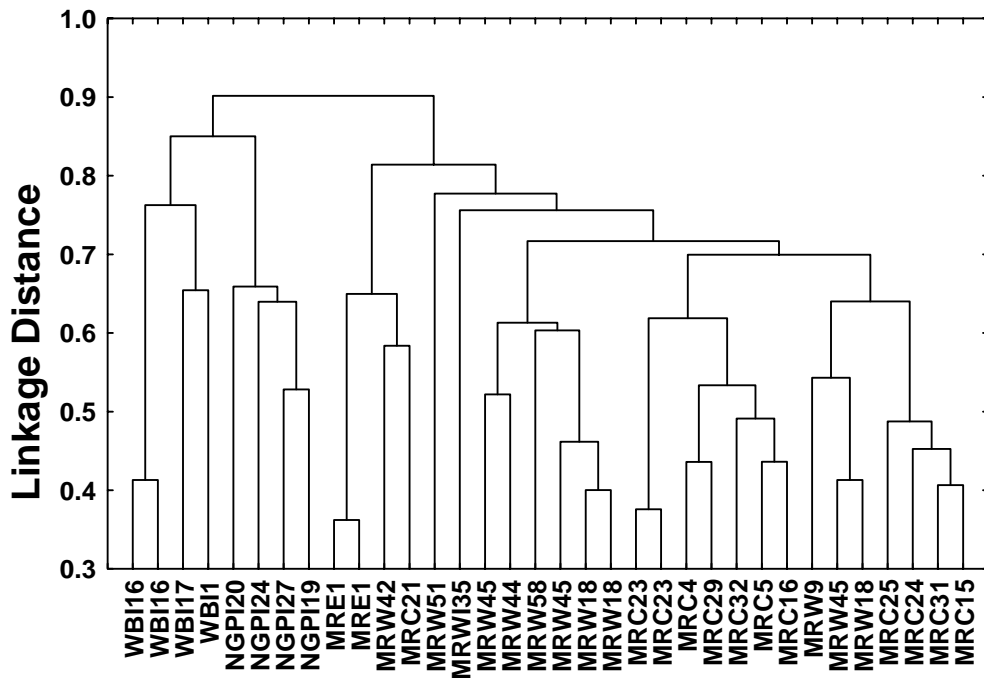
Two common formats are used for assessment reports: summary reports for making management decisions and those designed for more technical audiences. The goal of each is to highlight the objectives, scientific process, results, and final assessment. However, the first format is designed for use by managers and can also be valuable as a public information tool. The latter format is better for technical review and dissemination of scientific results to an audience of technical peers.

The ecosummary is an example of the first format (Figure 8-17). The style is simple, efficiently documents results, and assists a non-technical audience in making informed decisions. These reports are similar to executive summaries in content (between 1-4 pages). Simple color graphics can be used to enhance the presentation of findings. The purpose of the study should be clear, as well as the results and take home message. A summary of data, as well as technical information, can be attached as subsequent pages or an easy link to the information provided.

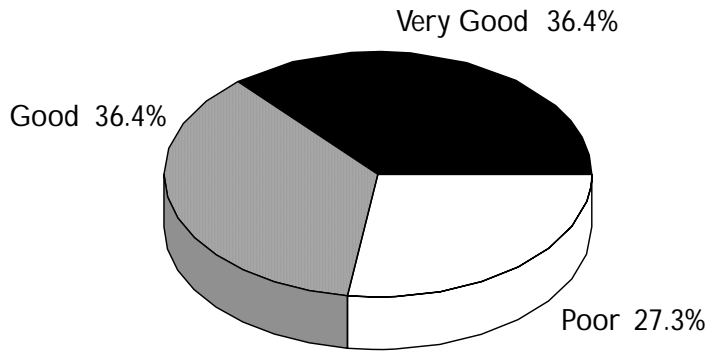
The second format for reporting is the scientific report, which is generally structured following peer-reviewed journal formats. These generally are reviewed by colleagues or non-agency peers to validate the technical quality of the work. Standard formats include an abstract or summary, followed by an introduction highlighting the technical foundation and outlining the study objectives, a methods section, a results section, and a discussion and conclusion section. Citation of relevant technical literature is necessary to support the validity of both the design and interpretation of the work. Preparation of these reports likely requires more effort than the summary report. However, this report includes all the supporting information, and is a more substantial defense of the work. Research to be published in journals will have to adhere to individual journal requirements.



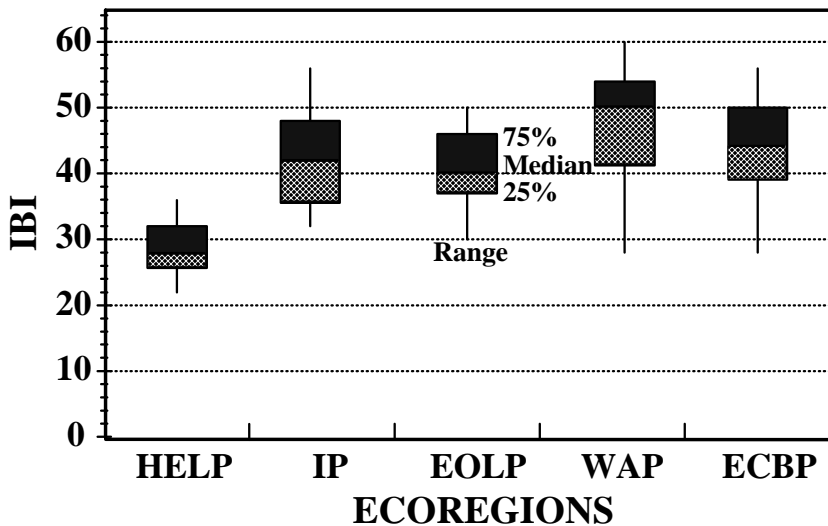
**FIGURE 8-8.** A bivariate scatter plot of an ordination used to support site classification. This figure for [Location] shows that grouping of sites based on taxonomic composition in ordination space reflects natural regional classes.



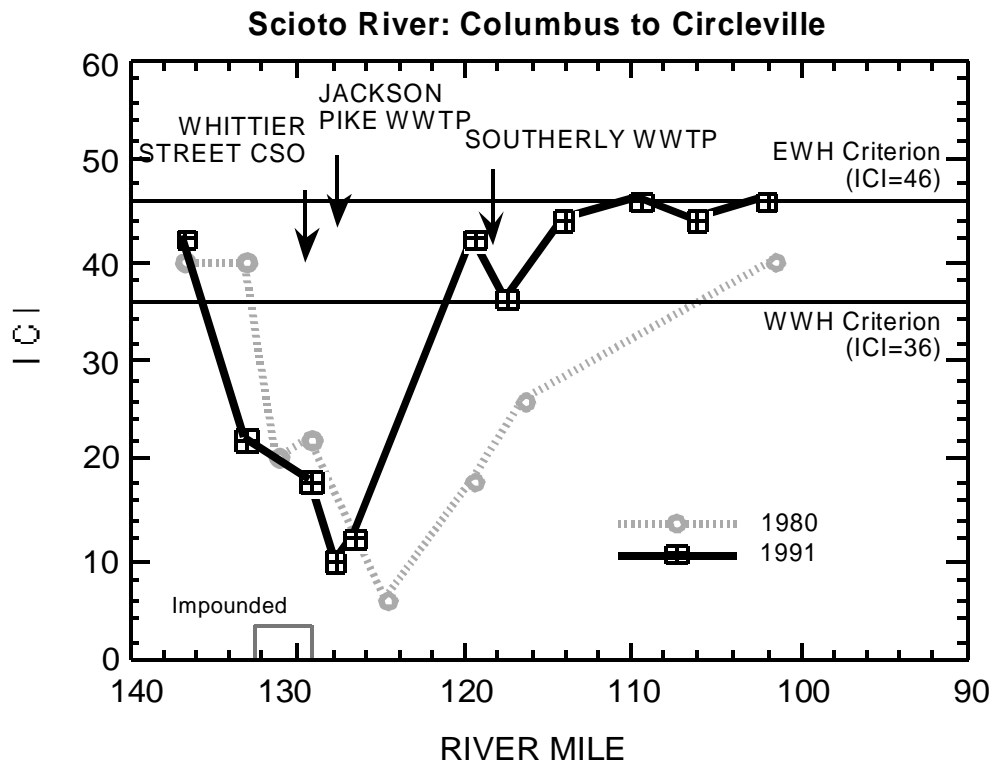
**FIGURE 8-9.** An example dendrogram, illustrating reference site clusters based on taxonomic composition. These figures are also used to support site classification.



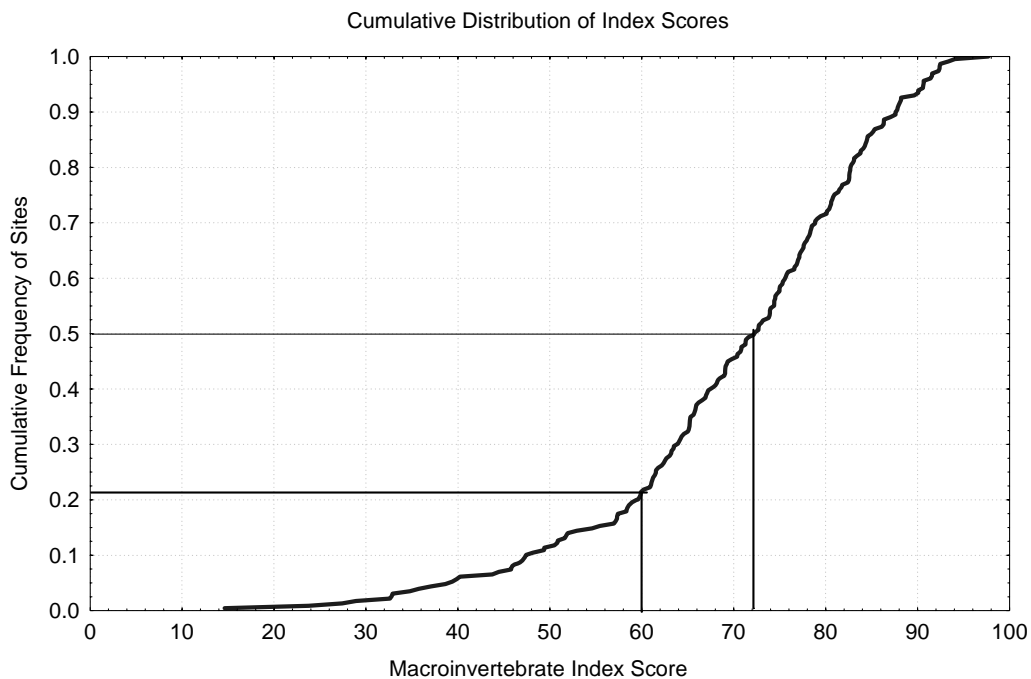
**FIGURE 8-10.** A pie chart, used to efficiently illustrate proportional information. This example shows the percent of stream miles in certain ecological condition categories within a watershed.



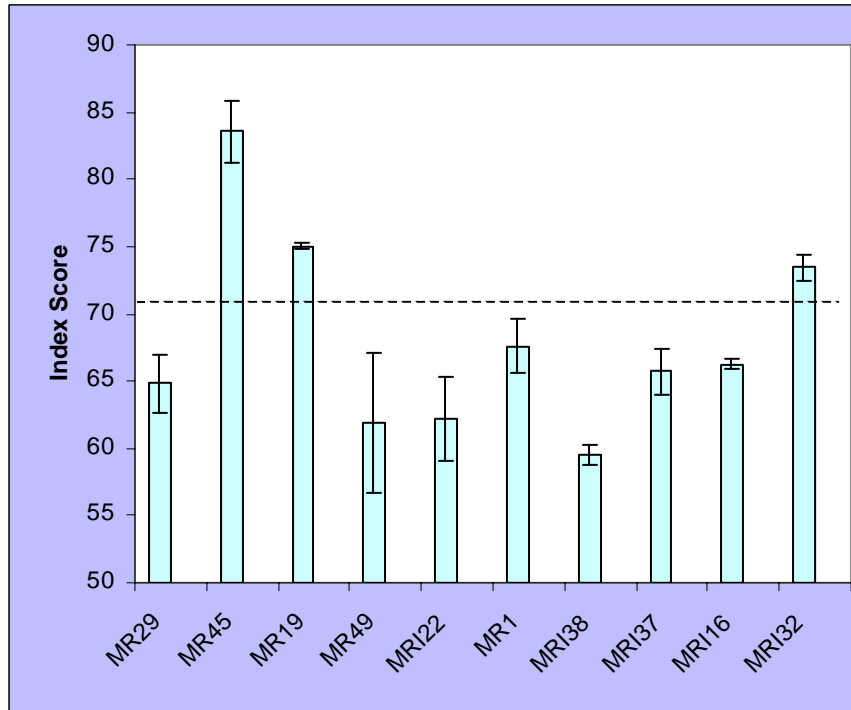
**FIGURE 8-11.** Box and whisker plots are used to illustrate differences in the distribution of values among different categories. Both central tendencies and a sense of variability can be conveyed. This particular figure illustrates differences in IBI scores among 5 ecoregions (contributed by Ohio EPA).



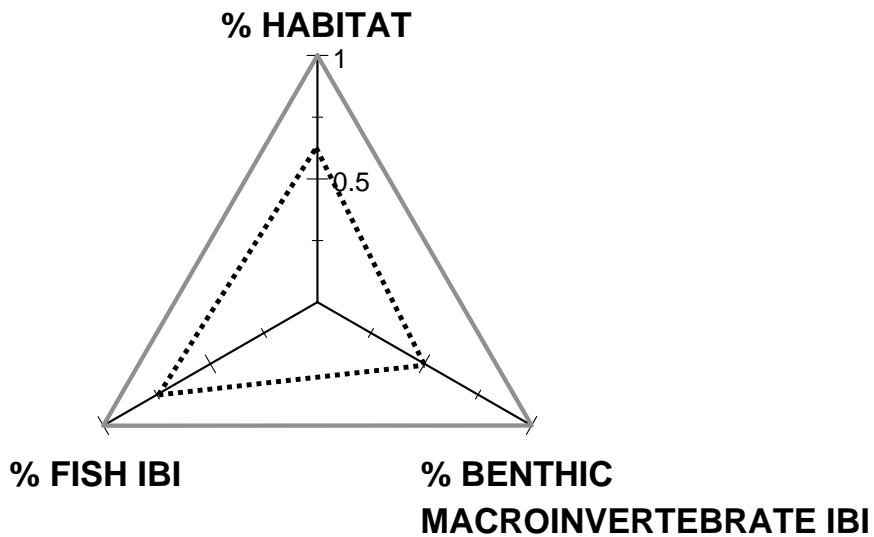
**FIGURE 8-12.** A line graph used to illustrate trends in the dependent variable relative to the independent variable. These are excellent tools for conveying temporal trends or trends along certain gradients. This example illustrates changes in a multimetric index along a river between two time periods (contributed by Ohio EPA).



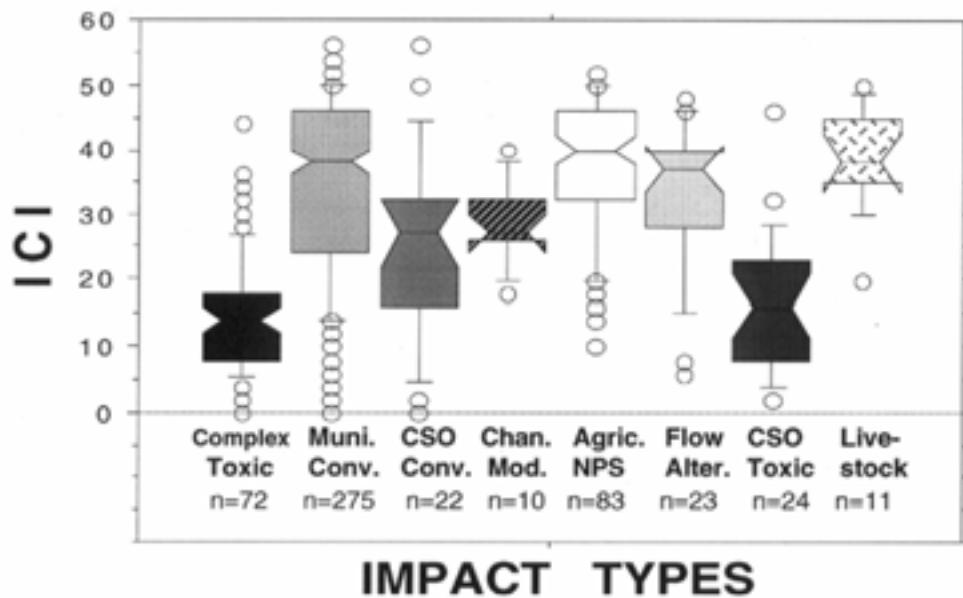
**FIGURE 8-13.** Cumulative frequency diagrams can be used to illustrate the ordered accumulation of observations from lowest to highest. These can be used to determine the percent of values exceeding any given value along the x-axis. This figure illustrates the median (50%) multimetric score and also indicates what percent of sites scored at or below 60 (21%).



**FIGURE 8-14.** A bar chart used to display the magnitude and variance of values for individual elements. This format can be used to rank relative scores. This example illustrates average multimetric scores and standard errors for several watersheds.



**FIGURE 8-15.** Sun ray plots are used to compare more than two endpoints simultaneously. Values can be scaled or compared to reference. This example simultaneously shows two multimetric indexes and a habitat score for a site relative to reference (1.0).



**FIGURE 8-16.** Box and whisker plots can also be used to illustrate the relative magnitude and variability associated with different variables on a common scale. This example illustrates multimetric values associated with different impacts (provided by Ohio EPA).

standard heading

assessment type

speedometer

standard heading

standard header text

locale map

site photo

variable text

standard footer

**BioReconnaissance Report (BioRecon):** A rapid, cost effective screening mechanism for identification of biological impairment.

**Introduction**  
 Spring Creek, located in Lee County, drains the area north of Bonita Springs into Estero Bay. The drainage basin consists of pine flatwoods with moderate residential development inland, with mangrove forest and denser residential development nearer the coast. The predominant land-use in the area is single family residential, encompassing approximately 50% of the drainage basin. Pine flatwoods, improved pasture, golf course, and a few commercial sites make up the other 50%. Spring Creek has been placed on the 303(d) list due to dissolved oxygen violations, and for excessive nutrient levels. Waterbodies on the 303(d)

list are required by EPA to have a Total Maximum Daily Load (TMDL) study performed on them. The purpose of the TMDL is to determine the amount of pollution reduction needed to restore the system to a condition suitable for its designated use. In this case, the designated use is for recreation and maintenance of a healthy, well-balanced aquatic community. DEP's South District Biology Section was requested to assess the status of selected waterbodies on the TMDL list that were placed on the list with "limited data". "Limited data" waterbodies were those with less than 10 observations in the STORET database, with the most recent observations occurring prior to 1990, or those with qualitative, non-point source survey data only.

**Results and Discussion**  
 Benthic macroinvertebrates communities, physical/chemical parameters, and nutrients were sampled in August of 1998. Macroinvertebrate

communities were sampled from in-stream habitats (using 4 discrete dip-net sweeps), field picked, and lab identified (the BioRecon procedure). Three metrics, consisting of total taxa richness, the Florida Index and total EPT taxa (Ephemeroptera, Plecoptera and Trichoptera), were calculated

and compared to existing thresholds to determine the community's health. The sample site was just upstream of the obviously estuarine portion of the creek. Spring Creek (with 28 taxa, 4 Florida Index points, and 4 EPTs) met two of the thresholds, but did not meet the Florida Index threshold (10). This indicates that the site may be impaired. Factors contributing to the marginal BioRecon scores included low water velocity (less than 0.1 m/sec), low dissolved oxygen (2.7 mg/L), suboptimal habitat, and possibly salt water influence.

One measured physical/chemical parameter or water quality variable did not meet the acceptable criteria for Class III waterbodies. Dissolved oxygen was only 2.7 mg/L, below the Class III standard of 5.0 mg/L, but only slightly lower than typical for streams in the region during the summer. Nutrient concentrations (nitrogen and phosphorus) were all below the median values for all Florida Streams.

**Conclusions**  
 Spring Creek failed one of three of the BioRecon metrics mainly due to low water velocity, low dissolved oxygen, suboptimal habitat, and possibly salt water influence. This is not a definite indicator of impairment. In light of this, and the reasonably good water chemistry values, it is recommended that Spring Creek be removed from the 303(d) list.

**FOR MORE INFORMATION, CONTACT:**  
 Albert S. Walton, Florida DEP South District  
 7481 Golf Course Blvd. Punta Gorda, FL 33982  
 (941) 575-5810 Walton\_A@fm1.dep.state.fl.us

FIGURE 8-17. Florida Department of Environmental Protection EcoSummary – an example summary report.