Comparing methods for inferring site biological condition from a sample of site biota

Kevin W. Brinck

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

University of Washington

2002

Program Authorized to Offer Degree: Quantitative Ecology and Resource Management

University of Washington

Graduate School

This is to certify that I have examined this copy of a master's thesis by

Kevin W. Brinck

and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the final examining committee have been made.

Committee Members:

_____

James J. Anderson

_____

James R. Karr

Date: _____

In presenting this thesis in partial fulfillment of the requirements for a Master's degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Any other reproduction for any purposes or by any means shall not be allowed without my written permission.

Signature_____

Date_____

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Dedication

To Iwelly ould Aly ould Birama, who I am sure thinks this is all quite silly.

1 — Introduction

1.1 – Background

Natural systems and urbanization

Human beings and our societies are embedded within and depend upon the natural world in which we live. The value of services provided by natural ecosystems has been estimated at 33 trillion US dollars (Costanza *et al.* 1997) – twice all the world's gross national products combined. Beyond their economic worth, people value natural systems for aesthetic and historical reasons.

And yet urban areas continue to expand. In the United States 19% of the land area is urbanized, compared to 9% in 1960 (Stoel 1999). The effect of urbanization is to disturb and degrade the environment, often eliminating the original natural system. However, there is a full range of impact between an undegraded ecosystem and elimination.

Measuring the quality of ecological systems

Given the value people place upon the natural world, and the effects of human activities upon it, there is a strong desire to reduce or avoid the negative consequences of urbanization and other forms of human impact. Resources are finite, however. There are several reasons that it would be useful to be able to measure the degree to which human activity has affected an ecosystem.

Conservation projects could allocate finite resources to select relatively unscathed systems. Programs attempting to improve or "restore" impacted systems could organize their efforts to target the most damaged. Scientific investigation, in either the context of "restoration" projects or a wider perspective could also benefit from a measure of the health of an ecological system.

Biological diversity on taxonomic and genetic scales has been used as a measure of ecosystem health, as a consistent, major effect of human activity is to reduce such diversity in natural systems. Folke *et al.* (1996) argue that including functional processes and system resilience make a more accurate measurement of ecosystem health, especially for activities on

the human time scale. Angermeier and Karr (1994) suggest that *biological integrity* is a better gauge of human impact on natural systems.

Following sections

The first chapter of this work continues with a discussion of biological integrity, followed by a review of methods used to measure stream quality with counts of benthic macroinvertebrates. Next are sections of definitions and mathematical techniques including brief description of some multivariate statistical methods. Finally comes a description of the B-IBI and the dataset used to compare multivariate techniques in this study.

Subsequent chapters will compare the results of mathematics-based multivariate techniques with the B-IBI. Chapter 2 considers patterns within the collection of taxon counts by site, and Chapter 3 uses additional measurements at sites to find pattern in the taxon counts. Finally, Chapter 4 compares the component metrics of three multimetric indexes.

1.2 – Biological integrity

Biological integrity

Biological integrity is a holistic property of ecological systems. Frey (1977) defined it as "the ability to support and maintain a balanced, integrated, adaptive community of organisms having a composition, diversity, and functional organization comparable with that of the natural habitats of the region." Biological integrity encompasses many processes and properties, from trophic interactions and evolution to diversity and abundance. These biological properties are a result of the physical characteristics of the habitat, from chemical conditions to energy inputs to habitat structure.

Biological condition

The biological condition of a system is the divergence from a state of biological integrity, usually expressed by comparing the system to a similar site that has been minimally influenced by human activity. As an integrative concept, a system's biological condition cannot be measured simply but instead must be inferred through multiple measurements of the presence and activity of its constituents.

Identifiable natural concepts such as organisms, species, and specific habitat types all interact through processes such as production, competition, and evolution to result in holistic properties. These holistic properties include long-term system stability – maintaining a state of biological integrity for long periods of time – and resilience – maintaining integrity when subject to external disturbance (Holling 1973). All organisms and properties are present in appropriate amounts in a site with a high biological condition, but the biological condition of a site exists as a function of all these things. The biological condition is not a measurable, physical property; instead it must be inferred from measurements of the biota that compose it.

How, then, does one measure the biological condition of a site? Theoretically, it would be possible to define a long list of species, physical states, and allowed changes over the course of natural fluctuations. Such a list would need to account for dozens or hundreds of species and observation over the range of time scales relevant to human activities – including decadal oscillations at the least.

Fundamentally, the decision of which places have a high biological condition and which places have a low biological condition comes down to an integrative evaluation by individual human beings. The necessary assumption is that all people have an internal standard of quality (Pirsig 1979). Given the above definition of biological integrity and a visit to the two sites, any reasonable person would agree that an undisturbed stream in Olympic National Park is very close to a state of biological integrity, and Thornton Creek in urban Seattle is very far from it.

A measurement of the divergence of a site's biological condition away from the state of biological integrity is often referred to as *disturbance*. The most common sort of disturbance in discussions of multimetric indexes is disturbance due to *human influence*.

Human influence

Human influence is another integrative concept that complements biological condition. Human beings going about their everyday lives affect the environment in many ways and on many different levels, almost always in a way that decreases the biological condition of

natural systems. Like biological condition, human influence is a holistic property that exists as a function of the physical effects of human activity on the environment. Similar to biological condition, human influence is not a physical characteristic to be measured directly, but instead it must be inferred through measurements of specific human activities.

While we have a better understanding of the factors and processes involved, the amount of human influence at a site is also subject to an individual's evaluation. The reasonable person in the preceding paragraph would also agree that the Thornton Creek watershed had been exposed to more human influence than the site in Olympic National Park.

Freshwater streams and benthic macroinvertebrates

Streams and larger watercourses are natural integrators of the landscape. Water runs across and through the ground connecting streams to their terrestrial surroundings (Minshall *et al.* 1985). A sample from a stream or river, then, contains a signal of events throughout the watershed (Karr and Chu 2000).

Using biological condition as the property to measure recommends direct examination of a site's biota. The biological condition of a site is a function of the organisms present at a site and their interactions with each other and their physical surroundings. Most forms of human influence affect the organisms indirectly, by modifying the physical properties of their environment to make them less habitable. The exact mechanisms linking a change in physical parameters (water temperature, for example) and the subsequent alteration in the biological community are often incompletely understood. This additional uncertainty suggests measuring the organisms themselves, rather than physical parameters, as the most direct reflection of a site's biological condition.

Measuring a site's biology integrates over time. Just as samples from a stream include a signal from the entire watershed, the population of an organism integrates over the life span of that organism (Karr 1991, Kerans and Karr 1994), as long as 3-4 years for some long-lived aquatic macroinvertebrates (Merritt and Cummins 1996). Fish are often longer-lived, but may not be as diverse. Streams in the central United States may support a community of 20-30 taxa, while Pacific Northwest streams seldom have more than 5-6 taxa (Moyle 1993).

Sampling benthic macroinvertebrates, mostly the aquatic, larval stages of insects, provides a more diverse community from which to sample.

1.3 – Natural-history vs. mathematics-based methods

Natural history based methods

In 1909, Kolkowitz and Marsson (1909) devised the saprobial index, which used aquatic invertebrates (mostly worms and wormlike larvae), chemical, and bacteriological measurements to produce a score gauging the amount of organic matter decaying in the environment. This index was the first use of invertebrates as a measure of environmental quality – in this case bio-chemical oxygen demand.

Chutter (1972) devised a metric of quality for streams in South Africa. Chutter assigned a weight or quality score for each taxon, and the average score for insects in a sample was used as a measure of stream quality. Hilsenhoff (1977, and 1982) developed a similar metric for Wisconsin streams, weighting taxa on a scale of 1-10. Taxa prevalent in degraded streams were assigned high weights and taxa found only in unimpaired streams received low weights. The weighted average of the sample was the stream score. The higher a stream's score, the more degraded it is.

The first multimetric index to measure biological integrity was based on fishes in the American Midwest (Karr *et al.* 1986, Ohio EPA 1987, Karr 1991)). Plafkin *et al.* (1989) included a multimetric index based on aquatic macroinvertebrates as part of the US-EPA's Rapid Bioassessment Protocols (RBP). Kerans and Karr (1994) developed a multimetric benthic macroinvertebrate index for Tennessee Valley rivers (B-IBI). Both Plafkin *et al.* and Kerans and Karr recommended that their proposed indexes be subjected to further testing and refinement.

The RBP index has been modified and adapted for use in Oregon streams (Mulvey *et al.* 1992). The B-IBI has been adapted for streams in the American Northwest and Japan (Kleindl 1995, Rossano 1995, Patterson 1996, Morley and Karr 2002). Doberstein *et al.* (2000) and Sovell (1999) have examined the effects of fixed-count sub-sampling, used in the RBP but not the B-IBI, and have found that the practice introduces a bias and reduces

precision in metric scores. Fore *et al.* (1994) examined the statistical properties of the B-IBI, and Fore *et al.* (1996) compared the discriminatory power of the RBP and B-IBI.

Multimetric indexes are not without critics. The metrics used in multimetric indexes are *top-down*, meaning the link between metric score and site quality is inferred without understanding the exact mechanism. Scrimgeour and Wicklum (1996) point out that top-down metrics cannot be tested scientifically. They and Suter (1993) both question the utility of biological integrity as a concept, and call for careful definitions. Others have made those definitions, distinguishing between biological integrity and the related concept of ecosystem health, and assert the concepts are needed to engage the public and provide a framework for policy and management (Meyer 1997, Fairweather 1999, Karr 1999).

Mathematics based methods

Mathematics based multivariate statistics are used for a number of purposes in ecology, including the multimetric goals of ordination and classification (Gittins 1985, Digby and Kempton 1987, Gower and Hand 1996). Specific methods applied to benthic macroinvertebrate counts include multiple regression (Nelson 1999), multivariate ANOVA (Faith *et al.* 1995), canonical correspondence analysis (Franquet *et al.* 1995), or combinations of several techniques (Boulton and Lake 1992).

Reynoldson *et al.* (1997) describe a pair of mathematics-based multivariate techniques designed to measure how closely a site approaches the reference condition. Two levels of multivariate analyses are used; the first to determine an appropriate reference site for a candidate site, the second to estimate the probability that the candidate site close enough to the reference to be considered the same. If the probability is high, the candidate site is judged as having a high biological condition, if the probability is low the site is considered impaired.

Field *et al.* (1982) describe a method of ordination that combines features of a multimetric index and a mathematics-based multivariate analysis. Mathematical multivariate techniques are used to identify groups of indicator taxa for distinguishing sites, which are then used to construct metrics.

Care must be taken in the use and interpretation of mathematics based multivariate statistics. In their review of multivariate analysis in ecology, James and McCulloch (1990) warn that the judging of the metrics produced is often based on their interpretability, an error they described as "dangerously close to circular reasoning". They also caution against confusion in statistical and biological language. Correlation is not causation, and producing a metric that "explains" 75% of variation in a system does not imply a cause-effect relationship. Both Karr and Martin (1981) and Stauffer *et al.* (1985) point out that for smaller datasets, multivariate analysis of completely random numbers may produce metrics that account for as much variation as significant metrics derived from analysis of real data, and therefore significant metrics may have no valid biological interpretation.

Similarities and differences

Both natural-history-based and mathematics-based multivariate methods attempt to accomplish the same task: promote understanding of a complex system by reducing the large number of variables (taxon counts) collected at a stream site to a smaller number of variables. Ideally, the smaller number of variables extracts the information needed for a specific goal. In the case of both the B-IBI and mathematics-based multivariate techniques that goal is to use the set of taxon counts from a site to infer the biological condition of the site. The differences arise in the criteria for combining variables in the reduction process.

Natural history based methods combine taxa based on the observed biology of the organism. These combinations may be based on taxonomy, niche, behavior, or life history of the taxa involved. Mathematics based methods define some function of the taxa counts – usually, but not always – related to squared differences or variance, and combine variables to minimize the value of this function.

1.4 – Definitions

Site

A site is a physical location of interest. In the data used for this study, sites were particular riffles in Puget Sound lowland streams.

Variable

A variable is a measure, or some transformation of a measure, of a property of a site. The value of a variable at a site is determined by taking a sample at that site. For the datasets discussed in this thesis the variables are the invertebrate taxa, and their values either are the number of individuals – raw counts or somehow transformed – of each taxon found at the site. In a multivariate context – whether natural-history-based like the B-IBI or mathematics-based like principal components analysis – variables are the original axes of the multi-dimensional measurement space.

Sample

Samples are measurements of one or more variables at a site. For the datasets discussed here samples are standardized collections of invertebrates taken from stream riffles. It is important to distinguish between statistical and biological samples. A statistical sample for designing a system to measure biological condition (the usage in this document) is the entire collection of invertebrates pertaining to a single site and time. These samples are usually composed of three replicates taken from the same stream riffle. For purposes of making an inference on a single stream (not the usage in this document), multiple collections from a riffle can each be considered a sample.

Data matrix

With samples and variables, one can define a matrix representation of the data matrix. The raw data matrix, **D** is a matrix with the number of rows equal to the number of variables (taxa) represented in the dataset, and number of columns equal to the total number of replicates. **D** itself is not used in any calculations. Instead the replicates from a site are combined – in different ways – to produce different transformations. The process is data formatting rather than a linear algebraic operation; Table 1-1 provides an example of the transformations detailed in Equations 1.8-10. Section 1.5, later in this chapter, provides a more detailed discussion of the transformations used.

The resulting transformations, **W**, **Y**, and **Z** or generically **X** (when the specific transformation is not important), are r □ c matrices where r equals the number of variables

(taxa) and c is the number of sites. Geometrically, **X** represents c points in an r-dimensioned space. (If c is less than r, the dimension of the space is c - 1; regardless of how many variables are measured at each point, three points can only define a plane.) An individual element of **X** can be addressed as $x_{i,j}$, where i is the row, and j is the column. In this study the rows (i) represent taxa and the columns (j) represent sites.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,c} \\ \vdots & \ddots & \vdots \\ x_{r,1} & \cdots & x_{r,c} \end{bmatrix} \tag{1.1}$$

Metric

A metric is a vector or line, or direction in the space represented by **X**. It can be visualized as a ray pointing from the origin. A metric can be identified as **m**, a vector of coefficients for each of the rows (variables) of **X**.

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_r \end{bmatrix} \tag{1.2}$$

Each m_i component of **m** can be treated as a weight or loading for its corresponding variable.

Score

A score (s_j) is the value for a metric at site j. The score for an individual site can be calculated by multiplying the value of each variable the metric's coefficient for that variable and summing

$$s_j = \sum_{i=1}^{r} m_i \cdot x_{i,j} \tag{1.3}$$

or the vector of metric scores for all c sites, **s**, could be calculated, in matrix notation, as

$$\mathbf{s} = \mathbf{m}^{\mathrm{T}} \cdot \mathbf{X} \tag{1.4}$$

where the "T" superscript indicates taking the transpose of **m**. Geometrically a score is a distance from the origin along a metric, and can be visualized as the projection of a site's coordinates in r-dimensional space onto the single dimension of the metric (Figure 1-1).

Index

An index is a collection of metrics. Selecting just a few metrics (< r) to focus on reduces the number of variables to be considered and simplifies interpretation of a dataset. By choosing metrics that discriminate across a gradient or gradients of interest, use of an index simplifies the process of interpreting a large number of variables. Identifying the gradient of interest and establishing criteria for good metrics allows the information relevant to a specific question to be extracted by multivariate techniques and summarized in the metrics of an index. It is possible for the individual scores of an index's metrics to be combined, adding them all together is common, and doing so in effect turns an index into another metric.

Scores for multiple metrics can be calculated as a single matrix algebra operation. For n metrics, **M** is an r × n matrix whose n columns are the **m** vectors for the metrics.

$$\mathbf{M} = \begin{bmatrix} m_{1,1} & \cdots & m_{1,n} \\ \vdots & \ddots & \vdots \\ m_{r,1} & \cdots & m_{r,n} \end{bmatrix} \tag{1.5}$$

S is an n × c matrix of metric scores, calculated by multiplying **M** and **X**.

$$\mathbf{S} = \mathbf{M}^{T} \cdot \mathbf{X} \tag{1.6}$$

Geometrically, if n is less than r, then **S** represents a subspace of **X**, and the vector of scores for each site is the projection of the site into **S**.

1.5 – Transformations

The raw data matrix **D** is not used for any calculations; instead one of three transformations is used, corresponding to the three mathematical groupings of B-IBI metrics (Table 1-1).

The transformations integrate information across the replicates for a site into a single column. $\mathbf{D}_j$ is the matrix of observations from the single site j. The $\mathbf{D}_j$ matrix will have r rows and c_j columns, where c_j is the number of replicates at the site. The transformation reduces $\mathbf{D}_j$ into an r by 1 matrix, a vector, and these vectors make up the columns of the transformed matrix.

For example, the presence/absence transformation produces **Z**, a matrix with r rows and c (the number of sites) columns. Each element of **Z** takes a value of 1 if the taxon was observed at the site, 0 if it was not. Mathematically

$$\mathbf{D}_j = \begin{bmatrix} d_{1,1} & \cdots & d_{1,c_j} \\ \vdots & \ddots & \vdots \\ d_{r,1} & \cdots & d_{r,c_j} \end{bmatrix} \tag{1.7}$$

and

$$\mathbf{Z}_{r,j} = \begin{cases} 1 \text{ if any of } \mathbf{D}_j \text{ row r are } > 0 \\ 0 \text{ otherwise} \end{cases} \tag{1.8}$$

In the B-IBI, the **Z** transformation is used for the long-lived and intolerant taxa richness metrics, which count how many long-lived or intolerant taxa, are found at a site. The value of the metric at each of the c sites is just the sum of the row values for long-lived or intolerant taxa.

The mean presence transformation produces **Y**, also an r by c matrix. The elements of **Y** are the fraction of replicates where the taxon was found.

$$\mathbf{Y}_{r,j} = \frac{\text{number of elements of } \mathbf{D}_j \text{ row r } > 0}{c_j} \tag{1.9}$$

The Ephemeroptera, Plecoptera, and Trichoptera taxonomic groups are larger than the long-lived and intolerant groups, so their richness metrics use the **Y** transformation to reduce noise by averaging across the replicates. Each row of the matrix contains a taxon's "mean presence", the fraction of a site's replicates that contained the taxon. The sum of the mayfly rows of **Y** produces the B-IBI Ephemeroptera richness score. This procedure is equivalent to counting the number of mayfly taxa in each of a site's replicates and taking the mean number of taxa as a site's Ephemeroptera richness score.

The relative abundance transformation is slightly more complicated, but also produces an r by c matrix. The elements of **W** are the relative abundances for a taxon across replicates; in

each replicate, the fraction of the sample composed of the taxon is calculated, and the final relative abundance score is the mean fraction across the site's replicates.

$$\mathbf{W}_{r,j} = \frac{1}{c_j} \sum_{i=1}^{c_j} \frac{d_{r,i}}{\sum_{k=1}^{r} d_{k,i}} \tag{1.10}$$

Once the **W** matrix has been calculated, the percent tolerant and percent predator metrics are simply the sum of the predator or tolerant rows of the matrix.

1.6 – Multivariate techniques

Geometric interpretation

Multivariate statistical methods attempt to facilitate the understanding of complex systems by reducing the number of variables that need to be considered. More specifically, they allow the information relevant to a specific question to be isolated and extracted from a large set of variables into just a few combinations of those variables, or *metrics*.

A geometric interpretation of the process assigns a dimension to each variable in the system. A simple two-variable system can be represented as a plane, and each possible state of the system is a point on that plane. A three-variable system is analogous to points in space. At four or more variables there is no physical analogue, but it is possible to describe a multi-variable system as analogous to a multidimensional space.

Choosing a metric (a line in a space) breaks an n-dimensional space into a one-dimensional space (the line) and an (n-1)-dimensional space. For example, imagine a three-variable system where individual observations are represented as points in space. Choosing a line in that space also defines a plane through the origin at right angles to that line. The original information represented by the positions of the points in space is divided into two parts: the projection of those points onto the line, and their projection onto the plane (Figure 1-1).

Choosing a second metric (line) in the remaining plane also specifies a third line through the origin at right angles to it. Looking at it another way, the two chosen metrics (the first two lines) define a plane, with a single line left over.

The plane produced by choosing two orthogonal metrics from the original three-dimensional system is often referred to as a sub-space (a sub-plane in this example). If the interesting behavior of the original system can be described in the sub-space, then the remaining dimension can be ignored, and the problem of understanding a system with 3 variables has been reduced to understanding a system with 2 variables.

If there is no unique best metric to be chosen from the original n-dimensional space, it might not make sense to choose successive metrics from the remaining (n-1)-dimensional space, but instead continue to identify metrics in the full n dimensions. There is no guarantee that these metrics will be orthogonal. The metrics might even not be linearly independent, though when choosing a small number of metrics from a large n-dimensional space that is less likely.

The decision of how to choose metrics from the original, multi-dimensional space in order to isolate the interesting behavior of the system in a reduced number of variables is what distinguishes different multivariate techniques. The datasets used in this study were collected from streams in a similar geographic and geologic setting to reduce variation from those sources. Streams were selected along a gradient of human influence driven mainly by urbanization, with the % impervious area in the watershed as a quantitative measure of human influence. The nature of the datasets suggest that correspondence analysis, which isolates the major gradients in a dataset, and canonical correlation, which uses an outside measurement to help identify metrics, would be useful in finding metrics to provide a signal of biological condition.

Correspondence analysis

Correspondence analysis chooses metrics by first performing a Chi-squared transformation of the original observations. This transformation non-linearly scales the value of each variable by how much it differs from the average value across all observations, corrected for the overall total of the variables at that observation. For any given metric in a multidimensional space, each observation can be projected onto that metric, producing a score for the observation in that metric. The variance of those scores (or, more correctly, the

sum of squared deviations from the mean) is the variation accounted for or "extracted by" that metric. Correspondence analysis chooses the metric that accounts for the most variation, removes that line, and then repeats the process for the remaining dimensions.

If correspondence analysis is performed upon taxon counts at sites, the results are metrics that best discriminate among sites by relative composition. A site with counts of 5, 5, and 10 would be mapped onto the same point as one with counts of 10, 10, and 20, and site whose composition was greatly different from the overall norm would be mapped to a point far from the origin.

Canonical correlation

Canonical correlation is a multivariate technique used when there are two sets of variables (and therefore two multi-dimensional spaces) for each observation. The data are not transformed. Canonical correlation chooses metrics by identifying a pair of metrics, one in each space. It then projects the observations onto the metrics to produce scores for each metric, and then chooses the pair of metrics that maximizes the correlation of scores. It removes the lines, and then repeats the process for the remaining dimensions.

1.7 – B-IBI

The Benthic Index of Biological Integrity also seeks to facilitate the understanding of complex systems by reducing the number of variables, so by the above definition it is also a multivariate technique. Specifically it measures macroinvertebrate taxon abundances (variables) at a number of sites (observation) and reduces the variables to a suite of 8-12 metrics.

The technique can be described, in the geometric terminology above, as first transforming the count data as per section 1.5 and then identifying candidate metrics (called *attributes* in IBI literature) based on knowledge of the biology of the organisms involved. Sites are projected onto candidate metrics to produce scores, and candidate metrics are evaluated by their ability to demonstrate a dose-response relationship with human influence. Other criteria, such as rank correlation with human influence, are also often used.

When an IBI metric is chosen, the original, n-dimensional space is **not** reduced to a line and an (n-1)-dimensional space; all metrics are chosen from the original n-dimensional space.

A second criterion for metrics was providing a useful signal across a range of studies, in the case of the B-IBI studies in Washington, Oregon, Wyoming, and Japan. Finally, metrics were included in the final index because a contrast between two metrics might be useful in diagnosing a specific form or magnitude of human influence. Using the B-IBI multivariate technique a problem with a large number of variables – where large is however many taxa were found in the system studied – is reduced to a problem with only ten variables. The method of deriving B-IBI metrics makes sure those ten variables contain information relevant to the question at hand: providing a signal of a site's biological condition.

Candidate metrics / Invertebrate terminology

B-IBI candidate metrics are sub-spaces of all taxa defined by biological affinity. Biological criteria for grouping taxa include:

Taxonomic: candidate Orders of invertebrates include *Diptera*, *Ephemeroptera*, *Hemiptera*, *Odonata*, *Plecoptera*, and *Trichoptera*. Larger Families used as candidates include the *Tipulidae* (Diptera) and *Heptageneiidae* (Ephemeroptera).

Feeding guild: most benthic macroinvertebrates can be classified by their manner of feeding and what they feed on. These guilds include *Predators*, which eat other animals. *Shredders* tear apart larger pieces of plant debris. *Collectors* can either gather benthic debris or filter it from the water column, and *Scrapers* survive by scraping periphyton from the stream substrate.

Invertebrate taxa are also classified as being generally *tolerant*, neutral, or *intolerant* of degraded stream conditions, focusing mostly on degradation caused by organic enrichment. They are also classified as being tolerant, neutral, or intolerant of sediment.

Clinger taxa are identified as organisms that spend much of their time clinging to cobble on the stream bottom.

1.8 – Dataset

The data used in this analysis are counts of benthic macroinvertebrates collected from Puget Sound lowland streams and then identified to genus by Bill Kleindl in 1994, Jeannie Udd in 1995 (Kleindl 1995), and Sarah Morley in 1997 (Morley and Karr 2002). The datasets were collected for studies focusing on the effects of urbanization on the biological condition of streams, and so avoided sites where the primary source of human disturbance took other forms, such as agriculture or logging.

1.9 – Tables

Table 1-1. Example: Transforming data to B-IBI style matrices

To produce the raw data, three distinct collections of invertebrates (replicates) are taken from a riffle in each of two streams, and three taxa (A, B, and C) are counted for each replicate. The Presence/Absence transformation records a 1 if a taxon is found in any of the site's sample, a 0 otherwise. The Relative Abundance transformation is the mean relative abundance across a site's replicates. The Mean Presence transformation is the fraction of a site's samples containing a taxon.

If taxa A and B are long-lived taxa, then the sum of rows A and B of the presence/absence transformation is the long-lived taxa richness score for each site. If taxa B and C are Ephemeroptera taxa, then the sums of rows B and C of the mean presence transformation are the Ephemeroptera taxa richness scores. If taxa A and C are classified as tolerant taxa, then summing rows A and C produces the fraction of tolerant taxa found at the sites. This procedure is equivalent to first calculating the relative fraction of taxa A and C in each replicate, then finding the mean across replicates.

Raw data

	Site 1			Site 2		
	1	2	3	1	2	3
A	0	1	0	12	10	7
B	0	0	0	4	6	8
C	1	0	2	2	1	1

Transformed data

	Presence/Absence		Relative Abundance		Mean Presence	
	Site 1	Site 2	Site 1	Site 2	Site 1	Site 2
A	1	1	0.33	0.56	0.33	1
B	0	1	0	0.36	0	1
C	1	1	0.67	0.08	0.67	1

	Site 2				Site 2			
	1	2	3		1	2	3	mean
A	**12**	**10**	**7**		**0.667**	**0.588**	**0.438**	**0.564**
B	4	6	8	□	0.222	0.353	0.500	0.358
C	**2**	**1**	**1**		**0.111**	**0.059**	**0.063**	**0.077**
	14/18	11/17	8/16		mean			
	0.778	0.647	0.500	□	0.642			0.642

1.10 – Figures

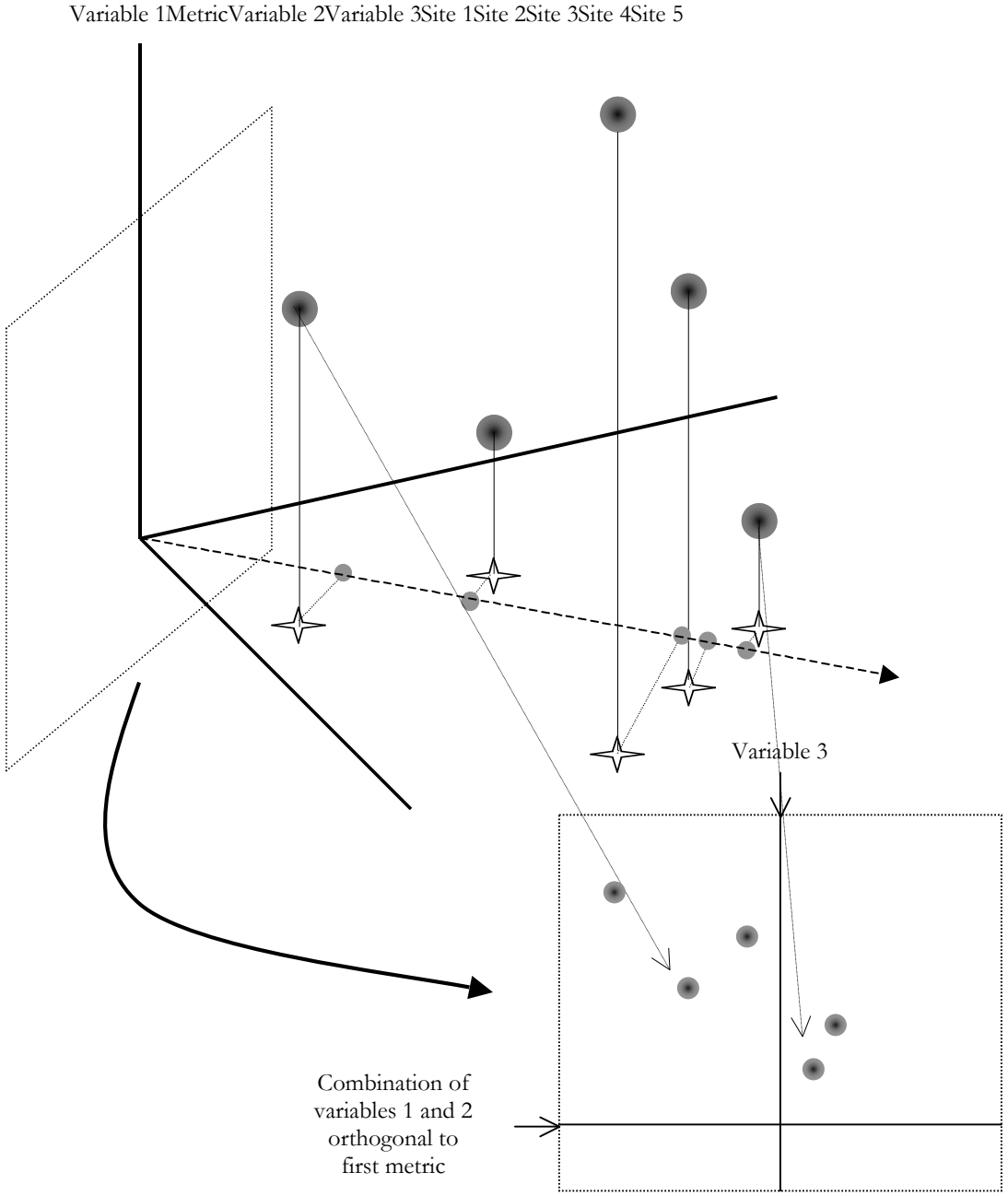

Figure 1-1. Projection of sites onto metrics to produce scores

Three variables are measured at five sites. The information in variables 1 and 2 is condensed into a single value by projecting each site onto a new dimension or metric that is a linear combination of variables 1 and 2. The score of a site in the metric is the distance of the projection from the origin. The distance between the true site coordinates and the projection onto the metric represents the loss of information in summarizing two variables in a single number. The site can then be projected onto a plane through the origin at 90° to the first metric.

2 — Gradients within the data

2.1 – Introduction

The B-IBI selects metrics that respond to human influence and so provide a signal of site biological condition. To identify these metrics, data are collected at sites along a gradient of human influence, and metric scores that follow a dose-response relationship to human influence (among other criteria) are kept as useful.

Correspondence analysis also identifies metrics that correlate with gradients in the dataset. The question investigated in this chapter is: Do the metrics identified by correspondence analysis match up with the metrics used by the B-IBI?

The metrics of the B-IBI can be represented in matrix algebra notation as vectors of taxa weights to be multiplied by a vector of taxon abundances or some transformation of taxon abundance (see section 1.5 – Transformations). Correspondence analysis and most other mathematics-based multivariate techniques produce metrics in the same form: a vector of weights for each taxon.

In this chapter I compare the vector representation of the B-IBI metrics with the metrics produced by correspondence analysis. I determine whether the metrics produced by the two different techniques are alike or not alike.

2.2 – Correspondence analysis

Description and history

Correspondence analysis (CA) is a multivariate technique designed to distinguish among sites based upon the relative values of a set of variables measured at those sites. Each variable should be a count of individuals by taxon to accord with the original derivation of the technique (Benzecri 1992, Chapter 1), but CA is sometimes used as an ordination technique on continuous variables, especially stabilizing transformations of the count data (Greenacre 1984).

Correspondence analysis is sometimes called "reciprocal averaging" in reference to the computational technique first used to perform it. Arbitrary coefficients are chosen for each

of the variables. Scores are computed for each site by finding the sum of the weighted variables. These site scores are used as site coefficients to calculate new scores for each variable, and the process is repeated until the relative proportions of the scores converge.

Mathematical formulation

Mathematically, reciprocal averaging has the effect of applying a chi-squared transformation to the original data matrix followed by singular value decomposition (Pielou 1984, Chapter 4.5). If **Y** is an intermediate placeholder, and **X** is a matrix of data (possibly the generic **X**, for one of the transformations defined in section 1.5), and defining **R** as a diagonal matrix of site totals and **C** as a matrix of variable totals, and $\mathbf{R}^{-\frac{1}{2}}$ as a matrix whose elements are the square root of the corresponding element of **R** then

$$\mathbf{Y} = \mathbf{R}^{-\frac{1}{2}} \cdot \mathbf{X} \cdot \mathbf{C} \tag{2.1}$$

is the matrix of χ^2transformed count data. The matrix $\mathbf{Y}^{\mathrm{T}} \cdot \mathbf{Y}$ can then be decomposed into the product of three matrices: $\mathbf{ULV}^{\mathrm{T}}$ (singular value decomposition Leon 1986, Chapter 7). **U** is a matrix of site scores, **V** is a matrix of variable scores, and **L** is a diagonal matrix of singular values. A vector of site scores can be obtained from $\mathbf{C}^{-\frac{1}{2}}\mathbf{VL}$ and variable scores from $\mathbf{R}^{-\frac{1}{2}}\mathbf{UL}$. Which scores are used depends on whether one is interested in interpreting site similarity based on the variable measurements, or on variable similarities based on their values at the sites. These equations are for balanced scores, in which sites and variables are given equal footing; scores can also be computed with an arbitrary relative weighting of sites vs. variables.

Uses

Correspondence analysis, and extensions of it, are often used in botany to identify patterns in plant species composition along a natural gradient (Whittaker 1967), using plant species abundance for variables at sites. Identifying trends in the biological condition of watersheds based on collections of benthic macroinvertebrates along an artificial gradient of human influence is a conceptually similar problem.

The B-IBI is designed with the opposite approach. Sites are selected to represent a gradient of human influence, and combinations of variables that produce a dose-response relationship to human influence are selected as metrics. Metrics derived from the two approaches can be compared. The comparison is complicated because CA identifies multiple orthogonal gradients within the data. In a dataset of sites selected to represent a gradient of human influence, at least some of the gradients identified by CA should be related to human influence. Correlation between B-IBI metrics and correspondence analysis derived metrics, especially if that correlation is consistent across time, would indicate that the CA-identified gradients are related to the human influence gradient targeted by the B-IBI.

2.3 – Questions

Do correspondence analysis and the Index of Biological Integrity produce metrics, which are alike?

The Benthic Index of Biological Integrity (B-IBI) uses knowledge of an ecosystem and the life history of specific taxa to define metrics. In the B-IBI based on studies in the northwestern United States and Japan, caddisflies, stoneflies, and mayflies are included separately because those three taxa respond differently to different types or levels of human disturbance. Long-lived taxa are grouped together and chosen as a metric because those organisms will respond to disturbance that occurs only every two or three years.

The exact rules for calculating B-IBI metric values from the raw taxa counts from a site can be expressed in the same matrix notation used in discussing correspondence analysis. For example: the Ephemeroptera taxa richness metric can be represented as vector of weights for each taxon in a sample. Each Ephemeroptera taxon would receive a weight of one, and all non-mayfly taxa would be weighted as zero. Multiplying this vector by a presence/absence (1/0) matrix of taxa by sites will produce the Ephemeroptera taxa richness scores, the number of mayfly taxa found at each site.

If a CA-identified gradient corresponds to a human influence gradient targeted by a B-IBI metric, then the metrics should be morphologically alike and the weights assigned to any particular taxon by the two methods would be correlated. This correlation would imply that the two methods have identified the same link between biota and biological condition. In the

absence of correlation I would conclude that the methods have not identified the same connection between biota and biological condition, or at least that the connection is not readily understood as a linear function of taxon abundances.

Does adding biological information to correspondence analysis produce metrics that are even more like those of the B-IBI?

Biological information

The B-IBI incorporates additional information that is not included in correspondence analysis of taxon counts. Life history attributes of the particular taxa and biological knowledge of the ecological processes occurring in the system are considered when candidate metrics are chosen for evaluation. By selectively combining taxa based on their phylogenetic proximity or similar life histories, the B-IBI incorporates a filter to reduce noise in and increase the signal of a site's biological condition.

Progressive addition of biological information

Some of this biological information can be progressively added to a correspondence analysis. Phylogenetic information can be added by combining taxa at the family or order level to create a new matrix for analysis. Feeding guild information can be added to the analysis by aggregating all the taxa belonging to the same guild before applying correspondence analysis. Combining taxa based upon taxonomic affiliation or feeding guild applies a filter to the original matrix of taxon counts, ideally reducing noise similar to the B-IBI.

Filtering the data in this fashion should make it more likely that the taxon weights produced by CA will be correlated with those representing B-IBI metrics. If there is an increase in correlation when biological information is added in this way to the correspondence analysis, then that increase will be a measure of the information added by considering the biology of the organisms. If there were no effect of biological information on the correlation between CA and B-IBI metrics, then, as above, I would conclude that the methods have not identified the same connection between biota and biological condition.

2.4 – Methods

The data sets (matrices of invertebrate counts by taxon and site) were transformed into the three mathematical groupings of the B-IBI metrics. In Table 2-1 two sites with three taxa are

transformed as an example. A Presence/Absence transformation (1 if a particular taxon was present at the site, 0 if it was not) is appropriate for Intolerant taxa, and Long-lived taxa. The sum of presence/absence values for all Long-lived taxa, therefore, produces the total number of long-lived taxa present at a site.

The Ephemeroptera, Plecoptera, and Trichoptera richness metrics go a step past simple presence/absence and look at the fraction of site's replicates containing the taxon. A Mean Presence transformation records that fraction for each site and taxon. The sum of those values for all Mayfly taxa therefore produces the B-IBI Ephemeroptera richness score.

A Relative abundance transformation (the across-replicate average the fraction of individuals belonging to a taxon) is appropriate for the percent tolerant and percent predator metrics, which measure the fraction of sample individuals belonging to their category. The B-IBI dominance metric is non-linear. It can be represented in matrix form but there is no gain in generality, as it requires a distinct ordering of taxa for each site; individual columns would have their rows in different orders.

Application of correspondence analysis to raw taxon counts

Correspondence analysis was performed on these three transformations to produce metrics (in the form of vectors of taxa loadings) which best discriminated among the sites. These metrics were compared with a matrix representation of the B-IBI metrics through inspection for morphological similarity and by calculating the column-by-column rank correlation.

Correlations were not calculated for the dominance metric or the total taxa richness metric. The Dominance metric cannot be represented in the same matrix form as the other metrics. The total taxa richness metric includes all taxa, so their coefficients are all the same (1) and correlation with a vector of identical values is undefined.

2.4.1 Comparison of ungrouped metrics to B-IBI metrics

Column by column correlation

The best possible matchings, defined as a one-to-one correspondence between each B-IBI metric and one of the CA metrics with the highest probabilities of being significant, were

identified for each set of metrics, and Bonferroni corrected post hoc p-values were calculated.

Hypotheses of no rank correlation

These p-values are for the null hypothesis of no rank correlation between the coefficients of two or four (depending upon the transformation) B-IBI metrics and the coefficients of the most significant two or four CA metrics from correspondence analysis of the appropriate transformation of the same dataset. These are not true hypothesis tests because the decision of which CA metric to correlate with which B-IBI metric was made after computing the rank correlation for each pair, and choosing the combination which produced the smallest p-value. The p-values are for the best possible match between CA and B-IBI metrics, according to the data.

Significant correlations between the B-IBI metrics and the correspondence analysis metrics would imply that the two techniques had identified the same metric, corresponding to a single link between the biota at the site and the site's biological condition. In the absence of correlation I would conclude the techniques identify different metrics and different connections between the taxon counts and site biological condition.

2.4.2 Comparison of grouped metrics to B-IBI metrics

Aggregation by biology

To investigate the effect of including biological information in a correspondence analysis, taxon counts were aggregated based on biological similarity. Counts were grouped taxonomically by family and order, and by life history traits based on their functional feeding guild classification. Correspondence analysis was performed on presence/absence, mean presence, and mean individuals transformations of these simplified matrices, and the coefficients produced were compared to the B-IBI metric coefficients. The three correspondence analyses of each year were compared with the B-IBI metrics by inspection for morphological similarity, and column-by-column rank correlation of B-IBI and CA-derived taxon loadings.

2.5 – Results

2.5.1 Comparison of ungrouped metrics to B-IBI metrics

The metrics produced by correspondence analysis of the ungrouped taxon counts were not similar to the B-IBI metrics. Comparison of the coefficients of the ungrouped correspondence analysis and B-IBI metrics (Table 2-3, Table 2-4 and Figure 2-1) revealed no obvious similarities.

The B-IBI coefficients are all either 0 or 1. Within the data matrix the taxa were (approximately) grouped by order, family, and genus, so closely related taxa were placed next to each other. This ordering results in a series of runs of zeros and ones in the matrix representation of the B-IBI metrics. No such pattern is evident in the correspondence analysis metrics. The general pattern in CA metrics was for most taxa to receive coefficients between –1 and 1, with occasional (< 10%) coefficients of ±2-5.

One significant correlation across all three years

Only one correlation, between the first CA metric from the Presence-Absence data matrix and Intolerant taxa richness, was significant across all three years (post hoc Bonferroni corrected $p = 0.09$). In general the Spearman's rank correlations for 72 comparisons of CA metrics to B-IBI metrics ranged from –0.31 to 0.35. Of the 72 p-values, 28% were less than 0.10 (Table 2-3). Aside from intolerant taxa richness there were no significant correlations across years. Even the post hoc best possible matchings between CA and B-IBI metrics for each year and transformation were not statistically significant.

2.5.2 Comparison of grouped metrics to B-IBI metrics

The metrics produced by correspondence analysis of the grouped taxa were closer to the B-IBI metrics than the CA metrics from ungrouped taxa, but were still not similar enough to be described as "like" the B-IBI metrics. The correlations of metric coefficients were larger, and there was greater morphological similarity (Table 2-5).

Grouping taxa by order, family, and feeding guild did produce some morphological similarity with certain B-IBI metrics. When grouping by family, all Ephemeroptera taxa received the same correspondence analysis coefficient; the B-IBI metric for Ephemeroptera taxa richness

has all ones for Ephemeroptera and zeros for other taxa, so there was some obvious visual similarity. Rank correlation provides a more detailed examination by including the relative magnitudes of the group coefficients.

As an example, Table 2-5 compares some metrics produced by correspondence analysis of the mean presence transformation of the 1994 data after aggregating by order. All of the Ephemeroptera taxa have the same coefficient, as for the B-IBI Ephemeroptera richness metric. However, the coefficients CA assigned to non-Ephemeroptera taxa are not all zeros. They are not even the same, but a variety of values both larger and smaller than the weight assigned to the Ephemeroptera taxa. Family aggregated results were comparable for Plecoptera taxa richness and Trichoptera taxa richness in the mean presence transformation, and grouping by feeding guild produced a CA metric that looked like the % Predator B-IBI metric.

Results of grouping taxa

In general, grouping taxa based on biological information before performing correspondence analysis increased the fraction of significant individual correlations. Grouping by order increased the fraction from 28% to 31%, and by family up to 57%. Grouping by feeding guild produced 43% of individual correlations significant at 0.10 (Table 2-6).

Two correlations, between Plecoptera taxa richness and Ephemeroptera taxa richness and the third and fourth correspondence analysis metrics of taxa grouped by Family were significant (post hoc Bonferroni correct p-values of 0.02 and 0.03 respectively). Otherwise there were no correlations significant across years. Only two of the best possible post hoc matchings between CA and B-IBI metrics were significant. In the 1995 and 1997 data sets the best possible p-values for correlating family aggregated, mean presence transformed CA metrics with the mean presence B-IBI metrics were 0.02 and 0.00. The corresponding best value in 1994 was 0.11.

2.6 – Discussion

Visual inspection

Visual inspection of vectors of numbers, even in organized as in Table 2-4 and Table 2-5, is a chancy way to decide whether the vectors are alike. Patterns of 1's and 0's in the vector representation of an IBI metric are easy to identify. If the pattern of coefficients in a CA vector follow a similar pattern, or a pattern of 0 and non-zero values, it might be easy to detect. A pattern of values ranging from 0.1 to 0.4 versus values from 0.3 to 0.5 (for example) would be much likelier results of correspondence analysis, and much more difficult for a human observer to notice. Coefficients should be examined and compared, perhaps with the aid of a graph (Figure 2-1), but straightforward inspection will miss subtler relationships. Interpreting a list of coefficients as a vector is not easy, and consequent difficulty of interpretation is a weakness of correspondence analysis and similar mathematics-based multivariate techniques.

Morphological similarity

The morphological similarities evident when comparing CA metrics produced from biologically grouped data and B-IBI metrics is not surprising. By definition B-IBI metrics focus on a group of organisms that are alike, so like taxa receive the same weighting in the metric. In the Ephemeroptera taxa richness metric, for example, all the Ephemeroptera taxa are given a weighting of 1.

Aggregating all similar organisms replaces multiple rows of taxa with a single, summed row. Aggregating by order, for instance, replaces a matrix of taxon counts, with a row for each taxon, with a matrix of order counts, with a row for each order. Correspondence analysis on such an aggregated matrix assigns a coefficient to each group. Assigning the group's coefficient to each member of that group will result in biological group members having the same coefficient, just as in the B-IBI. For instance, in the order example above, all Ephemeroptera taxa would be given the coefficient the correspondence analysis assigned to the Ephemeroptera row of the aggregated matrix.

CA coefficients assigned for each group were different, so even if, within a metric, a single group's taxa shared a value – just like the B-IBI metric – the other groups had widely varying coefficients, not at all like the B-IBI metric.

If there were no connection at all between the CA derived metrics and the B-IBI metrics one would expect only 10% of the correlations to be statistically significant at the 0.10 level (Zar 1996 p. 81). 28% of correlations tested produced p-values less than 0.10. While not a formal statistical test, this overabundance of significant correlations implies that there is some likeness between the two sets of metrics. However, with the possible exception of the first CA presence-absence component and intolerant taxa richness, the calculated correlations were not useful in divining the nature of that likeness.

Once the potential link between of the first CA presence-absence component and Intolerant taxa richness was identified, a more detailed perusal of the CA coefficients did not reveal any biological similarity to the B-IBI. In the B-IBI intolerant taxa richness metric, intolerant taxa have a coefficient of 1, and all other taxa have a coefficient of 0. In the CA-derived metric intolerant taxa did not receive unusually high or low loadings. There were 11 taxa classified as intolerant and only one of the 11 most extreme correspondence analysis metric coefficients was for an intolerant taxon.

Post-hoc p-values

The best p-values reported in Table 2-3 and Table 2-6 are all *post hoc*, meaning the hypotheses they pertain to were formulated after the analysis had been run. They are not true p-values, as their hypotheses were selected as being most likely to be rejected. Even so, none of the best p-values were significant at the 0.10 level, implying that the data are not consistent with such a simple, straightforward relationship between the B-IBI metrics and correspondence analysis metrics.

Grouping by family

Grouping taxa based on biological information before applying correspondence analysis increases the number of significant correlations with B-IBI metrics, up to 57% when grouping by family. This increase should not be interpreted as necessarily meaningful. Three

of the eight B-IBI metrics that were compared with CA are based on a taxon's family, so performing a CA on just family membership should be expected to increase the correlation between the two metrics. Still, the fractions of p-values less than 0.10 were still greater than the expected 10%.

Grouping by order

Grouping by Order produced two correlations that were significant in all three years. The B-IBI metrics involved in those correlations, Ephemeroptera and Plecoptera taxa richness, are based on order. As described above, grouping by family would be expected to increase correlation with such B-IBI metrics, so the increased significance with taxonomic grouping is not necessarily important. A meaningful correlation between the B-IBI metrics and CA metrics would be consistent across the three years of data. A more careful inspection for visual similarity (Table 2-3) did not show an interpretable pattern to the correlation between CA and B-IBI metrics or even between CA metrics across years.

The best p-values reported in Table 2-6 are also *post hoc*, so are not true p-values, merely a measure of strength of the best possible matching between CA and B-IBI metrics. Grouping by family produced two such significant p-values for the mean presence transformation. That transformation applies to the three family based B-IBI metrics (Ephemeroptera, Plecoptera, and Trichoptera richness) so, as above, increased significance is not surprising.

2.7 – Conclusions

Correspondence analysis strives to find metrics that best distinguish among a set of sites. The B-IBI metrics are deliberately constructed to distinguish among sites along a specific gradient of human influence. These similar goals may have resulted in similarities between the metrics, which in turn may be responsible for the excess of p-values reported in the correlation analysis.

Even in the presence of consistent correlation between CA and B-IBI metrics (which was not found in this study) the underlying biological interpretation must be the same before they can be identified as alike. Correspondence analysis seeks to distinguish among sites across all sources of variability, and in this study those sources did not align with obvious

biological properties of the variables. There would only be an exact correlation with the B-IBI in cases where the biological signal dominated total variation or perhaps in an extremely large sample size where the noise of the non-biological signals would average out. In contrast, the B-IBI metrics are designed to focus on a specific source of variation, and are chosen to correspond to biological similarities of the variables.

2.8 – Tables

Table 2-1. Example: Transforming data to B-IBI style matrices

To produce the raw data, three distinct collections of invertebrates (replicates) are taken from a riffle in each of two streams, and three taxa (A, B, and C) are counted for each replicate. The Presence/Absence transformation records a 1 if a taxon is found in any of the site's sample, a 0 otherwise. The Relative abundance transformation is the mean relative abundance across a site's replicates. The Mean Presence transformation is the fraction of a site's samples containing a taxon.

If taxa A and B are long-lived taxa, then the sum of rows A and B of the presence/absence transformation is the long-lived taxa richness score for each site. If taxa B and C are Ephemeroptera taxa, then the sums of rows B and C of the mean presence transformation are the Ephemeroptera taxa richness scores. If taxa A and C are classified as tolerant taxa, then summing rows A and C produces the fraction of tolerant taxa found at the sites. This procedure is equivalent to first calculating the relative fraction of taxa A and C in each replicate, then finding the mean across replicates.

Raw data

	Site 1			Site 2		
	1	2	3	1	2	3
A	0	1	0	12	10	7
B	0	0	0	4	6	8
C	1	0	2	2	1	1

Transformed data

	Presence/Absence		Relative Abundance		Mean Presence	
	Site 1	Site 2	Site 1	Site 2	Site 1	Site 2
A	1	1	0.33	0.56	0.33	1
B	0	1	0	0.36	0	1
C	1	1	0.67	0.08	0.67	1

	Site 2				Site 2			
	1	2	3		1	2	3	mean
A	**12**	**10**	**7**		**0.667**	**0.588**	**0.438**	**0.564**
B	4	6	8	□	0.222	0.353	0.500	0.358
C	**2**	**1**	**1**		**0.111**	**0.059**	**0.063**	**0.077**
	14/18	11/17	8/16		mean			
	0.778	0.647	0.500	□	0.642			0.642

Table 2-2. Metrics used in the B-IBI for Puget Sound Lowland Streams

Some metrics are expected to increase in value as human influence rises and biological condition drops, other metrics follow the opposite pattern. A selection of metrics, taken together, constitutes an index.

Metric	Response to increasing human influence
Total taxa	decrease
Ephemeroptera taxa richness	decrease
Plecoptera taxa richness	decrease
Trichoptera taxa richness	decrease
Dominance (top 3 taxa)	increase
Long-lived taxa richness	decrease
Intolerant taxa richness	decrease
Percent tolerant individuals	increase
Clinger taxa richness	decrease
Percent predator individuals	decrease

Table 2-3. Results of rank-correlation comparison of CA and B-IBI metrics

Of the 72 column-by-column rank correlations, 28% had p-values less than 0.10, implying that there is *some* connection between CA and B-IBI metrics. However, Bonferroni corrected multiple-comparisons – even *post-hoc* best possible matchings between CA and IBI metrics were not significant.

	Transformation	% p-values < 0.1	Best p-value
1994	Presence/Absence	0.25	0.61
	Mean Presence	0.25	0.75
	Mean Individuals	0.25	0.17
1995	Presence/Absence	0.25	0.66
	Mean Presence	0.31	0.66
	Mean Individuals	0.50	0.23
1997	Presence/Absence	0.50	0.25
	Mean Presence	0.19	0.48
	Mean Individuals	0.25	0.63
	Overall	0.28	

Table 2-4. Metric coefficients for B-IBI and correspondence analysis

Correspondence analysis was performed on the mean individuals transformation of the 1995 dataset. Coefficients for a selection of the taxa present in the dataset are shown below, along with the corresponding coefficients for the Percent Predator and Percent Tolerant B-IBI metrics. This table does not list coefficients for all taxa, it is intended as a sample of the results used to calculate the correlations reported in Table 2-6.

	B-IBI metrics		**Correspondence analysis of 1995 mean individuals transformation**					
Taxon	Percent Predator	Percent Tolerant	CA Axis 1	CA Axis 2	CA Axis 3	CA Axis 4	CA Axis 5	CA Axis 6
Pelecypoda	0	0	0.56	0.01	-0.51	-0.31	-0.76	0.79
Colembolla	0	0	0.45	0.18	-0.15	-0.32	-0.15	0.77
Gastropoda	0	0	0.21	0.01	0.46	-3.62	-1.96	-1.06
Nematoda	0	0	-0.12	0.07	0.08	0.42	0.02	-0.18
Amphipoda	0	1	0.11	0.32	-0.07	-0.73	-0.76	0.14
Isopoda	0	1	-0.95	0.02	0.08	0.76	-1.57	-0.24
Turbellaria	0	0	-1.72	-0.24	-2.33	-0.20	0.61	-0.08
Hirudinea	1	1	-3.98	-0.61	-4.96	-1.35	2.32	0.18
Oligochaeta	0	0	-0.43	0.59	0.54	0.31	0.08	0.28
Ampumixis	0	0	0.40	0.36	-0.08	-0.05	0.11	0.05
Cleptemis	0	1	0.19	-0.17	0.35	-1.56	-0.29	-0.16
Heterlimnius	0	0	0.48	0.33	-0.08	0.07	0.21	0.03
Narpus	0	0	-0.26	-0.33	0.28	0.70	-0.27	0.02
Optioservus	0	1	0.63	0.55	-0.41	0.35	0.17	-0.40
Zaitzevia	0	1	0.61	0.23	-0.17	-0.22	0.30	-0.13
Bezzia.Palpomyia	1	0	0.04	-0.81	0.17	0.38	0.09	0.07
Chironomidae	0	0	0.19	-0.70	-0.24	0.32	-0.28	-0.04
Dixa	0	0	1.12	0.64	-1.02	-0.61	-0.80	2.64
Dixella	0	0	-0.05	-0.44	0.48	-0.19	0.46	-0.41
Chelifera	1	0	0.24	0.25	-0.28	0.21	-0.34	-0.26
Hemerodromia	1	1	0.29	0.33	-0.02	-0.56	-0.37	-0.52
Glutops	1	0	-0.31	-1.12	0.38	0.44	-0.01	-0.42
Pericoma	0	0	-0.07	-1.33	0.56	0.34	0.51	-0.20
Simulium	0	0	-0.96	-0.57	-0.17	0.60	-1.29	-0.42
Antocha	0	0	0.64	0.58	-0.55	0.19	0.03	-0.10
Dicronata	1	0	0.48	0.06	-0.23	0.50	0.09	-0.38
Hexatoma	1	0	0.41	-0.83	-0.01	0.66	0.23	0.31
Prionocera	0	0	0.58	0.94	0.05	0.72	0.78	-0.27
Sialis	1	0	0.68	-0.30	-0.71	-0.24	-0.99	0.64
Baetis	0	1	-0.81	0.20	0.22	-0.05	-0.07	0.04
Acentrella	0	1	0.03	-0.35	0.41	-0.62	0.50	0.29

Table 2-5. Metric coefficients for B-IBI and order-aggregated CA

The mean presence transformation of the 1994 data was aggregated by order – counts of all taxa belonging to the same order were summed before performing correspondence analysis to produce coefficients for each order. Each taxon was assigned its appropriate order coefficient for comparison with the B-IBI metric coefficients in the table below. Taxa in the same order have the same coefficients for both the B-IBI and CA metrics, but beyond that likeness there is little similarity between the two. This table does not list coefficients for all taxa, it is intended as a sample of the results used to calculate the correlations reported in Table 2-6.

	B-IBI Metrics				**Correspondence analysis of 1994 mean presence transformation**			
Taxon	Ephem. Richness	Plecoptera Richness	Trichoptera Richness	Clinger Richness	CA Axis 1	CA Axis 2	CA Axis 3	CA Axis 4
Baetis	1	0	0	0	-0.15	-0.10	-0.03	0.06
Acentrella	1	0	0	0	-0.15	-0.10	-0.03	0.06
Drunella	1	0	0	1	-0.15	-0.10	-0.03	0.06
Serratella	1	0	0	1	-0.15	-0.10	-0.03	0.06
Cinygmula	1	0	0	1	-0.15	-0.10	-0.03	0.06
Ironodes	1	0	0	1	-0.15	-0.10	-0.03	0.06
Epeorus	1	0	0	1	-0.15	-0.10	-0.03	0.06
Paraleptophlebia	1	0	0	0	-0.15	-0.10	-0.03	0.06
Kathroperla	0	1	0	1	0.30	0.29	-0.17	-0.07
Suwallia	0	1	0	0	0.30	0.29	-0.17	-0.07
Sweltsa	0	1	0	1	0.30	0.29	-0.17	-0.07
Haploperla	0	1	0	0	0.30	0.29	-0.17	-0.07
Neaviperla	0	1	0	0	0.30	0.29	-0.17	-0.07
Paraperla	0	1	0	0	0.30	0.29	-0.17	-0.07
Diploperla	0	1	0	0	0.30	0.29	-0.17	-0.07
Isoperla	0	1	0	1	0.30	0.29	-0.17	-0.07
Skwala	0	1	0	1	0.30	0.29	-0.17	-0.07
Paracapnia	0	1	0	0	0.30	0.29	-0.17	-0.07
Brachycentrus	0	0	1	1	-0.03	0.05	0.01	0.01
Micrasema	0	0	1	1	-0.03	0.05	0.01	0.01
B.Lepidostoma	0	0	1	0	-0.03	0.05	0.01	0.01
Oligoplectrum	0	0	1	0	-0.03	0.05	0.01	0.01
Glossosoma	0	0	1	1	-0.03	0.05	0.01	0.01
Arctopsyche	0	0	1	0	-0.03	0.05	0.01	0.01
Cyrnellus	0	0	1	0	-0.03	0.05	0.01	0.01
Neureclipsis	0	0	1	0	-0.03	0.05	0.01	0.01
Polycentropus	0	0	1	1	-0.03	0.05	0.01	0.01
Rhyacophila	0	0	1	1	-0.03	0.05	0.01	0.01

Table 2-6. Correlation after aggregating based on biological information before CA

Grouping taxa based on biological similarity before performing correspondence analysis produces metrics that correlate better with B-IBI metrics. There is an increase in the number of significant individual comparisons progressing from no aggregation to grouping by order and family. Grouping by feeding guild produced a number of significant comparisons in between order and family.

		Aggregated by Order		Aggregated by Family		Aggregated by Feeding Guild	
Year	**Transform**	**% p < 0.1**	**Best p**	**% p < 0.1**	**Best p**	**% p < 0.1**	**Best p**
1994	Presence/ Absence	0.25	0.23	0.25	0.22	0.25	0.50
	Mean Presence	0.38	0.29	0.50	0.11	0.63	0.23
	Mean Individuals	0.00	0.26	0.25	0.31	0.25	0.11
1995	Presence/ Absence	0.00	0.69	0.50	0.49	0.00	0.70
	Mean Presence	0.38	0.87	0.63	0.02	0.44	0.10
	Mean Individuals	0.50	0.42	0.50	0.81	0.50	0.60
1997	Presence/ Absence	0.25	0.53	0.50	0.48	0.00	0.97
	Mean Presence	0.38	0.27	0.88	0.00	0.50	0.11
	Mean Individuals	0.00	0.57	0.25	0.71	0.50	0.14
mean	Presence/ Absence	0.17	0.48	0.42	0.40	0.08	0.72
	Mean Presence	0.38	0.48	0.67	0.04	0.52	0.15
	Mean Individuals	0.17	0.42	0.33	0.61	0.42	0.28
	Overall	0.31		0.57		0.43	

2.9 – Figures

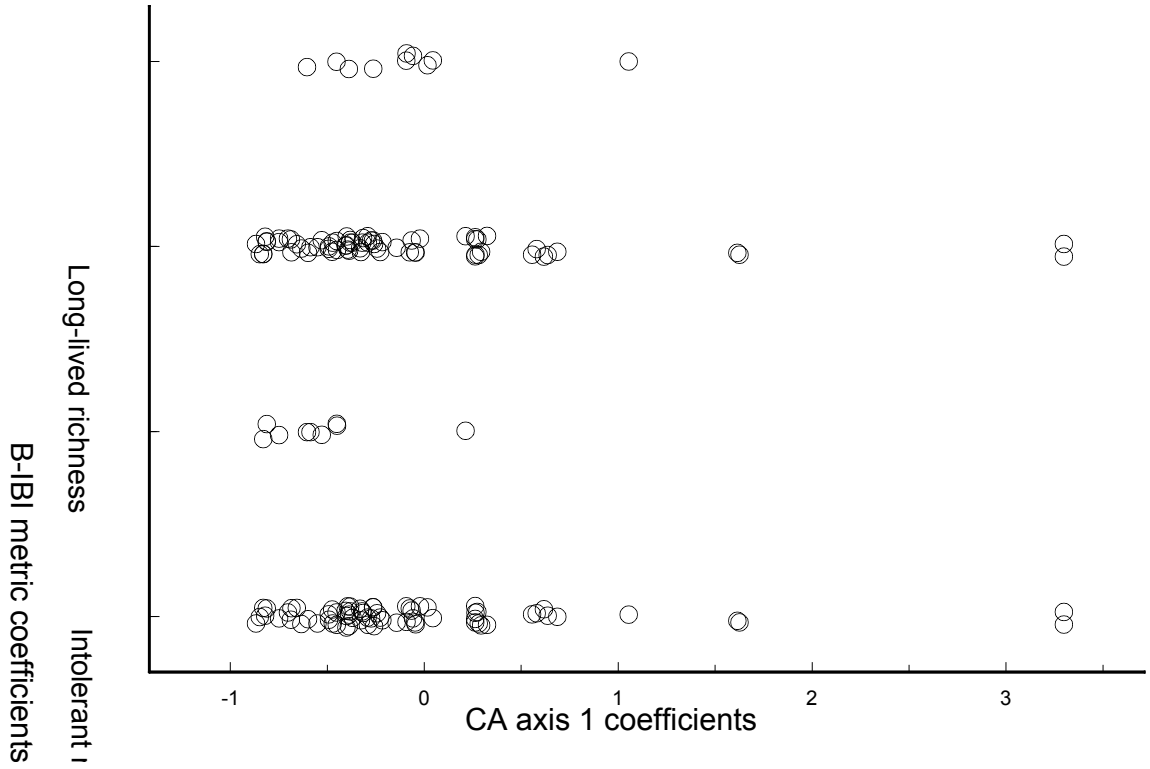

Figure 2-1 Plot of CA coefficients vs. B-IBI Long-lived and Intolerant coefficients
For the 1994 data set, the first correspondence analysis metric from the presence/absence transformation of the data produced coefficients for each taxon in the data set, here plotted against the corresponding coefficients for the B-IBI. Intolerant richness and Long-lived richness metrics. The B-IBI coefficients have been perturbed slightly from their true values of 0 or 1 to illuminate overlapping points. There is no visually obvious correlation between the two axes.

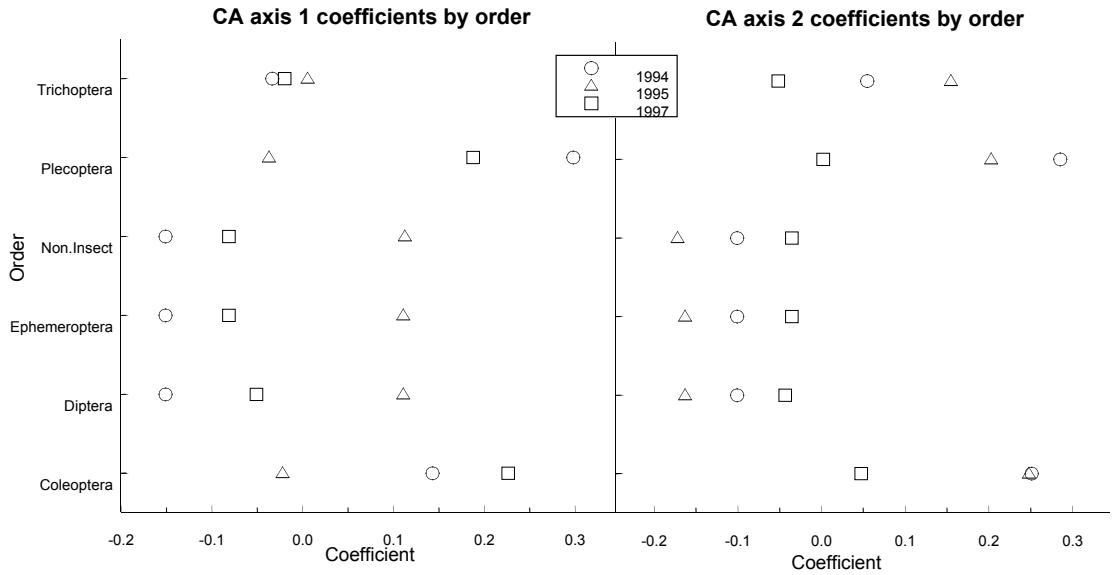

Figure 2-2. Top CA metric coefficients for order-aggregated mean presence data

Correspondence analysis applied to the order-aggregated mean presence transformation of the data produced different coefficients for each year in the data set. The two most significant metrics' coefficients had a significant rank correlation with the B-IBI Ephemeroptera and Plecoptera taxa richness metrics, but the different values, and orders of these coefficients across years prevents a biological interpretation.

3 — Concordance with other datasets

3.1 – Introduction

The method employed in devising the B-IBI selects metrics that respond to human influence and so provide a signal of site biological condition. It uses an external measurement of biological condition or human influence to identify those metrics. Metric scores that follow a dose-response relationship with the external measurement (among other criteria) are kept as useful.

Multiple regression finds the metric in a dataset that best correlates with another, response measurement. Canonical correlation finds the metrics in a dataset that best correlate with multiple external measurements simultaneously. How do the metrics produced by these two mathematics-based multivariate techniques compare with the metrics of the B-IBI?

Both multiple regression and canonical correlation produce metrics that maximize correlation with the external measurements. A metric must have other properties, however, to be useful as an indicator of biological condition. Both multiple regression and canonical correlation produce scores that correlate well with % impervious area (and by extension biological condition) in a set of sites. If those same metrics produce scores with little or no correlation to % impervious area at those same sites in another year, or with a different set of sites in the same year, they are not useful as indicators of biological condition.

The metrics of the B-IBI are, by definition, constant for all sites and years for the geographic region for which the index is calibrated.

In this chapter, I compare the correlation of metric scores produced by the B-IBI, multiple regression, and canonical correlation across three years of observations. I will also compare, across years, the metrics produced by multiple regression and canonical correlation, to see if they are alike enough to be useful as indicators of biological condition.

3.2 – Multiple regresssion

Description of multiple regression

While it is not usually considered such, multiple regression with a single response variable and many predictors is also a mathematically based multivariate technique. The response variable is modeled as an additive function of the predictor variables, or a transformation of the predictor variables.

Multiple regression assigns a coefficient to each of the predictor variables so that a weighted total, or "score" can be calculated for each site. A set of coefficients corresponds to a single line, or metric in the space of the predictor variables. The score for an site is the projection of the site onto the metric. The best metric minimizes the squared differences between the scores and the response variables.

Mathematical formulation of multiple regression

In multiple regression the best metric is traditionally called β, a vector of individual coefficients (β's) for each of the predictor variables. β can be estimated using the regression equation

$$\beta = \left(\mathbf{X}^{\mathbf{T}} \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^{\mathbf{T}} \cdot \mathbf{Y} \tag{3.1}$$

where **X** is the matrix of predictor variables by observation, and **Y** is the vector of responses (Neter *et al.* 1996).

3.3 – Canonical correlation

Description of canonical correlation

Canonical correlation can be considered an extension of multiple regression. It is used where there are two sets of variables measured for each observation. In addition to having multiple predictor variables, there are also multiple response variables.

Canonical correlation finds metrics in *each* set of variables. Scores are calculated in each set of variables by projecting the observations onto the appropriate metric. The best pair of metrics maximizes the correlation of scores. Once the best pair of metrics has been

identified, they are removed from the variable sets by projecting the observations into the remaining dimensions and the process is repeated.

In botany, both the abundances of plant taxa and a suite of physical environmental measurements (pH, moisture, altitude) are collected at a series of sites. Canonical correlation produces metrics, or combinations of plants that are closely associated with specific combinations of environmental variables.

Mathematical formulation of canonical correlation

If **X** is a matrix of m variables and **Y** is a matrix of p variables, both observed at the same set of sites, then we can imagine vectors of coefficients **a** and **b** so that

$$\mathbf{X}^* = \mathbf{a}^T \cdot \mathbf{x} = a_1 x_1 + a_2 x_2 + \cdots + a_m x_m \tag{3.2}$$

and

$$\mathbf{Y}^* = \mathbf{b}^T \cdot \mathbf{y} = b_1 y_1 + b_2 y_2 + \cdots + b_m y_m \tag{3.3}$$

where $\mathbf{X}^*$ and $\mathbf{Y}^*$ are vectors of site scores. The objective of canonical correlation is to find the vectors **a** and **b** so the correlation of $\mathbf{X}^*$ and $\mathbf{Y}^*$ ($\rho(\mathbf{X}^*, \mathbf{Y}^*)$) is maximized.

After Dillon and Goldstein (1984) we can define the variance-covariance matrices as

$$\begin{aligned} \Sigma_{XX} &\; E\left\{\left(\mathbf{X} - \mu_{\mathbf{X}}\right)\left(\mathbf{X} - \mu_{\mathbf{X}}\right)^T\right\} \\ \Sigma_{YY} &\; E\left\{\left(\mathbf{Y} - \mu_{\mathbf{Y}}\right)\left(\mathbf{Y} - \mu_{\mathbf{Y}}\right)^T\right\} \\ \Sigma_{XY} &\; E\left\{\left(\mathbf{X} - \mu_{\mathbf{X}}\right)\left(\mathbf{Y} - \mu_{\mathbf{Y}}\right)^T\right\} \end{aligned} \tag{3.4}$$

where μ_X and μ_Y are vectors of the means of each variable in **X** and **Y**. Now the correlation can be written as a function of **a** and **b**

$$\rho(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}^T \cdot \Sigma_{XY} \cdot \mathbf{b}}{\sqrt{\left(\mathbf{a}^T \cdot \Sigma_{XX} \cdot \mathbf{a}\right)\left(\mathbf{b}^T \cdot \Sigma_{YY} \cdot \mathbf{b}\right)}} \tag{3.5}$$

Because correlation is scale-invariant, we can normalize **a** and **b** so

$\mathbf{a}^T \cdot \Sigma_{XX} \cdot \mathbf{a} = \mathbf{b}^T \cdot \Sigma_{YY} \cdot \mathbf{b} = 1$, and $E\{\mathbf{X}^*\}$ and $E\{\mathbf{Y}^*\}$ are zero (the scores are centered around the origin). From there, minimizing the correlation is equivalent to solving

$$\left(\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \lambda\mathbf{I}\right)\cdot\mathbf{a} = \mathbf{0}$$
$$\left(\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \lambda\mathbf{I}\right)\cdot\mathbf{b} = \mathbf{0} \quad (3.6)$$

for **a** and **b**, where λ is an eigenvalue for the variance-covariance product matrices.

$$\left|\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \lambda\mathbf{I}\right| = 0$$
$$\left|\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \lambda\mathbf{I}\right| = 0 \quad (3.7)$$

There are two sets of eigenvectors for each eigenvalue, so

$$\mathbf{a} = \frac{\Sigma_{XX}^{-1}\Sigma_{XY}\mathbf{b}}{\sqrt{\lambda}}$$
$$\mathbf{b} = \frac{\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a}}{\sqrt{\lambda}} \quad (3.8)$$

and all that is needed is to solve for one of the variance-covariance characteristic equations (Equation 3.6) to obtain values for **a** and **b**.

Uses of canonical correlation

Canonical correlation and its extensions are used whenever it is desirable to establish a connection between two sets of measurements. In botany, both the abundances of plant taxa and a suite of physical environmental measurements (pH, moisture, altitude) are collected at a series of sites. Canonical correlation produces metrics, or combinations of plants that are closely associated with specific combinations of environmental variables. Identifying combinations of benthic macroinvertebrate counts that are associated with a measurement (or measurements) of site quality would be similar to the process to construct the B-IBI.

The B-IBI and canonical correlation both identify useful metrics as those whose scores correspond to measurements from an external set of data. In the B-IBI candidate metrics are proposed based on knowledge of the biology of the benthic macroinvertebrates. Generating candidate metrics in this fashion introduces another source of information, the biology of the organisms, to the B-IBI method that is not available to mathematics-based multivariate techniques. Once they have been generated, candidate metrics are evaluated by their response to a measure of human influence. Both multiple regression and canonical

correlation produce metrics that maximize the correlation between the metric scores and a measure, or combination of measures, of human influence. Neither multiple regression nor canonical correlation considers the biology of the organisms, only their presence/absence or abundance.

Question: Are metrics consistent across years?

The B-IBI defines metrics based on biological characteristics of the taxa involved. The metrics included are therefore constant. The fact that these metrics produce similar scores at sites with a similar biological condition across time and space is one of the properties that qualify them for inclusion in the B-IBI.

Multiple regression and canonical correspondence can be used to compute metrics whose scores are correlated with another measurement of human influence. If these metrics reflect a connection between the biota being sampled and the human influence at the site, then they should produce similar results (high correlation between score and human influence) for other sites, or for repeated samples from the same sites.

The reverse case is also true. If the multiple regression or canonical correspondence identifies a good metric (defined as maximizing correlation with the external measurements) for the same sites in two different years, but those two metrics are not alike, then the metrics have not identified a useful connection between the biology of the system and human influence.

A metric that produces useful scores (with a strong signal of biological condition) on a set of sites in one year, yet does not work for the same sites in another year, or other sites in the same year, is inconsistent. An inconsistent metric is not useful as a yardstick of biological condition; scores in the metric cannot be compared across time or space.

If multiple regression or correspondence analysis produces metrics that are consistent across multiple years, that consistency implies the metrics identify a useful connection between the biota and biological condition, suitable for use as an indicator of human influence and, by proxy, biological condition. In the absence of any correlation across years, I would conclude

that the technique has identified two different connections, perhaps a connection specific to the dataset, which is not useful as an indicator of biological integrity.

3.4 – Methods

3.5.1 Correlation of B-IBI scores across years

The consistency of B-IBI metrics was examined by computing B-IBI scores for all sites in three years of data, 1994, 1995, and 1997. To maintain uniformity with parallel examinations of multiple regression and canonical correlation metrics (see following sections), only the 46 (out of 104 total) taxa that were present in the datasets for all three years were used. This reduction in taxa would be expected to reduce the signal/noise ratio of B-IBI metric scores to % impervious area, and reduce correlation of B-IBI scores across years.

Individual B-IBI metric scores were calculated, using the reduced set of taxa, for all sites in all three years. Using the reduced set will also affect individual, un-scaled B-IBI scores slightly. Restricting the analysis to those taxa found in all three years eliminates rare taxa, which contain an important fraction of the sample information (Cao *et al.* 1998).

Cross-year correlations of scores were calculated for common sites. If the B-IBI metric scores are correlated across years, then the B-IBI technique is useful in deriving a stable, repeatable signal of biological condition. If the B-IBI metric scores are not correlated across years, I would conclude that the technique is not suitable for producing an indicator of biological condition.

Multiple regression of % impervious area and taxa richness vs. taxon counts

To evaluate multiple regression's ability to indicate biological condition the technique was used to regress both % impervious area and total taxa richness as linear functions of the 46 taxa common to all three years. Total taxa richness (in all 104 taxa) was used as a response variable because it is a consistent indicator of biological condition, even though it is not independent of the remaining taxa used as predictor variables.

When the original benthic macroinvertebrate data were collected, three replicates were taken from each site. Normally the replicates are combined (see section 1.5 – Transformations) but for multiple regression and canonical correlation the individual replicates were used to

provide a large enough sample size to fit coefficients for 46 different taxa. Treating the non-independent replicates as independent would be expected to bias estimated coefficients towards zero (Neter *et al.* 1996).

3.5.2 Correlation of regression metric scores across years

The consistency of multiple regression metrics was examined by computing site scores for all three years of data. The score for site j was calculated by averaging the taxon counts across replicates, multiplying by the appropriate fit coefficient (b_i), and summing across all t taxa.

$$s_j = \sum_{i=1}^{t} b_i \cdot \bar{x}_{i,j} \tag{3.9}$$

Cross-year score correlations were calculated for common sites. If the multiple regression metric scores are correlated across years, then multiple regression is effective at producing a signal of biological condition. If the multiple regression metric scores are not correlated across years, I would conclude that the technique is not able to reliably identify links between biota and biological condition.

3.5.3 Correlation of multiple regression metric coefficients across years

The multiple regression metrics generated for the 1994, 1995, and 1997 datasets were compared. Each metric was represented as a vector of coefficients, one coefficient for each of the 46 taxa used. The correlation coefficient was calculated for each pair of years. A high correlation would indicate that the same taxa tend to be given the same importance in determining the metric score across years; low correlation implies little or no similarity between the treatment of particular taxa across years.

Correlation of metric coefficients between years further implies that the metrics have identified the same connection between the biota and biological condition in each year. In the absence of correlation I would conclude that the multiple regression metrics had identified different connections in each year, or at least that the connection is too complicated to be represented as a linear combination of the taxon counts.

Rank correlation accounts for metric scaling

Scaling a metric, by multiplying it with a constant, might affect the Pearson's correlation coefficient with another, unscaled metric. Because it is the relative magnitudes of the coefficients that identify a metric (think of a vector in the X-Y plane: the line that goes from the origin to the point (2, 1) is the same as the line that goes from the origin to (4, 2), it just doesn't go as far) the Spearman rank correlation was also calculated to compare metrics. For the same reason, only the absolute value of the correlation coefficient should be considered when comparing the correlation of multiple regression metric coefficients (the line that goes from the origin to (-2, -1) is the same as the line from the origin to (2, 1), merely in the opposite direction). For example: the vectors (1, 2, 3, 4) and (-3, -6, -9, -12) have a strong relationship to each other, but would have a negative correlation coefficient.

3.5.4 Consistency of regression metrics across years

As a further check of the metrics' utility across years, the 1994 metrics were used to calculate scores for the sites in the 1995 and 1997 datasets. Like the correlation of B-IBI metric scores above, if the multiple regression metrics from the 1994 data produce site scores that are consistent across years, then the 1994 multiple regression metrics are a useful, repeatable signal of biological condition. If the scores are not correlated across years, I would conclude that the metrics are not a useful indicator of biological condition, as their efficacy at predicting % impervious area or total taxa richness is restricted to a single year.

Canonical correlation of taxon counts against % impervious area and total taxa

The ability of canonical correlation to indicate biological condition was examined by computing scores for both taxon metrics in all three years of data. The matrix of taxon counts was correlated with both % impervious area and total taxa richness simultaneously. % impervious area is the quantitative measure of human influence for each site, and total taxa richness at a site is another, well-used indicator of biological condition. Only the 46 taxa common to all three years were used for the canonical correlation, so total taxa richness contains some information on site biological condition not in the matrix of reduced taxa.

3.5.5 Correlation of canonical scores

Two response variables were used, so canonical correlation produces two metrics whose scores maximally correlate with combinations of the response variables, % impervious area and total taxa richness. Scores were computed for all sites in both metrics according to

$$s_j = \sum_{i=1}^{t} b_i \cdot \bar{x}_{i,j} \qquad (3.10)$$

where $c_{i,k}$ is the coefficient for taxon i in the k^{th} canonical correlation metric.

Cross-year score correlations were calculated for common sites. If the canonical metric scores are correlated across years, then canonical correlation is effective at producing a signal of biological condition. If the canonical metric scores are not correlated across years, I would conclude that the technique is not able to reliably identify links between biota and biological condition.

Canonical correlation scores are computed to maximally correlate with a combination of the response variables, rather than with an observed variable directly. Multiplying all of a metric's coefficients by a constant results in the same metric. Like comparisons of metric coefficients, when calculating the correlations of canonical scores, only the magnitude of the correlation coefficient should be considered. Two metrics, one of which assigns low scores to disturbed sites and high scores to undisturbed sites, and the other of which does the opposite, may be indicating the same information about biological condition, and would have a correlation coefficient close to -1. For an example, see the correlations between the second canonical metrics in 1995 and 1997 on Table 3-7.

3.5.6 Correlation of canonical metrics

The canonical correlation metrics generated for the 1994, 1995, and 1997 datasets were compared. Each metric was represented as a vector of coefficients, one coefficient for each of the 46 taxa used. The correlation coefficient was calculated for each pair of years. A high correlation would indicate that the same taxa tend to be given the same importance in determining the metric score across years; low correlation implies little or no similarity between the treatment of particular taxa across years.

If metric coefficients are correlated across years, that correlation implies that they have identified the same connection between the biota and biological condition in each year. In the absence of correlation, I would conclude that the canonical correlation metrics had identified different connections in each year.

Like the multiple regression coefficients, it is the relative magnitudes of the coefficients that specify the canonical correlation metrics. The Spearman rank correlation was also used to compare the canonical metrics. Again, only the absolute value of the correlation coefficients should be considered; high negative correlation coefficients are still strong correlation, albeit negative.

3.5.7 Consistency of canonical metrics across years

As a further check of utility across years, the 1994 metrics were used to calculate scores for the sites in the 1995 and 1996 datasets. If the canonical correlation metrics from the 1994 dataset produce site scores that are consistent across years, then the 1994 canonical correlation metrics are a useful signal of biological condition. If the scores do not correlate across years, I would conclude that the metrics are useful only within a single year, and are therefore not desirable as indicators of biological condition.

3.5 – Results

3.6.1 Correlation of B-IBI scores across years

Some of the B-IBI scores were highly correlated across years. Ephemeroptera, Plecoptera and Clinger richness had consistently high correlations ($\rho > 0.80$) indicating that they are consistent across years. Other metrics, such as dominance and percent tolerant were not a strongly correlated ($0.50 < \rho < 0.75$) and intolerant taxa richness was highly variable (ρ = .046, -.042, 0.944), indicating that they might be less useful as consistent indicators of biological condition (Table 3-1). The same pattern of more and less consistent metrics is seen in the Spearman (rank) correlations (Table 3-2).

Of the three years, 12 sites were in common in the 1994 and 1995 datasets, 11 between 1994 and 1997, and 6 sites in common between 1994 and 1997.

3.6.2 Correlation of regression metric scores across years

The multiple regression metric scores were an effective signal of biological condition. The correlations of scores across pairs of years were all high (minimum ρ = 0.70, half > 0.90) (Table 3-3 and Table 3-4). Multiple regression is effective at constructing metrics to reproduce both % impervious area and total taxa richness, two indicators of biological condition.

3.6.3 Correlation of regression metric coefficients across years

An examination of the actual multiple regression metrics derived for each year indicates that they are not suitable for use as an indicator of biological condition. The metric coefficients had little or no correlation (Pearson's or Spearman's) across years (Table 3-5 and Table 3-6). The metrics derived by multiple regression were distinctly different from year to year, implying the technique identified a different connection between the biota and the proxy for biological condition (either % impervious area or total taxa richness) in each year.

3.6.4 Consistency of regression metrics across years

Applying the same metric to multiple years also shows that the multiple regression metrics are not useful as indicators of biological condition. When the 1994 metrics were used to predict % impervious area and total taxa richness in 1995 and 1997, there was negligible correlation (3 of 4 < 0) between the scores in 1994 and the scores in the other two years (Table 3-3 and Table 3-4). This lack of consistency in across-year metric performance indicates that the multiple regression metrics are not useful indicators of biological condition, as they do not provide a useful signal outside of the dataset used to derive them.

3.6.5 Correlation of canonical scores

Canonical correlation is also successful at deriving metrics that signal biological condition. The correlations of scores across years were high (all $|r| > 0.80$, Table 3-7), indicating the technique can consistently derive metrics to link the biota and the indicators of biological condition.

Because there were two measures of site quality, two orthogonal metrics were computed for each year. The correlations between taxon scores and quality measures were .98, .90 and .86

for the first (strongest correlation) metric and .86, .81, and .82 for the second, in the '94, '95 and '97 datasets respectively.

3.6.6 Correlation of canonical metrics

Comparing the coefficients of the canonical metrics reveals that taxa were weighted differently in different years. These different weights indicate that the canonical correlation technique identified a different connection between the biota and biological condition in each year, and so the canonical metrics are not generally useful as indicators of biological condition. There was little or no correlation (Pearson's or Spearman's) of coefficients between years (Table 3-9 and Table 3-10).

3.6.7 Consistency of canonical metrics across years

When the canonical metrics derived for the 1994 dataset are used to compute scores for 1995 and 1997, there is little consistency in scores. This lack of consistency in scores shows that the canonical correlation metrics are not useful across years, and are therefore not desirable as indicators of biological condition.

3.6 – Discussion

Correlation of scores

The B-IBI, multiple regression, and canonical correlation are all effective techniques of deriving an indicator of biological condition from a vector of taxon counts. Generally the scores generated by all three techniques were similar at the same sites in different years, resulting in high correlations across years (Tables 3-4, 3-7 and 3-8). Intolerant taxa richness was a notable exception among the B-IBI scores, and is likely an artifact. Of the 113 taxa represented in all three datasets, only the 46 taxa present in all three years were used. This reduction has the effect of removing the rarest taxa, since uncommon taxa will be even less likely to be represented in all three datasets. Many of the rarer taxa are classified as intolerant, so removing them might have eliminated the signal of the Intolerant richness metric.

The multiple regression and canonical correlation metrics were computed using the same data used to calculate the scores. For multiple regression, the 1994 data was used to estimate

taxon coefficients to predict % impervious area, the 1995 data was used to estimate coefficients for % impervious area in 1995, and so on. Since the % impervious area for a site did not change among the three years, a high correlation of scores is to be expected.

The correlations are much lower when the metrics derived from 1994 were used to calculate scores in 1995 and 1997 (Tables 3-4, 3-7 and 3-8). The 1994 regression and canonical metrics do not do a good job of predicting a site's % impervious area or total taxa richness in 1995 or 1997. Some of the correlations are even negative, implying that sites with a high biological condition were assigned low scores, and vice versa. Clearly, the best metrics (as defined by multiple regression and canonical correlation) were different from year to year.

Correlation of metric coefficients

The B-IBI metrics are, by definition, identical from year to year. Multiple regression and canonical correlation metrics are defined by quantitative rules applied to specific datasets. The correlation of scores shows that all three methods are effective at deriving metrics to link taxon counts and the biological condition of a site.

The mere existence of a link, in a single set of sites or within a single year, does not imply that a mathematics-derived metric is useful for determining a site's biological condition. No correlation was found between the canonical correlation metrics computed from the 1994, 1995 and 1997 datasets. The metrics identified as best in each year were different, with a different importance attached to each taxon in each year.

The multiple regression metrics, whether trying to predict % impervious area or total taxa richness, were not correlated, with only a single correlation larger than 0.8, and half of the possible across-year correlations less than 0.5 (Table 3-5 and Table 3-6). Likewise the canonical correlation metrics were not correlated across years (Table 3-9 and Table 3-10) with all the correlation coefficients having an absolute value less than 0.5.

Conclusions

Both multiple regression and canonical correlation can, given a proxy measurement of a site's biological condition, identify combinations of taxon abundances that reproduce that proxy measurement. However, in this study the metrics they identified are specific to the

dataset used to calculate them. For example, multiple regression metrics whose scores closely followed % impervious area for the 1994 dataset (and therefore might be considered for use as a measure of biological condition) produced scores that were poorly correlated with % impervious area in the 1995 and 1997 datasets. A metric that was highly useful in the 1994 dataset was not useful in the 1995 or 1997 datasets. This disparity implies that the connections between taxon abundances and biological condition as identified in 1994 were not useful and consistent connections between the biota and biological condition. Instead, the 1994 connections were particular to the 1994 dataset.

In contrast, the B-IBI metric scores were not as highly correlated as the regression and canonical scores (Tables 3-2 vs. Tables 3-4, 3-7, and 3-8). This lower correlation may be due in part to the reduced set of taxa used to compute them, but can also be attributed to the method of selecting the B-IBI metrics. The scores of B-IBI metrics are not designed to maximize correlation with human influence, but instead produce a reliable and consistent correlation across time and space. As a result, the link between B-IBI metric scores and human influence is noisier, and the between-year correlations of scores lower than for the mathematics-based metric scores.

The B-IBI scores were not as highly correlated as the multiple regression or canonical scores, but they did produce generally high correlations, especially the Ephemeroptera, Plecoptera, Trichoptera, and Clinger taxa richnesses. These few metrics, at least, do seem to produce a reliable signal of biological condition across years. The same B-IBI metrics were used to calculate scores in all three years. As the same metric scores are correlated with biological condition across years, it is more likely that these B-IBI metrics correspond to or identify a useful connection between the biota and biological condition.

3.7 – Tables

Table 3-1. Pearson correlations of B-IBI metric scores

Across-year correlations of (unscaled) scores in the ten B-IBI metrics. There were 12 sites in common in the 1994 and 1995 datasets, 11 sites in common between 1994 and 1997, and 6 sites in common between 1995 and 1997.

	94-95 (12)	**94-97 (11)**	**95-97 (6)**
Taxa richness	0.595	0.652	0.894
Ephemeroptera richness	0.973	0.802	0.904
Plecoptera richness	0.837	0.820	0.965
Trichoptera richness	0.768	0.579	0.612
Intolerant richness	0.046	-0.042	0.944
Long-lived richness	0.566	0.700	0.891
Percent tolerant	0.748	0.578	0.892
Clinger richness	0.924	0.868	0.989
Percent predator	0.515	0.008	0.447
Dominance	0.619	0.523	0.707

Table 3-2. Spearman correlations of B-IBI metric scores

Across-year rank correlations of (unscaled) scores in the ten B-IBI metrics. There were 12 sites in common in the 1994 and 1995 datasets, 11 sites in common between 1994 and 1997, and 6 sites in common between 1995 and 1997.

	94-95 (12)	**94-97 (11)**	**95-97 (6)**
Taxa richness	0.449	0.425	0.754
Ephemeroptera richness	0.988	0.561	0.600
Plecoptera richness	0.586	0.809	1.000
Trichoptera richness	0.549	0.487	-0.029
Intolerant richness	0.136	0.027	0.775
Long-lived richness	0.577	0.615	0.971
Percent tolerant	0.615	-0.109	0.543
Clinger richness	0.888	0.692	0.943
Percent predator	0.671	0.409	0.429
Dominance	0.657	0.455	0.429

Table 3-3. Pearson correlations of regression metric scores

Linear regressions to predict % impervious area and total taxa richness as a function of taxon counts for each year, and their scores (predicted values) were correlated. The 1994 regressions were also used to calculate scores for the 1995 and 1997 datasets, and their correlations were much lower. There were 12 sites in common in the 1994 and 1995 datasets, 11 sites in common between 1994 and 1997, and 6 sites in common between 1995 and 1997.

	Year specific metrics			1994 metrics	
	94-95 (12)	**94-97 (11)**	**95-97 (6)**	**94-95 (12)**	**94-97 (11)**
% impervious area	0.990	0.828	0.895	-0.821	-0.263
Total taxa richness	0.934	0.780	0.915	-0.754	0.309

Table 3-4. Spearman (rank) correlations of regression metric scores

Linear regressions to predict % impervious area and total taxa richness as a function of taxon counts for each year, and their scores (predicted values) were correlated. The 1994 regressions were also used to calculate scores for the 1995 and 1997 datasets, and their correlations were much lower. There were 12 sites in common in the 1994 and 1995 datasets, 11 sites in common between 1994 and 1997, and 6 sites in common between 1995 and 1997.

	Year specific metrics			1994 metrics	
	94-95 (12)	**94-97 (11)**	**95-97 (6)**	**94-95 (12)**	**94-97 (11)**
% impervious area	0.965	0.755	0.943	-0.245	-0.482
Total taxa richness	0.874	0.700	0.943	-0.238	0.055

Table 3-5. Pearson correlations of regression metric coefficients

Linear regressions to predict % impervious area and total taxa richness as a function of taxon counts for each year. Pearson's correlation coefficients were calculated for the metric coefficients estimated in each year.

	94-95	**94-97**	**95-97**
% impervious area	0.587	0.429	0.811
Total taxa richness	0.013	0.087	0.641

Table 3-6. Spearman correlations of regression metric coefficients

Linear regressions to predict % impervious area and total taxa richness as a function of taxon counts for each year. Spearman's (rank) correlation coefficients were calculated for the metric coefficients estimated in each year.

	94-95	94-97	95-97
% impervious area	0.398	0.056	0.082
Total taxa richness	0.333	0.266	0.153

Table 3-7. Pearson correlations of canonical metric scores

Canonical correlations to predict % impervious area and total taxa richness simultaneously as a function of taxon counts for each year, and their taxon scores were correlated. The 1994 canonical analysis was also used to calculate scores for the 1995 and 1997 datasets, and their correlations were much lower. There were 12 sites in common in the 1994 and 1995 datasets, 11 sites in common between 1994 and 1997, and 6 sites in common between 1995 and 1997.

	Year specific metrics			1994 metrics	
	94-95 (12)	94-97 (11)	95-97 (6)	94-95 (12)	94-97 (11)
First metric	0.945	0.803	0.862	0.176	0.213
Second metric	-0.822	0.819	-0.908	0.483	0.562

Table 3-8. Spearman (rank) correlations of canonical metric scores

Canonical correlations to predict % impervious area and total taxa richness simultaneously as a function of taxon counts for each year, and their taxon scores were correlated. The 1994 canonical analysis was also used to calculate scores for the 1995 and 1997 datasets, and their correlations were much lower. There were 12 sites in common in the 1994 and 1995 datasets, 11 sites in common between 1994 and 1997, and 6 sites in common between 1995 and 1997.

	Year specific metrics			1994 metrics	
	94-95 (12)	94-97 (11)	95-97 (6)	94-95 (12)	94-97 (11)
First metric	0.895	0.755	0.657	0.273	-0.036
Second metric	-0.812	0.418	-0.657	0.266	-0.036

Table 3-9. Pearson correlations of canonical metric coefficients

Canonical correlations to predict % impervious area and total taxa richness simultaneously as a function of taxon counts for each year, and their taxon scores were correlated. Pearson's correlation coefficients were calculated for the metric coefficients estimated in each year.

	94-95	94-97	95-97
% impervious area	0.225	0.039	0.050
Total taxa richness	0.045	0.092	-0.491

Table 3-10. Spearman correlations of canonical metric coefficients

Canonical correlations to predict % impervious area and total taxa richness simultaneously as a function of taxon counts for each year, and their taxon scores were correlated. Spearman's (rank) correlation coefficients were calculated for the metric coefficients estimated in each year.

	94-95	94-97	95-97
% impervious area	0.356	-0.035	-0.041
Total taxa richness	-0.382	0.118	0.059

4 — Comparing multimetric indexes

4.1 – Introduction

The Benthic Index of Biological Integrity is not the only multimetric index in use. The Rapid Bioassessment Protocol (Plafkin *et al.* 1989) (RBP) includes an eight metric multimetric index for benthic macroinvertebrates. The original RBP was developed with data collected from streams in North Carolina. Mulvey *et al.* (1992) adapted the original RBP for use in Oregon, modifying some metrics, removing others, and adding new ones (Table 4-1, Table 4-2).

The metrics in the Benthic Index of Biological Integrity were selected by screening a pool of 38 candidate metrics for those displaying a response to the percent impervious area of a watershed as a proxy for human influence. The results were based on studies done in Washington (Kleindl 1995), Tennessee (Karr 1991), Oregon (Fore *et al.* 1996), Wyoming (Patterson 1996) and Japan (Rossano 1995). Metrics that showed a consistent response across these geographic regions were chosen for inclusion in the B-IBI (Karr and Chu 1998).

In contrast, the original RBP metrics were chosen from a pool of 13 metrics, winnowed to seven through use of a cluster analysis of the metric scores, with % Shredder added afterwards (Plafkin *et al.* 1989). On page 6-33 of (Plafkin *et al.* 1989 p. 6-33), Plafkin *et al.* states that "The few data acquired by this one pilot study do not constitute a rigorous analysis, nor are the results obtained by the cluster analysis intended to be a definitive validation of the rapid bioassessment technique." and expresses hope for further refinement, with larger datasets, in the future.

A useful metric provides information about the biological condition of the watershed from which a sample is taken. It should at least distinguish between undisturbed and highly degraded sites. Metrics which provide a signal across a range of site conditions are especially useful, as distinguishing among or ranking the condition of marginally disturbed sites is more difficult than recognizing the best and worst sites. The metrics used in multimetric indexes are chosen with the thought that certain groups of taxa (chosen by taxonomic similarity,

feeding strategy, or other common characteristics), provide a meaningful signal by virtue of their specific biology.

This line of reasoning is intuitive, and has been used in multimetric indexes to produce metrics whose scores bear a convincing relationship to site biological condition (Ohio EPA 1987, Plafkin *et al.* 1989, Kerans and Karr 1994). The benefit of using one group of taxa rather than another can be measured by comparing those biology-defined metrics to metrics generated without any reference to biology or site biological condition.

Metrics can be generated randomly by selecting taxa without consideration of their specific properties. If a candidate metric provides a better signal of biological condition than a random metric, it implies that the biological reasoning underlying the metric is sound. Furthermore, a biologically chosen metric that provides a better signal than many, many random metrics may be more useful than one that is better than merely "many" random metrics. Comparing the relative strength of metrics – after adjusting for the effects of the number of taxa included – might aid the designers of multimetric indexes in deciding which metrics to include or omit from their indexes.

4.2 – Theoretical model

Bernoulli distribution

When a random trial has two possible outcomes (success/failure, present/absent 1/0) with the process is said to follow a Bernoulli distribution with some probability, p, of one outcome and 1-p for the other outcome. If the random variable Y is distributed Bernoulli(p), then the probability that Y takes some value y is given by

$$P(Y = y) = p^{y} \cdot (1 - p)^{1-y} \tag{4.1}$$

where y can take values of 0 or 1 and p ranges from 0 to 1. The expected value of Y is p, and the variance of y is $p \cdot (1 - p)$.

Presence/Absence as a function of disturbance

Consider the matrix **X**, an r × c matrix with presence absence information for r taxa at c different sites. The individual elements, $x_{i,j}$ have a value of 0 if the taxon is not present at a site, and 1 if it is present. An individual x can be modeled as a Bernoulli random variable

$$x \sim \text{Bernoulli}(p) \tag{4.2}$$

where p is the probability of finding the taxon at that site.

If p is constant for some taxon, then that taxon is equally likely to be found at all sites. It is more useful to consider how p might change as a function of site condition

$$p = f(d, \mathbf{v}) \tag{4.3}$$

where d is the amount of human disturbance of a site away from a condition of biological integrity, and **v** represents a vector of other physical and biological parameters for a site. An obvious start would be to model p as a linear function of disturbance.

$$p = \alpha + \beta \cdot d \tag{4.4}$$

In this case α represents the probability of finding the taxon at a completely undisturbed site, and β controls how quickly that probability decreases (or increases) with increasing disturbance. Ephemeroptera richness is a biologically defined metric used in the B-IBI, which shows a linear response to degradation (Figure 4-1).

Models that are more complex are possible. The probability might follow a bent linear form

$$p = \begin{cases} \alpha_1 + \beta_1 \cdot d & d < \dfrac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1} \\ \alpha_2 + \beta_2 \cdot d & d \geq \dfrac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1} \end{cases} \tag{4.5}$$

where α_1 and β_1 describe the probability for low disturbance and α_2 and β_2 describe the probability of finding a taxon at higher disturbance. The B-IBI intolerant taxa richness metric follows a bent-linear response (Figure 4-2).

The bent-linear model could be adapted slightly, by allowing a linear decline in probability up to a certain level of disturbance (d_{crit}) with zero probability of finding the taxon beyond that point.

$$p = \begin{cases} \alpha + \beta \cdot d & d < d_{crit} \\ 0 & d \geq d_{crit} \end{cases} \tag{4.6}$$

Alternatively, the probability could be modeled as a simple step function with a constant probability of finding the taxon at low levels of disturbance and zero probability at high disturbance.

$$p = \begin{cases} p_{const} & d < d_{crit} \\ 0 & d \geq d_{crit} \end{cases} \tag{4.7}$$

Total taxa richness

Remembering that i represents taxa and j represents sites, then the total taxa richness at some site j is the sum of the $x_{i,j}$ for all r taxa at that site.

$$\text{total taxa richness}_j = \sum_{i=1}^{r} x_{i,j} \tag{4.8}$$

The expected value of the total taxa richness ($E\{\text{total taxa richness}\}$) is the sum of the expected values of the individual $x_{i,j}$'s

$$E\{\text{total taxa richness}_j\} = \sum_{i=1}^{r} p_{i,j} \tag{4.9}$$

where each individual probability is a function of the site's disturbance and other particular parameters.

$$p_{i,j} = f_i(d_j, \mathbf{v_j}) \tag{4.10}$$

If most individual taxon functions follow a decreasing trend with increasing disturbance then sites with low disturbance are expected to have higher total taxa richness than highly disturbed sites. There are groups for which the opposite holds – taxa that are more likely to be found at degraded sites – but for most taxa this is the pattern observed in nature.

Assuming independence, the variance of the sum of random variables is simply the sum of the variances of the individual variables

$$\mathrm{var}\{A + B\} = \mathrm{var}\{A\} + \mathrm{var}\{B\} \tag{4.11}$$

so the variance of total taxa richness is the sum of the variances of the individual Bernoulli distributed presence/absence values.

$$\mathrm{var}\{\text{total taxa richness}_j\} = \sum_{i=1}^{r} p_{i,j} \cdot (1 - p_{i,j}) \tag{4.12}$$

The presences or absences of two taxa are unlikely to be completely independent for all taxa, and the variance of the sum of two non-independent random variables is

$$\mathrm{var}\{A + B\} = \mathrm{var}\{A\} + \mathrm{var}\{B\} - 2 \cdot \mathrm{cov}\{A, B\} \tag{4.13}$$

However, in the absence of specific information, independence is a reasonable first approximation, and in a more informed model covariance terms could be treated in the same fashion as the site-specific **v** parameters in the section *Decreasing noise with a selection of taxa* below.

Richness metrics

Richness can also be calculated for a subset of taxa. Let **m** be a subset of **o**, the unknown but real set of all taxa that might be found in sites in the geographic region being investigated. If the set **m** has n members, then the richness score for that metric at site j (s_j) can be calculated as

$$s_j = \sum_{i=1}^{n} x_{i,j} \tag{4.14}$$

An example of calculating the Ephemeroptera taxa richness metric is given in Table 4-3. The expected value and variance of s_j are given by

$$E\{s_j\} = \sum_{i=1}^{n} p_{i,j} \tag{4.15}$$

and

$$\mathrm{var}\{s_j\} = \sum_{i=1}^{n} p_{i,j} \cdot (1 - p_{i,j}) \tag{4.16}$$

No response

If there is no response of taxa presence/absence to site conditions, and all taxa are equally likely to be found at all sites, then all $p_{i,j}$ equal some constant p and

$$E\{s_j\} = \sum_{i=1}^{n} p = n \cdot p \tag{4.17}$$

and

$$\mathrm{var}\{s_j\} = \sum_{i=1}^{n} p \cdot (1 - p) = n \cdot p \cdot (1 - p) \tag{4.18}$$

If p is held constant then the expected value and variance of the metric score are proportional to n, the number of taxa in the metric.

The coefficient of variation is the ratio of the standard deviation of a random variable to its expected value

$$\mathrm{c.v.}\{s_j\} = \frac{\sqrt{\mathrm{var}\{s_j\}}}{E\{s_j\}} = \frac{\sqrt{n \cdot p \cdot (1 - p)}}{n \cdot p} \tag{4.19}$$

Again, if p is held constant the expected score increases directly with n while the coefficient of variation *decreases* as a function of $\frac{1}{\sqrt{n}}$.

Linear response

It is more interesting to allow p to vary with site condition. If d_j is the disturbance at site j and all taxa follow the same linear response to site disturbance

$$p_{i,j} = \alpha + \beta \cdot d_j \tag{4.20}$$

then the expected value for s_j is

$$E\{s_j\} = \sum_{i=1}^{n} \alpha + \beta \cdot d_j \tag{4.21}$$

or

$$E\{s_j\} = n \cdot \alpha + n \cdot \beta \cdot d_j \tag{4.22}$$

and it can be seen that the slope of expected metric scores as a function of disturbance increases with n, the number of taxa in the metric.

If all taxa presence/absence probabilities have the same linear response to site disturbance the variance of s_j is given by

$$\mathrm{var}\{s_j\} = n \cdot (\alpha + \beta \cdot d_j) \cdot (1 - \alpha - \beta \cdot d_j) \tag{4.23}$$

so again the coefficient of variation decreases as $\frac{1}{\sqrt{n}}$, the number of taxa in the metric increases.

Signal and noise

Increasing signal with selection of taxa

If the probability of finding a taxon at a site is modeled as a linear function of disturbance, it is unlikely that the function is identical for all taxa, rather

$$p_{i,j} = \alpha_i + \beta_i \cdot d_j \tag{4.24}$$

with different intercepts and slopes for each taxon.

If the presence probabilities are a linear function of disturbance, then total taxa richness

$$E\{\text{total taxa richness}_j\} = \sum_{i=1}^{r} \alpha_i + \beta_i \cdot d_j \tag{4.25}$$

will also be a linear function of disturbance, and total taxa richness at a site provides a signal of disturbance at the site. The slope of the response to total taxa richness reflects the mean of the individual taxon β_i's. If some taxa do not follow the linear model, they will have the effect of adding noise to the overall linear response.

It would be desirable to identify a subset of total taxa so that the richness metric score for the subset has a greater signal than that of total taxa richness. In a linear model of presence/absence probability this would occur if the mean of the β_i parameters in the subset **m** ($\bar{\beta}_m$) is larger than $\bar{\beta}$, the mean of β_i parameters for all taxa.

$\bar{\beta}$ is not known exactly, but the mean of β parameters for a simple random sample of taxa will, in expectation, approach the true $\bar{\beta}$. For a particular set of taxa, **m**, with n taxa in it, the expected score at site j will be

$$E\{s_j\} = \sum_{i=1}^{n} \beta_i + \alpha_i \cdot d_j = \left(\sum_{i=1}^{n} \beta_i \right) + n \cdot \bar{\alpha}_m \cdot d_j \qquad (4.26)$$

If those n taxa are chosen as a random sample of taxa then $\bar{\beta}_m$ will be an estimate of $\bar{\beta}$ and

$$E\{s_j\} = \left(\sum_{i=1}^{n} \beta_i \right) + n \cdot \bar{\alpha} \cdot d_j \qquad (4.27)$$

If n taxa are chosen to provide an improved signal over total taxa richness, the slope can be compared to that obtained from a random sample of n taxa. If $n \cdot \bar{\alpha}_m$ is greater than $n \cdot \bar{\alpha}$ then the particular set **m** does provide a steeper slope and greater signal of disturbance, than taxa in general. To overcome the uncertainty introduced by estimating $\bar{\alpha}$ for all taxa with a finite sample, repeated random samples can be drawn to produce a population of responses of metrics with n taxa, and the proportion of those estimates smaller than $\bar{\beta}_m$ observed.

If the n taxa in subset **m** have a stronger response to disturbance than taxa in general, this represents a biological signal associated with those particular taxa. Choosing n taxa at random produces a response to disturbance without the biological signal associated with a particular group. If the two response strengths are not distinguishable, that sameness implies the additional biological signal of set **m** to disturbance is not significant. If the response strengths are different, that difference implies that the biological signal of the taxa in set **m** is significant.

Decreasing noise with a selection of taxa

If it is possible to reduce the variance introduced by the range of site conditions represented by **v** through the choice of taxa to include in the metric, then there is a set of taxa, or perhaps more than one set of taxa, that minimizes this variance. With a large (> 10) number of taxa, a random metric would be unlikely to happen upon the minimum variance

combination. The expected variance of a random metric would be Π, the overall mean variance. A useful candidate metric should have a variance less than Π.

Equation 4.3 states that the probability of finding a taxon at a site is a function of many parameters particular to a site outside of disturbance, collectively a vector, **v**, of specific site properties. This vector incorporates a site's geological setting, aspect, substrate, natural flow regime, connectivity to the surrounding ecosystem, and many other properties that will affect the probability of finding a taxon at that site.

The exact processes by which these parameters affect taxon presence/absence are, at best, complex and incompletely understood. In can be modeled simply, by supposing that there are K discrete classes of site (type 1, type 2, type 3, …), all with the same level of disturbance ($d_j = d$). The type of an individual site is not known, only that there are discrete types of site.

These site types all exist within the same broader category of sites meant to be compared. Variation is greatly reduced by properly classifying the sites to be examined beforehand, by the type of system sampled, geographic region, altitude and many others. Sites must be alike enough to be reasonably compared, but the practical objective of find a metric that is useful to compare many sites may mean there are distinct types within the set being studied.

Not all taxa are expected to be found at all types of site. If the response to disturbance is linear, the probability of finding a taxon at a site becomes

$$p_{i,j,k} = \begin{cases} \alpha_i + \beta_i \cdot d & \text{site is of type k} \\ 0 & \text{otherwise} \end{cases} \tag{4.28}$$

The expected value for the score in subset **m** containing n taxa must not encompass the fraction of sites that are of type k, and the number of taxa (out of n) that might be found at a type k site.

For a population of sites with identical disturbance d but distributed among K different site types, and making a simplifying assumption that all taxa have an identical response to disturbance, the probability of finding a taxon at site j of unknown type is

$$p = p_d \cdot p_{app} \tag{4.29}$$

where p_d is the probability of finding a taxon at the level of disturbance d and p_{app} is the probability that the taxon is appropriate for the site type. For a set of n taxa, p_{app} can be calculated as

$$p_{app} = f_{t,k} \cdot f_{s,k} \tag{4.30}$$

where $f_{t,k}$ is the fraction of taxa that are appropriate for site type k and $f_{s,k}$ is the fraction of streams that are of type k. Since the number of taxa in the metric, n, is known $f_{t,k}$ can be calculated as $\frac{n_k}{n}$, the number of taxa appropriate for site type k divided by the number of taxa.

The expected value of the richness score for a metric can be calculated as the sum over the distinct site types of the expected value at a site of that type times the probability that a site is that type

$$E\{s_j\} = \sum_{k=1}^{K} f_{s,k} \cdot E\{s_j \mid \text{site j is of type k}\} \tag{4.31}$$

since

$$E\{s_j \mid \text{site j is of type k}\} = \sum_{i=1}^{n_k} p_d = n_k \cdot p_d \tag{4.32}$$

then

$$E\{s_j\} = \sum_{k=1}^{K} f_{s,k} \cdot n_k \cdot p_d = \sum_{k=1}^{K} f_{s,k} \cdot f_{t,k} \cdot n \cdot p_d = n \cdot p_d \sum_{k=1}^{K} f_{s,k} \cdot f_{t,k} \tag{4.33}$$

The expected value of s_j^2 is given by

$$E\{s_j^2\} = n^2 \cdot p_d \left[\sum_{k=1}^{K} f_{s,k}^2 \cdot f_{t,k}^2 \right] \tag{4.34}$$

so the variance is

$$\text{var}\{s_j\} = n^2 \cdot p_d^2 \left[\sum_{k=1}^{K} f_{s,k}^2 \cdot f_{t,k}^2 \right] - \left[\sum_{k=1}^{K} f_{s,k} \cdot f_{t,k} \right]^2 \tag{4.35}$$

Again, metric score is expected to increase with metric size while the c.v. decreases as $\frac{1}{\sqrt{n}}$. The variance is a function of the distribution of sites among the type categories, and the distribution of type-appropriate taxa within the set. With two categories the variance is minimized when $f_{s,k} = f_{t,k}$. With more than two categories there is no exact solution, because a taxon might be appropriate for more than two site types. Minimizing the variance is further complicated by recognizing that the assumption of identical response to disturbance by all taxa is not reasonable. It is also likely that for some taxa the presence/absence probability will vary with site type, rather than simply falling to zero at inappropriate sites.

Still, one can propose that for some distribution of site types it is possible to choose a set of taxa so that the individual taxa responses to disturbance and their distribution of appropriateness among site types reduces variation in the metric score. The choice of taxa to include in this metric represents a reduction of noise through the properties of taxa included.

A random sample of n taxa would not be expected to have the individual responses to disturbance and distribution among types that minimizes the variance of metric score. The responses and distribution of a random sample would instead reflect the average response and variance of all taxa. Again, repeated random samples could be used to characterize the overall average variance of metric scores. If a particular set of taxa does not produce a lower variance, that sameness implies that the noise reduction produced by that particular set is not significant. If a particular set of taxa does produce a lower variance (reflected in a more significant regression coefficient, for instance) that significance implies that the choice of taxa is useful for extracting a signal of disturbance.

Considering the natural history of taxa should produce metrics with a stronger relationship to % impervious area than that of random metrics.

Random metrics

If, as a thought experiment, we toss out the assumption that different taxa respond differently to their environment and disturbance, then we are left with the null hypothesis that all taxa are equivalent; no particular taxon is more or less useful than any other for inclusion in a metric.

Computer simulation under this null hypothesis is easily done. Taxa can be randomly selected for inclusion in a metric, the data set transformed as appropriate (presence/absence, for example) and added together to produce scores for each site. I will call such a randomly generated metric a *random metric*. An example of a random metric is given in Table 4-4.

Biologically defined metrics

For a biologically defined metric, the individual m_i elements of a metric **m** are defined by their biology; for Ephemeroptera taxa richness the m_i's corresponding to Ephemeroptera taxa are 1, for all others 0. In a random metric the m_i's are assigned 0 or 1 at random.

Metric strength

A statistic measuring the degree of relationship between the random metric score at a site and the site's biological condition (or a measure of disturbance at the site, percent impervious area in the current case) can be calculated, and the process repeated a large number of times to produce a population of random metric statistics.

If the null hypothesis is not true, and taxa have different responses to environmental degradation, then a well-chosen metric will produce scores that have a clearer relationship to biological condition than that of most random metrics. The *strength* of a metric can be evaluated as the fraction of random metrics whose scores have a better relationship to biological condition.

Test statistics (w's) and metric strengths (p's)

If w is a statistic measuring the degree of relationship between metric score and biological condition, then a thousand randomly generated metrics produces a population of one thousand w's. If a candidate (biological) metric produces a w value larger than 920 of the randomly produce w's, then its strength is defined as $p = 0.08$. An alternative interpretation would be that the metric is in the 92^{nd} percentile. The strength, p, is not a p-value in the inferential statistical sense, but it does have similar interpretations.

Types of relationship between score and biological condition

Measuring the degree of relationship between a metric score and biological condition is not simple. IBI metrics, as mentioned previously, are selected after inspection of a graph of

metric scores against a biological condition proxy (and often a rank correlation), but graphical inspection is not amenable to computer simulation. In addition, useful metrics exhibit many response patterns to a range of biological condition. Some metrics are primarily useful for distinguishing between the best and worst sites, some display a linear or log-linear response to biological condition. Another common pattern is of a bent line, with little or no response below a certain point, and then sharp increase. The bent-line pattern is also an appropriate model for a response that follows an exponential or logarithmic curve.

Simulation requires that a computer algorithm be able to assess the degree of relationship between metric score and percent impervious area. Two-sample statistical tests can look for differences between the scores of sites with high and low biological condition (Figure 4-3). Linear regression can find straight line or log-linear fits (Figure 4-1), and non-linear regression techniques can be used to fit other functional forms (Figure 4-2).

Broken line

Another possibly useful metric shape to be considered is that of a *broken line* (Figure 4-4). The population of an individual taxon, or richness of a group of taxa, might follow an "s" shaped curve similar to that in the figure. Murray (1996) describes how a population under different levels of predation might have different equilibria. If "human influence" is substituted for "predation", the same pattern might appear in a plot of metric scores against human influence. Some solutions are unstable, so the observed population response to increasing human influence is a gradual decline from a high level, followed by a catastrophic drop to low levels. As human influence decreases, the metric would rise, slowly, until reaching another catastrophe point, at which it returns to the high level. In the middle section, where the two lines overlap, there are two stable metric scores for a given level of human influence. The upper level is accessible only from above, as human influence increases, while the lower level is only accessible from below, to sites with decreasing human influence.

None of the sites in the PSLS database have experienced a decrease in human influence, so there would not be any overlap between the high and low equilibria. The resulting pattern

would approximate a pair of lines representing the low and high states. Non-linear regression could be used to fit such a pattern, and test statistics (w's) collected for comparison.

Uncertainty in biological condition

A metric's score at a site responds to the biological condition at that site, and biological condition cannot be measured exactly. The percent impervious area is an uncertain measurement of the cumulative human influence, which is the proxy for biological condition, and that uncertainty will affect quantitative tests of metric response. For linear regression, uncertainty in the x-variables (% impervious area standing for human influence in this case) results in a bias of the slope towards zero (Neter *et al.* 1996).

If the test statistics for random metrics and candidate (biological) metrics are calculated with the same, uncertain measures of human influence, any biases will also be the same. The calculated strength (p) of a metric is relative to the population of random metrics. If the biases are the same for random and candidate metrics, then the metric strength is unaffected by those biases.

Interpretation of metric strength

If the strength (p) of a biologically defined metric is very small, the metric's scores have a closer relationship with % impervious area (and by proxy with human influence and biological condition) than the randomly defined metrics. The null hypothesis of equivalent taxa would be rejected in favor of believing the consideration of biology makes for a more useful metric.

Types of metric

There are three broad categories of metrics: metrics which look at the sum of the variables for the selected taxa (richness metrics) and metrics which look at the sum of the selected taxa variables relative to the total (percentage metrics) are used in all three indexes. Ratio metrics, which look at the ratio of richness for two selected groups of taxa, were used in the original RBP, but are not used in the Oregon modification or the B-IBI. Plafkin *et al.* (1989) included ratio metrics as an additional measure of community balance. Taking the ratio

random variables tends to inflate their variance (Casella and Berger 1990), and can produce singularities when the denominator is zero. For the Oregon modification of the RBP Mulvey *et al.* (1992) included the numerators and denominators of the original RBP ratio metrics as separate metrics.

Richness metrics are calculated with either the **Z** or **Y** transformed matrices (see section 1.5 – Transformations of this document for definitions of the transformations). For example, Ephemeroptera Taxa Richness is a sum metric, the number of mayfly taxa found at a site.

Percentage metrics are calculated with the **W** transformed matrix. For example, the % Predator Individuals taxa is a percentage metric, the fraction of a site's individuals classified as predators.

Ratio metrics are calculated as the ratio of two other metrics. They can use any of the **W**, **Y** or **Z** transformations. The RBP Scrapers/Collectors metric is a ratio metric, the number of individuals classified as scrapers divided by the number classified as collectors.

There are other metrics which cannot be fit into the above three categories. Random metrics as described above cannot be generated for the Shannon diversity and information theory based loss indexes. These metrics use all of the taxa, so random metrics cannot be generated by randomly drawing a sample of taxa.

Hilsenhoff metric

The Hilsenhoff metric relies upon assigning a weight ranging from 0 to 10 to each taxon. These weights represent how bad an organism is, "bad" meaning the organism is characteristic of degraded systems. An organism like *Hirudinea* worms receives a weight of 10, while *Caudatella* mayflies characteristic of undegraded sites are assigned a 1. Taxa found in moderately degraded systems, or common across a wide range of site quality have intermediate values (Hilsenhoff 1982).

The Hilsenhoff metric score at site j is simply the number of individuals in each taxa, multiplied by the appropriate weight, divided by the total number of individuals.

$$s_j = \frac{\sum_{i=1}^{r} h_r \cdot x_{r,j}}{\sum_{i=1}^{r} x_{r,j}} \tag{4.36}$$

where h_r is the Hilsenhoff weighting for taxon r. In matrix notation, the vector of site scores is

$$\mathbf{s} = \mathbf{h} \cdot \mathbf{X}_{mp} \tag{4.37}$$

where **h** is the vector of Hilsenhoff weightings. The Hilsenhoff metric can be compared to a *random weight metric*, in which the weights for each taxon are randomly chosen integers from 1 to 10.

Metric size

The size of a metric is the number of taxa included in the metric. All three multimetric indexes include the total taxa richness metric (Table 4-1, Table 4-2), which incorporates all taxa found in the samples. The size of the total taxa richness metric is the number of taxa that might be found in the region for which the index is calibrated. For a finite dataset, the size of (number of taxa in) the total taxa richness metric is the number of taxa in the dataset. No metric can be larger than total taxa richness, because no metric can include species that are not included in total taxa richness. For indexes that might include exotic vs. native as a way of grouping taxa this is not true, but for the benthic macroinvertebrate taxa used in this study there is no way of telling which taxa, if any, are exotic.

Equation 4.27 shows that if one assumes more taxa are likely to be found at sites with higher rather than lower biological condition, then the expected value for a metric score will increase with the size of the metric. Larger metrics will have larger responses to human influence.

Since the size of a metric affects the degree of relationship to % impervious area, size must be taken into account when assessing the strength of a candidate metric. To properly

calculate the strength (p) of some statistic (w) it must be compared with a pool of w's computed from random metrics of the same size as the candidate metric.

4.3 – Methods

Data

Thirty-four sites were selected from three years of PSLS data (1994, 1995, and 1997). These 34 sites were randomly divided into two data sets, set A and set B (Table 4-6). Known outliers were excluded, for instance lower Swamp creek has a riparian buffer zone and Coal creek has an old mine in its headwaters.

Candidate metrics

Forty-one candidate metrics were identified as possible to compute given the biological data available (Table 4-7). The size of each candidate was computed in each of the two datasets and for both datasets together. For instance, for Ephemeroptera taxa richness there were 16 Ephemeroptera taxa in set 1, 20 in set 2, and 21 taxa in both sets. There were 31 unique sizes ranging from 1 to 59.

Random metric scores

Seventy-six taxa were found at the sites in set 1, and 91 taxa at the sites in set 2, 102 taxa in both sets together. For each size of candidate metric, 1000 random metrics were generated from each set of taxa (see example in Table 4-4). For example, Ephemeroptera taxa richness was size 16 in set 1, 20 in set 2, and 21 taking both sets together. One thousand random metrics were generated by sampling 16 at a time from the taxa in set 1, 1000 by sampling 20 at a time from the taxa set 2, and 1000 by sampling 21 at a time from the union of both sets of taxa. One thousand is the recommended minimum number of simulations to accurately assign percentiles (Mathsoft 1999).

For a given random metric **m** of size n, scores were calculated by summing the appropriate transformation matrix (**W**, **Y**, or **Z**), over the n rows (taxa) included in **m** to produce a vector of scores for the sites (Figures 4-7 to 4-9).

Random metric scores were calculated as both sums and percentages, to provide populations for comparison of both sorts of biological metrics. Random ratio metrics, with appropriate

sized for metrics in the numerator and denominator, were generated separately for comparison with the biology-based ratio metrics in the RBP.

Random Hilsenhoff-style metrics

Hilsenhoff weights were available for only 82 of the 113 taxa in the PSLS database, as it was designed for invertebrate taxa found in Wisconsin streams. The 31 taxa without Hilsenhoff weights were ignored. The 31 taxa without Hilsenhoff scores were relatively rare taxa in the dataset; none were present at more than 15 sites in all three years of data, and no more than five were seen in any replicate. Seventy percent of the times these taxa were seen, only a single organism was found in the replicate. As the Hilsenhoff metric is a weighted by abundance average, their absence has a minimal effect on the final score.

Random Hilsenhoff-style metrics were generated by assigning random integers from 1 to 10 as weights for the remaining 82 taxa. Scores were calculated, using these weights, as per equation 4.36. One thousand sets of randomly weighted metrics were generated for comparison with the true Hilsenhoff metric.

Fit statistics

Fit statistics to measure the degree of relationship between metric score and % impervious area were calculated for each random metric in three basic shapes:

1) A non-parametric Wilcoxon test was used to decide if the metric could distinguish between the four sites with the lowest % impervious area in the watershed, and the four sites with highest % impervious area. The Wilcoxon test statistic, W (not the generic fit statistic, w, mentioned above) and its associated p-value were recorded for each random metric to create a pool for evaluation of the biology-based metrics (Figure 4-3).

2) A linear regression of score against % impervious area was run for each random metric. The slope, it's p-value, x and y-intercepts, and sum of squared errors were recorded for each.

3) A non-linear least-squares fit to a bent-line form was also run. To avoid singularities, the meeting angle between lines was constrained to be less than 170°; if the angle grew larger than 170° the result of the bent-line regression was replaced with a straight-line regression.

The slopes of the lines, sum of squared residuals, and the meeting angle of the two lines were recorded for each random metric.

Those few of the candidate biological metrics that scored highly to fit the broken-line pattern scored even higher for the high-low fit (Wilcoxon test), so the random metrics were not fit to a broken line.

4.3.1 Effect of sampling universe for random metrics

To assess the effect of the size of the sampling universe, random metrics generated for set A were used to calculate scores for sites in set B, and vice-versa. In addition, random metrics generated for the combination of sets A and B (sampling from taxa present at both sites) were used to generate scores for sites in both sets. A decrease in metric performance when comparing candidate metrics to random metrics generated from another set of sites would imply that the taxa present at the original set of sites were not representative, and that a larger pool should be used for comparison.

Populations of fit statistics for these random metric scores were computed for each of these across-metric sets.

4.3.2 Candidate metric strengths

The candidate metrics were each fit to the same patterns with a Wilcoxon, linear, and non-linear regression. Candidate metric strengths were calculated by finding the fraction of random metric fit statistics more extreme (meaning the upper tail, lower tail, or both tails as appropriate) than candidate metric fit statistics (Figure 4-9).

Both tails were used for the Wilcoxon test statistic, but the Wilcoxon p-value was calculated for a two-tailed alternate hypothesis, so the lower tail (small p-values) was used as "extreme." For the linear regression fit statistic, the slope, x, and y-intercepts were treated as two-tailed values and the F-test p-value was a extreme in the lower tail. In the bent-line regression the lower tail was used for the left slope, meeting angle and squared residuals, and the upper tail for the right slope.

Grades were calculated for each candidate metric. Of the ten fit statistics calculated for each candidate metric, the number in the most extreme 10% of the population of fit statistics for

randomly generated metrics were counted, and assigned as a grade, on a scale from 0 to 10, for each metric. The mean of the grades calculated for sets A and B, compared to the random metrics generated by drawing from the taxa in both sets was used as an overall grade.

4.3.3 Metric strengths by index

Grades were calculated for each of the multimetric indexes by calculating the mean grade of the component metrics for each index.

4.4 – Results

4.4.1 Effect of sampling universe for random metrics

The pattern of extreme vs. ordinary metric strengths is consistent across the two sets of sites and the three sets of random metrics (Table 4-10). Metrics that had extreme fit statistics in set A when compared to random metrics drawn from the taxa in set A also had extreme statistics when compared to the random metrics drawn from set B and from both sets. These metrics exhibited the same pattern when calculated for the set B and comparing against all three sets of random metrics.

4.4.2 Candidate metric strengths

Some candidate metrics had strengths that were consistently high, others had consistently low strengths. The clinger taxa richness metric consistently scored higher than random metrics with just as many taxa (Table 4-8 and Table 4-9). Other metrics, like Diptera richness do not seem to be extraordinary when compared to randomly generated metrics. Long-lived and intolerant taxa richness are two metrics that seem ordinary when comparing the Wilcoxon and linear regression fit statistics, but do have more extreme values for the bent-line regression. As a summary, the number of extreme fit statistics was counted to produce an overall grade for each metric (Table 4-11).

It is notable that the ratio metrics and the Hilsenhoff metric did not have any extraordinary fit statistics, in either set of sites, compared to any of the three sets of random taxa.

4.4.3 Metric strengths by index

The three multimetric indexes can be compared by calculating the fraction of extreme fit statistics of their component metrics (Table 4-12). When compared in this fashion both the B-IBI and the Oregon RBP score higher than the original RBP. The average fraction of extreme fit statistics for both RBP metrics is not as high as the mean of the candidate metrics not included in any metric. There is no inferential statistical basis for proclaiming any difference between indexes, they can only be ranked according to the mean grade of their component metrics: B-IBI > unused metrics > OR-RBP > RBP (Figure 4-10).

When ratio metrics are compared as a group to richness and percentage metrics (Table 4-12) they are seen to have a much smaller fraction of extreme fit statistics. Richness and percentage metrics, taken as a group, seem to have approximately the same fractions of extreme fit statistics, with a very slight edge going to richness metrics.

4.4.4 Metric size and strength

Finally, metric size seems to have an effect on the fraction of interesting fit statistics. Biologically defined candidate metrics that encompass ten or more taxa in the combined set of all taxa from both sets of sites have more extreme fit statistics than smaller metrics. This distinction arises despite the fact that the metrics are compared to random metrics with the same number of taxa.

4.5 – Discussion

Candidate metrics

If the null hypothesis of equivalent taxa were true, randomly generated metrics would be expected to do as well as biologically defined metrics in generating scores that correlate with the biological condition of a site. The fit statistics (w's) calculated for biology-based metrics would be randomly scattered in the population of w's generated from random metrics, and their strengths (p's) uniformly distributed between 0 and 1.

Overall, 31.2% of the calculated strengths were less than 0.10. In a hypothesis testing context, the null hypothesis of equivalence between randomly generated and biologically defined metrics would be rejected at an α of 0.10 or 0.05. It appears that, in general,

considering biological information when choosing a metric produces metrics with a stronger response to site biological condition.

Particular metrics

Certain metrics did especially well. Clinger taxa richness had strong fit strengths to all three of the patterns tested. Long-lived and intolerant taxa richness did well in distinguishing the best and worst sites (high-low) and the bent-line regression. EPT and Tolerant taxa richness distinguished themselves in the linear fit.

Other candidate metrics were not distinguishable from random metrics. The Hilsenhoff metric's poor performance may be because the particular weights used were calibrated for use in Wisconsin streams. The ratio metrics were also indistinguishable from random metrics. Other candidate biological metrics, such as Coleoptera richness and percent filterers, were also no better than random metrics at providing a signal of biological condition.

The fraction of significant metric strengths given in Table 4-12 indicate that all three multimetric indexes, taken as a whole, do better than a collection of randomly chosen metrics. The original RBP metrics, while better than random metrics, do not do as well as the rejected, biologically defined candidate metrics. The poor performance of the original RBP may be due to its inclusion of ratio metrics and Hilsenhoff metric, which did not distinguish themselves from random metrics.

Sampling universe for random metrics

Metric performance was consistent across the sets of sites and the three sets of random metrics. This consistency indicates that the taxa present in each set of sites were representative of the taxa in the dataset as a whole.

More than 30% of the candidate metric fit statistics evaluated were in the extreme 10% of their population. Purely random metrics would be expected to have only 10% of fit statistics in the most extreme 10%. This surplus of extreme fit statistics implies that biologically defined groups, even those not included in multimetric indexes, do better than random chance. It may be that any biologically defined group responds more similarly to some aspect of human influence. Using equation 4.25 as an example, a biologically defined group might

be more likely to have similar □'s and so produce a less-noisy linear response to degradation than a randomly selected group.

The candidate metrics are also not orthogonal, there is redundancy of information among them, so it is possible that some of the surplus in extreme values is the result of including the same information several times. EPT taxa richness, for example, includes the signal present in the Ephemeroptera, Plecoptera, and Trichoptera taxa richness. The clinger and predators categories overlap with the both the taxonomic categories and each other.

Metric size and strength

Another, surprising pattern was the tendency for larger metrics (those including more taxa) to have more extreme fit statistics, even when "extreme" is defined by comparison to random metrics of the same size. Equation 4.25 predicts that larger metrics will be better in general but Figure 4-11 shows that larger metrics tend to have more extreme fit statistics.

Conclusions

Considering the natural history of taxa does produce metrics with a stronger relationship to % impervious area than that of random metrics. Moreover, it appears that most (60%) of the biologically defined metrics do better than random metrics.

Treated as groups, the metrics included in the B-IBI and both versions of the RBP do better at providing a signal than random metrics. However, the metrics in the both versions of the RBP do not do as well as the biologically defined candidate metrics not included in any multimetric index, nor even as well the candidate metrics overall. The B-IBI did best as a group, with 36.5% of its metrics' fit statistics being in the extreme 10%. The Oregon RBP scored at 27.4%, just below than the overall 31.2%. The original RBP scored at 16.7%.

The strengths calculated in this study should not be taken as an absolute measure of a biology-based metric's worth. One should not reject a metric simply because it's calculated p's are all larger than 0.05. Rather, these strengths should be considered when choosing metrics for an index, and if two candidate metrics seem equally useful, their performance compared to random metrics should be an additional, supplementary criterion for making a choice.

Recommendations

If total taxa richness provides a signal of biological condition, then, in general, a metric that includes many taxa will provide a stronger signal than a metric with few taxa. Therefore, the number of taxa included in a candidate metric should be another thing to consider when comparing it with other candidates for use as an indicator. Defining statistics to measure the degree of response, and then comparing these statistics to those of random metrics can provide a method of evaluating the relative worth of two differently sized metrics. There are certainly other considerations, such as those detailed in (Karr and Chu 1998), but index formulators should also be aware of how many taxa are included in each metric.

Larger metrics provide a better signal of biological condition. A candidate metric should include enough taxa that members are expected to be found at most undisturbed sites in the region in question. Coleoptera richness, with 8 taxa in the dataset used, did not provide a clear signal of biological condition relative to random metrics with 8 taxa. EPT richness, with 59 taxa provided a much stronger signal, even in comparison to random metrics with 59 taxa.

4.6 – Tables

Table 4-1. Metrics used in the original Rapid Bioassessment Protocol (RBP III)

Some metrics are expected to increase in value as human influence rises and biological condition drops; other metrics follow the opposite pattern. A selection of metrics, taken together, constitutes an index.

Metric	Response to increasing human influence
Taxa richness	113
EPT	65
Dominance (1)	□
% Shredder	13
Scrapers/Collectors	11/8
EPT/Chironomids	65/1
Loss Index	□
Hilsenhoff	□

Table 4-2. Metrics used in the Oregon DEQ RBP multimetric index

Some metrics are expected to increase in value as human influence rises and biological condition drops, other metrics follow the opposite pattern. A selection of metrics, taken together, constitutes an index.

Metric	Response to increasing human influence
Total taxa	decrease
EPT taxa	decrease
% Dominance (1 taxon)	increase
% Shredder	increase
% Scraper	decrease
% Filterer	increase
% EPT taxa	decrease
% Chironomidae	increase
Shannon diversity index	decrease
Hilsenhoff biotic index	increase

Table 4-3. Example of calculating the Ephemeroptera richness metric

An example of nine taxa sampled at five sites. The table elements are 1 if the taxon was found at the site, 0 if it was not. The first three taxa in **bold** are Ephemeroptera taxa. Summing the values in those rows produces the Ephemeroptera taxa richness score for each site; the number of Ephemeroptera taxa found at that site.

	Site 1	Site 2	Site 3	Site 4	Site 5
Ephemerella	**1**	**0**	**1**	**0**	**0**
Epeorus	**1**	**1**	**0**	**0**	**0**
Paraleptophlebia	**0**	**1**	**0**	**0**	**0**
Kathroperla	1	0	1	0	1
Suwallia	1	0	1	1	0
Sweltsa	1	0	0	0	0
Clostoeca	1	1	1	0	0
Ecclisomyia	1	1	1	0	0
Onocosmoeus	0	0	1	1	0
Total	2	2	1	0	0

Table 4-4. Example of calculating a random richness metric

An example of nine taxa sampled at five sites. The table elements are 1 if the taxon was found at the site, 0 if it was not. Three of those nine taxa are chosen at random (in **bold**). Summing the values in those rows produces the richness score for that random metric at each site; the number of that random subset of taxa found at that site.

	Site 1	Site 2	Site 3	Site 4	Site 5
Ephemerella	1	0	1	0	0
Epeorus	1	1	0	0	0
Paraleptophlebia	0	1	0	0	0
Kathroperla	**1**	**0**	**1**	**0**	**1**
Suwallia	1	0	1	1	0
Sweltsa	1	0	0	0	0
Clostoeca	**1**	**1**	**1**	**0**	**0**
Ecclisomyia	1	1	1	0	0
Onocosmoeus	**0**	**0**	**1**	**1**	**0**
Total	2	1	3	1	1

Table 4-5. Number of taxa involved in individual multimetric metrics

All three multimetric indexes include Taxa Richness, involving all 113 taxa. Only the Plafkin RBP index uses ratio metrics. Dominance selects taxa based upon abundance rather than inherent properties. The Diversity, Community Loss, and Hilsenhoff metrics involve all taxa and are not computed in the fashion of other metrics. Eph. = Ephemeroptera, Plec. = Plecoptera, Trich. = Trichoptera, Intol. = Intolerant

PSLS B-IBI metrics	Number of Taxa	RBP III metrics	Number of Taxa	Oregon DEQ RBP metrics	Number of Taxa
Taxa Richness	113	Taxa richness	113	Taxa richness	113
Eph. Richness	15	EPT	65	EPT	65
Plec. Richness	22	Dominance (1)	□	Dominance (1)	
Trich. Richness	28	% Shredder	13	% Shredder	13
Intol. Richness	16	Scrapers/Collectors	11/8	% Scraper	11
LL Richness	11			% Filter	8
% Tolerant	16	EPT/Chironomids	65/1	% EPT	65
Cling. Richness	43		□	% Chironomid	1
% Predator	36	Loss Index	□	Diversity	
Dominance (3)	□	Hilsenhoff	□	Hilsenhoff	

Table 4-6. Two sets of sites used to compare candidate metrics to random metrics

Thirty-Four sites from three years of data collection were randomly divided into two sets of seventeen. Both sets covered the range of % impervious area in the watershed above the sampling site.

Set A				Set B			
Site	Year	% impv. Area	Taxa richness	Site	Year	% impv. Area	Taxa richness
Carey	1994	1.0	35	Carey	1995	1.0	34
Big Anderson	1994	1.2	32	Big Beef	1994	3.1	29
Big Anderson	1995	1.2	32	Rock	1995	3.1	39
Rock	1997	3.1	42	Rock	1994	3.2	38
Covington	1994	3.9	37	Covington	1995	3.9	34
Little Bear	1995	4.3	38	Little Bear	1997	4.3	51
Little Bear X	1995	4.4	27	Big Bear	1994	6.6	36
Big Bear	1995	6.6	27	Big Bear	1997	6.6	42
Little Bear B	1995	7.3	30	Jenkins	1997	13.1	33
Jenkins	1995	13.1	26	Swamp B	1995	24.9	28
North B	1995	26.2	23	North X	1995	25.7	30
North C	1995	26.3	27	Swamp C	1995	32.6	28
Swamp 2	1997	26.3	32	Juanita	1994	44.4	19
Swamp D	1995	31.5	26	Kelsey	1994	47.3	12
Juanita	1995	44.4	18	DM	1995	49.1	6
Kelsey	1995	47.3	8	Thornton	1995	52.5	12
Thornton	1994	52.5	9	Thornton 1A	1997	52.5	10

Table 4-7. List of candidate metrics tested

Forty-one candidate metrics were compared against randomly generated metrics of the same size in two sets of streams. Random metrics were generated from the taxa present in each set of streams, and the combination of both streams.

	Number of taxa			
Candidate metric	Set A	Set B	Both Sets	Indexes
Ephemeroptera richness	16	20	21	IBI
Plecoptera richness	13	12	13	IBI
Trichoptera richness	14	24	25	IBI
Intolerant richness	7	11	11	IBI
Long-lived richness	12	14	15	IBI
Percent tolerant	10	12	13	IBI
Clinger richness	36	38	40	IBI
Percent Predators	23	32	33	IBI
EPT richness	43	56	59	RBP, OR-RBP
Coleoptera richness	7	8	8	□
Diptera richness	13	17	19	□
Tipulidae richness	4	7	8	□
Non-insect richness	10	12	12	□
Heptageniidae richness	4	4	4	□
Tolerant richness	10	12	13	□
Sediment tolerant richness	5	6	6	□
Sediment intolerant richness	3	5	5	□
Predator richness	23	32	33	□
Scraper richness	6	12	12	□
Gatherer richness	22	22	23	□
Filterer richness	5	7	7	□
Omnivore richness	6	5	7	□
Percent Trichoptera	14	12	25	□
Percent Ephemeroptera	16	20	21	□
Percent Plecoptera	13	12	13	□
Percent clinger	36	38	40	□
Percent tolerant	10	12	13	□
Percent sediment tolerant	5	6	6	□
Percent predator	23	32	33	□
Percent shredder	6	12	12	RBP, OR-RBP
Percent scraper	8	12	12	OR-RBP
Percent gatherer	22	22	23	□
Percent filterer	5	7	7	OR-RBP
Percent non-insect	10	12	12	□
Percent Heptageniidae	4	4	4	□
Percent omnivore	6	5	7	□
Percent EPT	43	56	59	OR-RBP
Scraper/Filter abundance	8/5	12/7	12/7	RBP
EPT/Chironomid abundance	43/1	56/1	59/1	RBP
Percent Chironomidae	1	1	1	OR-RBP
Hilsenhoff score	-	-	-	RBP, OR-RBP

Table 4-8. Sample of Wilcoxon and linear fit statistics

The candidate metrics in Table 4-7 were fit to a high-low response, measured with a Wilcoxon test, and a linear response, measured with a linear regression. The W statistic, it's p-value, slope, F-test p-value, and Y and X intercepts were recorded for each. These statistics were compared to a population of statistics generated from 1000 random metrics of the same size. For each metric, the upper row contains the actual fit statistics observed, and the lower, shaded row contains the fraction of statistics from random metrics that were larger than the observed statistic. The metric size is the number of taxa included in the metric.

		Metric strength					
Candidate Metric	Size	W	W p-value	Slope	Slope p-value	Y intercept	X intercept
Ephemeroptera richness	16	2.19	0.03	-0.09	0.00	6.23	67.95
		0.36	0.90	0.89	**0.98**	0.26	0.83
Plecoptera richness	13	1.89	0.06	-0.08	0.00	4.91	58.48
		0.64	0.38	0.94	0.91	0.32	0.93
Trichoptera richness	14	2.19	0.03	-0.07	0.00	4.66	64.09
		0.30	0.89	0.80	0.77	0.51	0.86
Intolerant richness	7	2.10	0.04	-0.03	0.02	1.05	40.39
		0.32	0.70	0.51	0.50	0.93	0.95
Long-lived richness	12	25.00	0.06	-0.06	0.00	3.93	64.08
		0.10	0.48	0.78	0.70	0.53	0.85
Percent tolerant	10	13.00	0.20	0.00	0.01	0.16	-49.34
		0.75	0.71	0.07	0.87	0.34	0.68
Clinger richness	36	2.19	0.03	-0.23	0.00	15.41	67.55
		0.71	**0.98**	**1.00**	**0.95**	**0.02**	**0.97**
Percent predators	23	26.00	0.03	0.00	0.00	0.08	45.77
		0.00	0.95	0.75	**0.96**	**0.99**	0.52
EPT richness	43	2.18	0.03	-0.25	0.00	15.81	63.62
		0.79	0.17	**1.00**	**1.00**	0.21	**1.00**
Coleoptera richness	7	0.44	0.66	-0.03	0.03	2.75	85.30
		0.90	0.06	0.66	0.44	0.30	0.47
Diptera richness	13	2.18	0.03	-0.04	0.06	4.22	104.75
		0.34	0.56	0.34	0.17	0.55	0.29
Tipulidae richness	4	2.20	0.03	-0.01	0.21	0.95	86.64
		0.13	0.86	0.38	0.29	0.67	0.32
Non-Insect richness	10	-1.74	0.08	0.03	0.01	2.56	-83.24
		1.00	0.39	**0.00**	0.42	0.75	**0.97**
Heptageniidae richness	4	2.23	0.03	-0.04	0.00	2.26	53.60
		0.07	0.88	**0.98**	**0.98**	0.13	0.72

Table 4-9. Sample of bent-line regression strengths

The candidate metrics in Table 4-7 were fit to a bent line response. The slopes, sum of squared residuals, and meeting angle of the two line segments were recorded for each metric. These statistics were compared to a population of statistics generated from 1000 random metrics of the same size. For each metric, the strength, the fraction of statistics from random metrics that were larger than the observed statistic is reported. The metric size is the number of taxa included in the metric.

		Metric strength			
Candidate Metric	Size	Left Slope	Right Slope	SSR	Angle
Ephemeroptera richness	16	**0.02**	**0.03**	0.53	0.63
Plecoptera richness	13	0.08	0.05	0.49	0.76
Trichoptera richness	14	0.14	0.43	0.59	0.53
Intolerant richness	7	**0.99**	0.06	0.24	**0.01**
Long-lived richness	12	0.92	0.14	0.69	0.06
Percent tolerant	10	0.42	**0.03**	0.28	0.50
Clinger richness	36	0.85	0.40	0.51	0.74
Percent predators	23	0.42	0.82	0.33	**1.00**
EPT richness	43	**0.04**	0.06	0.06	0.61
Coleoptera richness	7	0.13	0.46	0.93	0.15
Diptera richness	13	**0.05**	0.91	0.42	0.77
Tipulidae richness	4	0.22	**0.99**	0.69	0.90
Non-Insect richness	10	0.24	0.10	0.06	0.67
Heptageniidae richness	4	0.58	0.48	0.78	0.77
Tolerant richness	10	0.92	0.51	0.37	0.15
Sediment tolerant richness	5	0.74	0.83	0.41	0.15
Sediment intolerant richness	3	0.87	0.27	0.77	0.60
Predator richness	23	0.36	0.27	0.57	0.33
Scraper richness	6	0.38	0.63	0.93	0.77
Gatherer richness	22	0.89	0.34	0.77	0.06
Filterer richness	5	0.29	0.31	0.65	0.17
Omnivore richness	6	0.15	0.67	0.46	0.78
Percent Trichoptera	14	0.88	0.59	0.85	0.08
Percent Ephemeroptera	16	0.20	0.24	0.87	0.48
Percent Plecoptera	13	0.43	0.83	0.65	0.26
Percent clinger	36	0.38	0.33	0.54	0.17
Percent tolerant	10	0.82	0.82	0.65	0.17
Percent sediment tolerant	5	0.44	0.11	0.25	**0.97**

Table 4-10. Number of metric fit statistics in the extreme 10%

The candidate metrics (Table 4-7) were fit to three different patterns of response, with ten different fit statistics calculated for each. Each set was compared to three different random metrics, set A was fit to random metrics drawing from the taxa present in set A, then random metrics from set B, and then random metrics drawing from both sets.

This table shows the number of fit statistics that were in the most extreme 10% when compared to a population fit statistics generated from 1000 random metrics. "Most extreme" means upper tail, lower tail, or two-tailed as appropriate. The number of extreme fit statistics is consistent within the metrics regardless of site and the sampling universe for the random metrics. This consistency indicates that the taxa present in each set of sites were representative of the taxa in the dataset as a whole.

	Number of extreme fit statistics					
	A in A	A in B	A in both	B in A	B in B	B in both
Ephemeroptera richness	2	4	3	2	1	2
Plecoptera richness	0	5	3	0	5	5
Trichoptera richness	0	1	2	0	0	2
Intolerant richness	0	2	5	0	4	5
Long-lived richness	0	0	3	0	0	1
Percent tolerant	0	2	0	0	3	5
Clinger richness	9	8	5	7	7	6
Percent predators	5	7	6	5	1	4
EPT richness	5	7	7	5	5	5
Coleoptera richness	0	0	0	0	0	1
Diptera richness	0	0	1	0	0	2
Tipulidae richness	0	0	1	0	5	5
Non-Insect richness	5	5	7	5	7	5
Heptageniidae richness	4	0	6	4	4	5
Tolerant richness	5	5	6	5	6	8
Sediment tolerant richness	0	0	3	0	0	1
Sediment intolerant richness	0	0	4	0	1	4
Predator richness	3	4	5	3	5	3
Scraper richness	5	4	4	5	0	2
Gatherer richness	0	2	3	0	2	4
Filterer richness	0	0	0	0	0	0
Omnivore richness	0	0	0	0	0	0
Percent Trichoptera	0	2	2	0	0	3

Table 4-10. Number of metric fit statistics in the extreme 10% (continued)

	Number of extreme fit statistics					
	A in A	A in B	A in both	B in A	B in B	B in both
Percent Ephemeroptera	0	0	1	0	0	0
Percent Plecoptera	0	2	2	0	2	4
Percent clinger	5	5	5	5	3	5
Percent tolerant	0	2	0	0	4	5
Percent sediment tolerant	0	0	1	0	0	2
Percent sediment intolerant	0	0	0	0	4	2
Percent predators	5	7	6	5	2	4
Percent shredders	0	0	1	0	0	4
Percent scrapers	3	3	7	3	1	5
Percent gatherers	0	2	0	0	3	5
Percent filterers	1	0	0	1	0	1
Percent non-insect	0	0	0	0	0	2
Percent Heptageniidae	3	0	7	3	5	5
Percent Omnivore	2	0	1	2	0	2
Percent EPT	6	5	6	7	0	1
Scraper/Filter abundance	0	0	0	0	0	0
EPT/Chironomid abundance	0	0	0	0	0	0
Percent Chironomidae	0	0	0	0	0	0
Hilsenhoff score	0	0	0	0	0	0

Table 4-11. Overall grades of candidate metrics

Each candidate metric was fit to three response shapes, and a total of ten fit statistics were calculated for each. Each fit statistic was compared to 1000 fit statistics calculated for random metrics, and the number (out of 10) of fit statistics in the extreme 10% of the population of random fit statistics was recorded as the grade for the candidate metric. This process was done twice (for sets A and B) and the mean grade reported in the table below.

Candidate metric	Grade	Candidate metric	Grade
Ephemeroptera richness	2.5	Omnivore richness	0
Plecoptera richness	4	Percent Trichoptera	2.5
Trichoptera richness	2	Percent Ephemeroptera	0.5
Intolerant richness	5	Percent Plecoptera	3
Long-lived richness	2	Percent clinger	5
Percent tolerant	2.5	Percent tolerant	2.5
Clinger richness	5.5	Percent sediment tolerant	1.5
Percent predators	5	Percent sediment intolerant	1
EPT richness	6	Percent predators	5
Coleoptera richness	0.5	Percent shredders	2.5
Diptera richness	1.5	Percent scrapers	6
Tipulidae richness	3	Percent gatherers	2.5
Non-Insect richness	6	Percent filterers	0.5
Heptageniidae richness	5.5	Percent non-insect	1
Tolerant richness	7	Percent Heptageniidae	6
Sediment tolerant richness	2	Percent Omnivore	1.5
Sediment intolerant richness	4	Percent EPT	3.5
Predator richness	4	Scraper/Filter abundance	0
Scraper richness	3	EPT/Chironomid abundance	0
Gatherer richness	3.5	Percent Chironomidae	0
Filterer richness	0	Hilsenhoff score	0

Table 4-12. Fraction of metrics in the most extreme 10%

The candidate metrics in Table 4-7 were fit to three different patterns of response, with ten different fit statistics calculated for each. This table shows the mean grade, for groups of metrics. The grade is the number out of the 10 fit statistics for each metric that were in the most extreme 10% of the population of random metric fit statistics. "Most extreme" means upper tail, lower tail, or two-tailed as appropriate.

	Mean grade		
	Set A in both	Set B in both	Overall mean
B-IBI metrics	3.5	3.8	3.7
RBP metrics	1.7	1.7	1.7
OR-RBP metrics	3.1	2.4	2.7
other metrics	2.9	3.3	3.1
Richness metrics	3.1	3.3	3.2
Percentage metrics	3.5	2.9	3.2
Ratio metrics	2.7	0.0	1.3
size < 10	0.0	2.4	1.2
size ≥ 10	3.0	5.9	4.5

4.7 – Figures

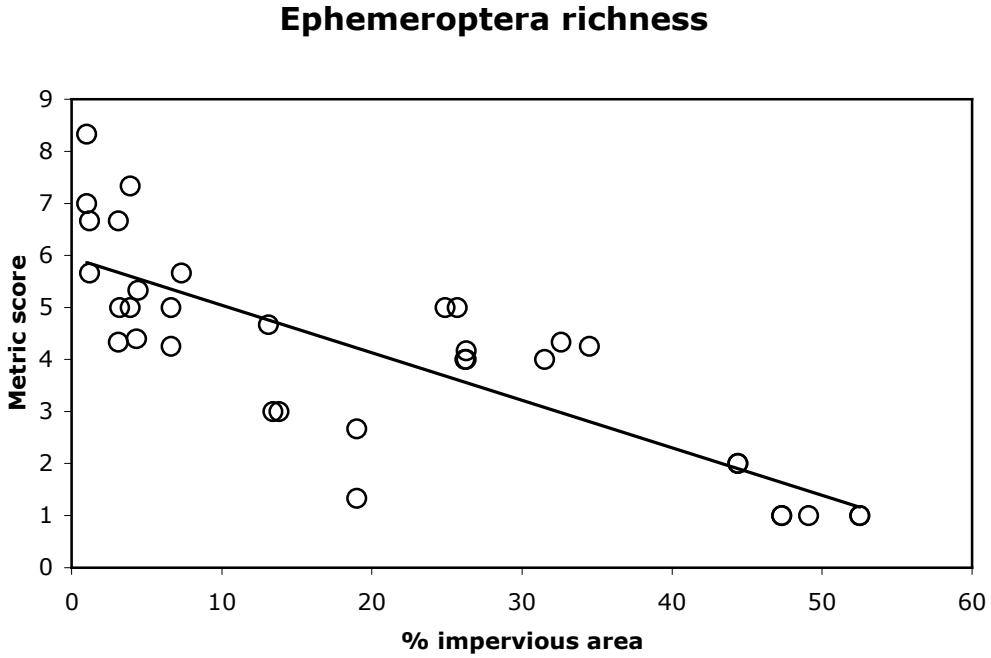

Figure 4-1. Linear gradient against biological condition

The Ephemeroptera richness metric has close to a straight-line relationship with % impervious area as a proxy for a site's biological condition.

Figure 4-2. Bent-line metric response to biological condition

The Intolerant richness metric features a bent-line response to % impervious area (a proxy for biological condition). There are few or no intolerant taxa in watersheds with more than 5% impervious area, and more at sites with very small % impervious area.

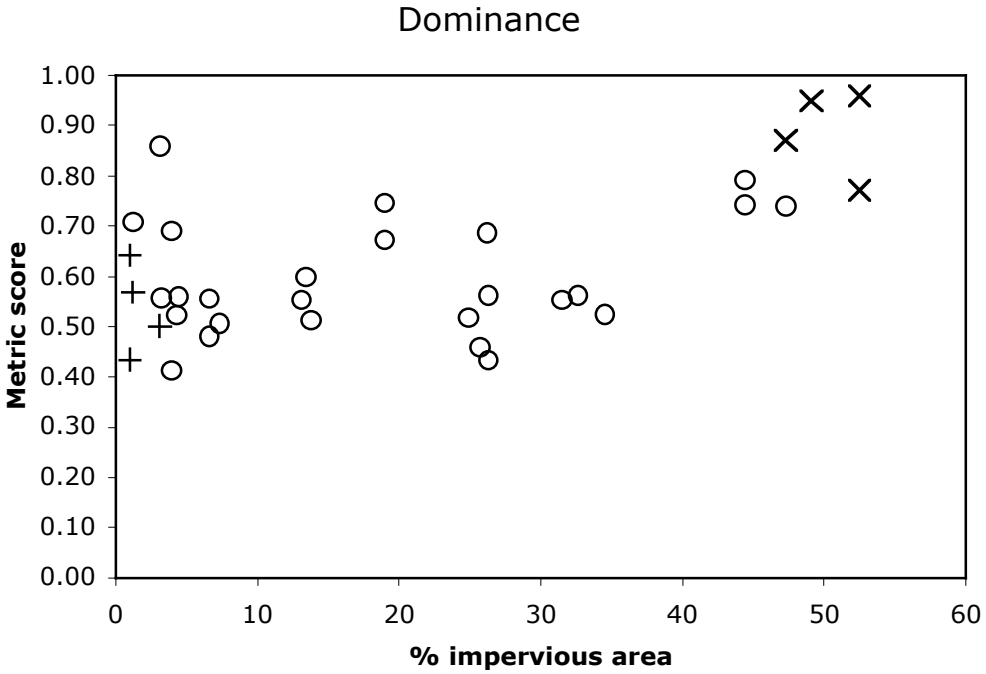

Figure 4-3. Metrics should distinguish between best and worst sites
The dominance (3) metric distinguishes between the best (+) and worst (□) sites.

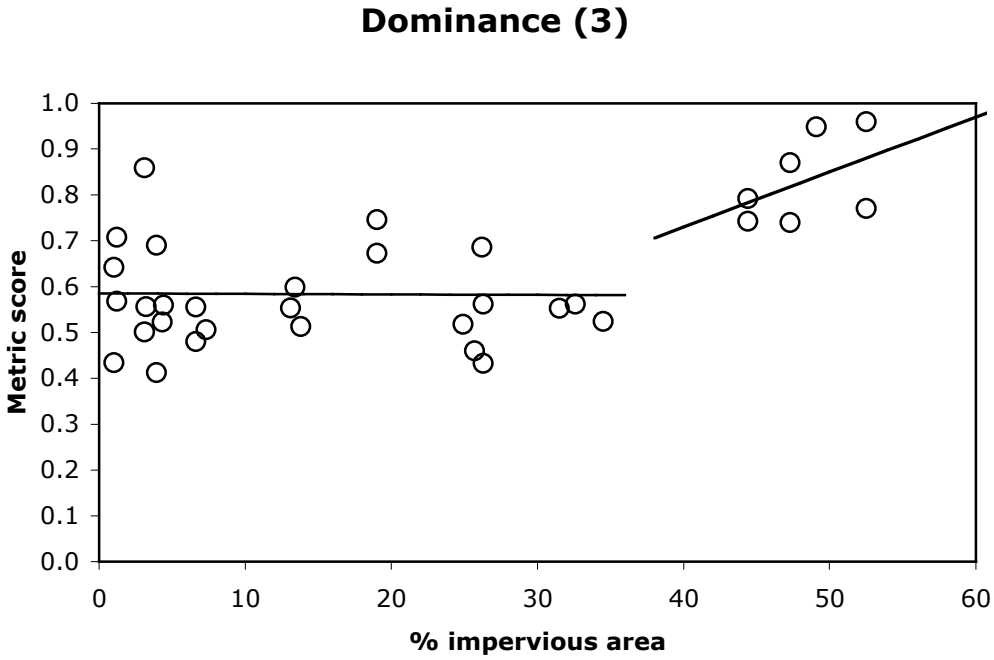

Figure 4-4. Broken-line metric response to biological condition

Here, as an example, the Dominance (3) metric is modeled as a broken line, with little or no trend for sites with low % impervious areas (the proxy for biological condition) which shifts to an increasing trend for more degraded sites.

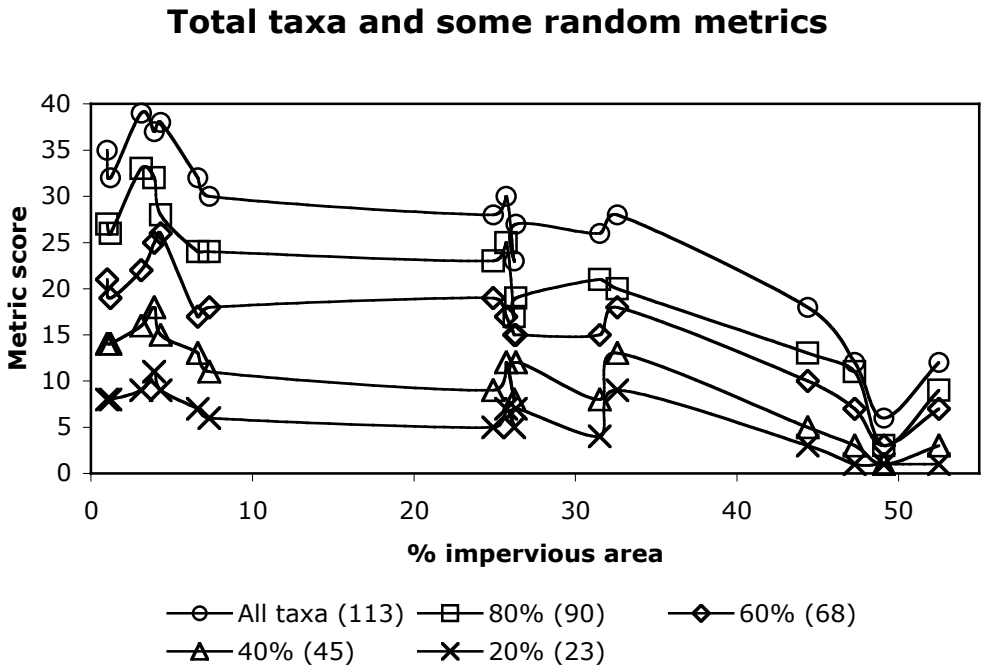

Figure 4-5. As metric size increases correlation with % impervious area increases
As the size of a random metric increases, the scores approach those of Total taxa richness. Random metrics including 20, 40, 60, and 80% of taxa in the dataset closely follow Total taxa richness, implying that the size of a metric must be considered when evaluating a metric.

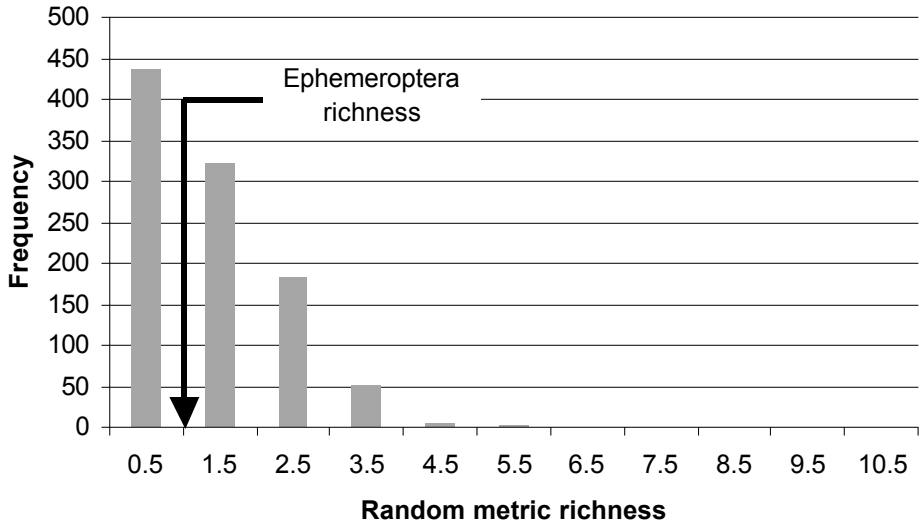

Figure 4-6. Histogram of random metric scores for Thornton Creek

1000 random metrics of size 16 were generated by randomly choosing 16 of the taxa represented in set 1 (Table 4-6). Taxa richness scores were computed for Thornton Creek (in 1994) for all of these random metrics. There are 16 Ephemeroptera taxa in set 1, so the observed Ephemeroptera taxa richness is included as a comparison.

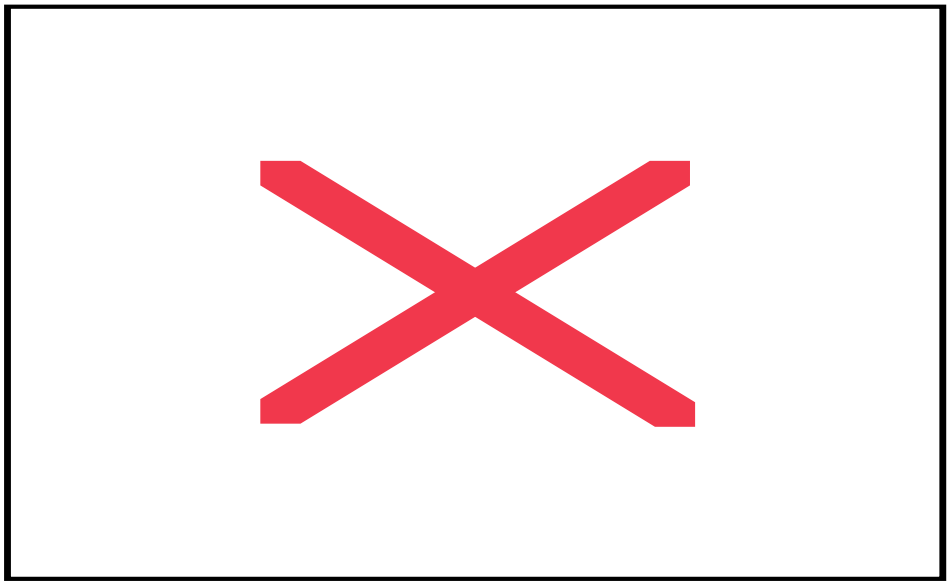

Figure 4-7. Histogram of random metric scores for North Creek (B)

1000 random metrics of size 16 were generated by randomly choosing 16 of the taxa represented in set 1 (Table 4-6). Taxa richness scores were computed for North Creek Site B (in 1995) for all of these random metrics. There are 16 Ephemeroptera taxa in set 1, so the observed Ephemeroptera taxa richness is included as a comparison.

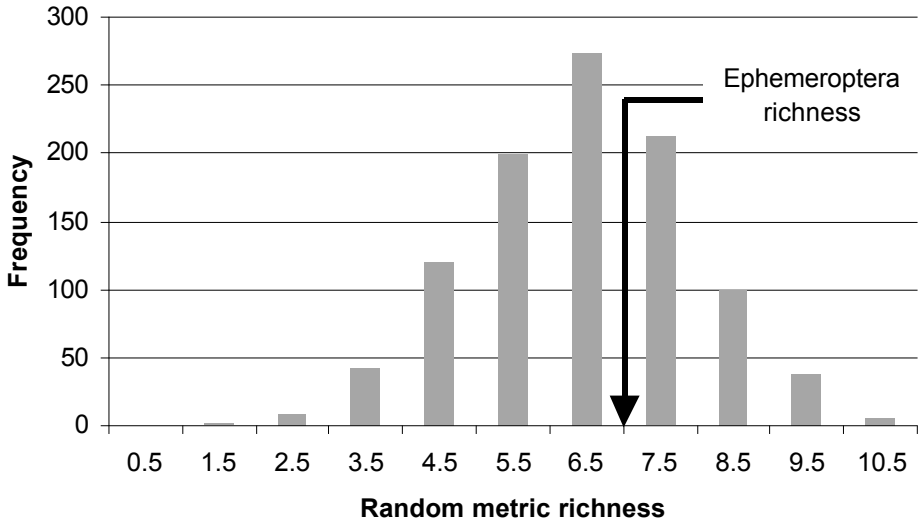

Figure 4-8. Histogram of random metric scores for Rock Creek

1000 random metrics of size 16 were generated by randomly choosing 16 of the taxa represented in set 1 (Table 4-6). Taxa richness scores were computed for Rock Creek (in 1997) for all of these random metrics. There are 16 Ephemeroptera taxa in set 1, so the observed Ephemeroptera taxa richness is included as a comparison.

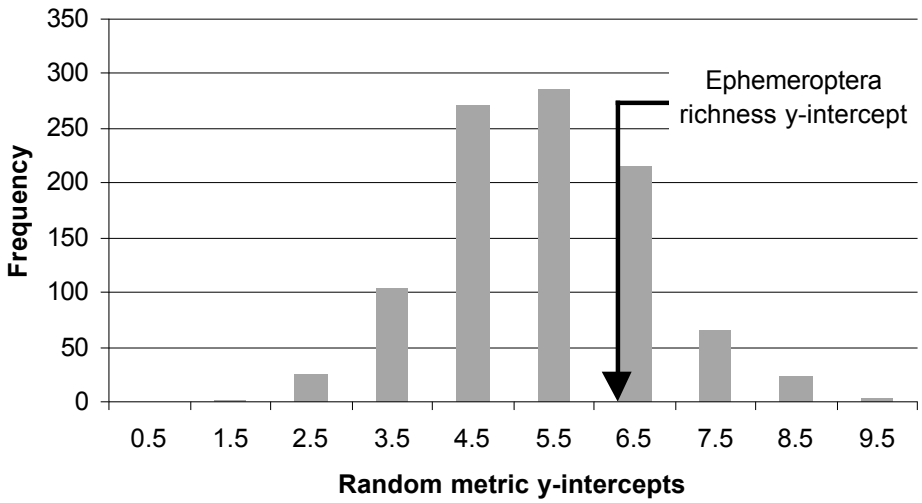

Figure 4-9. Histogram of random metric y-intercepts for set 1

1000 random metrics of size 16 were generated by randomly choosing 16 of the taxa represented in set 1 (Table 4-6). Linear regressions of taxa richness as a function of % impervious area were run for each random metric, and the y-intercepts were recorded. A regression of ephemeroptera taxa richness vs. % impervious area yielded a y-intercept of 6.23. 26% of the random metric y-intercepts were larger than 6.23, so the Ephemeroptera taxa richness y-intercept fit statistic has a strength of 0.26.

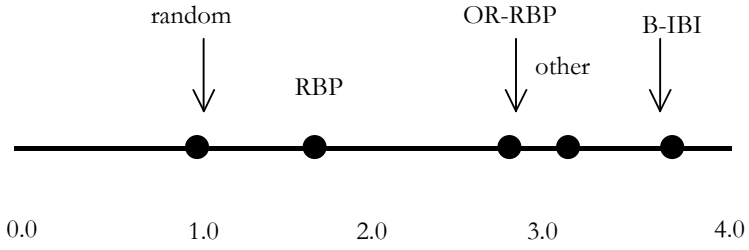

Figure 4-10. Mean grades of multimetric indexes

Grades were calculated for the three multimetric indexes by computing the mean grades of their component metrics. The B-IBI had the highest grade, followed by the mean grade of candidate metrics that were not included in any metric. The Oregon modification of the RBP and the original RBP were next. All three multimetric indexes received a higher grade than would be expected for random metrics.

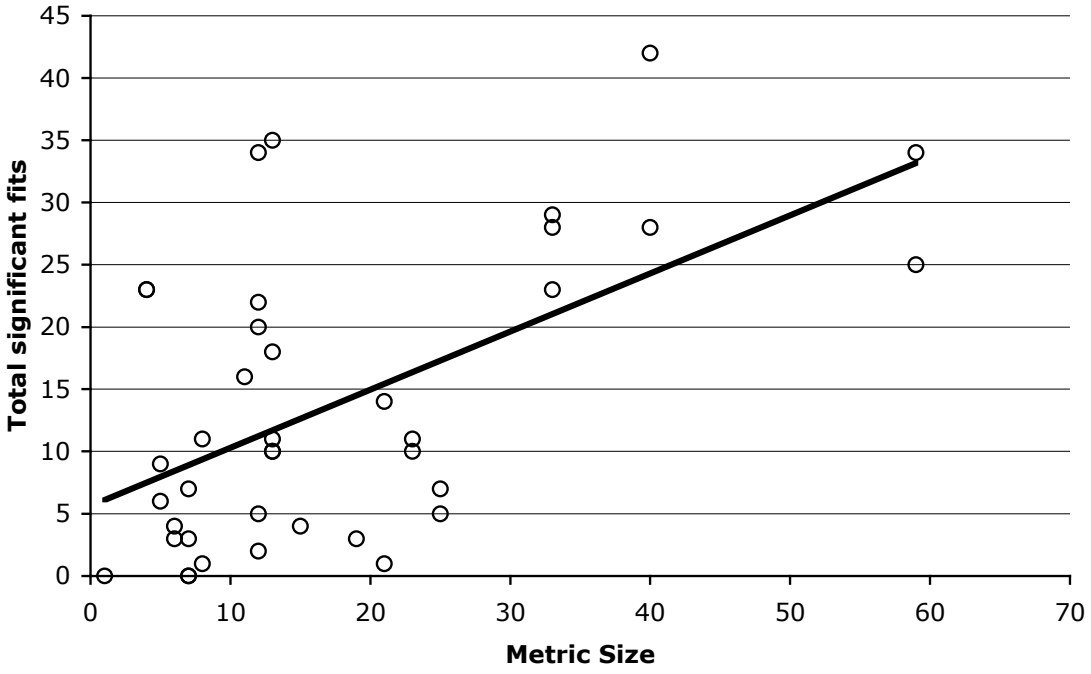

Figure 4-11. Larger metrics have more significant fit statistics

Candidate metrics were fit to three different shapes with ten different statistics. Large metrics – metrics that encompass a larger number of taxa – have more extreme fit statistics when compared to the fit statistics of random metrics of the same size. The line indicates the linear regression fit ($p < 0.001$, $F_{1,37}$).

5 — Conclusions and recommendations

5.1 – Framework

Multivariate techniques attempt to simplify the interpretation of complex systems by reducing the number of variables to be considered. They do this by identifying new, synthetic variables, called metrics, which are linear combinations of the original variables. These metrics simplify interpretation of the original system by isolating the specific behavior of interest and reducing the variation (noise) that is not related to the behavior of interest.

The Index of Biological Integrity method of proposing candidate metrics based on knowledge of the system's biology and examining the metric response to a gradient of interest (Karr and Chu 1998) has been used in studies in the Northwestern United States and Japan to identify the metrics in the Benthic Index of Biological Integrity (B-IBI).

Mathematics-based multivariate techniques such as correspondence analysis, canonical correlation, and multiple regression identify metrics whose response optimizes some function of the variables in the original, complex system.

Metrics are used to produce scores. A metric score for a set of variables measured at a site is computed as a sum of the products of each variable and its corresponding metric coefficient. Geometrically the metric score is the projection of a site in the higher-dimensioned space represented by the original variables onto the line corresponding to the metric.

I wish to distinguish between the metrics themselves and the scores they produce when applied to a particular dataset. The metrics can be defined as a vector of coefficients, with one coefficient for each variable in the original, complex system. The scores are calculated for a specific set of observations of the variables, from a sample of a single site at a specific time.

For the purpose of inferring the biological condition of a site from a sample of the site's biota, we want metrics whose scores provide a signal of site biological condition, and we

identify such metrics by examining their response to a measured proxy for biological condition.

5.2 – Performance of mathematics-based multivariate metrics

Conclusions

The metrics generated by correspondence analysis produce scores that are able to discriminate among sites along a gradient of human influence in all three datasets (1994, 1995, and 1997). The B-IBI metric scores were also able to distinguish among those sites. The B-IBI and correspondence analysis metrics themselves were not alike. The correspondence analysis metrics derived in each dataset were also unlike each other; in each year different metrics were generated to best accomplish the same goal of discriminating sites along a gradient. The lack of consistency across time indicates that the connection between the sampled biota and the major gradients were different for each dataset.

Adding biological information to the correspondence analysis procedure by first aggregating taxa based on their biological properties did not increase their similarity to the B-IBI metrics. The addition did not even increase the similarity of correspondence analysis metrics derived for the datasets collected in different years.

Both multiple regression and canonical correlation produced metrics whose scores were maximally correlated with measurements of human influence. The correlation of scores calculated for sites across years was higher than the correlation of B-IBI metric scores across years. The actual metrics derived by these techniques were quite different, with little or no correlation of the coefficients across years. When the metrics derived from the 1994 dataset – by definition of the technique designed to maximize correlation between scores and human influence in 1994 – were used to produce scores for the variables measured in the 1995 and 1997 datasets, those scores were unrelated to human influence in 1995 and 1997. Again, this lack of temporal consistency indicates that the connection identified by mathematics-based multivariate techniques between the biota and human influence (and, by proxy, biological condition) was different for each dataset.

Recommendations

Before a biological interpretation is assigned to a metric generated by a mathematics-based multivariate technique, that metric should be tested for consistency in behavior for datasets collected across space and time. If the metric of interest is not consistent, that difference implies that different biological processes are being identified in each dataset. If that is the case, then any inferences made from one dataset cannot be generalized.

Consistency of behavior in datasets collected in different places and at different times – subject to being similar enough that their underlying biological processes are the same – is a prerequisite to assigning a biological interpretation to a metric. The absence of such consistency implies that different processes are operating in each dataset, which would in turn imply different governing biological properties.

5.3 – Random metrics as a baseline

Conclusions

Some biologically defined metrics provide a better signal of biological integrity than others do. Random metrics – generated without consideration of biology – provide a baseline for measuring the effect of the particular aspect of biology used to define a metric.

Multimetric indexes use different criteria for choosing which metrics to include. Using a comparison to random metrics to determine the efficacy of their component metrics, an average grade can be calculated for a multimetric index. The three indexes examined in this study can be ranked, the Benthic Index of Biological Integrity > Oregon Rapid Bioassessment Protocol > Rapid Bioassement Protocol (III).

The grades used to produce this ranking measure the relative value of the biological information included in each index.

Recommendations

The number of taxa included in a biologically defined metric (the "size" of the metric) should be considered when evaluating candidate metrics. If total taxa richness is accepted as providing a signal of biological condition, then in general a biologically defined metric that

includes many taxa will provide a clearer signal of biological condition than a metric with few taxa.

Comparison of the performance of candidate metrics to that of randomly generated metrics as described in this study provides another tool for evaluating metrics. In the absence of other, overriding criteria a metric that performs much better than random metrics of the same size is to be preferred to a metric that does not distinguish itself against random metrics of the same size.

5.4 – Speculation

Some random metrics provided a clearer signal of biological condition than biologically defined metrics within a single dataset.

Even confining consideration to the 113 taxa present in the datasets used in this study, the universe of possible linear metrics is very, very large. Even confining oneself to simple linear combinations, where the coefficients are all either 0 or 1 (as in the B-IBI metrics) the number of possibilities is 2^{113}.

Given the objective of finding metrics to provide a signal of biological condition it is possible to define a search algorithm that would generate metrics, test them for the strength of their signal of biological condition, reject the ones that did poorly and keep the ones that did well for further consideration.

As mentioned above, a metric must also be consistent in providing a signal of biological condition in datasets taken at different places and at different times. An evolutionary search algorithm might generate metrics, select winners, and then test these winners further by examining their performance in other datasets.

If this data exploration procedure found a metric that provided a clear signal of biological condition (or some other ecological property of a site) across space and time, this metric would provide insight into the nature of the biological processes governing the system.

This technique would be almost the opposite of that used in formulating the B-IBI. That method begins with insight into the biological processes, and uses that knowledge to guide the search for metrics.

BIBLIOGRAPHY

Angermeier, P. L., and J. R. Karr. 1994. Biological integrity versus biological diversity as policy directives. BioScience 44:690-697.

Benzecri, J. P. 1992. *Correspondence Analysis Handbook*. Marcel Dekker, Inc., New York.

Boulton, A. J., and P. S. Lake. 1992. The ecology of two intermittent streams in Victoria Australia II. Comparisons of faunal composition between habitats, rivers and years. Freshwater Biology 27:99-121.

Cao, Y., D. D. Williams, and N. E. Williams. 1998. How important are rare species in aquatic community ecology and bioassessment? Limnology and Oceanography 43:1403-1409.

Casella, G., and R. L. Berger. 1990. *Statistical Inference*. Duxbury Press, Belmont, California.

Chutter, F. M. 1972. An empirical biotic index of the quality of water in South African streams and rivers. Water Resources 6:19-20.

Costanza, R., R. d'Arge, R. deGroot, S. Farber, M. Grasso, B. Hannon, K. Linberg, S. Naeem, R. V. O'Neill, J. Paruelo, R. G. Rashim, P. Sutton, and M. v. d. Belt. 1997. The value of the world's ecosystem services and natural capital. Nature 387:253-260.

Digby, P. G. N., and R. A. Kempton. 1987. *Multivariate Analysis of Ecological Communities*. Chapman and Hall, London.

Dillon, W. R., and M. Goldstein. 1984. *Multivariate Analysis: methods and applications*. John Wiley & Sons, New York.

Doberstein, C. P., J. R. Karr, and L. L. Conquest. 2000. The effect of fixed-count subsampling on macroinvertebrate biomonitoring in small streams. Freshwater Biology 44:355-371.

Fairweather, P. G. 1999. State of environment indicators of 'river health': exploring the metaphor. Freshwater Biology 42:211-220.

Faith, D. P., P. L. Dostine, and C. L. Humphrey. 1995. Detection of mining impacts on aquatic macroinvertebrate communities: Results of a disturbance experiment and the design of a multivariate BACIP monitoring programme at Coronation Hill, Northern Territory. Australian Journal of Ecology 20:167-180.

Field, J. G., K. R. Clarke, and R. M. Warwick. 1982. A practical strategy for analysing multispecies distribution patterns. Marine Ecology Progress Series 8:37-52.

Folke, C., C. S. Holling, and C. Perrings. 1996. Biological diversity, ecosystems, and the human scale. Ecological Applications 6:1018-1024.

Fore, L. S., J. R. Karr, and L. L. Conquest. 1994. Statistical properties of an index of biological integrity used to evaluate water resources. Canadian Journal of Fisheries and Aquatic Sciences 51:1077-1087.

Fore, L. S., J. R. Karr, and R. W. Wisseman. 1996. Assessing invertebrate responses to human activities: evaluating alternative approaches. Journal of the North American Benthological Society 15:212-231.

Franquet, E., S. Doledec, and D. Chessel. 1995. Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. Hydrobiologa 300-301:425-431.

Frey, D. 1977. Biological integrity of water - an historical approach. Pages 127-144 *in* R. K. Ballantine and L. J. Guarraia, editors. The integrity of water. U.S. Environmental Protection Agency, Washington, D.C.

Gittins, R. 1985. *Canonical analysis: a review with applications in ecology*. Springer-Verlag, Berlin.

Gower, J. C., and D. J. Hand. 1996. *Biplots*. Chapman & Hall, London.

Greenacre, M. J. 1984. *Theory and Applications of Correspondence Analysis*. Harcourt Brace Jovanovich, Publishers, London.

Hilsenhoff, W. L. 1977. Use of arthropods to evaluate water quality of streams. Pages 16 *in*. Department of Natural Resources, Madison, Wisconsin.

Hilsenhoff, W. L. 1982. Using a biotic index to evaluate water quality in streams. Pages 22 *in*. Department of Natural Resources, Madison, Wisconsin.

Holling, C. S. 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics 4:1-23.

James, F. C., and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or pandora's box? Annual Review of Ecological Systems 21:129-166.

Karr, J. R. 1991. Biological integrity: A long-neglected aspect of water resource management. Ecological Applications 1:66-84.

Karr, J. R. 1999. Defining and measuring river health. Freshwater Biology 41:221-234.

Karr, J. R., and E. W. Chu. 1998. *Restoring Life in Running Waters: Better Biological Monitoring.* Island Press, Washington D.C.

Karr, J. R., and E. W. Chu. 2000. Sustaining living rivers. Hydrobiologia 422-423:1-14.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser. 1986. Assesment of biological integrity in running waters: a method and its rationale. Illinois Natural History Survey Special Publication 5.

Karr, J. R., and T. E. Martin. 1981. Random numbers and principal components: further searches for the unicorn. Pages 20-24 *in* D. E. Capen, editor. *The use of multivariate statistics in studies of wildlife habitat*. United States Forest Service General Technical Report RM-87.

Kerans, B. L., and J. R. Karr. 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. Ecological Applications 4:768-785.

Kleindl, W. J. 1995. A Benthic Index of Biotic Integrity for Puget Sound Lowland Streams, Washington, USA. Master of Science. University of Washington, Seattle, Washington.

Kolkowitz, R., and M. Marsson. 1909. Ecology of animal saprobia. Internal Review of Hydrobiology and hydrogeography 2:126-153.

Leon, S. J. 1986. Linear Algebra With Applications, 2 edition. MacMillan Publishing Company, New York.

Mathsoft, Inc. 1999. S-PLUS 2000 documentation. Seattle.

Merritt, R. W., and K. W. Cummins. 1996. *An Introduction to the Aquatic Insects of North America, 3rd edition*. Kendall/Hunt Publishing Company, Dubuque, Iowa.

Meyer, J. L. 1997. Stream health: Incorporating the human dimension to advance stream ecology. Journal of the North American Benthological Society 16:439-447.

Minshall, G. W., K. W. Cummins, R. C. Petersen, C. E. Cushing, D. A. Bruns, J. R. Sedell, and R. L. Vannote. 1985. Developments in stream ecosystem theory. Canadian Journal of Fisheries and Aquatic Sciences 42:1045-1155.

Morley, S. A., and J. R. Karr. 2002. Assessing and restoring the health of urban streams in the Puget Sound basin. Conservation Biology In press.

Moyle, P. B. 1993. *Fish*. University of California Press, Berkley.

Mulvey, M., L. Caton, and R. Hafele. 1992. Oregon nonpoint-source monitoring protocols and stream bioassesment field manual for macroinvertebrates and habitat assessment. Oregon Department of Environmental Quality Laboratories, Portland, Oregon.

Murray, J. D. 1996. *Mathematical Biology*, 2 edition. Springer-Verlag, New York.

Nelson, S. M. 1999. Leaf pack breakdown and macroinvertebrate colonization: bioassessment tools for a high-altitude regulated system? Environmental Pollution 110:321-329.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. *Applied Linear Regression Models, 3rd edition*. Irwin, Chicago.

Ohio EPA, O. E. E. P. 1987. Biological criteria for the protection of aquatic life. Vols. 1-3. Division of Water Quality, Ohio Environmental Protection Agency, Columbus, Ohio.

Patterson, A. J. 1996. The effect of recreation on biotic integrity of small streams in Grand Teton National Park. MS. University of Washington, Seattle.

Pielou, E. C. 1984. *The Interpretation of Ecological Data*. John Wiley & Sons, New York.

Pirsig, R. M. 1979. *Zen and the Art of Motorcycle Maintenance*. Bantam Books, New York, N.Y.

Plafkin, J. L., M. T. Barbour, K. D. Porter, S. K. Gross, and R. M. Hughes. 1989. Rapid Bioassessment Protocols for Use in Streams and Rivers: Benthic Macroinvertebrates and Fish. U.S. Environmental Protection Agency, Assessment and Watershed Protection Division, Washington, D.C.

Reynoldson, T. B., R. H. Norris, V. H. Resh, K. E. Day, and D. M. Rosenberg. 1997. The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. Journal of the North American Benthological Society 16:833-852.

Rossano, E. M. 1995. Development of an index of biological integrity for Japanese streams (IBI-J). MS. University of Washington., Seattle.

Scrimgeour, G. J., and D. Wicklum. 1996. Aquatic ecosystem health and integrity: problems and potential solutions. Journal of the North American Benthological Society 15:254-261.

Sovell, L. A., and B. Vondracek. 1999. Evaluation of the fixed-count method for Rapid Bioassessment Protocol III with benthic macroinvertebrate metrics. Journal of the North American Benthological Society 18:420-426.

Stauffer, D. F., E. O. Garton, and R. K. Steinhorst. 1985. A comparison of principal components from real and random data. Ecology 66:1693-1698.

Stoel, T. B. 1999. Reining in urban sprawl. Environment 41:6-11, 29-33.

Suter, G. W. 1993. A critique of ecosystem health concepts and indexes. Environmental Toxicology and Chemistry 12:1533-1539.

Whittaker, R. H. 1967. Gradient analysis of vegetation. Biological Reviews 49:207-264.

Zar, J. H. 1996. *Biostatistical Analysis, 3rd edition*. Prentice Hall, Upper Saddle River, New Jersey.